Some Steps towards Experimental Design for Neural Network Regression

Richard Kodzo Avuglah

Some Steps towards Experimental Design for Neural Network Regression

Richard Kodzo Avuglah

Vom Fachbereich Mathematik

der Technischen Universität Kaiserslautern

zur Verleihung des akademischen Grades

Doktor der Naturwissenschaften

(Doctor rerum naturalium, Dr. rer. nat.)

genehmigte Dissertation

1. Gutachter: Prof. Dr. Jürgen Franke

2. Gutachter: Prof. Dr. Jens-Peter Kreiß

Datum der Disputation: 7. Juni 2011

D 386

Some Steps towards Experimental Design for Neural Network Regression

Richard Kodzo Avuglah

Approved dissertation

by the Department of Mathematics

at the University of Kaiserslautern

for awarding the degree

Doctor of Natural Sciences

(Doctor rerum naturalium, Dr. rer. nat.)

First referee: Prof. Dr. Jürgen Franke

Second referee: Prof. Dr. Jens-Peter Kreiß

Date of Public Defense: 7th June, 2011

D 386

Abstract

We discuss some first steps towards experimental design for neural network regression which, at present, is too complex to treat fully in general. We encounter two difficulties: the nonlinearity of the models together with the high parameter dimension on one hand, and the common misspecification of the models on the other hand.

Regarding the first problem, we restrict our consideration to neural networks with only one and two neurons in the hidden layer and a univariate input variable. We prove some results regarding locally $D-$optimal designs, and present a numerical study using the concept of maximin optimal designs.

In respect of the second problem, we have a look at the effects of misspecification on optimal experimental designs.

Dedication

To my lovely wife Yayra and priceless children Esinam and Etornam.

Acknowledgments

All glory and honor goes to God Almighty because without his abundant grace and love nothing would have been possible.

I would like to express my deep and sincere gratitude to my supervisor, Professor Dr. Jürgen Franke, Chair for Applied Mathematical Statistics, University of Kaiserslautern, for the opportunity and continuous support for my PhD study and research, for his patience, motivation, enthusiasm, and immense knowledge. His outstanding guidance helped me during the time of research and writing of this thesis. I wish to also thank him for providing extra funding through the University of Kaiserslautern for my study when it was needed.

My profound gratitude goes to Professor Dr. Jens-Peter Kreiß, University of Braunschweig, for accepting to be the second referee for my thesis.

I owe a lot of gratitude to my second supervisor, Dr. Alex Sarishvili, Department of Systems, Analysis and Prognosis (SYS), Fraunhofer Institute for Industrial and Financial Mathematics (ITWM), Kaiserslautern, for his kind support, guidance, constructive and valuable comments and suggestions throughout my work.

I am very grateful to Dr. Patrick Lang, Head of Department of Systems, Analysis and Prognosis (SYS), Fraunhofer Institute for Industrial and Financial Mathematics (ITWM), Kaiserslautern, for the offer to join their research department and the important extensive discussions and contributions concerning my work. I also give thanks to my colleagues and workers at ITWM for their kind support and the lovely interactions.

Many thanks must go to my colleagues and workers at the Department of Mathematics at the University of Kaiserslautern in general and the Statistics Group in particular not forgetting our awesome Secretary, Frau Beate Siegler; for the immense contributions and wonderful working atmosphere and environment. In addition, I would in particular like to thank Dr. Joseph Tadjuidje Kamgaing, Mr. Mark Kimathi, Mr. Oliver Tse and Mr. Uditha Prabhath Liyanage for their special supports. I also appreciate the many assistance and special academic and social programs of the International School of Graduate Studies (ISGS). Big hugs to the lovely students and workers I met through the ISGS.

Special thanks to the German Academic Exchange Service (DAAD) for the scholarship which enabled me pursue the Mathematics in Industry and Commerce (MIC) PhD program; and also many thanks to members of DAAD-Freundeskreis in Kaiserslautern for the many programs we enjoyed together.

I am deeply grateful to the Ghanaian community in Kaiserslautern for the many good times we have had together.

I warmly thank Pastor Philip Burton, his family and my entire church family at City Mission Kaiserslautern for the sincere love they have always given to my wife, children and I during our stay.

Finally, I owe my loving thanks to my wife Yayra for her amazing love and support especially through the hard times. Her soothing words always made me feel better and urged me on. Therefore, it is not surprising that I dedicate this thesis to her and our miracle twin babies. Since their birth last September, Esinam and Etornam have brought a lot of energy, joy and blessings to our home. I wish to thank my lovely mum, stepmothers, siblings, in-laws and all friends. Without their encouragement, understanding, support and prayers it would have been impossible for me to finish this work.

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Motivation

The starting point of this thesis was the desire for some guidelines for experimental design in neural network regression, motivated by some practical problems in the context of industry cooperation at the Fraunhofer Institute for Industrial Mathematics (Fraunhofer ITWM).

Let us consider the case of a one-dimensional regression only, where we have real-valued data Y_j depending on a real-valued x_j :

$$Y_j = m(x_j) + \varepsilon_j \quad j = 1, \ldots, n,$$

with independent identically distributed (i.i.d.) residuals ε_j with mean $\mathrm{E}\,\varepsilon_j = 0$.

In neural network regression, the unknown regression function

$$m(x) = \mathrm{E}\{Y_j | x_j = x\}$$

is approximated by a feedforward neural network with, say, one hidden layer with H hidden neurons. This means that $m(x)$ is approximated by a function

of the following parametric form:

$$m(x, \theta) = v_0 + \sum_{h=1}^{H} v_h \psi(w_{0h}) + w_{1h} x$$

for some given activation function $\psi(u)$. Here, we consider the popular logistic activation function

$$\psi(u) = \frac{1}{1 + e^{-u}}$$

which is of sigmoid form, i.e. it looks like the distribution function. The parameter vector $\theta = (v_0, \ldots, v_H, w_{01}, \ldots, w_{0H}, w_{11}, \ldots, w_{1H})^T$ consists of the network weights. For a survey on neural networks, compare e.g. Haykin (1999) or Anders (1997). Furthermore, papers on optimal experimental design for neural networks are found in the engineering and machine learning literature, but they focus on numerical studies and algorithms, compare e.g. Cohn (1996), Choueiki and Mount-Campbell (1999) and Witczak (2006). We are looking for a theoretical basis for those methods.

Given a prescribed sample size n, the experimental design problem consists of the choice of $x_1, \ldots, x_n \in [a, b]$ such that the regression function $m(x)$ may be estimated as good as possible on the interval $[a, b]$. Of course, we have to be precise about what we mean by "as good as possible."

For the original problem of choosing optimal designs for neural network regression, it turned out to be much too ambitious for two reasons:

i) Even if the neural network output function $m(x, \theta)$ describes the data-generating mechanism exactly, i.e. $m(x, \theta_0) = m(x)$ for some θ_0, the structure of the functions $m(x, \theta)$ for general H is much too complicated. The current literature on optimal design for nonlinear regression is still concerned with much simpler regression functions and low-dimensional

parameters, compare, e.g. Dette and Pepelyshev (2008) and Dette et al. (2006).

ii) Additionally, in neural network regression, one does not usually assume that the model is completely correct, i.e. we only have $m(x) \approx m(x, \theta_0)$ for some θ_0 and large enough H. So we have to deal with optimal design in misspecified regression models which has also not been investigated a lot in the literature. Some first steps have been done in the context of robustness of design, compare, section 2.4.

So in this thesis we can only do some first steps towards a theory of experimental design for neural network regression. The outline of the thesis is as follows:

In chapter 2, we give a review of some literature on experimental designs and also the introduction of some concepts which we shall need later. We start with the classical optimal design problem for linear regression models. In section 2.2, we have a look at nonlinear regression models. Then, in section 2.3, we consider sequential optimal designs which may be appropriate for nonlinear regression in particular, since one chooses the design points one after the other and may exploit preliminary estimates of the parameters since, in general, the optimal design will be local, i.e. depending on the unknown true parameter value. We close in section 2.4 with a survey on misspecified models in the context of experimental designs.

In Chapter 3, we consider neural network regression with $H = 1$ or 2 hidden neurons only. We follow Dette and Pepelyshev (2008) by focusing on locally $D-$optimal designs, proving some results and having a look at some simulations. We conclude the chapter with a numerical study which concerns the

concepts of locally $D-$optimal designs and maximin designs.

Finally, in chapter 4, we study the effect of model misspecification on experimental design in general without referring to neural network regression in particular. We prove convergence of parameter estimates and asymptotic normality including formulas for the error covariance matrix which is a major tool in judging the quality of an estimate and, therefore, in choosing good designs.

Chapter 2

A Survey of Optimal Design Problems

2.1 Classical Regression Optimal Designs

Regression is a statistical tool used for obtaining information on a response variable Y that depends on a (possibly vector valued) variable x. When the variable x is under the control of an experimenter, he may like to know the values of x where it is "best" to observe the response Y. Usually, the experimenter is constrained by resources such as money, time and the number of observations he can take. The optimal regression design problem is about choosing levels of x and allocating observations at x so as to optimize specified criteria related to various constraints. There is a vast number of criteria in the experimental design literature. The choice of criteria would depend on the objective of the experiment.

Following Kiefer and Wolfowitz (1960), we suppose that $z_1, z_2 \ldots, z_p$ are p given linearly independent functions on a space Ω and are continuous in a

topology in which Ω is compact. The space Ω will usually be a closed compact set in a Euclidean space of a particular dimension. In the linear regression setting, we assume that at each point x in Ω the experimenter observes a random response variable Y_j, given a vector of predictors x_j for a subject j; $j = 1, 2, \ldots, n$ assuming a model of the form

$$Y_j = m(x_j, \theta) + \varepsilon_j \tag{2.1}$$

where $m(x_j, \theta) = z^T(x_j)\theta$, where θ is the $p \times 1$ column vector of p unknown real parameters, $z^T(x)$ consists of a $p \times 1$ column vector of p regressor $X = z(x) = (z_1(x_1), z_2(x_2), \ldots, z_p(x_n))^T$ and ε_j are random errors such that they are uncorrelated and have constant variance σ^2. The least squares estimate of the vector of model parameters is given by

$$\hat{\theta} = \mathrm{I}^{-1} X^T Y \tag{2.2}$$

where the response, $Y = (Y_1, Y_2, \ldots, Y_n)^T$, and the information matrix, $\mathrm{I} = X^T X$. Thus, the information matrix depends on the design vector x through the matrix X, which is also called the design matrix.

Suppose an experimenter would like to conduct an experiment whose response Y satisfies (2.1). When the total number of observations to be taken is n, the objective of the optimal regression designs is to choose optimal values of $x_1, x_2, \ldots, x_n$ not necessarily distinct, from a design space Ω such that certain criteria are satisfied. In experimental designs, it is vital to distinguish between *discrete* and *continuous* designs. Wynn (1970), Kiefer (1961) and Adewale and Wiens (2009) are among the authors that helped to establish the distinction.

An $n-$tuple of points $x_1, x_2, \ldots, x_n$ not necessarily distinct, from the design

space Ω is an *exact* or *discrete* design. Thus, the exact or discrete design, denoted by D_n corresponds to a discrete probability measure ξ on Ω which is formed by attaching masses which are integral multiples of n^{-1} to each point in D_n. A design measure, referred to merely as a measure, is a probability measure, denoted by ξ, on Ω. The probability measure ξ on Ω is also called *approximate* or *continuous* design. Specifically, ξ is a member of the set Ξ, of all measures defined on the Borel field $\mathcal{B}$ generated by the open sets of Ω and such that

$$\int_{\Omega} \xi(dx) = 1.$$

It is assumed that $\mathcal{B}$ contains all one-point sets. Finding exact designs is an integer optimization problem- optimization in a discrete domain - which is, in general, analytically intractable. The intractability of the exact problem led to the development of Kiefer's "approximate theory." With approximate theory comes mathematical convenience such that the various optimizations which are otherwise unwieldy in the exact theory become tractable through convex theory. However, the resulting designs from approximate theory are not directly implementable. They need to be approximated by exact designs. The books by Fedorov (1972), Silvey (1980) and Pukelsheim (1993) are classical references on this subject.

We denote the information matrix of θ corresponding to the design ξ as $\mathrm{I}(\xi)$. The $p \times p$ matrix $\mathrm{I}(\xi)$ is assumed here to be positive definite and for a measure ξ on Ω it can be written as

$$\mathrm{I}(\xi) = \int_{\Omega} z(x) z^T(x) d\xi(x).$$

Furthermore, from these definitions,

$$X^T X = n\mathrm{I}(\xi) \tag{2.3}$$

where X is an $n \times p$ design matrix.

Optimal designs are usually obtained by optimizing functions of the information matrix, $\mathrm{I}(\xi)$. The most intensively studied design criterion is the D-optimality criterion (Silvey (1980)) and it is the design ξ^* that maximizes the determinant of the information matrix. That is,

$$\xi^* = \arg\max_{\xi \in \Xi} \det\{\mathrm{I}(\xi)\}.$$

This design minimizes the determinant of the variance-covariance matrix of the estimates of the model parameters.

Other criteria that have been studied in the literature include the G-optimality criterion, the design minimizing the maximum (over the design space) variance of the predicted response (Kiefer and Wolfowitz (1960)). That is,

$$\xi^* = \min_{\xi \in \Xi} \max_{x \in \Omega} \{z^T(x)\mathrm{I}^{-1}(\xi)z(x)\}.$$

The Q-optimality criterion, also known as I-optimal criterion seeks the design minimizing the integrated (or average) variance of the estimated response over the design space;

$$\xi^* = \min_{\xi \in \Xi} \int_{\Omega} z^T(x)\mathrm{I}^{-1}(\xi)z(x)\, dx.$$

The A-optimality criterion seeks the design minimizing the trace of the variance-covariance matrix;

$$\xi^* = \arg\min_{\xi \in \Xi} \mathit{trace}\{\mathrm{I}^{-1}(\xi)\}.$$

The E-optimality criterion seeks the design minimizing the maximum eigenvalue (λ) of the variance-covariance matrix of model estimates;

$$\xi^* = \arg\min_{\xi \in \Xi} \lambda_{\max}\{\mathrm{I}^{-1}(\xi)\}.$$

The c-optimality criterion seeks the design minimizing the variance of a given linear combination of parameter estimates. For a fixed vector c, the c-optimal design is given by

$$\xi^* = \arg\min_{\xi\in\Xi}\{c^T \mathrm{I}^{-1}(\xi)c\}.$$

Kiefer and Wolfowitz (1960) presented extensive results on D- and G-optimality, including the celebrated Equivalence Theorem. The Equivalence Theorem established that a design is D-optimal if and only if it is G-optimal.

Theorem 2.1.1 (General Equivalence Theorem (Kiefer and Wolfowitz (1960)))**.** *A measure ξ^* is $D-$optimum if ξ is chosen such that*

$$\det\{\mathrm{I}(\xi^*)\} = \sup_{\xi\in\Xi}\det\{\mathrm{I}(\xi)\}. \tag{2.4}$$

Let

$$d(x,\xi) = z^T(x)\mathrm{I}^{-1}(\xi)z(x). \tag{2.5}$$

A measure ξ^* is $G-$optimum if ξ is chosen such that

$$\sup_{x\in\Omega} d(x,\xi^*) = \inf_{\xi\in\Xi}\sup_{x\in\Omega} d(x,\xi). \tag{2.6}$$

The integral with respect to ξ of $d(x,\xi)$ is p; hence, $\sup_{x\in\Omega} d(x,\xi) \geq p$. Thus, a sufficient condition for ξ to satisfy (2.6) is

$$\sup_{x\in\Omega} d(x,\xi) = p. \tag{2.7}$$

(2.4), (2.6) and (2.7) are equivalent wherever $\mathrm{I}(\xi)$ is nonsingular.

From the above theorem, we note that the design that maximizes $\det\{\mathrm{I}(\xi)\}$ also minimizes the maximum value of $z^T(x)\mathrm{I}^{-1}(\xi)z(x)$ over the design space

Ω.

We also note that from (2.3) and (2.5),

$$d(x, \xi) = nz^T(x)(X^T X)^{-1} z(x). \tag{2.8}$$

2.1.1 Standard Designs

Experimental designs are chosen based on the objectives of the experiment and the number of factors to be investigated. Screening designs for instance may be used if the aim of the experiment is to select or screen out the few important main effects from the many less important ones. Comparative designs are employed when we have one or several factors under investigation, but the main aim of our experiment is to make a conclusion whether a factor, in the presence of, and/or in spite of the existence of the other factors, is significant. That is, whether or not there is a significant change in the response for different levels of that factor. Response surface method (RSM) designs are used when we intend to estimate interaction and even quadratic effects, and therefore also have an idea of the (local) shape of the response surface we are investigating. They are used to find improved or optimal process settings and also used to make a product or process more robust against external and non-controllable influences. "Robust" means relatively insensitive to these influences. If you have factors that are proportions of a mixture and you want to know what the "best" proportions of the factors are so as to maximize (or minimize) a response, then you need a mixture design. Furthermore, regression designs are used if we want to model a response as a mathematical function (either known or empirical) of a few continuous factors and we desire "good" model parameter estimates (i.e., unbiased and minimum variance). Below are other standard or classical designs which are

employed for various experimental objectives.

2.1.1.1 Full factorial designs

An experimental design with all possible combinations of high and low levels (or '+1' and '-1') of all the input factors is called a full factorial design. In other words, a design in which every setting of every factor appears with every setting of every other factor is a full factorial design. As an example, if there are k factors, each at 2 levels, a full factorial design has 2^k runs. When the number of factors is 5 or greater, a full factorial design requires a large number of runs and is not very efficient. Fractional factorial design or a Plackett-Burman design is a better choice for 5 or more factors.

2.1.1.2 Fractional factorial designs

A factorial experiment in which only an adequately chosen fraction of the treatment combinations required for the complete factorial experiment is selected to be run. Considering a full factorial design of k factors, each of 2 levels as above, even if the number of factors in a design is small, the runs specified for a full factorial can quickly become very large. For example, $2^6 = 64$ runs is for a two-level, full factorial design with six factors. To this design we need to add a good number of center point runs and we can thus quickly run up a very large resource requirement for runs with only a modest number of factors. This problem is solved by using only a fraction of the runs specified by the full factorial design. Which runs to keep and which to leave out is the subject of interest here. In general, we pick a fraction such as $1/2, 1/4$, etc. of the runs called for by the full factorial. Various strategies are used to ensure an appropriate choice of runs. Thus, a carefully chosen fraction of the runs may be all that is necessary.

2.1.1.3 Plackett-Burman(PB)

Plackett and Burman (1946) described the construction of very economical designs with the run number a multiple of four rather than a power of 2. Plackett-Burman (PB)designs are very efficient screening designs when only main effects are of interest. These designs are used for screening experiments because, in a PB design, main effects are, in general, heavily confounded with two-factor interactions. The PB design in 12 runs, for example, may be used for an experiment containing up to 11 factors. With a 20-run design we can run a screening experiment for up to 19 factors, up to 23 factors in a 24-run design, and up to 27 factors in a 28-run design. PB designs even exist for design runs higher than 28. These Resolution III designs are known as Saturated Main Effect designs because all degrees of freedom are utilized to estimate main effects.

These designs do not have a defining relation since interactions are not identically equal to main effects. With the 2^{k-p}_{III} designs, a main effect column X_i is either orthogonal to X_iX_j or identical to $\pm X_iX_j$. For Plackett-Burman designs, the two-factor interaction column X_iX_j is correlated to every X_k (for k not equal to i or j). However, these designs are very useful for economically detecting large main effects, assuming all interactions are negligible when compared with the few important main effects.

2.1.1.4 Central Composite Design

A Box-Wilson Central Composite Design, commonly called 'a central composite design,' contains an embedded factorial or fractional factorial design with center point that is augmented with a group of 'star points' that allow estimation of curvature. If the distance from the center of the design space to

a factorial point is 1 unit for each factor, the distance from the center of the design space to a star point is $\pm\alpha$ with $|\alpha| > 1$. The precise value of α as well as the number of center point runs the design contains, depends on certain properties desired for the design and on the number of factors involved. A central composite design always contains twice as many star points as there are factors in the design. The star points represent new extreme values (low and high) for each factor in the design.

There are three types of central composite designs. These depend on where the star points are placed:

1. Circumscribed (CCC): This is the original form of the central composite design. The star points are at some distance α from the center based on the properties desired for the design and the number of factors in the design. These designs have circular, spherical, or hyper-spherical symmetry and require 5 levels for each factor. Augmenting an existing factorial or resolution V fractional factorial design with star points can produce this design.

2. Inscribed (CCI): This is a scaled down CCC design with each factor level of the CCC design divided by α to generate the CCI design. When true limits for factor settings are specified, the CCI design uses the factor settings as the star points and creates a factorial or fractional factorial design within those limits. This design also requires 5 levels of each factor.

3. Face Centered (CCF): The star points are at the center of each face of the factorial space, so $\alpha = \pm 1$. This type requires three levels of each factor. Augmenting an existing factorial or resolution V design with

appropriate star points can also produce this design.

2.1.1.5 Box-Behnken Design

This is an independent quadratic design in that it does not contain an embedded factorial or fractional factorial design. In this design the treatment combinations are at the midpoints of edges of the process space and at the center. It is an alternative choice for fitting quadratic models that requires three levels of each factor.

2.1.1.6 Latin Square

A Latin square is an $n \times n$ array filled with n different Latin letters, each occurring exactly once in each row and exactly once in each column.

2.1.2 DOE Terminology

The following are some definitions for some of the basic terms used in design of experiment.

2.1.2.1 Design

A set of experimental runs which allows you to fit a particular model and estimate your desired effects.

2.1.2.2 Balanced Design

An experimental design where all cells (i.e. treatment combinations) have the same number of observations.

2.1.2.3 Design Matrix

A matrix description of an experiment that is useful for constructing and analyzing experiments.

2.1.2.4 Effect

This is how changing the settings of a factor changes the response. The effect of a single factor is also called a main effect.

2.1.2.5 Treatment factors and their levels

A treatment is a specific combination of factor levels whose effect is to be compared with other treatments. Although the term treatment factor might suggest a drug in a medical experiment, it is used to mean any substance or item whose effect on the data is to be studied. The levels are the specific types or amounts of the treatment factor that will actually be used in the experiment.

2.1.2.6 Orthogonality

Two vectors of the same length are orthogonal if the sum of the products of their corresponding elements is zero. An experimental design is orthogonal if the effects of any factor balance out (sum to zero) across the effects of the other factors.

2.1.2.7 Randomization

A schedule for allocating subjects or experimental material to treatments such that the conditions in one run neither depend on the conditions of the previous run nor predict the conditions in the subsequent runs. The impor-

tance of randomization cannot be over stressed. Randomization is necessary for conclusions drawn from the experiment to be correct, unambiguous and defensible. It is to prevent systematic and personal biases from being introduced into the experiment by the experimenter.

2.1.2.8 Rotatability

A design is rotatable if the variance of the predicted response at any point x depends only on the distance of x from the design center point. A design with this property can be rotated around its center point without changing the prediction variance at x. Rotatability is a desirable property for response surface designs (i.e. quadratic model designs).

2.1.2.9 Blocking

The experimental conditions under which an experiment is run should be representative of those to which the conclusions of the experiment are to be applied. For inferences to be broad in scope, the experimental conditions should be rather varied. However, an unfortunate consequence of increasing the scope of the experiment is an increase in the variability of the response. Blocking is a technique that can often be used to help deal with this problem.

To block an experiment is to divide, or partition, the observation into groups called blocks in such a way that the observations in each block are collected under relatively similar experimental conditions. If blocking is done well, then comparisons of two or more treatments are made more precisely than similar comparisons from an unblocked design. Blocking also isolates a systematic effect and prevents it from obscuring the main effects.

