On Convergence of the Maximum Likelihood Estimator in Adaptive Designs

Dissertation

zur Erlangung des akademischen Grades

doctor rerum naturalium
(Dr. rer. nat.)

von Fritjof Freise

geb. am 20.03.1982 in Melle

genehmigt durch die Fakultät für Mathematik

der Otto-von-Guericke-Universität Magdeburg

Gutachter: Prof. Dr. Rainer Schwabe

Priv.-Doz. Dr. Jürgen Dippon

eingereicht am: 04.11.2015

Verteidigung am: 11.02.2016

Acknowledgments

First and foremost, I would like to thank my supervisor Prof. Dr. Rainer Schwabe. Without his support and encouragement, I would not have finished this work, and possibly not even thought about engaging a doctorate in the first place.

I am very grateful for my time at the Institute for Mathematical Stochastics and would like to thank all its members. In the past years it became a second home. Not because of hours and hours I have spent in the office, but because of the people, open doors and the informal atmosphere, which was and is a big support for me. This naturally includes former colleagues, like Dr. Tobias Mielke for example, and especially the heart and soul of the institute: Kerstin Altenkirch.

There are and have been several other people at the Faculty of Mathematics and the University, who influenced and supported me during the past years. Especially two of them I would like to mention here: Prof. Dr. Herbert Henning and Dr. Peter Dröse.

I am very grateful for the support of my family and friends, who have always been there: Thank you very much.

Last but not least, a most special thanks and my deepest gratitude, which cannot be put into words, goes to my dearest friend Nadja Malevich: Вялікі дзякуй!

Summary

In this thesis convergence of the maximum likelihood estimator for binary response models is considered, when the design is carried out adaptively. Adaptively means, that the choice of the next design point, i.e. the value of the control variable at which the next observation will take place, is based on prior observations. In our case the dependence is through the maximum likelihood estimate of the parameter in the model.

Since the methods used for independent observations cannot be easily generalized to this situation, the dependence of design and observations makes it difficult to analyze the behavior of both. The convergence of the estimator is not necessarily assured, not even the existence of a finite estimate. From the point of view of design also the question arises, whether the adaptive design is "optimal" in some sense. We will consider these problems and start with the search for conditions, under which the maximum likelihood estimator eventually exists and converges.

A natural way to tackle the dependence is to consider the sequence of estimators as a recursion. This can be studied using ordinary differential equations or related constructs, as has been exemplified for stochastic approximation algorithms. Therefore the trajectories of the estimator are split into a "mean part" and "perturbations", which might be deterministic or random. We show, that the mean behavior of the sequence of estimators can be described by solutions of ordinary differential equations. The limit points of the sequence follow from the corresponding asymptotic behavior of said solutions.

As an application an adaptive version of the "Wynn algorithm" is studied. In classical design theory the original nonadaptive version of this vertex direction method is known to converge to the optimal design. Results are presented concerning the convergence of the adaptive design and of the estimator when this design is used. Finally its properties are investigated in simulation studies.

Zusammenfassung

Diese Arbeit beschäftigt sich mit der fast sicheren Konvergenz des Maximum-Likelihood-Schätzers in Modellen mit binären Beobachtungen bei adaptiven Designs. Adaptives Design bedeutet hier, dass Designpunkte für neue Beobachtungen auf Grundlage früherer Schätzungen des Modellparameters bestimmt werden. Der adaptive Ansatz sorgt dafür, dass die Beobachtungen im allgemeinen nicht mehr als unabhängige Zufallsvariablen modelliert werden können und macht daher Untersuchungen des asymptotischen Verhaltens komplizierter. Bevor die Konvergenz des Schätzers betrachtet werden kann, stellt sich zudem die Frage, ob zumindest asymptotisch endliche Schätzungen existieren. Aus Sicht der Versuchsplanung ist zudem von Interesse, ob das entstehende Design „optimal“ ist.

Die rekursive Natur des Problems motiviert Methoden aus dem Bereich der Stochastischen Approximation und Stochastischen Algorithmen zu verwenden. Indem wir die