2.1.2.10 Replication

Performing the same treatment combination more than once. Replication allows an estimate of the random error independent of any lack of fit error. There is a difference between replication" and repeated measurements." For example, suppose four subjects are each assigned to a drug and a measurement is taken on each subject. The result is four independent observations on the drug. This is replication." On the other hand, if one subject is assigned to a drug and then measured four times, the measurements are not independent. We call them repeated measurements."

2.1.2.11 Resolution

A term which describes the degree to which estimated main effects are aliased (or confounded) with estimated 2-level interactions, 3-level interactions, etc. In general, the resolution of a design is one more than the smallest order interaction that some main effect is confounded (aliased) with. If some main effects are confounded with some 2-level interactions, the resolution is III. Full factorial designs have no confounding and are said to have resolution "infinity". For most practical purposes, a resolution V design is excellent and a resolution IV design may be adequate. Resolution III designs are useful as economical screening designs.

2.1.2.12 Screening Designs

A DOE that identifies which of many factors have a significant effect on the response. Typically screening designs have more than five factors.

2.1.2.13 Scaling or Coding Factor Levels

Transforming factor (input) levels so that the high value becomes +1 and the low value becomes −1.

2.1.2.14 Experimental Units

These are the "material" to which the levels of the treatment factor(s) are applied. For example, in agriculture these would be individual plots of land, in medicine they would be human or animal subjects, in industry they might be batches of raw material, factory workers, etc. If an experiment has to be run over a period of time, with the observations being collected sequentially, then the times of the day can also be regarded as experimental units.

2.2 Nonlinear Optimal Designs

According to Khuri and Cornell (1996), a model $Y_j = m(x_j, \theta) + \varepsilon_j$ is said to be nonlinear if at least one of its parameters appears nonlinearly. For example, the models

$$Y_j = \theta_1 e^{-\theta_2 x_j} + \varepsilon_j \tag{2.9}$$

$$Y_j = \theta_1 + \theta_2 e^{-\theta_2 x_j} + \varepsilon_j \tag{2.10}$$

$$Y_j = \frac{1}{\theta_1 + \theta_2 x_j} + \varepsilon_j \tag{2.11}$$

$$Y_j = \left(\frac{\theta_1}{\theta_1 - \theta_2}\right)(e^{-\theta_2 x_j} - e^{-\theta_1 x_i}) + \varepsilon_j \tag{2.12}$$

The term *partially nonlinear* is used to describe a model in which some of the parameters are linear and some are nonlinear, such as models (2.9) and (2.10). Khuri and Cornell (1996) call a model *intrinsically linear* if:

1. it can be reduced to a linear model by a suitable re-parameterization of the model. For example, the nonlinear model

$$E(Y_j) = \theta_1 + e^{\theta_2} x_j \tag{2.13}$$

can be reduced to a linear model, $E(Y_j) = \theta_1 + \gamma_1 x_j$, by transforming $\gamma_1 = e^{\theta_2}$;

2. the nonlinear model is reduced to a linear form by applying a transformation to the model itself. For example, if we consider the model in (2.9), then a natural logarithmic transformation can reduce $E(Y_j)$ to the linear form $\ln[E(Y_j)] = \ln(\theta_1) - \theta_2 x_j$ provided $\theta_1 > 0$.

Such a transformation can change the structure and distribution of the error term associated with the model. To explain this, let Y and ε be the observed response and random error, respectively, for model (2.9). Then

$$\begin{aligned} \ln\,(Y) &= \ln\,[\eta(x) + \varepsilon] \\ &= \ln\,[\eta(x)] + \ln\,\left[1 + \frac{\varepsilon}{\eta(x)}\right] \end{aligned} \tag{2.14}$$

The error term for the transformed model is now $\ln\left[1 + \varepsilon/\eta(x)\right]$, which in general has the distribution different from that of ε. For example, if ε satisfies the usual assumptions of normality, independence, and homogeneity of variance, the error term for model (2.14) will have a non-normal distribution which depends on x through $\varepsilon(x)$. Thus, the variance of this error term cannot be assumed to be constant as in the original model. Consequently, even if the mean $\eta(x)$ in a nonlinear model can be reduced to a linear form by a proper transformation, such a transformation should be used only if it can be demonstrated that the aforementioned assumptions with respect to the transformed model are not severely violated. Nonlinear models have been used in many fields, particularly in biological and chemical sciences where

the growth of a particular organism, or the yield that results from a chemical reaction, can be depicted by a nonlinear model. Draper and Smith (1981) and Chaudhuri and Mykland (1993) had listed several examples.

2.2.1 Generalized linear Models (GLMs)

Another example (or class) of nonlinear models is generalized linear models which are quite frequently used in clinical or epidemiological studies where the data violate the assumptions of a linear model. In standard general linear model, the responses are assumed to be continuous (quite often, normally distributed) with uncorrelated errors and homogeneous variances. Introduced by Nelder and Wedderburn (1972), GLMs are a unified class of regression models for discrete and continuous response variables, and have been used routinely in dealing with observational studies.

A generalized linear model consists of three (3) components:

1. The elements (or observations) $y_1, y_2, \ldots, y_n$ of a response vector y, with respective means $\mu_1, \mu_2, \ldots, \mu_n$ are distributed independently according to a certain probability distribution considered to belong to the exponential family whose probability density (or mass) function is given by

$$m(y, \theta, \phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right], \tag{2.15}$$

where $a(\cdot), b(\cdot)$ and $c(\cdot)$ are specific (known) functions; θ is a natural location parameter, and ϕ is often called dispersion parameter. The function $a(\phi)$ is frequently of the form $a(\phi) = \phi \cdot \omega$ where ω is a known constant. The binomial, Poisson, gamma, probit and normal distributions are members of this family. For some common members of the

family, $\phi = 1-$ like the binomial and Poisson- except in situations of over-dispersion. The most prominent member of the exponential family is the normal distribution. The probability density function for a normal random variable y with parameters μ and σ is given by

$$\begin{aligned} m(y;\mu,\sigma) &= \exp\{-[y-\mu]^2/2\sigma^2\} \cdot \frac{1}{\sigma\sqrt{2\pi}} \\ &= \exp\left\{(y\mu - \mu^2/2)/\sigma^2 - \frac{1}{2}[y^2/\sigma^2 + \ln(2\pi\sigma^2)]\right\}. \end{aligned}$$

This density function is of the form given in equation (2.15) with $\theta = \mu, b(\theta) = \mu^2/2, a(\phi) = \phi, \phi = \sigma^2$, and $c(y,\phi) = -\frac{1}{2}[\frac{y^2}{\sigma^2} + \ln(2\pi\sigma^2)]$. The location parameter and the natural scale parameters here are respectively, μ and σ^2 as expected.

For the Poisson distribution, the probability function is given by

$$\begin{aligned} m(y;u) &= \frac{e^{-\mu}\mu^y}{y!} \\ &= \exp[y\ln\mu - \mu - \ln(y!)]. \end{aligned}$$

As a result, $\theta = \ln\mu, b(\theta) = e^\theta$, and $c(y,\phi) = -\ln(y!)$. Thus, the location parameter is μ and the scale parameter is $\phi = 1$.

For any distribution in the form of (2.15), the mean and variance of the response variable y are respectively given by

$$E(y) = \mu = \frac{db(\theta)}{d\theta} = b'(\theta) \tag{2.16}$$

and

$$\begin{aligned} \mathrm{Var}(y) &= \frac{d^2b(\theta)}{d\theta^2}a(\phi) = a(\phi)b''(\theta) \\ &= \frac{d\mu}{d\theta}a(\phi), \end{aligned}$$

where primes denote differentiation with respect to the canonical parameter θ.

Let Var_μ be the variance of the response, y, apart from $a(\phi)$; Var_μ denotes the dependence of the variance of the response on its mean. Thus,

$$\text{Var}_\mu = \frac{\text{Var}(y)}{a(\phi)} = \frac{d\mu}{d\theta}.$$

As a result, we have

$$\frac{d\theta}{d\mu} = \frac{1}{\text{Var}_\mu}. \tag{2.17}$$

2. A linear regression function, or linear predictor, in n control variables $x_1, x_2, \ldots, x_n$ of the form

$$\eta = z^T(x)\theta, \tag{2.18}$$

where $z(x) = (z_1(x), z_2(x), \ldots, z_p(x))^T$ are p regressors depending on a vector of n control (input) variables $x = (x_1, x_2, \ldots, x_n)$. θ is an unknown parameter vector of order $p \times 1$ and $z^T(x)$ is the transpose of $z(x)$.

3. A link function $g(\mu)$ which relates η in (2.18) to the mean response μ so that

$$\eta_j = g(\mu_j), \;\; j = 1, 2, \ldots, n,$$

where $g(\cdot)$ is a monotone differentiable function. The term link is derived from the fact that the function is the link between the mean and the linear predictor. The expected response is

$$E(y_j) = g^{-1}(\eta_j) = g^{-1}[z^T(x)\theta].$$

When g is the identity function and the response has the normal distribution, we obtain the special class of linear models. Thus, in multiple

linear model

$$\mu_j = \eta_j = z^T(x)\theta \quad j = 1, 2, \ldots, n$$

suggests a special case in which $g(\mu_j) = \mu_j$, and thus the link function used is the identity link. There are many possible choices of the link function. If we choose

$$\eta_j = \theta_j \tag{2.19}$$

then we say that n_j is the canonical link.

Also, the variance $\sigma_j^2 (j = 1, 2, \ldots, n)$ is a function of the mean μ_j. The mean response, $\mu(x)$, at a point x in a region of interest, R, is given by

$$\mu(x) = g^{-1}[z^T(x)\theta] = g^{-1}[\eta(x)], \tag{2.20}$$

where $\eta(x)$ is the linear predictor in (2.18), and g^{-1} is the inverse function of the g. An estimate of $\mu(x)$ is obtained by replacing θ in (2.20) with $\hat{\theta}$, the maximum likelihood estimate of θ, that is

$$\hat{\mu}(x) = g^{-1}[z^T(x)\hat{\theta}]. \tag{2.21}$$

2.2.2 Dependency of the Information Matrix

The Fisher information associated with a nonlinear experiment is typically a complex nonlinear function of the unknown parameter of interest. As a result, we face an awkward situation. Designing an efficient experiment will require knowledge of the parameter, but the purpose of the experiment is to generate data to yield parameter estimates. Cochran (1973) described this dependency: "You tell me the value of θ, and I promise to design the best experiment for estimating θ." Bates and Watts (1988) also remarked on page 129 of their book : "It is awkward to specify initial estimates ... before an experimental design can be obtained, since, after all, the purpose of the

experiment is to determine parameter estimates." The following are some approaches that have been outlined by many authors including Adewale and Wiens (2009) and Chaudhuri and Mykland (1993) for handling this dependency problem.

The easiest and earliest approach is to adopt a best guess of the parameter values. Given best guesses for parameter values, the nonlinear design problem becomes amenable to the theory of optimal design for linear models. Chernoff (1953) dubbed this design locally optimal design. An obvious practical drawback of this approach, noted by several authors, is that the choice of the best guesses for the parameters may be far from the true parameters and the behavior of the locally optimal design may be quite sensitive to even small perturbations in the parameter value.

An approach that has been used to remedy the non-robustness of the locally optimal design is a Bayesian paradigm. In the Bayesian approach a prior distribution, say $\pi(\theta)$, is assumed on the unknown parameters. The Bayesian optimal design is the design optimizing the expectation of the criterion of interest, where expectation is taken with respect to the assumed prior distribution. That is, if we let $\Psi(\mathrm{I}(\xi,\theta))$ be a function of $\mathrm{I}(\xi,\theta)$,

$$E_\theta \Psi(\mathrm{I}(\xi,\theta)) = \int \Psi(\mathrm{I}(\xi,\theta))\pi(\theta)d\theta.$$

The prior distribution is usually interpreted as the experimenter's prior belief in the adequacy of the model over a specified range of parameter values. Chaloner and Larntz (1989), Chaloner and Verdinelli (1995) and others have studied Bayesian designs.

An alternative to the Bayesian paradigm is the minimax (or maximin) approach used by Sitter (1992). The approach assumed that there is range of plausible values for unknown parameters. That is, $\theta \in \Theta$, where Θ is a range of specific (not represented by distribution) parameter values the experimenter beliefs are plausible. The minimax optimal design is the design minimizing the maximum (over the range of the parameters) of the criterion, that is,

$$\min_{\xi} \max_{\theta \in \Theta} \Psi(\mathrm{I}(\xi, \theta)).$$

This approach is robust in the sense that it produces the optimal design with the least loss when the parameters take the worst possible value within their ranges. These least favorable parameter values are those that maximize the loss (King and Wong (2000); Dette et al. (2003)).

Sequential design is another strategy that has been used in dealing with parameter-dependency of design criteria. In sequential design, the experiment is done in stages. The fundamental idea behind such a strategy is to divide the resources (e.g., time, money, and human power) into small groups and to split the entire experiment into several steps or stages. At each step or stage an experiment is carried out using only a single portion of the divided resources. Analysis is carried out at the end of stage. Parameter estimates from a previous stage are used as best guesses for the current design i.e. updating the parameter estimates by using the available data to efficiently design the next step. Sequential design can be described as progressive locally optimal design. Sinha and Wiens (2002) are among authors that have taken this approach to nonlinear design.

2.2.3 Least Squares Estimates

Ratkowsky (1983a) discussed the least squares (LS) estimate of the parameter θ by considering the nonlinear model

$$Y_j = m(x_j, \theta) + \varepsilon_j. \tag{2.22}$$

Just like in linear models, the least squares estimate of the parameter θ is obtained by minimizing the function

$$S(\theta) = \sum (Y_j - m(x_j, \theta))^2 \tag{2.23}$$

Writing S in place of $S(\theta)$ to simplify the notation, the minimum of S may be obtained by differentiating (2.23) with respect to θ, setting the derivative equal to zero, as follows:

$$\frac{\partial S}{\partial \theta} = -2 \sum (Y_j - m(x_j, \theta))(\log x_j) m(x_j, \theta) = 0$$

and attempting to solve for θ, the solution to which is denoted $\hat{\theta}$. However, this does not lead to explicit expression for $\hat{\theta}$. Instead, the resulting rearranged equation

$$\sum Y_j (\log x_j) m(x_j, \hat{\theta}) = \sum (\log x_j) m(x_j, \hat{\theta})^2 \tag{2.24}$$

can yield the LS estimate $\hat{\theta}$ only by an iterative procedure starting from some assumed value of $\hat{\theta}$. This procedure can be very complex.

Khuri and Cornell (1996) mention several methods for computing the least squares estimates which include the the most widely used Gauss- Newton method and its modified version by Hartley (1961), the steepest descent method, and the method developed by Marquardt (1963) and finally the derivative-free Gauss-Newton algorithm developed by Ralston and Jennrich

(1978). All the aforementioned methods require that initial values be specified for the nonlinear model's parameters. The convergence of any of these methods to the least squares estimates and the rate of convergence heavily depend on the choice of initial values but Ratkowsky (1983a) described procedures for obtaining good initial values of the parameters. Lawton and Sylvestre (1971) also introduced a method whereby the specification of initial values is required only for those parameters which appear nonlinearly in the model.

2.3 Sequential Optimal Designs

A sequential D-optimal design scheme is described by Wynn (1970) in the following procedure by making use of equation (2.8).

Let D_{n_0} be a discrete design with n_0 points, $x_1, \ldots, x_{n_0}$ which is admissible in the sense that $X_{n_0}^T X_{n_0}$ is non-singular. From $x_1, \ldots, x_{n_0}$, by successive addition of points, he generates a sequence of designs such that in the limit the associated measures become $D-$optimum. Thus, he first finds a point $x_{n_0+1} \in \Omega$ which maximizes the variance function obtained by using D_{n_0}; that is choose x_{n_0+1} such that

$$\sup_{x \in \Omega} d(x, \xi_{n_0}) = d(x_{n_0+1}, \xi_{n_0}).$$

He then forms a new design D_{n_0+1}, with $n_0 + 1$ points by adding x_{n_0+1} to D_{n_0} and continues the process to obtain a sequence $D_{n_0} \subset D_{n_0+1} \subset \cdots \subset D_n \subset \cdots$, where D_n is obtained from D_{n-1} by adding a point of maximum variance, over Ω, of the estimated response mean obtained from using D_{n-1}. The following theorem contains the basic result of his paper which concerns the sequence of associated measures $\{\xi_n\}_{n_0}^{\infty}$.

Theorem 2.3.1 (Wynn (1970)). *As* $n \to \infty$, $\lim \det\{\mathrm{I}(\xi_n)\} = \det\{\mathrm{I}(\xi^*)\}$, *where* ξ^* *is a* $D-$*optimum measure.*

Chaudhuri and Mykland (1993) investigated the designing of nonlinear experiments that allowed them construct efficient estimates of parameters. The experiments considered were in two stages: a static design in the initial stage, followed by a fully adaptive sequential stage in which the design points were chosen sequentially, exploiting a $D-$ optimality criterion and using parameter estimates based on available data. Their methodology is as follows:

2.3.0.1 Description of Sampling Scheme

Suppose resources available allow altogether n trials in the experiment, n_1 of these trials are performed in the initial static stage and the remaining $n - n_1$ of the trials are performed in a sequential manner.

Let $x_1, x_2, \ldots, x_{n_1}$ be the first n_1 design points, $Y_1, Y_2, \ldots, Y_{n_1}$ the responses observed after the initial experiment is carried out and $\theta^*_{n_1}$the estimate of θ based on $(Y_1, x_1), \ldots, (Y_{n_1}, x_{n_1})$. For each j such that $n_1 + 1 \leq j \leq n$, the design point x_j, which belongs to the sequential stage of the experiment, will be chosen in such a way that the determinant of the total Fisher Information $\sum_{r=1}^{j} \mathrm{I}(x_r, \theta^*_{j-1})$ is maximized. Here θ^*_{j-1} is an estimate of θ based on $(Y_1, x_1), \ldots, (Y_{j-1}, x_{j-1})$, the data available prior to the jth trial.

Two conditions that play a crucial role in implementing the scheme and studying its performance are:

Condition 2.3.1. *The design space* Ω *is a compact metric space.*

Condition 2.3.2. *It is possible to express* $\mathrm{I}(x,\theta)$ *in the form* $\mathrm{I}(x,\theta) = \{V(x,\theta)\}\{V(x,\theta)\}^T$, *where* V *is the* R^d-*valued function that is jointly continuous in* θ *and* x.

2.3.0.2 Asymptotic Optimality

We discuss the asymptotic optimality of the chosen design whose performance depends on the choice of n_1, the initial design points $x_1, x_2, \ldots, x_{n_1}$ and the estimates, θ_j^*'s. Sufficient conditions to ensure the convergence of the chosen design to the $D-$optimal one as $n \to \infty$ are as follows:

Condition 2.3.3. *(Choice of initial design). As* $n \to \infty, n_1 \to \infty$. *Further, the initial design points* $x_1, x_2, \ldots, x_{n_1}$, *are chosen in such a way that the smallest eigenvalue of the matrix*

$$\frac{1}{n_1}\sum_{j=1}^{n_1}\mathrm{I}(x_j,\theta)$$

remains bounded away from 0 *as* $n \to \infty$ *for any* $\theta \in \Theta$.

Condition 2.3.4. *(The relative size of the initial experiment). The fraction* $\frac{n_1}{n} \to 0$ *as* $n \to \infty$.

Condition 2.3.5. *(A consistency condition). For any* $\varepsilon > 0$,

$$\max_{n_1 \le j \le n} P_\theta(|\theta_j^* - \theta| > \varepsilon) \to 0$$

as $n \to \infty$.

Condition 2.3.6. *(A stability condition). For* $n_1 < k < n$, *let* U_k *denote the product of the determinants*

$$\prod_{j=n_1+1}^{k} \det\left\{\sum_{r=1}^{j}\mathrm{I}(\theta_{j-1}^*, x_r)\right\} \det\left\{\sum_{r=1}^{j}\mathrm{I}(\theta_j^*, x_r)\right\}^{-1}.$$

Then, for any $\varepsilon > 0$, $\max_{n_1 < k < n} P_\theta(U_k > 1+\varepsilon) \to 0$ *as* $n \to \infty$.

Lemma 2.3.1. *The function* $\mathrm{h(A)} = -\log\{\det(\mathrm{A})\}$, *where* A *is the symmetric* $d \times d$ *positive definite matrix, is a strictly convex function. In other words, for* $0 < \alpha < 1$ *and two positive definite matrices* A *and* B *such that* $\mathrm{A} \neq \mathrm{B}$, *we have*

$$\mathrm{h}\{\alpha\mathrm{A} + (1-\alpha)\mathrm{B}\} < \alpha\mathrm{h(A)} + (1-\alpha)\mathrm{h(B)}.$$

Lemma 2.3.2. *Let* $\{\theta_n\}$ *be a sequence of points in* Θ *such that, as* $n \to \infty$, $\theta_n \to \theta \in \Theta$. *Let* ξ_n^* *be the locally* $D-$*optimal design associated with* θ_n *and* ξ^* *be associated with* θ. *Then under conditions (2.3.1) and (2.3.2), the matrix*

$$\int_\Omega \mathrm{I}(x,\theta_n)\xi_n^*(dx) \underset{n\to\infty}{\longrightarrow} \int_\Omega \mathrm{I}(x,\theta)\xi^*(dx),$$

provided

$$\int_\Omega \mathrm{I}(x,\theta)\xi^*(dx)$$

is nonsingular.

Fact 2.3.1. *Let* $1 < n_1 < n$ *be integers* (n_1 *may be a function of* n) *such that* $\frac{n_1}{n} \underset{n\to\infty}{\longrightarrow} 0$. *Then as* $n \to \infty$, *the sum* $\sum_{j=n_1}^{n} \frac{1}{n}$ *diverges to infinity.*

Theorem 2.3.2. *Assume Conditions (2.3.1) to (2.3.6) and* ξ^* *is a locally* $D-$*optimal design at* θ. *If design points are chosen following the scheme at the sequential stage of the experiment, then*

$$\frac{1}{n_1}\sum_{j=1}^{n_1} \mathrm{I}(x_j,\theta) \xrightarrow{p} \int_\Omega \mathrm{I}(x,\theta)\xi^*(dx)$$

as $n \to \infty$.

Now in order to discuss the behavior of the maximum likelihood estimate $\hat{\theta}_n$, Chaudhuri and Mykland (1993) introduced the following conditions on the model $m(x,\theta)$. The parameter space is assumed to be an open convex subset of R^d. We will write $|\cdot|$ to denote the usual Euclidean norm of vectors and matrices.

Condition 2.3.7. *The support of* $m(x,\theta)$ *does not depend on* θ *or* x. *Further, for every fixed* $x \in \Omega$ *and* $y \in \mathbb{R}$, $\log m(x_i,\theta)\}$ *is thrice continuously differentiable in* θ.

Condition 2.3.8. *Let* $\nabla \log m(x,\theta) = \mathrm{G}(x,\theta)$ *be the gradient vector obtained by computing the first-order partial derivatives of* $\log m(x,\theta)\}$ *with respect to* θ. *Then* $\mathrm{G}(x,\theta)$ *satisfies*

$$\int_{\mathbb{R}} \mathrm{G}(x,\theta) m(x,\theta)\mu(dy) = 0$$

and

$$\sup_{x\in\Omega} \int_{\mathbb{R}} |\mathrm{G}(x,\theta)|^{2+t} m(x,\theta)\mu(dy) < \infty$$

for some $t > 0$.

Condition 2.3.9. *Let* $\mathrm{H}(x,\theta)$ *denote the* $d \times d$ *Hessian matrix of* $\log m(x,\theta)$ *obtained by computing the second-order partial derivatives with respect to* θ. *Then* $\mathrm{H}(x,\theta)$ *satisfies*

$$\int_{\mathbb{R}} \mathrm{H}(x,\theta) m(x,\theta)\mu(dy) \;=\; -\int_{\mathbb{R}} \{\mathrm{G}(x,\theta)\}\{\mathrm{G}(x,\theta)\}^T m(x,\theta)\mu(dy) = -\mathrm{I}(x,\theta),$$

and

$$\sup_{x\in\Omega} \int_{\mathbb{R}} |\mathrm{H}(x,\theta)|^2 m(x,\theta)\mu(dy) < \infty.$$

Condition 2.3.10. *For every* $\theta \in \Theta$, *there is an open neighborhood* $N(\theta)$ *of* θ *and a nonnegative random variable* $K(x,\theta)$ *such that*

$$\sup_{x\in\Omega} \int_{\mathbb{R}} K(x,\theta) m(x,\theta)\mu(dy) < \infty,$$

and each of the third-order partial derivatives of $\log m(x,\theta')\}$ *with respect to* θ' *is dominated by* $K(x,\theta)$ *for all* $\theta' \in N(\theta)$.

Theorem 2.3.3 (Chaudhuri and Mykland (1993))**.** *Assume that in addition to conditions assumed in Theorem (2.3.2), Conditions (2.3.7) to (2.3.10) hold. Then there is a consistent choice of the maximum likelihood estimate* $\hat{\theta}_n$ *of* θ *such that, as* $n \to \infty$*, the distribution of* $n^{1/2}(\hat{\theta} - \theta)$ *converges weakly to a* $d-$*dimensional normal distribution with zero mean and* $\left\{\int_\Omega \mathrm{I}(x,\theta)\xi^*(dx)\right\}^{-1}$ *as the variance-covariance matrix.*

Corollary 2.3.1. *Suppose that all the conditions assumed in Theorems (2.3.2) and (2.3.3) hold and let* $\hat{\theta}$ *be consistent choice of the maximum likelihood estimate. Then, as* $n \to \infty$*, the estimated average Fisher information*

$$\frac{1}{n}\sum_{j=1}^{n} \mathrm{I}(\hat{\theta}_n, x_j)$$

converges in probability to the $D-$*optimal Fisher information*

$$\int_\Omega \mathrm{I}(\hat{\theta}_n, x_j)\xi^*(dx).$$

Further, the asymptotic distribution of

$$\left\{\sum_{j=1}^{n} \mathrm{I}(\hat{\theta}_n, x_j)\right\}^{1/2} (\hat{\theta}_n - \theta)$$

is $d-$*variate normal with zero mean and the* $d \times d$ *identity matrix as the variance-covariance matrix.*

Chaudhuri and Mykland (1993) have shown that sequential design in general parametric nonlinear settings, including GLMs, could lead to fully efficient designs and asymptotically efficient maximum likelihood estimators. The work of Dror and Steinberg (2008) is similar to that of Chaudhuri and Mykland (1993) with few differences. Dror and Steinberg (2008) were concerned with small samples, and thus, rapid progress toward efficient design, whereas Chaudhuri and Mykland (1993) emphasized only asymptotic properties. While Chaudhuri and Mykland (1993) gave only general conditions

for initial designs , which could be quite large, Dror and Steinberg (2008) provide an algorithm for efficient design beginning with the first observation.

The methodology of Dror and Steinberg (2008) used a Bayesian methods to jump start the sequential process to achieve good initial small-sample designs, taking advantage of computationally efficient representation of the posterior distribution of the coefficients. The local $D-$optimality criterion for a particular parameter vector θ is $|\mathrm{I}(\theta, \xi)|$, where $|A|$ denotes the determinant of the matrix A. Following Chaloner and Larntz (1989), Dror and Steinberg (2008) began with a proper prior for the parameters in the model. The Bayesian $D-$optimality criterion of Chaloner and Larntz (1989) is

$$\phi(d) = \int \log(|\mathrm{I}(\theta, \xi)|) d\pi(\theta), \tag{2.25}$$

where $\pi(\theta)$ is the prior distribution on θ.

Their algorithm can be run in a fully sequential mode, adding one new site at each step, or in a group-sequential mode, adding a fixed number of sites. The number of sites added are usually determined by practical issues in running the experiment and so is set by the user. The augmentation strategy also ensures that enough design points are used in order that the information matrices will be nonsingular. The implementation of the fully Bayesian approach is based on the posterior distribution of θ which is computed based on the data at hand. The exact posterior distribution which is used as a basis to find the next design point requires substantial computation at each iteration of the design. So they used an alternative approach. The posterior is represented by using a large (say, $N = 10,000$) discrete set of random vectors sampled from the prior, $\theta_1, \ldots, \theta_N$. The likelihood $L(\theta_u)$ for each of these vectors at any stage of the experiment is then computed and normal-

ized across the sample to generate weight $r_u = \frac{L(\theta_u)}{\sum_{v=1}^{N} L(\theta_v)}$. Functionals of the posterior are then estimated as weighted summaries of the vectors sampled from our prior. e.g. the posterior mean vector can be estimated as $\sum r_u \theta_u$. This is essentially an important sampling scheme, with the prior serving as the base sampling distribution and the important weights coming from the fact that the posterior provided by the prior is proportional to the likelihood.

The fully Bayesian approach for adding points to an existing design is implemented by using $\phi(d)$ from (2.25), averaging at each step with respect to the posterior distribution. This average is then approximated by using the criterion

$$\phi_1(d) = \sum_{u=1}^{N} r_u \log|\mathrm{I}(\theta_u, \xi)|. \tag{2.26}$$

Optimizing this criterion is not trivial.

They therefore suggested to replace the average by $\log|\mathrm{I}(\theta_u; \xi)|$ at a single point for an even faster computation. The posterior median for each of the parameters is evaluated and again the weighted representation of the posterior is used to estimate the median. This gives the criterion

$$\phi_2(d) = \log|\mathrm{I}(\tilde{\theta}, \xi)|, \tag{2.27}$$

where $\tilde{\theta}$ is the median of θ. The algorithm involves using the computationally fast approximate design criterion ϕ_2 to produce a limited set of candidate points. The better, but more computationally intensive ϕ_1 is then used to evaluate this small set. The algorithm also provides a simple fix for early stages in the experiment when the information matrix is singular. Since the singularity of the information matrix is a function only of the regression

matrix X, it is sufficient to check for the singularity at the posterior median.

2.4 Misspecified Models

For nonlinear models, most of the authors we have already mentioned and also Fedorov (1972), Ford and Silvey (1980) have explored the construction of optimal designs while assuming that the nonlinear model (including GLM) of interest is correctly specified. The expository article Ford et al. (1989) hinted that in the context of nonlinear models, as in the case of linear models, the misspecification of the model itself is of serious concern. They asserted that "indeed, if the model is seriously in doubt, the forms of design that we have considered may be completely inappropriate." Adewale and Wiens (2006) and Adewale and Wiens (2009) have developed criteria that generate robust designs and use such criteria for the construction of designs that insure against possible misspecification in the models. While Adewale and Wiens (2006) dealt with linear models, Adewale and Wiens (2009) discussed logistic models. We now present a summary of their work.

Suppose an experimenter is faced with a set $\Omega = \{x_j\}_{j=1}^{N}$ of possible design points from which he is interested in choosing n, not necessarily distinct, points at which to observe response Y. The experimenter makes $n_j \geq 0$ observations at x_j such that $\sum_{j=1}^{N} n_j = n$ The design problem is how to choose $n_1, \ldots, n_N$ in an optimal manner. Alternatively, the objective is to choose a probability distribution $\{p_j\}_{j=1}^{N}$ with $p_j = \frac{n_j}{n}$, on the design space Ω. the resulting design is said to be integer valued.

Adewale and Wiens (2006) considered the model:

$$Y_j = m(x_j) + \varepsilon_j \tag{2.28}$$

where $m(x, \theta) = E(Y|x) = z^T(x)\theta$ and $z^T(x)$ consists of p regressors $z(x) = (z_1(x), z_2(x), \ldots, z_p(x))^T$.

The experimenter believes that the mean response $E(Y|x)$ may be approximated by $z^T(x)\theta$ but since $E(Y|x) = z^T(x)\theta$ is just an approximation to the true model, the "best " θ_0 for predicting the mean response is defined to be the minimizer of the average-squared error of the approximation:

$$\theta_0 = \arg\min_t \frac{1}{N} \sum_{j=1}^{N} \left(E[Y|x_j] - z^T(x)t\right)^2. \tag{2.29}$$

We define $f(x) = E[Y|x] - z^T(x)\theta_0$, so that the model becomes

$$Y_{ij} = z^T(x)\theta_0 + f(x_i) + \varepsilon_{ij}, \quad i = 1, 2, \ldots, N \; j = 1, 2, \ldots, n_i \tag{2.30}$$

where ε_{ij} is the random error associated with the jth observation chosen at the ith design point and $\text{var}(\varepsilon_{ij}) = \sigma^2$.

From (2.29), Adewale and Wiens (2006) defined

$$\mathcal{F} = \left\{ f : \frac{1}{N} \sum_{i=1}^{N} z(x_i) f(x_i) = 0, \;\; \frac{1}{N} \sum_{i=1}^{N} f^2(x_i) \leq \tau^2 \right\} \tag{2.31}$$

as the class of contamination functions $f(x)$. The first condition in $\mathcal{F}$ says that f and z are orthogonal. The second condition is to ensure that the bias in the least-squares estimate $\hat{\theta}$ remains within bounds by placing a bound on the misspecification.

Adewale and Wiens (2006) defined the loss I as the average mean-squared error (amse) of $\hat{Y} = z^T(x)\hat{\theta}$ as the estimate of $E(Y|x)$:

$$\begin{aligned} L &= \frac{1}{N}\sum_{j=1}^{N} E\left\{\hat{Y}(x_i) - E[Y|x_j]\right\}^2 \qquad (2.32) \\ &= \frac{1}{N}\sum_{j=1}^{N}\left(E[\hat{Y}(x_j)] - z^T(x_i)\theta_0\right)^2 + \frac{1}{N}\sum_{i=1}^{N}\mathrm{var}[\hat{Y}(x_j)] + \frac{1}{N}\sum_{j=1}^{N} f^2(x_j). \end{aligned}$$

Fang and Wiens (2000) used a minimax approach to construct integer-valued designs. The optimal design in the minimax sense is the design that minimizes the maximum, over the misspecification neighborhood $\mathcal{F}$, value of the loss. The minimax approach aims to obtain the best design for the worst possible case of model misspecification. Adewale and Wiens (2006) introduce new criteria for robust designs which they claim may have more intuitive appeal to practitioners. Rather than minimizing the maximum loss they instead choose the design which minimizes the average value of the loss over the misspecification neighborhood. The averaging requires a parameterization of $\mathcal{F}$. This approach can be seen as a generalization of the approach employed by Läuter (1974) and Läuter (1976). While Läuter accommodated model uncertainty in the choice of design by averaging design criterion functions over a finite set of plausible models, Adewale and Wiens (2006) have an infinite set of plausible models as defined above. While Läuters criterion is based on variance only, in the spirit of Box and Lucas (1959), Adewale and Wiens (2006) based their design criteria on possible bias engendered by the model misspecification as well as on variance.

Given the misspecification neighborhood, Adewale and Wiens (2006) sought integer-valued designs that minimize the average (over $\mathcal{F}$) value of the loss. Let $\{pj = n_j/n\}_{j=1}^{N}$ be an integer-valued design on Ω, P the $N \times N$ diagonal

matrix with diagonal elements $\{p_j\}$, X the $N \times p$ matrix, assumed to be of full rank, with rows $z^T(x_1), \ldots, z^T(x_N)$. Define $f = (f(x_1), \ldots, f(x_N))^T$. In this notation, the amse defined in (2.32) can be written as

$$\begin{aligned} L &= \frac{1}{N}\Big\{\frac{\sigma^2}{n}\mathrm{tr}\left[(X^TPX)^{-1}X^TX\right] \\ &\quad + f^TPX(X^TPX)^{-1}X^TX(X^TPX)^{-1}X^TPf + f^Tf\Big\}. \end{aligned}$$

Adewale and Wiens (2006) noted that assuming the design is feasible for the full parameter vector θ, or equivalently that it has a minimum of p distinct support points x_i in Ω such that the vectors $z(x_i)$ are linearly independent. This implies the nonsingularity of X^TPX.

Averaging is carried out using the singular value decomposition $X = U_{N\times p}\Lambda_{p\times p}V^T_{p\times p}$, with $U^TU = V^TV = \mathrm{I}_p$ and Λ diagonal and invertible. U is augmented by $\tilde{U}_{N\times(N-p)}$ such that $[U : \tilde{U}]_{N\times N}$ is orthogonal. Then from (2.31), there is an $(N-p)\times 1$ vector c with $\|c\| \le 1$, satisfying $f(= f_c) = \tau\sqrt{N}\tilde{U}c$, and then

$$\begin{aligned} L &= \frac{1}{N}\Big\{\frac{\sigma^2}{n}\mathrm{tr}\left[(U^TPU)^{-1}\right] \\ &\quad + \tau^2 N\mathrm{tr}\left[\tilde{U}^TPU(U^TPU)^{-2}U^TP\tilde{U}cc^T\right] + \tau^2 Nc^Tc\Big\}. \end{aligned} \tag{2.33}$$

Fang and Wiens (2000) gives details of this development. Adewale and Wiens (2006) define their design criterion as I, with f integrated over c :

$$\begin{aligned} L_{\mathrm{ave}} &= \frac{\sigma^2}{nN}\mathrm{tr}\left[(U^TPU)^{-1}\right] \\ &\quad + \tau^2\int_{\|c\|\le 1}\Big(\mathrm{tr}\left[\tilde{U}^TPU(U^TPU)^{-2}U^TP\tilde{U}cc^T\right] + c^Tc\Big)dc. \end{aligned} \tag{2.34}$$

Adewale and Wiens (2006) hence formulated the following theorem:

Theorem 2.4.1. *Define*

$$\kappa_{N,p} = \frac{\pi^(N-p)/2}{((N-p)/2+1)\Gamma((N-p)/2} = \int\limits_{\|c\|\le 1} c^Tcdc.$$

The average of I, the amse over the misspecification neighborhood $\mathcal{F}$, is given by $I_{ave} = (\sigma^2/n + \tau^2\kappa_{N,p})\mathcal{L}_{ave}$, where

$$\mathcal{L}_{ave} = \rho\frac{tr\left[(U^TPU)^{-1}\right]}{N} + (1-\rho)\left(1 + \frac{tr\left[(U^TPU)^{-2}(U^TP^2U)\right]}{N-p}\right) \quad (2.35)$$

for $\rho = \sigma^2/n/(\sigma^2/n + \tau^2\kappa_{N,p})$.

Chapter 3

Robust and Efficient Designs

3.1 Introduction

Exponential regression models or Sigmoidal growth curves are widely used tools for analyzing data from processes arising in various fields such as biology, chemistry, pharmacokinetics or microbiology. Dette et al. (2006) and Dette and Pepelyshev (2008) mention a few examples. In microbiology these models are usually applied for describing growth and death of microorganisms, dose-response analysis and risk assessment (Coleman and Marks (2010)), and kinetics of metabolite production. These models are also incorporated in the numerous models in predictive microbiology for describing effects of temperature (Geeraerd et al. (2010)). Typical applications also include subject areas such as biology (see Lawdaw and DiStefano III (2010)), pharmacokinetics (see Liebig (1988) or Krug and Liebig (2010)) or toxicology (Becka et al. (1993); Becka and Urfer (1996)).

An appropriate choice of the experimental conditions can improve the quality of statistical inference substantially. The goal of an optimal or efficient

experimental design usually is the maximization of a real-valued function ϕ of the Fisher information matrix, or the minimization of the generalized inverse of this matrix. This is usually referred to as optimality criterion. There are numerous optimality criteria proposed in the literature to discriminate between competing designs. We restrict ourselves to the famous $D-$optimality criterion, where the determinant of the Fisher information is maximized by the design ξ, thus minimizing the (first order approximations of the) volume of the ellipsoid of concentration for the parameter θ. When a confidence ellipsoid for θ is constructed based on the asymptotic covariance matrix, its content is proportional to $[\det \mathrm{I}(\xi, \theta)]^{-1/p}$, which is minimized by a $D-$optimal design.

When a model has been specified, locally optimal designs which were proposed by Chernoff (1953) are the oldest and simplest to determine. When the model is nonlinear, the implementation of local optimal designs in practice requires a prior guess (nominal value) for the unknown parameter, which is rarely available in real experiments according to Dette et al. (2006), thus making practical implementation difficult. These nominal value typically comes from pilot studies, experts' opinion or related studies from the literature. Many authors including Melas (1978) and Han and Chaloner (2003) concentrate on locally optimal designs, where it is assumed that a preliminary guess for the unknown parameter is available (see Chernoff (1953); Silvey (1980)). A locally optimal design can be verified to be optimal using an equivalence theorem. Equivalence theorems are available when the design is a convex of the information matrix (Pukelsheim (1993)) and allows one to easily verify a design optimality by plotting the directional derivative of the criterion evaluated at that design over the design interval.

It is well known that locally optimal designs can depend on the prior guess or nominal value sensitively. This means that small misspecification in the nominal value can result in a very different optimal design. Consequently, a locally optimal design constructed under one set of nominal values can become inefficient when another set of nominal values is assumed.

To avoid this problem, several authors use a Bayesian approach to obtain robust designs (see Mukhopadhyay and Haines (1995); Dette and Neugebauer (1997) or Han and Chaloner (2003)). The Bayesian methodology requires the specification of a prior distribution for the nonlinear parameters in the models. Moreover, because statistical inference based on a local optimal design might be very sensitive with respect to a misspecification of this preliminary guess, as an alternative for the construction of robust designs, standardized maximin optimal designs were introduced by Dette (1995) and Müller and Pázman (1998) as another way to avoid the dependence on the guesses or nominal values. In the simplest case, they maximize the minimum of efficiencies that may arise from misspecification of the nominal values.

The method used by Dette and Pepelyshev (2008), which is based on the $D-$ optimality criterion, determines a design which maximizes a minimum of D-efficiencies (see also Müller (1995); Dette (1997); Imhof (2001)). Equivalently, minimax optimal designs seek to minimize the worst possible loss from misspecification of the nominal values. In either the minimax or maximin approach, we need to specify a plausible region for all possible values of the model parameters so that we may optimize within this region. This is usually accomplished by specifying a plausible interval (range) for each

unknown parameter of the model. We are motivated by the fact that in some cases practitioners will have difficulties to specify a single best guess or prior distribution for the unknown parameter, especially if this is multidimensional. Consequently, maximin or minimax optimal designs can be appealing in practice. However, according to Dette, the construction of minimax or maximin optimal design for nonlinear models is notoriously difficult and they defy analytical description, except for the simplest problems. Wong (1992) provided an overview of theoretical design issues for minimax optimality criteria and Dette (1995) provided yet another compelling rationale for use of such optimal designs in practice. Note that Bayesian and maximin are two different concepts. While the maximin approach addresses the worst case scenario by definition, Bayesian designs consider an average over the parameter space.

The maximin approach is started by assigning an index to each model parameter of interest to form the index set, say $J = \{1, 2, 3, \ldots, p\}$ if all the p model parameters are of interest. For a given set of nominal values, we define a standardized maximin optimal design as one that maximizes the minimum of efficiencies over the index set J. In practice, for a given set of parameters of interest, we first determine the locally optimal design for estimating each of the parameters in the index set J and the variances of all these parameter estimators. The standardized maximin optimal design sought is the one that provides th maximal minimum of efficiencies among a class of all designs on the design interval.

In many experiments, we may be constrained to use only a fixed maximal number of time points. This may arise because it is impractical to sample at

a new point or simply because of budget limits. This means that if we are only allowed s time points, then we must search within the class of designs with s points. We call the resulting design a $s-$point standardized maximin optimal design. Dette pointed out that such designs are typically easier to find numerically than the standardized maximin optimal designs.

The standardized maximin optimal design still depends on the nominal values. One may extend the above optimization by specifying a plausible interval for each parameter. A second maximin approach which is a clear natural extension of the first is used. The plausible region now comprises (i) the set J and (ii) the plausible interval for each parameter. The resulting optimal design is called a robust design because the design maximizes the minimum of the set of efficiencies of estimated parameters in the set J and, for each parameter, over each of its possible values in the plausible interval.

Even though, Dette et al. (2006) and Dette and Pepelyshev (2008) have considered some sigmoidal and exponential models using the maximin approach and found it to be very useful, not much attention has been paid to the problem of designing experiments for these models. We therefore wish to consider further models in this area.

Let us consider the nonlinear regression model

$$Y_j = m(x_j, \theta) + \varepsilon_j \quad j = 1, \ldots, n; \tag{3.1}$$

where $m(x_j, \theta) = E_\theta(Y_j | x_j = x)$, $\varepsilon_j \sim$ i.i.d. $N(0, \sigma^2)$, $x_j \in \Omega$ is explanatory variable, $\Omega \subset \mathbb{R}$ a compact design space, $\theta \in \Theta \subseteq \mathbb{R}^p$ an unknown parameter vector with p parameters.

Without loss of generality we let $\sigma^2 = 1$ and also assume that $m(x, \theta)$ is differentiable with respect to θ with continuous derivatives $g(x, \theta) = \frac{\partial}{\partial \theta} m(x, \theta) = (g_1(x, \theta), \ldots, g_p(x, \theta))$ for all $\theta \in \Theta$.

Definition 1. *Following Kiefer (1974), we define an (approximate) experimental design ξ with finite support $x_1, \ldots, x_n \in \Omega$, $x_i \neq x_j \, (i \neq j)$ and masses (weights) $w_1, \ldots, w_n > 0$, $\sum_{j=1}^{n} w_j = 1$ as a probability measure*

$$\xi = \begin{pmatrix} x_1 & \ldots & x_n \\ w_1 & \ldots & w_n \end{pmatrix}$$

on the interval or design space Ω.

The support points which are also referred to as design points give the locations where observations have to be taken, while the associated masses (weights) correspond to the relative proportions of the total observations to be taken at the particular points.

According to O'Brien (1995), the design problem for the nonlinear model (3.1) typically involves choosing an $n-$point design, ξ, to estimate some function of the above $p-$dimensional parameter vector, θ, with high efficiency. He stated that the design points, x_j are not necessarily distinct.

If the distribution of Y_j in (3.1) is normal, the matrix

$$\mathrm{I}(\xi, \theta) = \int_{\Omega} g(x, \theta) g^T(x, \theta) \mathrm{d}\xi(x)$$

is called the information matrix of the design ξ. If ξ puts masses $\frac{n_j}{n}$ at the points $x_j (j = 1, \ldots, n)$, then the experimenter takes observations, n_j at each x_j, and the information matrix is proportional to the asymptotic matrix of

the maximum likelihood estimator for θ.

Since the model $m(x,\theta)$ is nonlinear in the parameters, the information matrix $\mathrm{I}(\xi,\theta)$ which is usually a function of the unknown parameter θ and consequently an optimal design, maximizing (or minimizing) $\phi(\mathrm{I}(\xi,\theta))$ will depend on θ.

Now let the integrand of $\mathrm{I}(\xi,\theta)$ be given by the expression

$$F(x,\theta) = g(x,\theta)g^T(x,\theta),$$

where $g(x,\theta) = (g_1(x,\theta),\ldots,g_p(x,\theta))^T$ is the gradient of the regression function $m(x,\theta)$ with respect to θ. That is,

$$\begin{aligned} g(x,\theta) &= \frac{\partial}{\partial\theta}m(x,\theta) \\ &= \left(\frac{\partial}{\partial\theta_1}m(x,\theta),\ldots,\frac{\partial}{\partial\theta_p}m(x,\theta)\right)^T \\ &= (g_1(x,\theta),\ldots,g_p(x,\theta))^T. \end{aligned}$$

Now from our general regression model (3.1), we consider the following exponential regression models whose regression functions are as follows:

$$m(x,\theta) = \frac{\theta_3}{1+\theta_1 e^{\theta_2 x}} \tag{3.2}$$

with

$$g(x,\theta) = \left(-\frac{\theta_3 e^{\theta_2 x}}{(1+\theta_1 e^{\theta_2 x})^2}, -\frac{\theta_3\theta_1 x e^{\theta_2 x}}{(1+\theta_1 e^{\theta_2 x})^2}, \frac{1}{1+\theta_1 e^{\theta_2 x}}\right)^T;$$

$$m(x,\theta) = \frac{\theta_3}{1+\theta_1 e^{-\theta_2 x}} \tag{3.3}$$

with

$$g(x,\theta) = \left(-\frac{\theta_3 e^{-\theta_2 x}}{(1+\theta_1 e^{-\theta_2 x})^2}, \frac{\theta_3\theta_1 x e^{-\theta_2 x}}{(1+\theta_1 e^{-\theta_2 x})^2}, \frac{1}{1+\theta_1 e^{-\theta_2 x}}\right)^T;$$

$$m(x,\theta) = \theta_4 + \frac{\theta_3}{1+\theta_1 e^{\theta_2 x}} \tag{3.4}$$

with

$$g(x,\theta) = \left(-\frac{\theta_3 e^{\theta_2 x}}{(1+\theta_1 e^{\theta_2 x})^2}, -\frac{\theta_3 \theta_1 x e^{\theta_2 x}}{(1+\theta_1 e^{\theta_2 x})^2}, \frac{1}{1+\theta_1 e^{\theta_2 x}}, 1 \right)^T;$$

and

$$m(x,\theta) = \theta_4 + \frac{\theta_3}{1+\theta_1 e^{-\theta_2 x}} \tag{3.5}$$

with

$$g(x,\theta) = \left(-\frac{\theta_3 e^{-\theta_2 x}}{(1+\theta_1 e^{-\theta_2 x})^2}, \frac{\theta_3 \theta_1 x e^{-\theta_2 x}}{(1+\theta_1 e^{-\theta_2 x})^2}, \frac{1}{1+\theta_1 e^{-\theta_2 x}}, 1 \right)^T.$$

In many applications the systematic part of the response is known to be monotonic increasing in x. Nonlinear regression models with this property are called growth models. The simplest growth model is the exponential growth model $m(x,\theta) = \theta_1 e^{-\theta_2 x}$, but pure exponential growth is usually short-lived. A more generally useful growth curve is the logistic curve like (3.3) which produces a symmetric growth curve which asymptotes to θ_3 as $x \to \infty$ and to zero as $x \to -\infty$. Of the two other parameters, θ_1 determines horizontal position or 'take-off point', and θ_2 controls steepness.

3.1.1 Maximum Likelihood Estimation

If it is assumed that ε_j in (3.1) is normally distributed with mean zero and variance σ^2, that successive values of the stochastic term ε_j are independent and that the values for x are predetermined, then it is possible to write the log-likelihood function for Y_j, using (3.2), as

$$\log L = -\frac{\log(\sigma^2)}{2} - \log(2\pi) - \frac{1}{2\sigma^2}\varepsilon_j^2, \tag{3.6}$$

where

$$\varepsilon_j = Y_j - \frac{\theta_3}{1+\theta_1 e^{\theta_2 x}}.$$

Differentiating log L with respect to $\theta_1, \theta_2, \theta_3$ and σ^2 gives the following first partial derivative expression for each observation:

$$\frac{\partial \log L}{\partial \theta_1} = -\frac{1}{\sigma^2}\left(\varepsilon_j \theta_3 [1 + \theta_1 e^{\theta_2 x}]^{-2} e^{\theta_2 x}\right) \tag{3.7}$$

$$\frac{\partial \log L}{\partial \theta_2} = -\frac{1}{\sigma^2}\left(\varepsilon_j \theta_3 [1 + \theta_1 e^{\theta_2 x}]^{-2} x \theta_1 e^{\theta_2 x}\right) \tag{3.8}$$

$$\frac{\partial \log L}{\partial \theta_3} = \frac{1}{\sigma^2}\left(\varepsilon_j [1 + \theta_1 e^{\theta_2 x}]^{-1}\right) \tag{3.9}$$

$$\frac{\partial \log L}{\partial \sigma^2} = -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}\varepsilon_j^2. \tag{3.10}$$

Writing the maximum likelihood estimators of $\theta_1, \theta_2, \theta_3$ and σ^2 as $\tilde{\theta}_1, \tilde{\theta}_2, \tilde{\theta}_3$ and $\tilde{\sigma}^2$, it is evident that $\tilde{\theta}_1, \tilde{\theta}_2$ and $\tilde{\theta}_3$ can be derived from (3.7)-(3.9) independently of equation (3.10). These are the least squares equations and their solutions require numerical optimization. The properties of maximum likelihood estimation ensure that, in large samples, $\tilde{\theta}_1, \tilde{\theta}_2, \tilde{\theta}_3$ and $\tilde{\sigma}^2$ are normally distributed with mean ($\theta_1, \theta_2, \theta_3$ and σ^2) and a variance-covariance matrix found by differentiating equations (3.7)-(3.10) again. This double differentiation will produce 16 columns of derivatives, some of which will be identical in pairs, with the length of each column equaling the sample size. If the expected value for each observation in each column is taken and the totals from each column placed 4×4 matrix, then the negative of this matrix, when inverted, equals the asymptotic covariance matrix. The second partial derivatives are located down the diagonal and cross-partials off the diagonal in this 4×4 matrix.

3.2 Locally D-Optimal Designs

Optimal designs typically maximize some convex function of $\mathrm{I}(\xi, \theta)$ or minimize some convex function of $\mathrm{I}^{-1}(\xi, \theta)$. For example, designs which maximize

the determinant $|\mathrm{I}(\xi,\theta)|$ of $\mathrm{I}(x,\theta)$ are called $D-$optimal. The term "locally" is used to emphasize that the design is based on an initial estimate of the parameter vector θ.

We let ξ_θ^* denote a locally $D-$optimal design with respect to θ, i.e. a design which maximizes the determinant under the assumption that θ is the 'true' parameter. One measure of the "distance" between ξ and ξ_θ^* is $D-$efficiency. Pukelsheim (1993) and Atkinson (1992) defined the $D-$efficiency of ξ (with respect to the locally $D-$optimal design) as

$$\mathrm{eff}_D(\xi,\theta) = \left(\frac{\det\ \mathrm{I}(\xi,\theta)}{\det\ \mathrm{I}(\xi_\theta^*,\theta)}\right)^{\frac{1}{p}}. \tag{3.11}$$

To verify these locally $D-$optimal designs, we employ an analogue of Kiefer and Wolfowitz's General Equivalence Theorem (i.e. Theorem (2.1.1)) given by White (1973) for the nonlinear model.

3.2.1 Analogue of General Equivalence Theorem For The Nonlinear Model.

Let ξ be a member of the set Ξ of all measures and defined on the Borel field, $\mathcal{B}$ generated by the open sets of Ω and such that

$$\int_\Omega \xi(dx) = 1.$$

A design measure ξ^* is called $D-$optimum if

$$\det\{\mathrm{I}(\xi^*,\theta)\} = \max_{\xi\in\Xi} \det\{\mathrm{I}(\xi,\theta)\} \tag{3.12}$$

for θ taking its true value.

Let the variance function of $m(x,\theta)$ for the given ξ be given by

$$d(x,\xi,\theta) = g^T(x,\theta)\mathrm{I}^{-1}(x,\theta)g(x,\theta) \tag{3.13}$$

where $\mathrm{I}(\xi,\theta)$ is as usual nonsingular. A generalized inverse is used whenever $\mathrm{I}(\xi,\theta)$ is singular.

A design measure ξ^* is called $G-$optimum if

$$\sup_{x\in\Omega} d(x,\xi^*,\theta) = \min_{\xi\in\Xi} \sup_{x\in\Omega} d(x,\xi,\theta). \tag{3.14}$$

for θ taking its true value.

Theorem 3.2.1 (White (1973))**.** *The following conditions on a design measure ξ are equivalent:*

(i) ξ is $D-$optimum,

(ii) ξ is $G-$optimum,

(iii) $\sup_{x\in\Omega} d(x,\xi,\theta) = p.$

As in Kiefer and Wolfowitz's General Equivalence Theorem, (2.1.1), this analogous theorem (3.2.1) of White (1973) establishes the equivalence between locally $D-$optimal designs and $G-$optimal designs; which are those designs which minimize the maximum (over all $x \in \Omega$) of the variance function in (3.13). Also a corollary to this theorem states that the variance function in (3.13) evaluated using $D-$optimal design achieves its maximum value at the support points of this design.

Now Silvey (1980) gives the following important lemma which we shall later make use of to prove Theorem 3.2.2.

Lemma 3.2.1. *If $\Omega \in \mathbb{R}^p$ and spans $\mathbb{R}^p$, and if a $D-$optimal design measure is supported on p points, then it puts a probability of p^{-1} at each of them.*

Proof. For the linear case generally, Silvey (1980) states that if ξ is a design measure then

$$\mathrm{I}(\xi) = \int_\Omega z(x)z^T(x)\xi(dx) = \sum_{j=1}^{n} w_j z(x_j)z^T(x_j) = X^T D_\xi X \tag{3.15}$$

where X is the $n\times p$ matrix whose jth row is $z^T(x_j)$ and D_ξ is $\mathrm{diag}(w_1,\ldots,w_n)$. When $n = p$,

$$\det \mathrm{I}(\xi) = (\det X)^2 \prod_{j=1}^{p} w_j, \tag{3.16}$$

and for nonsingular X this is maximized, subject to $w_j \geq 0$ and $\sum w_j = 1$, by $w_j = n^{-1}, j = 1,\ldots,p$.

We now formulate an analogue of this proof for the nonlinear case:

According to O'Brien (1995), if we consider our nonlinear model (3.1), the information matrix is given by

$$\mathrm{I}(\xi,\theta) = V^T D_\xi V, \tag{3.17}$$

where V is the $n \times p$ Jacobian of m and D_ξ is $\mathrm{diag}(w_1,\ldots,w_n)$ as above. Hence, again when $n = p$,

$$\det \mathrm{I}(\xi,\theta) = (\det V)^2 \prod_{j=1}^{p} w_j, \tag{3.18}$$

and for nonsingular V this is maximized, subject to $w_j \geq 0$ and $\sum w_j = 1$, by $w_j = n^{-1}, j = 1,\ldots,p$. □

Now, we consider in detail model (3.5) where

$$m(x,\theta) = \theta_4 + \frac{\theta_3}{1+\theta_1 e^{-\theta_2 x}}.$$

This function corresponds to the output function of a feedforward neural network with one hidden neuron and activation function $\psi(u) = (1 + e^{-u})^{-1}$ as

$$\begin{aligned} m(x, \theta) &= v_0 + \frac{v_1}{1 + e^{-(w_{01} + w_{11} x)}} \\ &= v_0 + v_1 \psi(w_{01} + w_{11} x) \end{aligned}$$

with $\theta_4 = v_0, \theta_3 = v_1, \theta_2 = w_{11}$ and $\theta_1 = e^{-w_{01}}$. We observe that $\theta_1 > 0$ which we assume henceforth.

Moreover, the parameters are not identifiable. For example, we have

$$\psi(x) = \frac{1}{1 + e^{-x}} = 1 - \frac{1}{1 + e^{x}} = 1 - \psi(-x).$$

That is, parameters $\theta_1 = 1, \theta_2 = 1, \theta_3 = 1, \theta_4 = 0$ and $\theta_1 = 1, \theta_2 = 1, \theta_3 = -1, \theta_4 = 1$ give rise to the same function. Due to this property of the activation function: $\psi(-x) = 1 - \psi(x)$, we have in general that

$$m(x, \theta) = \theta_4 + \theta_3 - \frac{\theta_3}{1 + \frac{1}{\theta_1} e^{\theta_2 x}}.$$

That is, $(\theta_1, \theta_2, \theta_3, \theta_4)$ and $(\theta_1^{-1}, -\theta_2, \theta_3, \theta_3 + \theta_4)$ define the same function. To avoid this non-identifiability we assume henceforth that $\theta_2 > 0$, compare Rüger and Ossen (1997) for a discussion of that issue for general number of neurons.

Another popular activation function in neural network regression is the hyperbolic tangent, $\psi(u) = \tanh(u)$. The following results for the logistic activation function may be used more or less directly for that case too by exploiting the close relationship between $D-$optimal designs for the two cases given in Theorem 2 of Witczak (2006).

Now in Theorem 3.2.2, we present some results on locally $D-$optimal designs with respect to several parameter combinations and also give locally $D-$optimal designs on different design spaces for model (3.5). Obviously, analogous results hold for model (3.4). We remark the results are similar to Theorem 2.6 of Dette and Pepelyshev (2008) and use similar ideas for the proof.

Theorem 3.2.2. *Assume model (3.5) with a parameter set Θ chosen such that $\theta_1, \theta_2 > 0$.*

(a) The locally $D-$optimal design does not depend on θ_3 and θ_4. If we let $x_j(\theta_1, \theta_2, x_{max})$ denote a support of a locally $D-$optimal design on the interval $[0, x_{max}]$, then

$$x_j(\theta_1, r\theta_2, x_{max}) = \frac{1}{r} x_j(\theta_1, \theta_2, r x_{max})$$

for any $r > 0$. The weights of the locally $D-$optimal designs do not depend on the factor r.

(b) The locally D-optimal 4-point designs on the interval $[0, x_{max}]$ are uniquely determined and have equal masses at the four points $0 = x_1 < x_2 < x_3 < x_4 = x_{max}$.

(c) Any locally $D-$optimal design consisting of $k \geq 4$ points $x_1, \ldots, x_k$ includes the boundary points $x_1 = 0$ and $x_k = x_{max}$.

Proof. (a) Recalling

$$g(x, \theta) = \left(-\frac{\theta_3 e^{-\theta_2 x}}{(1 + \theta_1 e^{-\theta_2 x})^2}, \frac{\theta_3 \theta_1 x e^{-\theta_2 x}}{(1 + \theta_1 e^{-\theta_2 x})^2}, \frac{1}{1 + \theta_1 e^{-\theta_2 x}}, 1 \right)^T$$

of (3.5) we see straightforward that $g(x,\theta)g^T(x,\theta)$ as well as $\mathrm{I}(\xi,\theta)$ does not depend on θ_4. Hence, the locally $D-$optimal design does not depend on θ_4.

By the elementary properties of determinant, for any $n \times n$ matrix A and any scalar k, $|kA| = k^n|A|$. Hence, $|\mathrm{I}(\xi,\theta_1,\theta_2,\theta_3,\theta_4)| = \theta_3^4 |\mathrm{I}(\xi,\theta_1,\theta_2,1,1)|$. Therefore, the locally $D-$optimal designs do not depend on the parameters θ_3 and θ_4.

If we let $F(x,\theta_1,\theta_2) = g(x,\theta_1,\theta_2,1,1)g^T(x,\theta_1,\theta_2,1,1)$,
then

$$\begin{aligned} \det \int_0^{x_{\max}} F(x,\theta_1,\theta_2)\mathrm{d}\xi(x) &= \frac{1}{r^2}\det \int_0^{x_{\max}} F(rx,\theta_1,\theta_2)\mathrm{d}\xi(x) \quad (3.19) \\ &= \frac{1}{r^2}\det \int_0^{rx_{\max}} F(u,\theta_1,\theta_2)\mathrm{d}\xi(u/r). \end{aligned}$$

This identity is proved as follows for model (3.5).

First of all,

$$g(x,\theta_1,\theta_2,\theta_3,\theta_4) = \left(-\frac{\theta_3 e^{-\theta_2 x}}{(1+\theta_1 e^{-\theta_2 x})^2}, \frac{\theta_1\theta_3 x e^{-\theta_2 x}}{(1+\theta_1 e^{-\theta_2 x})^2}, \frac{1}{1+\theta_1 e^{-\theta_2 x}}, 1\right)^T,$$

$$g(x,\theta_1,r\theta_2,1,1) = \left(-\frac{e^{-r\theta_2 x}}{(1+\theta_1 e^{-r\theta_2 x})^2}, \frac{\theta_1 x e^{-r\theta_2 x}}{(1+\theta_1 e^{-r\theta_2 x})^2}, \frac{1}{1+\theta_1 e^{-r\theta_2 x}}, 1\right)^T,$$

$$g(rx,\theta_1,\theta_2,1,1) = \left(-\frac{e^{-r\theta_2 x}}{(1+\theta_1 e^{-r\theta_2 x})^2}, \frac{r\theta_1 x e^{-r\theta_2 x}}{(1+\theta_1 e^{-r\theta_2 x})^2}, \frac{1}{1+\theta_1 e^{-r\theta_2 x}}, 1\right)^T.$$

This implies,

$$g(x,\theta_1,r\theta_2,1,1) = Ag(rx,\theta_1,\theta_2,1,1),$$

where

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \frac{1}{r} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Therefore,

$$\int_0^{x_{\max}} F(x, \theta_1, r\theta_2, 1, 1)\mathrm{d}\xi(x) = A \int_0^{x_{\max}} F(rx, \theta_1, \theta_2, 1, 1)\mathrm{d}\xi(x)A^T.$$

This proves the first equality of (3.19), noting that $\det A = \frac{1}{r}$. From Dette and Pepelyshev (2008), the second equality in (3.19) is a direct consequence of the definition of the Stieltjes integral.

(b) Standard arguments of Silvey (1980), given in Lemma 3.2.1 above, show that the weights of a locally $D-$optimal design ξ consisting of 4 points $x_1 < \cdots < x_4$ have to be equal, i.e. $\frac{1}{4}$. Then, we have

$$\det \mathrm{I}(\xi, \theta) = \left(\frac{1}{4}\right)^4 \left[\tilde{\phi}(x_1, x_2, x_3, x_4)\right]^2$$

with

$$\tilde{\phi}(x_1, x_2, x_3, x_4) = \det\left(g(x_1, \theta), g(x_2, \theta), g(x_3, \theta), g(x_4, \theta)\right).$$

We use the same kind of arguments as the in the proofs of Lemma 2.4 and 2.5 of Dette and Pepelyshev (2008). First, we remark that the components of the vector

$$g(x, \theta) = \left(-\frac{\theta_3 e^{-\theta_2 x}}{(1 + \theta_1 e^{-\theta_2 x})^2}, \frac{\theta_3 \theta_1 x e^{-\theta_2 x}}{(1 + \theta_1 e^{-\theta_2 x})^2}, \frac{1}{1 + \theta_1 e^{-\theta_2 x}}, 1\right)^T$$

form a Chebyshev system; proven as Corollary 3.2.1 below. This implies that $\tilde{\phi}(x_1, x_2, x_3, x_4)$ does not vanish and, due to continuity, therefore always has the same sign for all $0 \leq x_1 < x_2 \leq x_3 < x_4 < x_{\max}$ (compare Zalik (1978)).

Hence, $\det \mathrm{I}(\xi, \theta)$ is always positive.

Now let us consider $\psi_1(x) = \tilde{\phi}(x, x_1, x_2, x_3)$ with fixed x_1, x_2, x_3. Since a determinant is linear in the values of the first column, we get, with ψ' denoting the derivative w.r.t x,

$$\psi_1'(x) = \det\left(g'(x,\theta), g(x_1,\theta), g(x_2,\theta), g(x_3,\theta)\right).$$

An elementary calculation shows that

$$g'(x,\theta) = Q(x,\theta)\left(\frac{\theta_3\theta_2}{\theta_1}q(x,\theta), \theta_3(1-\theta_2 x q(x,\theta)), \theta_2, 0\right)^T$$

with $Q(x,\theta) = \frac{\theta_1 e^{-\theta_2 x}}{(1+\theta_1 e^{-\theta_2 x})^2} > 0$ and $q(x,\theta) = \frac{1-\theta_1 e^{-\theta_2 x}}{1+\theta_1 e^{-\theta_2 x}}$.

Using these abbreviations, we also have

$$g(x_j,\theta) = \left(-\frac{\theta_3}{\theta_1}Q(x_j,\theta), \theta_3 x_j Q(x_j,\theta), \frac{1}{1+\theta_1 e^{-\theta_2 x_j}}, 1\right)^T, \quad j = 1,2,3.$$

Let

$$D_{1,3} = \det\begin{pmatrix} -\frac{\theta_3}{\theta_1}Q(x_1,\theta) & -\frac{\theta_3}{\theta_1}Q(x_2,\theta) & -\frac{\theta_3}{\theta_1}Q(x_3,\theta) \\ x_1\theta_3 Q(x_1,\theta) & x_2\theta_3 Q(x_2,\theta) & x_3\theta_3 Q(x_3,\theta) \\ 1 & 1 & 1 \end{pmatrix},$$

$$D_{1,2} = \det\begin{pmatrix} -\frac{\theta_3}{\theta_1}Q(x_1,\theta) & -\frac{\theta_3}{\theta_1}Q(x_2,\theta) & -\frac{\theta_3}{\theta_1}Q(x_3,\theta) \\ \frac{1}{1+\theta_1 e^{-\theta_2 x_1}} & \frac{1}{1+\theta_1 e^{-\theta_2 x_2}} & \frac{1}{1+\theta_1 e^{-\theta_2 x_3}} \\ 1 & 1 & 1 \end{pmatrix},$$

$$D_{1,1} = \det\begin{pmatrix} x_1\theta_3 Q(x_1,\theta) & x_2\theta_3 Q(x_2,\theta) & x_3\theta_3 Q(x_3,\theta) \\ \frac{1}{1+\theta_1 e^{-\theta_2 x_1}} & \frac{1}{1+\theta_1 e^{-\theta_2 x_2}} & \frac{1}{1+\theta_1 e^{-\theta_2 x_3}} \\ 1 & 1 & 1 \end{pmatrix}$$

be the determinants of the adjoints w.r.t. to the non-zero elements of the first column of ψ_1', such that

$$\begin{aligned}\frac{1}{Q(x,\theta)}\psi_1'(x) &= \frac{\theta_3\theta_2}{\theta_1}q(x,\theta)D_{1,1} - \theta_3(1-\theta_2 x q(x,\theta))D_{1,2} + \theta_2 D_{1,3} \\ &= \frac{\theta_3^2\theta_2}{\theta_1}q(x,\theta)D^*_{1,1} - \frac{\theta_3^2}{\theta_1}(1-\theta_2 x q(x,\theta))D^*_{1,2} + \frac{\theta_3^2\theta_2}{\theta_1}D^*_{1,3}\end{aligned}$$

where $D^*_{1,1}, D^*_{1,2}$ and $D^*_{1,3}$ do not depend on θ_3. A lengthy and tedious argument, using the Chebyshev property of the sets of functions $\{Q(x,\theta), xQ(x,\theta), 1\}, \{xQ(x,\theta), (1+\theta_1 e^{-\theta_2 x})^{-1}, 1\}$ and $\{Q(x,\theta), (1+\theta_1 e^{-\theta_2 x})^{-1}, 1\}$ which can be shown by the same arguments as for the full set in Corollary 3.2.1, shows that $\psi_1'(x) < 0$ for $0 \le x < x_1 < x_2 < x_3 \le x_{\max}$ i.e. $\psi_1(x)$ is decreasing in x.

Analogously, with

$$\psi_4(x) = \tilde{\phi}(x_1, x_2, x_3, x), \qquad 0 \le x_1 < x_2 < x_3 < x_4 \le x_{\max},$$

with fixed x_1, x_2, x_3, we get $\psi_4'(x) = -\psi_1'(x)$, and, hence, $\psi_4'(x) > 0$ for $0 \le x_1 < x_2 < x_3 < x \le x_{\max}$. i.e. $\psi_4(x)$ is increasing in x.

Consequently, the boundary points 0 and $x_{\max}$ are both part of the locally $D-$optimal 4-point design by the same argument as in the proof of Lemma 2.4 of Dette and Pepelyshev (2008).

(c) We use the same kind of arguments as in the proof of Lemma 2.5 of Dette and Pepelyshev (2008). We let ξ be the $k-$point design with weights $w_1, w_2, \ldots, w_k$ at the support points $x_1 < x_2 < \cdots < x_k$. Due to the Cauchy-Binet formula,

$$\det \mathrm{I}(\xi,\theta) = \sum_{1\le i_1<i_2<i_3<i_4\le k} w_{i_1}w_{i_2}w_{i_3}w_{i_4}\left[\tilde{\phi}(x_{i_1}, x_{i_2}, x_{i_3}, x_{i_4})\right]^2.$$

From part (b) of the proof, we know then that det $\mathrm{I}(\xi, \theta)$ is decreasing with respect to the smallest support point x_1 and increasing with respect to the largest x_k. Hence, any $D-$optimal design includes the boundary points 0 and x_{max}.

□

An important consequence of Theorem 3.2.2(a) is that it is not necessary to calculate locally $D-$optimal designs for all combinations of the parameters $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)$ and $x_{\max}$. In many cases locally $D-$optimal designs on different design spaces or with respect to a different specification of the parameters can easily be calculated by a non-linear transformation. For example if x_j of $D-$optimal design on the interval $[0, x_{\max}]$ are known, when $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)$ are known, the points $\frac{1}{r}x_j$ are the support points of the locally $D-$ optimal design on the interval $[0, rx_{\max}]$ when $\theta = (\theta_1, r\theta_2, \theta_3, \theta_4)$. Therefore, if the locally $D-$optimal designs for θ_1, θ_2 and $x_{\max}$ are known, then the locally optimal designs for any θ_1, θ_2 and any design space can easily be derived.

We later give some numerical results for the models (3.2)- (3.5) in Tables (3.3) - (3.6) in section 3.4.

Lemma 3.2.2. *For $\lambda \neq 0$, the functions $1, x, e^{-\lambda x}, e^{\lambda x}$ form a Chebyshev system on the interval $[0, x_{max}]$ for any $x_{max} > 0$, i.e. any linear combination*

$$\gamma(x) = \alpha_1 + \alpha_2 x + \alpha_3 e^{-\lambda x} + \alpha_4 e^{\lambda x}$$

has at most 3 roots in $[0, x_{max}]$ except for the trivial case of $\alpha_1 = \cdots = \alpha_4 = 0$.

Proof. We have to distinguish several cases.
i) $\alpha_3 = \alpha_4 = 0$. Here $\gamma(x) = \alpha_1 + \alpha_2 x$ has at most 1 root.

ii) α_3 and α_4 have the same sign. Then,

$$\gamma''(x) = \alpha_3 \lambda^2 e^{-\lambda x} + \alpha_4 \lambda^2 e^{\lambda x}$$

is either positive or negative for all $x \geq 0$, and, hence, $\gamma(x)$ is either convex or concave and has at most 2 roots.

iii) α_3 and α_4 have different signs. Then,

$$\gamma'''(x) = -\alpha_3 \lambda^3 e^{-\lambda x} + \alpha_4 \lambda^3 e^{\lambda x}$$

is either positive or negative for all x, i.e. $\gamma''(x)$ is either increasing or decreasing in $[0, x_{\max}]$ and has at most one root, say x_0, in that interval. So, $\gamma(x)$ is either convex or concave in $[0, x_{\max}]$ or convex on one side of x_0 and concave on the other side, and it can have at most 3 roots. □

This almost immediately implies the desired result that the coordinate functions of $g(x, \theta)$ form a Chebyshev system. We have to assume that $\theta_1, \theta_2, \theta_3 \neq 0$, since otherwise those functions would not be linearly independent. However, this only excludes the trivial cases where the regression function m would be constant but not genuine sigmoid.

Corollary 3.2.1. *For* $\theta_1, \theta_2, \theta_3 \neq 0$,

$$-\frac{\theta_3 e^{-\theta_2 x}}{(1+\theta_1 e^{-\theta_2 x})^2}, \frac{\theta_3 \theta_1 x e^{-\theta_2 x}}{(1+\theta_1 e^{-\theta_2 x})^2}, \frac{1}{1+\theta_1 e^{-\theta_2 x}}, 1$$

form a Chebyshev system on the interval $[0, x_{max}]$ *for any* $x_{max} > 0$.

Proof. We have to show that any non-trivial linear combination of the four functions has at most 3 roots in $[0, x_{\max}]$. By multiplying with the positive

factor $(1+\theta_1 e^{-\theta_2 x})^2$, this is equivalent to, including the non-vanishing factors $-\theta_3$ and $\theta_3\theta_1$ into the coefficients α_1 and α_2.

$$\begin{aligned} 0 &= \alpha_1 e^{-\theta_2 x} + \alpha_2 x e^{-\theta_2 x} + \alpha_3(1+\theta_1 e^{-\theta_2 x}) + \alpha_4(1+\theta_1 e^{-\theta_2 x})^2 \\ &= (\alpha_1 + \alpha_3\theta_1 + 2\alpha_4\theta_1)e^{-\theta_2 x} + \alpha_2 x e^{-\theta_2 x} + \alpha_4\theta_1^2 e^{-2\theta_2 x} + (\alpha_3+\alpha_4) \\ &= \beta_1 e^{-\theta_2 x} + \beta_2 x e^{-\theta_2 x} + \beta_3 e^{-2\theta_2 x} + \beta_4 \\ &= e^{-\theta_2 x}\left(\beta_1 + \beta_2 x + \beta_3 e^{-\theta_2 x} + \beta_4 e^{\theta_2 x}\right) \end{aligned}$$

for appropriately defined $\beta_1, \ldots, \beta_4$. As $e^{-\theta_2 x} > 0$ for all x, this can happen for at most 3 values of x in $[0, x_{\max}]$ by Lemma 3.2.2. □

Models (3.4) and (3.5) correspond to a regression function represented by a feedforward neural network with one neuron in the only hidden layer, where the activation function is the logistic one. We now consider the model

$$m(x,\theta) = \theta_7 + \frac{\theta_6}{1+\theta_4 e^{-\theta_5 x}} + \frac{\theta_3}{1+\theta_1 e^{-\theta_2 x}} \tag{3.20}$$

which corresponds to a feedforward neural network with two neurons in the hidden layer. For this function we get the gradient with respect to θ as

$$g(x,\theta) = \left(g_1^T(x,\theta), g_2^T(x,\theta), 1\right)^T$$

where

$$g_1(x,\theta) = \left(-\frac{\theta_3 e^{-\theta_2 x}}{(1+\theta_1 e^{-\theta_2 x})^2}, \frac{\theta_3\theta_1 x e^{-\theta_2 x}}{(1+\theta_1 e^{-\theta_2 x})^2}, \frac{1}{1+\theta_1 e^{-\theta_2 x}}\right)^T;$$

and

$$g_2(x,\theta) = \left(-\frac{\theta_6 e^{-\theta_5 x}}{(1+\theta_4 e^{-\theta_5 x})^2}, \frac{\theta_6\theta_4 x e^{-\theta_5 x}}{(1+\theta_4 e^{-\theta_5 x})^2}, \frac{1}{1+\theta_4 e^{-\theta_5 x}}\right)^T.$$

We conclude this section by proving an analogous version of part (a) of Theorem 3.2.2 for the more complicated model (3.20).

Theorem 3.2.3. *The locally $D-$optimal design in model (3.20) does not depend on θ_3, θ_6 and θ_7. If we let $x_j(\theta_1, \theta_2, \theta_4, \theta_5, x_{max})$ denote a support of a locally $D-$optimal design on the interval $[0, x_{max}]$, then*

$$x_j(\theta_1, r\theta_2, \theta_4, r\theta_5, x_{max}) = \frac{1}{r} x_j(\theta_1, \theta_2, \theta_4, \theta_5, rx_{max})$$

for any $r > 0$. The weights of the locally $D-$optimal design do not depend on the factor r.

Proof. (i) Since we have for model (3.20), $g(x, \theta) = \left(g_1^T(x, \theta), g_2^T(x, \theta), 1\right)^T$, the integrand of $\mathrm{I}(\xi, \theta)$ is of the form

$$g(x,\theta)g^T(x,\theta) = \begin{pmatrix} g_1(x,\theta)g_1^T(x,\theta) & g_1(x,\theta)g_2^T(x,\theta) & g_1(x,\theta) \\ g_2(x,\theta)g_1^T(x,\theta) & g_2(x,\theta)g_2^T(x,\theta) & g_2(x,\theta) \\ g_1^T(x,\theta) & g_2^T(x,\theta) & 1 \end{pmatrix}.$$

Recalling that

$$g_1(x,\theta) = \frac{1}{(1+\theta_1 e^{-\theta_2 x})^2}(-\theta_3 e^{-\theta_2 x}, \theta_3\theta_1 x e^{-\theta_2 x}, 1+\theta_1 e^{-\theta_2 x})^T$$

and

$$g_2(x,\theta) = \frac{1}{(1+\theta_4 e^{-\theta_5 x})^2}(-\theta_6 e^{-\theta_5 x}, \theta_6\theta_4 x e^{-\theta_5 x}, 1+\theta_4 e^{-\theta_5 x})^T,$$

we immediately see that $g(x, \theta)g^T(x, \theta)$, and by extension therefore $\mathrm{I}(\xi, \theta)$ does not depend on θ_7 at all. Hence, the locally $D-$optimal design does not depend on θ_7.

As a next step, we show that the optimal design ξ^* does not depend on θ_3 and θ_6 as well. For that purpose, let

$$F(x,\theta) = g(x, \theta_1, \theta_2, 1, \theta_4, \theta_5, 1, 1)g^T(x, \theta_1, \theta_2, 1, \theta_4, \theta_5, 1, 1)$$

which does not depend on θ_3, θ_6 and θ_7.

Using the particular form of $g_1(x,\theta)$ and $g_2(x,\theta)$ as parts of $g(x,\theta)$; and using the abbreviation $a = \theta_3$ and $b = \theta_6$; and setting

$$B = \begin{pmatrix} a^2 & a^2 & a & ab & ab & a & a \\ a^2 & a^2 & a & ab & ab & a & a \\ a & a & 1 & b & b & 1 & 1 \\ ab & ab & b & b^2 & b^2 & b & b \\ ab & ab & b & b^2 & b^2 & b & b \\ a & a & 1 & b & b & 1 & 1 \\ a & a & 1 & b & b & 1 & 1 \end{pmatrix},$$

we immediately get

$$g(x,\theta)g^T(x,\theta) = B \odot F(x,\theta)$$

where $\odot$ denotes the Hadamard product, i.e. the element wise product of two matrices. Since B does not depend on x, we also have

$$\mathrm{I}(\xi,\theta) = \mathrm{B} \odot \int \mathrm{F}(x,\theta)\mathrm{d}\xi(x). \tag{3.21}$$

An elementary, but lengthy and tedious calculation shows that for any permutation $(i_1,\ldots,i_7)$ of $(1,\ldots,7)$ we have

$$(B_{1i_1} \cdot \ldots \cdot B_{7i_7}) = a^4b^4. \tag{3.22}$$

This relationship can be checked by a MATLAB program which we put at subsection 3.2.2.

By the basic definition of the determinant of $\mathrm{I} = \mathrm{I}(\xi,\theta)$,

$$|\mathrm{I}| = \sum_{\Pi=(i_1,\ldots,i_7)} (\operatorname{sgn}\Pi)\mathrm{I}_{1i_1} \cdot \ldots \cdot \mathrm{I}_{7i_7} \tag{3.23}$$

where the summation runs over all permutations Π of $(1, \ldots, 7)$.

Using equations (3.21) and (3.22), we see that in each summand, a and b show up in the same factor a^4b^4. We therefore, finally have

$$|\mathrm{I}(\xi, \theta)| = \theta_3^4\theta_6^4 \left| \int F(x, \theta)\mathrm{d}\xi(x) \right|.$$

i.e. the locally $D-$optimal design ξ^* does not depend on θ_3 and θ_6 since $\mathrm{F}(x, \theta)$ depends only on $\theta_1, \theta_2, \theta_4$ and θ_5. This finishes the proof of the first part.

(ii) As in the proof of Theorem 3.2.2, we have

$$g(x, \theta_1, r\theta_2, 1, \theta_4, r\theta_5, 1, 1) = Ag(rx, \theta_1, \theta_2, 1, \theta_4, \theta_5, 1, 1)$$

where A is a diagonal matrix with diagonal entries $1, \frac{1}{r}, 1, 1, \frac{1}{r}, 1, 1$. This follows immediately from the explicit formulas for g given above. Therefore, we have

$$\int_0^{x_{\max}} F(x, \theta_1, r\theta_2, 1, \theta_4, r\theta_5, 1, 1)\mathrm{d}\xi(x) = A \int_0^{x_{\max}} F(rx, \theta)\mathrm{d}\xi(x)A^T.$$

Since $\det A = \frac{1}{r^2}$, we get

$$\begin{aligned} \det \int_0^{x_{\max}} F(x, \theta_1, r\theta_2, 1, \theta_4, r\theta_5, 1, 1)\mathrm{d}\xi(x) &= \frac{1}{r^4}\det \int_0^{x_{\max}} F(rx, \theta)\mathrm{d}\xi(x) \\ &= \frac{1}{r^4}\det \int_0^{rx_{\max}} F(u, \theta)\mathrm{d}\xi(u/r) \end{aligned}$$

where the last equality follows by substitution setting $u = rx$. The second assertion of the theorem relating locally $D-$ optimal designs on $[0, x_{\max}]$ and $[0, rx_{\max}]$ follows immediately. □

Unfortunately, we were not able to show the Chebyshev property of the 7 functions which form the coordinates of $g(x, \theta)$ in the model (3.20). Therefore, we cannot show the analogue of part (b) of the Theorem 3.2.2 though we believe it to be true.

3.2.2 MATLAB Program

```
function C = optdesnn2

% checks the claim in Proof of Theorem 3.2.3 that the matrix B satisfies
% the condition
% B(1,p(1))...B(7,p(7)) = a^4 * b^4 for all permutations p

    v = [ 1  2  3  4  5  6  7 ];

    P = perms(v); % 5040 x 7 matrix containing all permutations of v
    BE = [20 20 10 11 11 10 10;
          20 20 10 11 11 10 10;
          10 10  0  1  1  0  0;
          11 11  1  2  2  1  1;
          11 11  1  2  2  1  1;
          10 10  0  1  1  0  0;
          10 10  0  1  1  0  0];
% BE(i,j) = 10 * m+n if B(i,j) = a^m * b^n

% to show: S = BE(1,p(1)) · ... · BE(7,p(7)) = 44 for all permutations
% p
% 5040 x 1 vector for those sums over all those permutations: S

S = zeros(5040,1);

for   z= 1 : 5040;
p = P(z,:);
```

S(z) = BE(1, p(1)) ∗ BE(2, p(2)) ∗ BE(3, p(3)) ∗ BE(4, p(4)) ∗ BE(5, p(5)) ∗ BE(6, p(6)) ∗ BE(7, p(7));
end;
C = (S == 44); % vector with entry C(i)=1 if S(i)=44 and =0 else
% running the program results in sum(C)=5040, i.e. the claimed condition
% is true

3.3 Standardized Maximin $D-$optimal Designs

Clearly, the $D-$optimal designs in section (3.2) depend on the model parameters that we try to estimate and so they are locally optimal. To remove the dependence on the nominal values, Dette et al. (2006) introduced a concept of robust optimality criterion and define ξ^* as a standardized maximin $D-$optimal (with respect to Θ) if it maximizes the minimal (worst) $D-$efficiency calculated over a certain range for the parameter θ, thus protecting the experiment against the worst case scenario. That means that ξ^* maximizes (over ξ) the expression

$$\min_{\theta\in\Theta} \text{ eff}_D(\xi,\theta) = \min_{\theta\in\Theta} \left[\left(\frac{\det \text{I}(\xi,\theta)}{\det \text{I}(\xi^*_\theta,\theta)}\right)^{\frac{1}{p}}\right], \tag{3.24}$$

where the parameter space $\Theta \subset \mathbb{R}^p$ is a given set of possible (plausible) values for the unknown parameter θ which has to be specified in advance by the experimenter. In practice, the set Θ is a Cartesian product of the intervals specified for each parameter.

Following Dette and Pepelyshev (2008), we compute standardized maximin designs by maximizing the optimality criterion within the class of all $k-$point designs on the given design space. Here k is typically the minimal number of

points required for estimation of all parameters in the model. We employ the Nelder-Mead algorithm in the MATLAB package for optimization. After the optimal $k-$point standardized maximin design is found, we consider the class of all $k+1-$points designs and find an optimal design within this class and repeat the procedure. At each iteration, we increase the number of points by one, until no reduction in the criterion value is observed. The value of k for all our models was $k = 4$.

An advantage of this approach compared to the Bayesian set-up is that it is not required to specify a prior distribution for the unknown parameter θ, which is not possible in all circumstances. The only "prior knowledge" needed to use the standardized maximin $D-$optimality criterion is an approximate range Θ for the parameter θ.

Dette et al. (2006) noted that the optimality criterion (3.24) is not differentiable and as a consequence the problem of determining standardized maximin $D-$optimal designs is not trivial. This difficulty is also reflected in the following equivalence theorem for this type of optimality criterion which gives a characterization of standardized maximin $D-$optimal designs.

Theorem 3.3.1. *(Dette and Pepelyshev (2008)) A design ξ^* is standardized maximin $D-$optimal with respect to Θ if and only if there exists a probability distribution (prior) π^* supported on the set $\mathcal{N}(\xi^*) \subseteq \Theta$*

$$\mathcal{N}(\xi^*) = \left\{ \tilde{\theta} \in \Theta | eff_D(\xi^*, \tilde{\theta}) = \min_{\theta \in \Theta} \ eff_D(\xi^*, \theta) \right\} \tag{3.25}$$

such that the inequality

$$d(\xi^*, x) = \int_{\mathcal{N}(\xi^*)} g^T(x, \theta) I^{-1}(\xi^*, \theta) g(x, \theta) d\pi^*(\theta) \leq p \tag{3.26}$$

holds for all $x \in \Omega$, *where* $g(x,\theta) = (g_1(x,\theta), \ldots, g_p(x,\theta))^T$ *has previously been defined above. Moreover, there is an equality in (3.26) for all support points of the design* ξ^*.

The distribution π^* is called least favorable prior. The definition of the standardized maximin $D-$optimality criterion requires the knowledge of the local $D-$optimal design ξ_θ^*, or at least knowledge of the value of the optimal determinant $\det \mathrm{I}(\xi_\theta^*, \theta)$.

Obtaining the standardized maximin $D-$optimal designs is a lot more difficult than the local $D-$optimal design. This is due to the fact that the number of support points in the standardized maximin $D-$optimal designs are not necessarily equal to the number of parameters in the regression models. Once again if we consider the representations of the information matrices for the models (3.2)- (3.5), the $D-$efficiency (3.11) of our designs depend on only the parameters θ_1 and θ_2. We therefore use the notation $\mathrm{eff}_D(\xi, \theta_1, \theta_2)$ for the efficiency and

$$\min_{\theta_1,\theta_2 \in \Lambda} \mathrm{eff}_D(\xi, \theta_1, \theta_2)$$

for the optimality criterion (3.24), where Λ is an interval in the positive real line. i.e. $\Lambda = [[\theta_{11}, \theta_{12}] \in \theta_1, [\theta_{21}, \theta_{22}] \in \theta_2]$. The standardized maximin $D-$optimal design (for the set Λ) is denoted by ξ_Λ^*.

We use standardized maximin D-optimal designs in our models for similar reasons given by Dette and Pepelyshev (2008):

1. They are very efficient for a rather broad range of the non-linear parameters in the model.

2. They have approximately between $80 - 90\%$ D-efficiency.

3. They often advise the experimenter to take observations at a large number of different locations. For this reason these designs can also be used for testing the postulated models against models with more than four parameters by means of a goodness-of-fit test.

As we have already stated, Dette and Pepelyshev (2008) advocates the use of numerical methods in all cases of practical interest in determining the standardized maximin $D-$optimal designs since it is a very hard problem. For our numerical calculation we first considered the standardized maximin optimal 4-point designs. The optimality of the best 4-point designs was checked by the application of Theorem (3.3.1). If the optimality of the minimally supported design could be established, the procedure is terminated. Otherwise, we increase the number of support points and determine the standardized maximin optimal design within the class of all 5-point designs. This procedure is repeated until it terminates. This usually happens after a few steps. We considered standardized maximin $D-$optimal designs for models (3.4) and (3.5).

We present some results in Tables (3.7) and (3.8) in section 3.4.

3.4 Numerical Results and Discussion

To demonstrate the potential benefits of using maximin $D-$optimal designs for the analysis of our models, we re-design an experiment and re-analyze the data presented in Ratkowsky (1983b). The data which is listed in Table 3.1 shows the water content of bean root cells (Y) vrs the distance from tip (x).

x	Y
0.5	1.3
1.5	1.3
2.5	1.9
3.5	3.4
4.5	5.3
5.5	7.1
6.5	10.6
7.5	16.0
8.5	16.4
9.5	18.3
10.5	20.9
11.5	20.5
12.5	21.3
13.5	21.2
14.5	20.9

Table 3.1: The water content of been root cells (Y) versus the distance from tip (x).

We fitted model (3.5) to this data and obtained the following parameter estimates and their corresponding 95% confidence and results of goodness of fit.

Statistical Analysis of the data in Ratkowsky (1983b)

θ	Estimate	LCB	UCB	Goodness of fit	
θ_1	93.33	-7.537	194.2	SSE	5.19
θ_2	0.6977	0.5422	0.8532	R-square	0.9945
θ_3	20.4	18.7	22.09	Adj. R-square	0.993
θ_4	0.8845	-0.3293	2.098	RMSE	0.6869

Table 3.2: Parameter estimates, lower and upper confidence bounds (LCB & UCB), the sum-of-squares-error (SSE), (Adjusted) R-square values and the root-mean-square-error (RMSE).

The design used in the experiment is uniform design with 15 observations on the interval $[0.5, 14.5]$ while the maximin $D-$optimal design with respect to the intervals $[\theta_{11}, \theta_{12}] = [0.4, 0.8]$ and $[\theta_{21}, \theta_{22}] = [0.4, 0.8]$ is supported at only five points by

$$\begin{pmatrix} 0.5 & 1.3008 & 2.7496 & 5.1232 & 14.5 \\ 0.2448 & 0.1842 & 0.1656 & 0.1590 & 0.2464 \end{pmatrix} \tag{3.27}$$

and this design has a minimal $D-$efficiency of 91.92%. This makes the maximin $D-$optimal design cost effective and highly recommendable.

θ_1	θ_2	x_1	x_2	x_3
0.2	0.1	0	4.4488	10
0.2	1	0	1.1364	10
0.2	5	0	0.2274	9.5852
5	0.1	0	5.9663	10
5	1	0	2.4664	10
5	5	0	0.4938	9.7546
0.1	5	0	0.2142	9.4454
0.2	0.2	0	3.7802	10
2	2	0	0.9397	10

(a) Initial Design space $[1, 9]$.

θ_1	θ_2	x_1	x_2	x_3
0.2	0.1	0	4.4502	10
0.2	1	0	1.1366	10
0.2	5	0	0.2274	8.1607
5	0.1	0	6.0187	10
5	1	0	2.4665	10
5	5	0	0.4938	8.3335
0.1	5	0	0.2142	8.3283
0.2	0.2	0	3.7804	10
2	2	0	0.9397	10

(b) Initial Design space $[4, 7]$.

Table 3.3: Locally D-optimal designs for $m(x, \theta) = \frac{\theta_3}{1+\theta_1 e^{-\theta_2 x}}$ in space $[0, 10]$.

Table 3.3 shows locally $D-$optimal 3-point designs for model (3.3) on the interval $[0, 10]$ for various choices of parameters, θ_1 and θ_2. We remark that by Theorem 2.6 of Dette and Pepelyshev (2008), the design does not depend on θ_3.

We need a MATLAB program of Dette and Pepelyshev (2008) which requires the simplification of an initial interval. We choose $[1, 9]$ and $[4, 7]$ and it turns out that the choice does not have much influence on the final result. Nevertheless, a few of the results seem to correspond to the local optima of the target function. Therefore, working with various initial values and using the best final result may be desirable.

θ_1	θ_2	x_1	x_2	x_3	x_4
0.2	0.1	0	2.3006	6.6678	10
0.2	1	0	0.5451	1.9867	10
0.2	5	0	1.0062	9.0280	10
5	0.1	0	3.1344	7.5256	10
5	1	0	1.3461	3.1571	10
5	5	0	0.8188	9.8442	10
0.1	5	0	0.9998	9.0044	10
0.2	0.2	0	1.8712	5.9285	10
2	2	0	0.4729	1.3117	10

(a) Initial Design space $[1, 9]$.

θ_1	θ_2	x_1	x_2	x_3	x_4
0.2	0.1	0	2.3007	6.6679	10
0.2	1	0	0.5451	1.9867	10
0.2	5	0	4.0574	6.9996	10
5	0.1	0	3.1345	7.5256	10
5	1	0	1.3461	3.1571	10
5	5	0	4.0494	7.0694	10
0.1	5	0	4.0988	7.6554	10
0.2	0.2	0	1.8712	5.9285	10
2	2	0	0.4729	1.3117	10

(b) Initial Design space $[4, 7]$.

Table 3.4: Locally D-optimal designs for $m(x, \theta) = \theta_4 + \frac{\theta_3}{1+\theta_1 e^{-\theta_2 x}}$ in space $[0, 10]$.

θ_1	θ_2	x_1	x_2	x_3
0.2	0.1	0	5.2075	10
0.2	1	0	1.3482	3.1625
0.2	5	0	0.9688	9.4219
5	0.1	0	3.7204	10
5	1	0	0.5455	1.9885
5	5	0	1.0227	9.2642
0.1	5	0	0.8950	9.7753
0.2	0.2	0	4.9421	10
2	2	0	1.0671	0.3146

(a) Initial Design space $[1, 9]$.

θ_1	θ_2	x_1	x_2	x_3
0.2	0.1	0	5.2082	10
0.2	1	0	1.3482	3.1625
0.2	5	0	1.4562	9.7070
5	0.1	0	8.6898	10.6219
5	1	0	0.5455	1.9885
5	5	0	1.3939	9.5438
0.1	5	0	0.8950	9.7753
0.2	0.2	0	4.9439	10
2	2	0	1.0671	0.3146

(b) Initial Design space $[4, 7]$.

Table 3.5: Locally D-optimal designs for $m(x, \theta) = \frac{\theta_3}{1+\theta_1 e^{\theta_2 x}}$ in space $[0, 10]$.

θ_1	θ_2	x_1	x_2	x_3	x_4
0.2	0.1	0	3.1345	7.5255	10
0.2	1	0	1.3461	3.1571	10
0.2	5	0	1.0500	8.5501	10
5	0.1	0	2.3007	6.6679	10
5	1	0	0.5451	1.9867	10
5	5	0	0.9043	9.6610	10
0.1	5	0	1.0068	8.5496	10
0.2	0.2	0	3.2599	7.4676	10
2	2	0	0.3146	1.0671	10

(a) Initial Design space $[1, 9]$.

θ_1	θ_2	x_1	x_2	x_3	x_4
0.2	0.1	0	3.1345	7.5256	10
0.2	1	0	1.3461	3.1571	10
0.2	5	0	4.2344	7.1941	10
5	0.1	0	2.3006	6.6679	10
5	1	0	0.5451	1.9867	10
5	5	0	3.5624	7.5103	10
0.1	5	0	4.0024	7.1760	10
0.2	0.2	0	3.2598	7.4676	10
2	2	0	0.3146	1.0671	10

(b) Initial Design space $[4, 7]$.

Table 3.6: Locally D-optimal designs for $m(x, \theta) = \theta_4 + \frac{\theta_3}{1+\theta_1 e^{\theta_2 x}}$ in $[0, 10]$.

Tables (3.4)-(3.6) show the same kind of numerical results as Table (3.3), but now for 4-point designs for model (3.5), 3-point designs for model (3.2) and 4-point designs for model (3.4). We remark that the numerical minimization confirms the theoretical result of Theorem 3.2.2 where we have shown that the boundary points always belong to the optimal design. Also, the results strongly suggest that our models (3.2) and (3.3) are supported at only three points while models (3.4) and (3.5) are supported at only four points. The number of design points for the locally $D-$optimal designs coincide with the number of parameters in the respective models. We verify the optimality of these derived designs within the class of designs by using Theorem (3.2.1). We illustrate this with two examples in Figures (3.1) and (3.1) using particular choices of ($\theta_1 = 0.2, \theta_2 = 0.1$ and $\theta_1 = 0.2, \theta_2 = 1$) of parameters for the case of Table 3.6(a), taking the variance in equation (3.13) of Theorem 3.2.1. We see that the maxima are assumed at the 4 points of the optimal design.

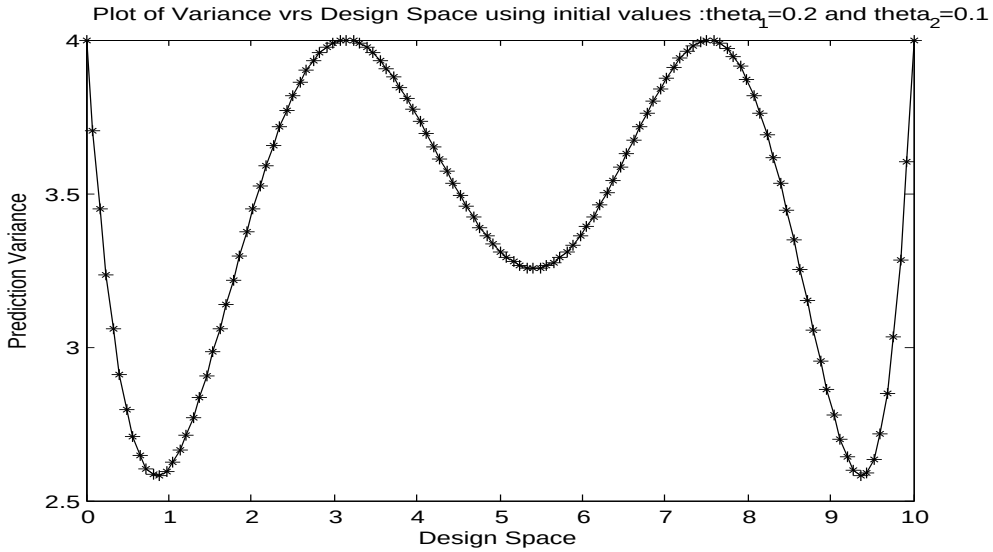

Figure 3.1: Plot of Variance vrs Design Space using initial values:$\theta_1 = 0.2$ and $\theta_2 = 0.1$.

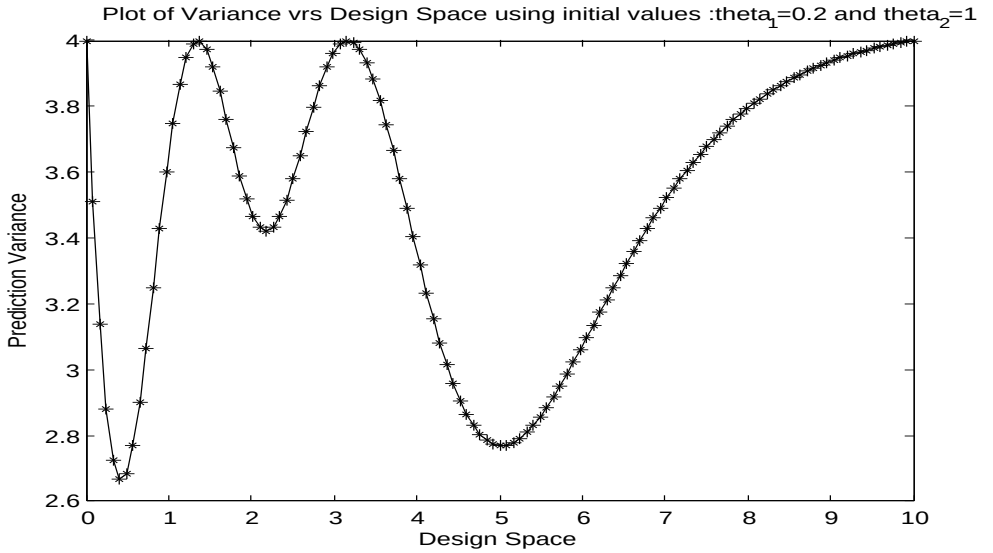

Figure 3.2: Plot of Variance vrs Design Space using initial values:$\theta_1 = 0.2$ and $\theta_2 = 1$.

θ_{11}	θ_{12}	θ_{21}	θ_{22}	x_1	x_2	x_3	x_4	x_5	w_1	w_2	w_3	w_4	w_5	min eff
0.8	1.2	0.8	1.2	0	0.7551	2.3585	10		0.25	0.25	0.25	0.25		0.9615
0.6	1.4	0.6	1.4	0	0.5730	1.5584	3.2930	10	0.2338	0.1776	0.1636	0.1538	0.2712	0.9073
0.4	1.6	0.4	1.6	0	0.6594	1.8174	4.3396	10	0.2428	0.1829	0.2097	0.1544	0.2103	0.8890
0.2	2.0	0.4	1.8	0	0.4632	1.7429	4.9499	10	0.2259	0.1549	0.2143	0.1430	0.2620	0.8072
0.1	1.2	0.5	1.9	0	0.6510	2.3868	5.8119	10	0.2482	0.1849	0.1380	0.1653	0.2637	0.8510
0.3	1.7	0.4	1.5	0	0.5841	1.8543	4.7041	10	0.2071	0.2053	0.1908	0.1798	0.2170	0.8598
0.5	1.5	0.5	1.5	0	0.6462	1.7427	3.8369	10	0.2527	0.1734	0.1937	0.1537	0.2266	0.9062
0.7	1.1	0.4	1.3	0	0.8134	2.5815	10		0.25	0.25	0.25	0.25		0.9545
0.9	1.1	0.9	1.1	0	0.7508	2.3359	10		0.25	0.25	0.25	0.25		0.9903
0.2	0.8	0.3	1.2	0	0.9677	2.8243	5.9427	10	0.2453	0.1792	0.2010	0.1577	0.2167	0.9131

Table 3.7: Maximin D-optimal designs for $m(x,\theta) = \theta_4 + \frac{\theta_3}{1+\theta_1 e^{-\theta_2 x}}$ in space $[0, 10]$.

θ_{11}	θ_{12}	θ_{21}	θ_{22}	x_1	x_2	x_3	x_4	x_5	w_1	w_2	w_3	w_4	w_5	min eff
0.8	1.2	0.8	1.2	0	0.7551	2.3585	10		0.25	0.25	0.25	0.25		0.9615
0.6	1.4	0.6	1.4	0	0.5730	1.5584	3.2930	10	0.2338	0.1776	0.1636	0.1538	0.2712	0.9073
0.4	1.6	0.4	1.6	0	0.6594	1.8174	4.3396	10	0.2428	0.1829	0.2097	0.1544	0.2103	0.8890
0.2	2.0	0.4	1.8	0	0.4562	1.6750	4.8324	10	0.1965	0.1645	0.2209	0.1623	0.2559	0.8144
0.1	1.2	0.5	1.9	0	0.3703	1.7743	4.9222	10	0.1997	0.1128	0.2999	0.1356	0.2520	0.7354
0.3	1.7	0.4	1.5	0	0.5841	1.8543	4.7041	10	0.2071	0.2053	0.1908	0.1798	0.2170	0.8598
0.5	1.5	0.5	1.5	0	0.6462	1.7427	3.8369	10	0.2527	0.1734	0.1937	0.1537	0.2266	0.9062
0.7	1.1	0.4	1.3	0	0.8651	2.6638	10		0.25	0.25	0.25	0.25		0.9525
0.9	1.1	0.9	1.1	0	0.7508	2.3359	10		0.25	0.25	0.25	0.25		0.9903
0.2	0.8	0.3	1.2	0	1.6690	4.8258	10		0.25	0.25	0.25	0.25		0.8145

Table 3.8: Maximin D-optimal designs for $m(x,\theta) = \theta_4 + \frac{\theta_3}{1+\theta_1 e^{\theta_2 x}}$ in space $[0, 10]$.

Tables (3.7) and (3.8) are concerned with the maximin $D-$optimal designs discussed in section 3.3. They don't depend on single values θ_1, θ_2 but rather on prior intervals for them, where $[\theta_{11}, \theta_{12}] \ni \theta_1$ and $[\theta_{21}, \theta_{22}] \ni \theta_2$ denote those ranges.

It turns out here that the 4-point designs do not always seem to be optimal but rather in the majority of cases we need 5 points. Though we did not prove that, it seems that still the boundary points 0 and 10 always belong to the support of the optimal design. In the 4-point designs, the optimal weights are equal just like for the locally $D-$optimal designs, but in the cases where 5-point designs are better, the weights differ. We get the same behavior for both models (3.4) and (3.5).

The last column of both tables gives the minimal $D-$efficiency defined in equation (3.11). As Dette and Pepelyshev (2008) already remarked for the simpler models (3.2) and (3.3), the minimal $D-$efficiency is pretty close to 1. So we do not lose much by the maximin approach.

Initial parameters				Design points and weights										EAMSE for various sample sizes (N)			
θ_{11}	θ_{12}	θ_{21}	θ_{22}	x_1	x_2	x_3	x_4	x_5	w_1	w_2	w_3	w_4	w_5	$N = 100$	$N = 250$	$N = 500$	$N = 1000$
0.8	1.2	0.8	1.2	0	0.7551	2.3585	10		0.25	0.25	0.25	0.25		0.0409	0.0195	0.0030	0.0012
0.7	1.1	0.4	1.3	0	0.8134	2.5815	10		0.25	0.25	0.25	0.25		0.0647	0.0512	0.0195	0.0115
0.2	0.8	0.3	1.2	0	0.9677	2.8243	5.9427	10	0.2453	0.1792	0.2010	0.1577	0.2167	0.1036	0.0874	0.0819	0.0662

Table 3.9: Expected AMSE values using maximin optimal designs when the assumed model and the data generating model are both $m(x, \theta) = \theta_4 + \frac{\theta_3}{1+\theta_1 e^{-\theta_2 x}}$.

Initial parameters				Design points and weights										Std dev. values for various sample sizes (N)			
θ_{11}	θ_{12}	θ_{21}	θ_{22}	x_1	x_2	x_3	x_4	x_5	w_1	w_2	w_3	w_4	w_5	$N = 100$	$N = 250$	$N = 500$	$N = 1000$
0.8	1.2	0.8	1.2	0	0.7551	2.3585	10		0.25	0.25	0.25	0.25		0.0807	0.0553	0.0158	0.0038
0.7	1.1	0.4	1.3	0	0.8134	2.5815	10		0.25	0.25	0.25	0.25		0.0944	0.0888	0.0558	0.0432
0.2	0.8	0.3	1.2	0	0.9677	2.8243	5.9427	10	0.2453	0.1792	0.2010	0.1577	0.2167	0.0889	0.0896	0.0904	0.0867

Table 3.10: Standard deviation values using maximin optimal designs when the assumed model and the data generating model are both $m(x, \theta) = \theta_4 + \frac{\theta_3}{1+\theta_1 e^{-\theta_2 x}}$.

Initial parameters				Design points and weights										EAMSE for various sample sizes (N)			
θ_{11}	θ_{12}	θ_{21}	θ_{22}	x_1	x_2	x_3	x_4	x_5	w_1	w_2	w_3	w_4	w_5	$N = 100$	$N = 250$	$N = 500$	$N = 1000$
0.8	1.2	0.8	1.2	0	0.7551	2.3585	10		0.25	0.25	0.25	0.25		0.2281	0.2326	0.2251	0.2272
0.7	1.1	0.4	1.3	0	0.8134	2.5815	10		0.25	0.25	0.25	0.25		0.2171	0.2230	0.2187	0.2335
0.2	0.8	0.3	1.2	0	0.9677	2.8243	5.9427	10	0.2453	0.1792	0.2010	0.1577	0.2167	0.1913	0.1836	0.1914	0.1828

Table 3.11: Expected AMSE values using maximin optimal designs when the assumed model is $m(x,\theta) = \theta_4 + \frac{\theta_3}{1+\theta_1 e^{-\theta_2 x}}$ and the data generating model is $m(x,\theta) = \theta_1 - \theta_2 e^{-\theta_3 x^{\theta_4}}$.

Initial parameters				Design points and weights										Std dev. values for various sample sizes (N)			
θ_{11}	θ_{12}	θ_{21}	θ_{22}	x_1	x_2	x_3	x_4	x_5	w_1	w_2	w_3	w_4	w_5	$N = 100$	$N = 250$	$N = 500$	$N = 1000$
0.8	1.2	0.8	1.2	0	0.7551	2.3585	10		0.25	0.25	0.25	0.25		0.1689	0.1617	0.1521	0.1441
0.7	1.1	0.4	1.3	0	0.8134	2.5815	10		0.25	0.25	0.25	0.25		0.1715	0.1652	0.1570	0.1501
0.2	0.8	0.3	1.2	0	0.9677	2.8243	5.9427	10	0.2453	0.1792	0.2010	0.1577	0.2167	0.1664	0.1650	0.1630	0.1555

Table 3.12: Standard deviation values using maximin optimal designs when the assumed model is $m(x,\theta) = \theta_4 + \frac{\theta_3}{1+\theta_1 e^{-\theta_2 x}}$ and the data generating model is $m(x,\theta) = \theta_1 - \theta_2 e^{-\theta_3 x^{\theta_4}}$.

Initial parameters		Design points and weights										AMSE for various sample sizes (N)			
θ_{31}	θ_{32}	x_1	x_2	x_3	x_4	x_5	w_1	w_2	w_3	w_4	w_5	$N = 100$	$N = 250$	$N = 500$	$N = 1000$
0.6	1.0	0	0.44	2.08	10		0.25	0.25	0.25	0.25		0.4650	0.3651	0.2858	0.2100
0.5	1.0	0	0.48	2.23	10		0.25	0.25	0.25	0.25		0.5629	0.4451	0.3634	0.2589
0.1	1.0	0	0.50	1.92	5.27	10	0.24	0.19	0.19	0.16	0.22	0.7962	0.6280	0.5609	0.4490

Table 3.13: Expected AMSE values using maximin optimal designs when the assumed model and the data generating model are both $m(x, \theta) = \theta_1 - \theta_2 e^{-\theta_3 x^{\theta_4}}$.

Initial parameters		Design points and weights										EAMSE for various sample sizes (N)			
θ_{31}	θ_{32}	x_1	x_2	x_3	x_4	x_5	w_1	w_2	w_3	w_4	w_5	$N = 100$	$N = 250$	$N = 500$	$N = 1000$
0.6	1.0	0	0.44	2.08	10		0.25	0.25	0.25	0.25		0.4061	0.3164	0.2749	0.1941
0.5	1.0	0	0.48	2.23	10		0.25	0.25	0.25	0.25		0.5960	0.4803	0.4061	0.2969
0.1	1.0	0	0.50	1.92	5.27	10	0.24	0.19	0.19	0.16	0.22	0.4832	0.3554	0.3014	0.2826

Table 3.14: Standard deviation values values using maximin optimal designs when the assumed model and the data generating model are both $m(x, \theta) = \theta_1 - \theta_2 e^{-\theta_3 x^{\theta_4}}$.

Initial parameters		Design points and weights										AMSE for various sample sizes (N)			
θ_{31}	θ_{32}	x_1	x_2	x_3	x_4	x_5	w_1	w_2	w_3	w_4	w_5	$N = 100$	$N = 250$	$N = 500$	$N = 1000$
0.6	1.0	0	0.44	2.08	10		0.25	0.25	0.25	0.25		3.2513	3.2922	1.9990	1.4438
0.5	1.0	0	0.48	2.23	10		0.25	0.25	0.25	0.25		1.3485	1.5270	1.1740	1.0979
0.1	1.0	0	0.50	1.92	5.27	10	0.24	0.19	0.19	0.16	0.22	1.9129	1.8467	1.5391	1.2053

Table 3.15: Expected AMSE values using maximin optimal designs when the assumed model is $m(x,\theta) = \theta_1 - \theta_2 e^{-\theta_3 x^{\theta_4}}$ and the data generating model is $m(x,\theta) = \theta_4 + \frac{\theta_3}{1+\theta_1 e^{-\theta_2 x}}$.

Initial parameters		Design points and weights										AMSE for various sample sizes (N)			
θ_{31}	θ_{32}	x_1	x_2	x_3	x_4	x_5	w_1	w_2	w_3	w_4	w_5	$N = 100$	$N = 250$	$N = 500$	$N = 1000$
0.6	1.0	0	0.44	2.08	10		0.25	0.25	0.25	0.25		2.9524	2.7060	2.0320	1.4766
0.5	1.0	0	0.48	2.23	10		0.25	0.25	0.25	0.25		1.6617	1.6888	1.5047	1.3627
0.1	1.0	0	0.50	1.92	5.27	10	0.24	0.19	0.19	0.16	0.22	1.9970	2.0457	1.8277	1.6985

Table 3.16: Standard deviation values using maximin optimal designs when the assumed model is $m(x,\theta) = \theta_1 - \theta_2 e^{-\theta_3 x^{\theta_4}}$ and the data generating model is $m(x,\theta) = \theta_4 + \frac{\theta_3}{1+\theta_1 e^{-\theta_2 x}}$.

Tables 3.9 - 3.16 study the effect of misspecification. We distinguish between the data generating model on one hand and the assumed model on the other hand. The assumed model is the basis for calculating the optimal design in the maximin sense. We use the root of the average mean squared error as a performance measure, which is used as an approximation of the integrated mean squared error.

Tables (3.9), (3.10), (3.13) and (3.14) correspond to correctly specified situations where we find the expected behavior. i.e. the error becomes smaller with sample size. we use our model (3.5) and, for comparison, a model with

$$m(x, \theta) = \theta_1 - \theta_2 e^{-\theta_3 x^{\theta_4}}$$

which has been discussed by Dette and Pepelyshev (2008). In the misspecified of Tables (3.11) and (3.12), the effect of assuming a wrong model is rather bad. The errors decrease slowly if at all with sample size N. Misspecification the other way round, shown in Tables (3.15) and (3.16) is relatively more well-balanced. Here the errors decrease with sample size with reasonable rate, but they are much larger than in their respective correctly specified cases.

What can we learn from this numerical study? If in doubt about the regression model, then choosing an optimal design for a wrong model may not be a good idea. In such cases, it would probably be better to use an equidistant design or another simple design which spreads the observations more or less homogeneously over the whole design space to catch the unknown shape of the regression function. An alternative would be to develop a theory for optimal design under misspecification which is still lacking. We do some first steps in that direction in the next chapter.

Chapter 4

Optimal Designs in Misspecified Models

Most papers on experimental design assume that the underlying model describes the data-generating process exactly. In this section, we want to start a discussion on how model misspecification may be taken into account.

To have a concrete situation in mind, let us assume that the given data $z_j = (Y_j, x_j)$; $z_1, z_2, \ldots, z_n$ i.i.d., are generated by the following correct (true) model

$$\text{true:} \qquad Y_j = m(x_j) + \varepsilon_j \tag{4.1}$$

with ε_j i.i.d. $(0, \sigma_\varepsilon^2)$, and independent of x_j; $j = 1, \ldots, n$. $m(x)$ is a completely arbitrary regression function.

The data are fitted with an assumed parametric model

$$\text{assumed:} \qquad Y_j = m(x_j, \theta) + \varepsilon_j \quad \text{for some } \theta \in \Theta. \tag{4.2}$$

We talk of misspecification if $m(x) \neq m(x, \theta)$ for all $\theta \in \Theta$. On the other hand, the model is correctly specified if $m(x) = m(x, \theta_0)$ for some $\theta_0 \in \Theta$.

If we are working with regression models based on feedforward neural networks, then we are typically confronted with misspecification, since the output functions of neural networks are usually only approximations for the unknown regression function $m(x)$. If the networks are large enough, then the approximations are good but not perfect.

In the correctly specified case, experimental design looks for optimal designs which in some sense allow for the most precise estimation of the true parameter value θ_0 from a given sample of size n. We have discussed such situations in the previous chapters. In a misspecified situation we do not have a true parameter. Hence, we have to ask what then should the goal of experimental design be. Given the assumed model, but allowing for misspecification, we formulate this goal in a rather general form which still guarantees enough freedom to look at various approaches by making the vague formulation *as much information as possible* precise and by choosing the method of fit, i.e. of estimating the parameter of the assumed model.

Goal: Choose the design $x_1, \ldots, x_n$, to get as much information as possible about the true regression function $m(x)$ from fitting the assumed regression function $m(x, \theta)$ to the data.

One possibility to make this goal precise is the following: We choose some distance D between the function $m(x)$ which we want to estimate and the function which we get from estimation based on the misspecified model where

the latter depends on the chosen design $\xi_n = (x_1, \ldots, x_n)$. Let $\hat{\theta}(\xi_n)$ denote the parameter estimate in model (4.2). Then, the estimated function is $\hat{m}_{\xi_n}(x) = m(x, \hat{\theta}(\xi_n))$. We can then formulate an optimal design as a solution to

$$ED(m, \hat{m}_{\xi_n}) = \min_{\xi_n}.$$

Before we continue our discussion, we give a short survey of estimation in misspecified models in the next two sections.

4.1 Consistency

If we estimate the parameter in model (4.2) by a maximum likelihood approach, but the model does not hold, then we are dealing with a quasi (or pseudo) maximum likelihood (QML) estimate which is a special case of an M-estimate.

Definition 2 (M-Estimates)**.** *Let $z_j = (Y_j, x_j)$; $j = 1, \ldots, n$ be given, $z = (z_1, \ldots, z_n)$, and let $Q_n : \mathbb{R}^{2n} \times \Theta \to \mathbb{R}$ be a measurable function. $\hat{\theta}_n$ is an M-estimate of a parameter θ of the distribution of z if*

$$\hat{\theta}_n = arg \min_{\theta \in \Theta} Q_n(z, \theta). \tag{4.3}$$

If in model (4.2), we pretend additionally that the residuals ε_j are Gaussian, we get as the QML estimate of θ

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{j=1}^{n} (Y_j - m(x_j, \theta))^2$$

which is an M-estimate with

$$Q_n(z, \theta) = \frac{1}{n} \sum_{j=1}^{n} q(z_j, \theta), \qquad q(z_j, \theta) = (Y_j - m(x_j, \theta))^2 .$$

In case of a **random design**, where $z_1, \ldots, z_n$ are i.i.d., we get under appropriate assumptions from a law of large numbers that

$$Q_n(z, \theta) \to E\ q(z_j, \theta) = E\ \left(Y_1 - m(x_1, \theta)\right)^2$$

and, by well-known standard arguments for M-estimates, which we will also use below in proving Theorem 4.1.2,

$$\hat{\theta}_n \to \theta_0 = \arg\min_{\theta \in \Theta} E\left(Y_1 - m(x_1, \theta)\right)^2 = \arg\min_{\theta \in \Theta} E\ q(z_1, \theta),$$

i.e. $\hat{\theta}_n$ is a consistent estimate of θ_0.

If we assume model (4.1) with ε_j independent of x_j and having mean 0 and finite variance σ_ε^2 we have

$$E\ \left(Y_1 - m(x_1, \theta)\right)^2 = E\ \left(m(x_1) - m(x_1, \theta)\right)^2 + \sigma_\varepsilon^2.$$

Therefore,

$$\theta_0 = \arg\min_{\theta \in \Theta} E\left(m(x_1) - m(x_1, \theta)\right)^2,$$

i.e. θ_0 minimizes the L^2-distance (w.r.t. the distribution of the x_j) between the functions $m(x)$ and $m(x, \theta)$. In this sense, θ_0 can be interpreted as the best parameter for approximating $m(x)$ by $m(x, \theta)$.

However, in experimental design, we want to choose the x_j by ourselves, i.e. we have to deal with a **deterministic design**. Therefore, we have to modify the standard consistency argument for M-estimates in the i.i.d. case and make appropriate assumptions. We need the following general result on the consistency of M-estimates (compare Theorem 5.7 of van der Vaart (1998)).

Theorem 4.1.1. *van der Vaart (1998)*

If for some deterministic function $q(\theta)$ of θ, $Q_n(z,\theta)$ satisfies

$$\sup_{\theta\in\Theta} |Q_n(z,\theta) - q(\theta)| \to 0 \quad (\textit{in probability}),$$

$$\inf_{\theta:\|\theta-\theta_0\|\geq\delta} q(\theta) > q(\theta_0) \quad \textit{for all } \delta > 0,$$

then a sequence of M-estimates $\hat{\theta}_n$ given by (4.3) converges in probability to θ_0.

Assuming model (4.1), we have for deterministic $x_1, \ldots, x_n$

$$\begin{aligned} Q_n(z,\theta) &= \frac{1}{n}\sum_{j=1}^{n}(Y_j - m(x_j,\theta))^2 \qquad (4.4)\\ &= \frac{1}{n}\sum_{j=1}^{n}(m(x_j) - m(x_j,\theta) + \varepsilon_j)^2 \\ &= \frac{1}{n}\sum_{j=1}^{n}(m(x_j) - m(x_j,\theta))^2 + \frac{2}{n}\sum_{j=1}^{n}\varepsilon_j(m(x_j) - m(x_j,\theta)) + \frac{1}{n}\sum_{j=1}^{n}\varepsilon_j^2. \end{aligned}$$

Since ε_j are i.i.d. with mean 0 and finite variance σ_ε^2, we have from the law of large numbers

$$\frac{1}{n}\sum_{j=1}^{n}\varepsilon_j^2 \to \mathrm{E}\,\varepsilon_j^2 = \sigma_\varepsilon^2 \quad \text{(in probability)}.$$

Let $\xi_n = (x_1, \ldots, x_n)$ denote the design as well as the empirical measure with respect to $x_1, \ldots, x_n$ such that we may write

$$\frac{1}{n}\sum_{j=1}^{n}(m(x_j) - m(x_j,\theta))^2 = \int (m(x) - m(x,\theta))^2 \xi_n(\mathrm{d}x)$$

Let us furthermore assume that $x_1, \ldots, x_n \in [a,b]$ for some finite interval.

The crucial assumption, which is rather common in experimental design, is about the limiting behavior of ξ_n for $n \to \infty$:

A1. There exists a probability measure ξ on $[a, b]$ such that

$$\xi_n \to \xi \quad \text{(weakly)}$$

If $m(x)$ and $m(x, \theta)$ are continuous in x on $[a, b]$, then we have

$$\int (m(x) - m(x, \theta))^2 \xi_n(\mathrm{d}x) \to \int (m(x) - m(x, \theta))^2 \xi(\mathrm{d}x) = e(\theta).$$

In order to apply Theorem 4.1.1, we need this convergence to be uniform in θ, and we need $e(\theta)$ to have a unique global minimum in θ_0 :

A2.

$$\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{j=1}^{n} (m(x_j) - m(x_j, \theta))^2 - \int (m(x) - m(x, \theta))^2 \xi(\mathrm{d}x) \right| \longrightarrow 0$$

in probability.

A3.

$$\inf_{\theta : \|\theta - \theta_0\| \geq \delta} e(\theta) > e(\theta_0) \quad \text{for all } \delta > 0.$$

We also need smoothness of the regression functions $m(x, \theta)$ as functions of the parameter:

A4. $m(x, \theta)$ is continuous in x and Lipschitz continuous in θ uniformly in $x \in [a, b]$.
That is

$$|m(x, \theta) - m(x, \eta)| \leq L \parallel \theta - \eta \parallel \quad \text{for all } x \in [a, b]; \quad \theta, \eta \in \Theta$$

for some Lipschitz constant L.

Theorem 4.1.2. *Assume model (4.1) with continuous $m(x)$ and i.i.d. ε_j having mean 0 and variance $\sigma_\varepsilon^2 < \infty$. Let $\hat{\theta}_n$ denote the Gaussian QML-estimate based on model (4.2), i.e. the M-estimate corresponding to*

$$Q_n(z,\theta) = \frac{1}{n}\sum_{j=1}^{n}(Y_j - m(x_j,\theta))^2.$$

Let the design $\xi_n = (x_1,\ldots,x_n)$ satisfy A.1, and let A2.-A4. be satisfied. Then, with

$$\begin{aligned} q(\theta) &= \int (m(x) - m(x,\theta))^2 \xi(\mathrm{d}x) + \sigma_\varepsilon^2 \\ &= e(\theta) + \sigma_\varepsilon^2, \end{aligned}$$

we have

$$\begin{aligned} \hat{\theta}_n \to \theta_0 &= \underset{\theta\in\Theta}{arg\ min}\ q(\theta) \\ &= \underset{\theta\in\Theta}{arg\ min}\ e(\theta). \end{aligned}$$

Proof. (a) We have to check the conditions of Theorem 4.1.1. A3. guarantees that the second assumption of that theorem holds. We only have to check the first one. From (4.4) and the triangular inequality, we have

$$\begin{aligned} \sup_{\theta\in\Theta}|Q_n(z,\theta) - q(\theta)| &\le \sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{j=1}^{n}\left(m(x_j) - m(x_j,\theta)\right)^2 - \int\left(m(x) - m(x,\theta)\right)^2\xi(\mathrm{d}x)\right| \\ &+ 2\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{j=1}^{n}\varepsilon_j\left(m(x_j) - m(x_j,\theta)\right)\right| \\ &+ \left|\frac{1}{n}\sum_{j=1}^{n}\varepsilon_j^2 - \sigma_\varepsilon^2\right| \end{aligned}$$

The first term converges to 0 by assumption A2., and also the last term by

the law of large numbers. For the second term, we have

$$\begin{aligned}\operatorname{var}\left(\frac{1}{n}\sum_{j=1}^{n}\varepsilon_j\left(m(x_j)-m(x_j,\theta)\right)\right) &= \frac{1}{n^2}\sum_{j=1}^{n}\operatorname{var}\left(\varepsilon_j\left(m(x_j)-m(x_j,\theta)\right)\right)\\ &= \frac{1}{n^2}\sum_{j=1}^{n}\left(m(x_j)-m(x_j,\theta)\right)^2\sigma_\varepsilon^2\\ &\sim \frac{1}{n}\int\left(m(x_j)-m(x_j,\theta)\right)^2\xi(\mathrm{d}x)\sigma_\varepsilon^2\longrightarrow 0\end{aligned}$$

using the independence of the ε_j and assumption A2.
Since $\mathrm{E}\,\varepsilon_j = 0$, and, hence, the mean of the second term is 0, we have

$$\frac{1}{n}\sum_{j=1}^{n}\varepsilon_j\left(m(x_j)-m(x_j,\theta)\right)\underset{p}{\longrightarrow}0.$$

However, we need the convergence to be uniform in θ, which we shall show in the second part of the proof. This will finish the proof of the 1st condition of Theorem 4.1.1.

b) We use the abbreviation $g(x,\theta) = m(x) - m(x,\theta)$. Since Θ is compact, we have for any $\triangle > 0$ a $K \geq 1$; $\theta_1, \ldots, \theta_K \in \Theta$ such that for any $\theta \in \Theta$ there is a $k \leq K$ with $\|\theta - \theta_k\| < \triangle$. Then, using that for arbitrary positive random variables $U, V, U_1, \ldots, U_K$ and for $\delta > 0$, we have

$$\begin{aligned}\operatorname{pr}\left(U+V>\delta\right) &\leq \operatorname{pr}\left(U>\frac{\delta}{2}\right)+\operatorname{pr}\left(V>\frac{\delta}{2}\right)\\ \operatorname{pr}\left(\sup_{k\leq K}U_k>\delta\right) &\leq \sum_{k=1}^{K}\operatorname{pr}\left(U_k>\delta\right).\end{aligned}$$

Restricting the suprema always to $\theta, \eta, \cdots \in \Theta$, we get

$$\operatorname{pr}\left(\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{j=1}^{n}\varepsilon_j\,g(x_j,\theta)\right|>\delta\right)=\operatorname{pr}\left(\sup_{k\leq K}\sup_{\|\theta-\theta_k\|<\triangle}\left|\frac{1}{n}\sum_{j=1}^{n}\varepsilon_j g(x_j,\theta)\right|>\delta\right)$$

$$= \text{pr}\left(\sup_{k\le K}\sup_{\|\theta-\theta_k\|<\triangle}\left|\frac{1}{n}\sum_{j=1}^{n}\varepsilon_j\left(g(x_j,\theta)-g(x_j,\theta_k)\right)+\frac{1}{n}\sum_{j=1}^{n}g(x_j,\theta_k)\right|>\delta\right) \tag{4.5}$$

$$\le \text{pr}\left(\sup_{\|\theta-\eta\|<\triangle}\left|\frac{1}{n}\sum_{j=1}^{n}\varepsilon_j\left(g(x_j,\theta)-g(x_j,\eta)\right)\right|>\frac{\delta}{2}\right)+\sum_{k=1}^{K}\text{pr}\left(\left|\frac{1}{n}\sum_{j=1}^{n}g(x_j,\theta_k)\right|>\frac{\delta}{2}\right). \tag{4.6}$$

For any given K, the second term can be made as small as we like by choosing n large enough since by the argument at the end of part a) of the proof,

$$\frac{1}{n}\sum_{j=1}^{n}g(x_j,\theta_k)\to 0 \quad \text{in probability.}$$

It remains to show that the first term of (4.6) becomes small if we choose $\triangle$ and K appropriately and let $n\to\infty$. We have, for $\|\theta-\eta\|<\triangle$, from assumption A4.,

$$\begin{aligned}\left|\frac{1}{n}\sum_{j=1}^{n}\varepsilon_j\left(g(x_j,\theta)-g(x_j,\eta)\right)\right| &= \left|\frac{1}{n}\sum_{j=1}^{n}\varepsilon_j\left(m(x_j,\eta)-m(x_j,\theta)\right)\right| \\ &\le \frac{1}{n}\sum_{j=1}^{n}|\varepsilon_j|L\triangle.\end{aligned}$$

Therefore, we have for the first term of (4.6),

$$\begin{aligned}&\text{pr}\left(\sup_{\|\theta-\eta\|<\triangle}\left|\tfrac{1}{n}\textstyle\sum_{j=1}^{n}\varepsilon_j\left(g(x_j,\theta)-g(x_j,\eta)\right)\right|>\tfrac{\delta}{2}\right)\\ &\qquad\le \text{pr}\left(\frac{1}{n}\sum_{j=1}^{n}|\varepsilon_j|>\frac{\delta}{2\triangle L}\right)\\ &\qquad= \text{pr}\left(\frac{1}{n}\sum_{j=1}^{n}|\varepsilon_j|-\text{E}\,|\varepsilon_j|>\frac{\delta}{2\triangle L}-\text{E}\,|\varepsilon_j|\right)\longrightarrow 0\end{aligned}$$

for $n\to\infty$ by the law of large numbers for $|\varepsilon_1|,|\varepsilon_2|,\ldots$ if we choose $\triangle$ small enough such that $\frac{\delta}{2\triangle L}>\text{E}\,|\varepsilon_j|$. Since δ may be chosen arbitrarily, the assertion follows. □

We remark that assumption A4. on the assumed regression model is not particularly strong. It is, for example, satisfied if $m(x,\theta)$ is continuously differentiable in θ with derivative which is bounded for $\theta \in \Theta$; $x \in [a,b]$. This holds, for example, for linear regression models where $m(x,\theta) = \theta_1 f_1(x) + \cdots + \theta_d f_d(x)$ with $f_1, \ldots, f_d$ bounded on $[a,b]$.

4.2 Asymptotic Normality

In this section, we limit our discussion to the case of a one-dimensional parameter θ to simplify notation. The general multi-dimensional case could be handled in exactly the same manner. $m'(x,\theta), \ldots$ denotes the partial derivative of $m(x,\theta), \ldots$ w.r.t. θ. Let θ_0 be defined as in Theorem 4.1.2. We assume

A5. $m(x,\theta)$ is twice continuously differentiable w.r.t θ for all $\theta \in \Theta$, $m'(x,\theta_0)$ is continuous in $x \in [a,b]$; and $m''(x,\theta)$ is Lipschitz continuous in θ uniformly in x, i.e. for some constant L

$$|m''(x,\theta_1) - m''(x,\theta_2)| \leq L \cdot |\theta_1 - \theta_2| \quad \text{for all } x.$$

Since θ_0 is the minimizer (point of minimum) of $q(\theta)$, we have

$$q'(\theta_0) = -2 \int (m(x) - m(x,\theta_0)) m'(x,\theta_0) \xi(dx) = 0. \tag{4.7}$$

We conclude from our assumptions on the design $\xi_n = (x_1, \ldots, x_n)$ that

$$\frac{1}{n} \sum_{j=1}^{n} (m(x_j) - m(x_j,\theta_0)) m'(x_j,\theta_0) \to \int (m(x) - m(x,\theta_0)) m'(x,\theta_0) \xi(dx) = 0.$$

For asymptotic normality, we need a certain rate of this convergence. Therefore, we assume

A6. $\frac{1}{n}\sum_{j=1}^{n}(m(x_j) - m(x_j,\theta_0))m'(x_j,\theta_0) = o(\frac{1}{\sqrt{n}})$,

This assumption is not too strong. Assume, for instance, that ξ_n is an equidistant design on the interval $[a,b]$ and that the integrand $(m(x) - m(x,\theta_0))m'(x,\theta_0)$ is Lipschitz continuous in x. Then, Kabajah concluded from a result of Wals and Sewell (1937) that A6. holds even with a rate $O(1/n)$ instead of $o(1/\sqrt{n})$ (compare Corollary 2.2 of Kabajah (2010)).

Similarly, we need

A7. For $a(x,\theta) = (m'(x,\theta))^2 - (m(x) - m(x,\theta))m''(x,\theta)$ we have uniformly in θ

$$\frac{1}{n}\sum_{j=1}^{n} a(x_j,\theta) \to \int a(x,\theta)\xi(dx) = A(\theta).$$

Theorem 4.2.1. *Let $\hat{\theta}_n$ and θ_0 be as in Theorem 4.1.2, and let the assumptions of that theorem be satisfied. Furthermore, assume A5., A6., A7., and that the third absolute moment of the residuals is finite: $E\ |\varepsilon_j|^3 = \gamma_\varepsilon < \infty$. Then,*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \underset{L}{\to} \mathcal{N}\left(0, \frac{B(\theta_0)}{A^2(\theta_0)}\right)$$

where $A(\theta)$ is given in A7., and

$$B(\theta) = \int |m'(x,\theta)|^2\xi(dx)$$

Proof. a) Since the M-estimate $\hat{\theta}_n$ is the minimizer of

$$Q_n(z,\theta) = \frac{1}{n}\sum_{j=1}^{n}(Y_j - m(x_j,\theta))^2,$$

we have $Q_n'(z,\hat{\theta}_n) = 0$ with

$$\begin{aligned} Q_n'(z,\theta) &= -\frac{2}{n}\sum_{j=1}^{n}(Y_j - m(x_j,\theta))m'(x_j,\theta) \\ &= -\frac{2}{n}\sum_{j=1}^{n}(m(x_j) - m(x_j,\theta))m'(x_j,\theta) - \frac{2}{n}\sum_{j=1}^{n}\varepsilon_j m'(x_j,\theta). \quad (4.8) \end{aligned}$$

Let us use the mean value theorem $f(a) = f(b) + f'(c)(a-b)$ for $c \in [a,b]$ with $f(\hat{\theta}_n) = Q_n'(z,\hat{\theta}_n), a = \hat{\theta}_n$ and $b = \theta_0$. Therefore linearizing around θ_0, we have

$$0 = Q_n'(z,\hat{\theta}_n) = Q_n'(z,\theta_0) + Q_n''(z,\theta_n^*)(\hat{\theta}_n - \theta_0)$$

for some $\theta_n^* \in [\hat{\theta}_n, \theta_0]$. From here we get

$$\hat{\theta}_n - \theta_0 = -\frac{Q_n'(z,\theta_0)}{Q_n''(z,\theta_n^*)}$$

and

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -\frac{\sqrt{n}Q_n'(z,\theta_0)}{Q_n''(z,\theta_n^*)}. \quad (4.9)$$

We now discuss the asymptotic behavior of the numerator and denominator of (4.9) separately.

b) We first have a look at

$$-\frac{1}{2}Q_n'(z,\theta_0) = \frac{1}{n}\sum_{j=1}^{n}(m(x_j) - m(x_j,\theta_0))m'(x_j,\theta_0) + \frac{1}{n}\sum_{j=1}^{n}\varepsilon_j m'(x_j,\theta_0),$$

compare (4.8). From our assumptions on the design and the regression functions $m(x)$ and $m(x,\theta_0)$, we have immediately that the first term on the right-hand side converges to

$$\int (m(x) - m(x,\theta_0))m'(x,\theta_0)\xi(dx) = 0$$

by (4.7). For the second term, we have from the same kind of argument and from independence of the residuals

$$\begin{aligned} n \operatorname{var}(\frac{1}{n}\sum_{j=1}^{n} \varepsilon_j m'(x_j,\theta_0)) &= \frac{1}{n}\sum_{j=1}^{n} \sigma_\varepsilon^2 (m'(x_j,\theta_0))^2 \\ &\to \sigma_\varepsilon^2 \int (m'(x,\theta_0))^2 \xi(dx) < \infty \end{aligned}$$

since $m'(x,\theta_0)$ is continuous and therefore bounded on the finite interval $[a,b]$. We conclude $Q_n'(z,\theta_0) \to 0$ in mean-square and, hence, in probability. Therefore, the asymptotic mean of $Q_n'(z,\theta_0)$ is 0.

Now, we want to show asymptotic normality of the numerator of (4.9). We want to apply Lyapunov's central limit theorem. Let $Z_{jn} = (Y_j - m(x_j,\theta_0))m'(x_j,\theta_0)$. The Z_{jn} are independent with mean and variance

$$\mu_{jn} = (m(x_j) - m(x_j,\theta_0))m'(x_j,\theta_0), \qquad \sigma_{jn}^2 = \sigma_\varepsilon^2 (m'(x_j,\theta_0))^2$$

and third moment

$$\gamma_{jn} = E\ |Z_{jn} - \mu_{jn}|^3 = E\ |\varepsilon_j m'(x_j,\theta_0)|^3 = \gamma_\varepsilon |m'(x_j,\theta_0)|^3.$$

We have to check the Lyapunov condition:

$$\begin{aligned} \rho_n &= \frac{\sum_{j=1}^n \gamma_{jn}}{\left(\sum_{j=1}^n \sigma_{jn}^2\right)^{\frac{3}{2}}} \\ &= \frac{\gamma_\varepsilon}{\sqrt{n}\sigma_\varepsilon^3} \frac{\frac{1}{n}\sum_{j=1}^n |m'(x_j,\theta_0)|^3}{(\frac{1}{n}\sum_{j=1}^n |m'(x_j,\theta_0)|^2)^{\frac{3}{2}}} \\ &\sim \frac{\gamma_\varepsilon}{\sqrt{n}\sigma_\varepsilon^3} \frac{\int |m'(x,\theta_0)|^3 \xi(dx)}{(\int |m'(x,\theta_0)|^2 \xi(dx))^{\frac{3}{2}}} \end{aligned}$$

Since the right-hand side converges to 0, the Lyapunov condition is satisfied, and we conclude that

$$\frac{\sum_{j=1}^n (Z_{jn} - \mu_{jn})}{\sqrt{\sum_{j=1}^n \sigma_{jn}^2}} \xrightarrow{L} \mathcal{N}(0,1)$$

in distribution. Using Slutsky's Lemma (Lemma 2.8. of van der Vaart (1998)) and

$$\frac{1}{n}\sum_{j=1}^{n}\sigma_{jn}^2 \to \int |m'(x,\theta_0)|^2\xi(dx) = B(\theta_0),$$

we get

$$\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\varepsilon_j m'(x,\theta_0) \xrightarrow{L} \mathcal{N}(0, B(\theta_0)),$$

and, applying assumption A6.,

$$\frac{1}{\sqrt{n}}\sum_{j=1}^{n}(Y_j - m(x_j,\theta_0))m'(x,\theta_0) = \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\varepsilon_j m'(x,\theta_0) + \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\mu_{jn} \xrightarrow{L} \mathcal{N}(0, B(\theta_0)).$$

Therefore, we finally have from (4.8)

$$-\sqrt{n}Q_n'(z,\theta_0) \xrightarrow{L} \mathcal{N}(0, 4B(\theta_0)). \tag{4.10}$$

c) Next let us consider the denominator in (4.9). For all θ,

$$\begin{aligned} Q_n''(z,\theta) &= -\left[\frac{2}{n}\sum_{j=1}^{n}(Y_j - m(x_j,\theta))m'(x_j,\theta)\right]' \\ &= -\frac{2}{n}\sum_{j=1}^{n}Y_j m''(x_j,\theta) + \frac{2}{n}\sum_{j=1}^{n}(m'(x_j,\theta))^2 + m(x_j,\theta)m''(x_j,\theta) \\ &= -\frac{2}{n}\sum_{j=1}^{n}\varepsilon_j m''(x_j,\theta) + \frac{2}{n}\sum_{j=1}^{n}a(x_j,\theta) \end{aligned} \tag{4.11}$$

with

$$a(x,\theta) = (m'(x,\theta))^2 - (m(x) - m(x,\theta))m''(x,\theta).$$

The first part of (4.11) converges to 0 in probability uniformly in θ by a uniform law of large numbers which we shall discuss below (compare Corollary 4.2.1). The second or deterministic part of (4.11) converges uniformly in θ to $2A(\theta)$ by assumption A7.

Since by definition, θ_n^* is a point between $\hat{\theta}_n$ and θ_0 and since from the

consistency result of Theorem 4.1.2, $\hat{\theta}_n \underset{p}{\to} \theta_0 \ (n \to \infty)$, we have $\theta_n^* \to \theta_0$ in probability.

Together we get, using the continuity of $A(\theta)$ which follows from assumption A5.,

$$Q_n''(z, \theta_n^*) \to 2A(\theta_0) \tag{4.12}$$

in probability. Combining (4.9),(4.10) and (4.12) and using Slutsky's Lemma again, we have

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \underset{L}{\to} \frac{\mathcal{N}(0, 4B(\theta_0))}{2A(\theta_0)} = \mathcal{N}\left(0, \frac{B(\theta_0)}{A^2(\theta_0)}\right).$$

□

In the previous proof we have used the following result which is a corollary of the uniform law of large numbers Theorem 3 of Andrews (1992).

Corollary 4.2.1. *Under the assumptions of Theorem 4.2.1*

$$\frac{1}{n}\sum_{j=1}^{n} m''(x_j, \theta)\varepsilon_j \to 0$$

uniformly in θ in probability

Proof. We prove this result by using Theorem 3 of Andrews (1992) and checking its assumptions. Let $V_{j,\theta} = m''(x_j, \theta)\varepsilon_j$ which are independent, but not identically distributed with common mean $EV_{j,\theta} = 0$. Our goal now is to show

$$\frac{1}{n}\sum_{j=1}^{n} V_{j,\theta} \to 0 \quad \text{uniformly in } \theta \in \Theta.$$

Boundedness (BD): This condition of Andrews (1992) follows immediately from the assumed compactness of Θ.

Pointwise Weak Law of Large Number (P-WLLN): As

$$\begin{aligned}\operatorname{var}\left(\frac{1}{n}\sum_{j=1}^{n} V_{j,\theta}\right) &= \frac{1}{n^2}\sum_{j=1}^{n}\operatorname{var}\ V_{j,\theta} \\ &= \frac{1}{n^2}\sum_{j=1}^{n}\left(m''(x_j,\theta)\right)^2\cdot\sigma_\varepsilon^2 \\ &\sim \frac{C(\theta)}{n}\sigma_\varepsilon^2 \underset{n\to\infty}{\longrightarrow} 0\end{aligned}$$

with

$$C(\theta) = \int \left(m''(x,\theta)\right)^2 \xi(dx).$$

As the $V_{j,\theta}$ have mean 0, we conclude a pointwise weak law of large numbers

$$\frac{1}{n}\sum_{j=1}^{n} V_{j,\theta} \underset{p}{\longrightarrow} 0.$$

Weak Lipschitz (W-LIP) Condition: From A5., we have

$$|m''(x_j,\theta^*)\varepsilon_j - m''(x_j,\theta)\varepsilon_j| \le |\varepsilon_j| L|\theta^* - \theta|,$$

and, as the ε_j are i.i.d., we have $\frac{1}{n}\sum_{j=1}^{n} E|\varepsilon_j| = E|\varepsilon_1|$, and, therefore, the condition $\sup_{n\ge 1}\frac{1}{n}\sum_{j=1}^{n} E|\varepsilon_j| < \infty$ is trivially fulfilled.

From Theorem 3(a) of Andrews (1992), *BD, P-WLLN* and *W-LIP* imply the uniform weak law of large numbers, i.e.

$$\sup_{\theta\in\Theta}\frac{1}{n}\sum_{j=1}^{n} V_{j,\theta} \underset{p}{\longrightarrow} 0$$

□

4.3 Forecasting in misspecified linear models

In this section, we consider only misspecified linear models with, for sake of simplicity, $[a, b] = [0, 1]$.

$$Y_j = m(x_j, \theta) + \varepsilon_j \qquad \theta \in \mathbb{R}^p, \;\; x_j \in [0, 1], \tag{4.13}$$

with, for some given vector of functions $f = (f_1, \ldots, f_p)^T$,

$$m(x_j, \theta) = \sum_{k=1}^{p} f_k(x_j)\theta_k = f^T(x_j)\theta.$$

4.3.1 The case of correct specification

When our model (4.13) is correctly specified, i.e. there is a true parameter θ_0 for which $m(x) = m(x, \theta_0)$, assuming ε_j is i.i.d. $\mathcal{N}(0, \sigma^2)$, the least squares estimate $\hat{\theta}_n$ of θ_0 equals the maximum likelihood estimate. Also,

$$\mathcal{L}\left(\sqrt{n}(\hat{\theta}_n - \theta_0)\right) = \mathcal{N}_p\left(0, \sigma_\varepsilon^2 (X^T X)^{-1}\right)$$

where $X = (f_j(x_i))_{i=1,\ldots,n, j=1,\ldots,p}$ is the $n \times p$ design matrix and $(X^T X)^{-1}$ is the covariance matrix. Therefore, the variability of the estimate $\hat{\theta}_n$ is determined by $(X^T X)^{-1}$ which, as a function of the design $\xi_n = (x_1, \ldots, x_n)$, should be small in an appropriate sense to obtain a good design. More precisely, for D-optimal designs, $\det(X^T X)^{-1}$ should be small or $\det(X^T X)$ should be large.

Instead of looking at the precision of the estimate $\hat{\theta}_n$, we could look at the performance of forecasts as a design criterion. Let us assume that we shall observe an additional pair (t, Y_t), and we are asked to forecast the observation Y_t given t. The best predictor of Y_t given t would be the expectation EY_t, but that depends on the unknown $m(t)$ which coincides with $m(t, \theta_0)$

in the correctly specified case. The latter is estimated by $m(t, \hat{\theta}_n)$ using only the already available $Y_1, \ldots, Y_n$. Therefore we predict Y_t by $\widehat{EY}_t = f^T(t)\hat{\theta}_n$.

Given the available data $\{(x_j, Y_j), j = 1, 2, \ldots, n\}$ the least squares estimate $\hat{\theta}_n$ of θ_0 is

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} \sum_{j=1}^{n} \left(Y_j - m(x_j, \theta)\right)^2 = (X^T X)^{-1} X^T Y \tag{4.14}$$

with $Y = (Y_1, \ldots, Y_n)^T$, which implies

$$E\hat{\theta}_n = (X^T X)^{-1} X^T EY$$

with $EY = (m(x_1), \ldots, m(x_n))^T$. Using the notation

$$\mu(\xi_n) = \frac{1}{n} X^T EY = \begin{pmatrix} \frac{1}{n}\sum_{j=1}^{n} f_1(x_j) m(x_j) \\ \vdots \\ \frac{1}{n}\sum_{j=1}^{n} f_p(x_j) m(x_j) \end{pmatrix} = \begin{pmatrix} \int f_1(x) m(x) \xi_n(dx) \\ \vdots \\ \int f_p(x) m(x) \xi_n(dx) \end{pmatrix}$$

and $\frac{1}{n} X^T X = \mathrm{I}(\xi_n)$, we can write this as

$$E\hat{\theta}_n = \mathrm{I}^{-1}(\xi_n)\mu(\xi_n), \tag{4.15}$$

where $\mathrm{I}(\xi_n)$ is a $p \times p$ information matrix.

For the estimate $\widehat{EY}_t = f^T(t)\hat{\theta}_n$ of EY_t we get correspondingly

$$\begin{aligned} \mathrm{var}\left(\widehat{EY}_t\right) &= \mathrm{var}\left(f^T(t)\hat{\theta}_n\right) = f^T(t)\mathrm{cov}(\hat{\theta}_n) f(t) \\ &= f^T(t)\sigma_\varepsilon^2 (X^T X)^{-1} f(t) = \frac{\sigma_\varepsilon^2}{n} f^T(t)\mathrm{I}^{-1}(\xi_n) f(t). \end{aligned} \tag{4.16}$$

In the correctly specified case, where $\hat{\theta}_n$ is an unbiased estimate of θ_0 and, therefore,

$$E(\widehat{EY}_t) = f^T(t) E\hat{\theta}_n = f^T(t)\theta_0 = EY_t,$$

the mean squared error (mse) of $\widehat{EY}_t$ is also given by

$$\mathrm{mse}\left(\widehat{EY}_t\right) = \frac{\sigma_\varepsilon^2}{n} f^T(t)\mathrm{I}^{-1}(\xi_n) f(t).$$

4.3.2 The case of misspecification

If the model is misspecified, i.e. where (4.1) holds, but $m(x) \neq m(x,\theta)$ for all θ, the covariance matrix of the data vector Y is still σ^2 times the identity matrix such that the covariance matrix of the least-squares estimate given by (4.14) is still $\sigma^2(X^TX)^{-1}$. Therefore, the variance of the forecast $\widehat{EY}_t$ is still of the form (4.16).

However, due to misspecification, a bias is introduced in the calculation of the mean-squared forecasting error mse $\left(\widehat{EY}_t\right)$.

$$\begin{aligned}\text{bias}\left(\widehat{EY}_t\right) &= E(f^T(t)\hat{\theta}_n) - m(t) \\ &= f^T(t)E\hat{\theta}_n - m(t) \\ &= f^T(t)\mathrm{I}^{-1}(\xi_n)\mu(\xi_n) - m(t)\end{aligned}$$

by (4.15). Therefore, we get for the mean-square forecasting error

$$\text{mse}\left(\widehat{EY}_t\right) = \frac{\sigma_\varepsilon^2}{n} f^T(t)\mathrm{I}^{-1}(\xi_n)f(t) + \left[f^T(t)\mathrm{I}^{-1}(\xi_n)\mu(\xi_n) - m(t)\right]^2$$

Example: We consider the case of a one-dimensional parameter ($p = 1$), where

$$Y_j = \theta f_1(x_j) + \varepsilon_j, \;\; \theta \in \mathbb{R}, \;\; X = \begin{pmatrix} f_1(x_1) \\ \vdots \\ f_1(x_n) \end{pmatrix},$$

$$X^TX = \sum_{j=1}^n f_1^2(x_j), \;\; \mathrm{I}(\xi_n) = \frac{1}{n}X^TX.$$

Therefore, the mean-square forecasting error is in this case

$$\text{mse}\left(\widehat{EY}_t\right) = \frac{\sigma_\varepsilon^2}{n}\frac{f_1^2(t)}{\frac{1}{n}\sum_{j=1}^n f_1^2(x_j)} + \left[\frac{\frac{1}{n}\sum_{j=1}^n f_1(x_j)m(x_j)}{\frac{1}{n}\sum_{j=1}^n f_1^2(x_j)} f_1(t) - m(t)\right]^2.$$

Suppose that θ_0 is the parameter vector for which $\theta_0 f_1(x)$ will be the best approximation of the function $m(x)$, i.e. from the results of section 4.1

$$\theta_0 = \arg\ \min_{\theta} \int (m(x) - f_1(x)\theta)^2\, \xi(dx).$$

By setting the derivative of the function to be minimized w.r.t. θ to 0, we get immediately

$$\theta_0 = \frac{\int f_1(x) m(x) \xi(dx)}{\int f_1^2(x) \xi(dx)}.$$

For the least-squares estimate, we have

$$\begin{aligned}
\hat{\theta}_n &= (X^T X)^{-1} X^T Y \\
&= \frac{1}{\sum_{j=1}^n f_1^2(x_j)} \sum_{j=1}^n f_1(x_j) Y_j \\
&= \frac{1}{\frac{1}{n}\sum_{j=1}^n f_1^2(x_j)} \left(\frac{1}{n} \sum_{j=1}^n f_1(x_j) Y_j \right) \\
&= \frac{1}{\int f_1^2(x) \xi_n(dx)} \left[\int f_1(x) m(x) \xi_n(dx) + \frac{1}{n} \sum_{j=1}^n f_1(x_j) \varepsilon_j \right].
\end{aligned}$$

For consistency, i.e. for $\hat{\theta}_n \to \theta_0$, we **only** need assumption A1 and continuity of $m(x), f_1(x)$ as functions of x, as then

$$\int f_1(x) m(x) \xi_n(dx) \to \int f_1(x) m(x) \xi(dx)$$

and

$$\int f_1^2(x) \xi_n(dx) \to \int f_1^2(x) \xi(dx),$$

and the latter also implies

$$\begin{aligned}
\operatorname{var}\left[\frac{1}{n} \sum_{j=1}^n f_1(x_j) \varepsilon_j \right] &= \frac{\sigma_\varepsilon^2}{n^2} \sum_{j=1}^n f_1^2(x_j) \\
&= \frac{\sigma_\varepsilon^2}{n} \int f_1^2(x) \xi_n(dx) \underset{n\to\infty}{\longrightarrow} 0,
\end{aligned}$$

and, therefore, using $\mathrm{E}\,\varepsilon_j = 0$ and Chebyshev's inequality

$$\frac{1}{n}\sum_{j=1}^{n} f_1(x_j)\varepsilon_j \to 0 \quad \text{(in probability)}.$$

Bibliography

A.J. Adewale and D.P. Wiens. New criteria for robust integer-valued designs in linear models. *Computational Statistics and Data Analysis*, 51:723–736, 2006.

A.J. Adewale and D.P. Wiens. Robust designs for misspecified logistic models. *Journal of Statistical Planning and Inference*, 139:3–15, 2009.

U. Anders. *Statistische Neuronale*. Vahlen, München, 1997.

D.W.K. Andrews. Generic uniform convergence. *Econometric Theory*, 8: 241–257, 1992.

A.C. Atkinson. *Optimum Experimental Designs*. Clarendon Press, 1992.

M. Bates, D and D.G. Watts. *Nonlinear Regression, Analysis and its Applications*. John Wiley, New York, 1988.

M. Becka and W. Urfer. Statistical aspects of inhalation toxicokinetics. *Environ. Ecol. Statist.*, 3:51–64, 1996.

M. Becka, H.M. Bolt, and W. Urfer. Statistical evaluation of toxicokinetic data. *Environmentrics*, 4:311–322, 1993.

G.E.P. Box and H.L. Lucas. Design of experiments in nonlinear situations. *Biometrika*, 46:77–90, 1959.

K. Chaloner and K. Larntz. Optimal bayesian design applied to logistic regression experiments. *J. Statist. Plann. Inference*, 21:191–208, 1989.

K. Chaloner and I. Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10:273–304, 1995.

P. Chaudhuri and P. Mykland. Nonlinear experiments : optimal design and inference based on likelihood. *J. Amer. Statist. Assoc.*, 88:538–546, 1993.

H. Chernoff. Locally optimal designs for estimating parameters. *The Annals of Mathematical Statistics*, 24:586–602, 1953.

M.H. Choueiki and C.A. Mount-Campbell. Training data development with $d-$optimality criterion. *IEEE Trans. Neural Networks*, 10:56–63, 1999.

W.G. Cochran. Experiments for nonlinear functions(r.a. fisher memorial lecture). *Journal of American Statistical Association*, 68:771–781, 1973.

D.A. Cohn. Neural network exploration using optimal experimental design. *Neural Networks*, 9:1071–1083, 1996.

M. Coleman and H. Marks. Topics in dose-response modeling. *J. Food Protection*, 61:1550–1559, 2010.

H. Dette. Designing experiments with respect to "standardized" optimality criteria. *Journal of Royal Statistical Society, Ser. B.*, 59:97–110, 1995.

H. Dette. Designing experiments with respect to 'standardized' optimality criteria. *J. Roy. Statist. Soc.*, 59, 1997.

H. Dette and H.-M. Neugebauer. Bayesian $d-$optimal designs for exponential regression models. *J. Statist. Plann. Inference*, 60:331–349, 1997.

H. Dette and A. Pepelyshev. Efficient experimental designs for sigmoidal growth models. *Journal of Statistical Planning and Inference*, 138:2–17, 2008.

H. Dette, L. Haines, and L. Imhof. Bayesian and maximin optimal designs for heteroscedastic regression models. *Canadian Journal of Statistics*, 33: 221–241, 2003.

H. Dette, I.M. Lopez, O. Rodriguez, and A. Pepelyshev. Maximin efficient design of experiment for exponential regression models. *Journal of Statistical Planning and Inference*, 136, 2006.

N.R. Draper and H. Smith. *Applied Regression Analysis*. John Wiley, New York, 2nd ed. edition, 1981.

H.A. Dror and D.M. Steinberg. Sequential experimental designs for generalized linear models. *Journal of the American Statistical Association*, 103: 288–298, 2008.

Z Fang and D.P Wiens. Integer-valued, minimax robust designs for estimation and extrapolation in heteroscedastic, approximately linear models. *J. Amer. Statist. Assoc.*, 95:807–818, 2000.

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

I. Ford and S.D. Silvey. A sequentially constructed design for estimating a nonlinear parametric function. *Biometrika*, 67:381–388, 1980.

I. Ford, D.M. Titterington, and C.P. Kitsos. Recent advances in nonlinear experimental design. *Technometrics*, 31:49–60, 1989.

A.H. Geeraerd, C.H. Herremans, and J.F. Van Impe. Structural model requirements to describe microbial inactivation during a mild heat treatment. *Internat. J. Food Microbiol.*, 59, 2010.

C. Han and K. Chaloner. D- and c-optimal designs for exponential regression models used in viral dynamics and other applications. *Journal of Staistical Planning and Inference*, 115, 2003.

H.O. Hartley. The modified gauss-newton method for the fitting of nonlinear regression functions by least squares. *Technometrics*, 3:269–280, 1961.

S. Haykin. *Neural Networks: A comprehensive foundation.* Prentice-Hall, 1999.

L.A. Imhof. Maximin designs for exponential growth models and heteroscedastic polynomial models. *Ann. Statist.*, 29:561–576, 2001.

H. Kabajah. *Local Smoothers with Regularization.* PhD thesis, Dept. of Mathematics, University of Kaiserslautern, 2010.

A.I. Khuri and J.A. Cornell. *Response Surfaces.* Marcel Dekker, New York, 1996.

J. Kiefer. Optimum experimental designs v, with applications to systematic and rotatable designs. In *Proceedings of the Fourth Berkeley Symposium Vol.1*, pages 381–405, 1961.

J. Kiefer and J. Wolfowitz. The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12:363–366, 1960.

J.C. Kiefer. General equivalence theory for optimum designs (approximate theory). *Ann. Statist.*, 2:849–879, 1974.

J. King and W.-K. Wong. Minimax $d-$optimal designs for the logistic model. *Biometrics*, 56:1263–1267, 2000.

H. Krug and H.-P. Liebig. Static regression models for planning greeenhouse production. *Acta. Horticulturae*, 230:427–433, 2010.

E. Läuter. Experimental design in a class of models. *Math. Operationsforschung Statist.*, 5:379–396, 1974.

E. Läuter. Optimal multipurpose designs for regression models. *Math. Operationsforschung Statist.*, 7:51–68, 1976.

E.W. Lawdaw and J.J. DiStefano III. Multiexponential multicompartmental and noncompartmental modeling. ii. data analysis and statistical considerations. *Amer. J. Physiol.*, 246:665–677, 2010.

W.H. Lawton and E.A. Sylvestre. Elimination of linear parameters in nonlinear regression. *Technometrics*, 13:461–467, 1971.

H.-P. Liebig. Temperature integration by kohlrabi growth. *Acta Horticulturae*, 230:371–380, 1988.

D.W. Marquardt. An algorithm for least squares estimation of nonlinear parameters. *J. Soc. Industrial Appl. Math*, 11:431–441, 1963.

V.B. Melas. Optimal designs for exponential regression. *Math. Operat. Forsch. Statist. Ser. Statist.*, 1978.

S. Mukhopadhyay and L.M. Haines. Bayesian $d-$ optimal designs for the exponential growth model. *J. Statist. Plann. Inference*, 44 (3):385–397, 1995.

C.H. Müller and A Pázman. Applications of necessary and sufficient conditions for maximum efficient design. *Metrika*, 48:1–19, 1998.

Ch. H. Müller. Maximin efficient designs for estimating nonlinear aspects in linear models. *J. Statist. Plann. Inference*, 44:117–132, 1995.

J.A. Nelder and R.W.M. Wedderburn. Generalized linear models. *Journal of Royal Statistical Society Ser. A*, 135:370–384, 1972.

T.E. O'Brien. Optimal design and lack of fit in nonlinear regression models. *Statistical Modelling*, pages 201–206, 1995.

R.L. Plackett and J.P. Burman. The design of optimal multifactorial experiments. *Biometrika*, 33, 1946.

F. Pukelsheim. *Optimal Design of Experiments*. Wiley, New York, 1993.

M.L. Ralston and R.I. Jennrich. Dud, a derivative-free algorithm for nonlinear least squares. *Technometrics*, 20:7–14, 1978.

D.A. Ratkowsky. *Nonlinear Regression Modelling*. Marcel Dekker, New York, 1983a.

D.A. Ratkowsky. *Nonlinear regression*. Dekker, 1983b.

S.M. Rüger and A. Ossen. The metric structure of weight space. *Neural Processing Letters*, 5:6372, 1997.

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

S. Sinha and D.P. Wiens. Robust sequential designs for nonlinear regression. *Canadian Journal of Statistics*, 30:601–618, 2002.

R.R. Sitter. Robust designs for binary data. *Biometrics*, 48:1145–1155, 1992.

A.W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, Cambridge, 1998.

J.L. Wals and W.E. Sewell. Note on degree of approximation to an integral by riemann. *The American Mathematical Monthly*, 44:155–160, 1937.

L.V. White. An extension of general equivalence theorem to nonlinear models. *Biometrika*, 60:345, 1973.

M. Witczak. Toward the training of feedforward neural networks with the $d-$optimum input sequence. *IEEE Trans. Neural Networks*, 17:357–373, 2006.

W.K. Wong. A unified approach to the construction of minimax designs. *Biometrika*, 79(3):611–619, 1992.

H.P. Wynn. The sequential generation of $d-$optimal experimental designs. *The Annals of Mathematical Statistic*, 41, No. 5:1655–1664, 1970.

R.A. Zalik. A characterization of tchebycheff systems. *J. Approx. Theory*, 22:356–359, 1978.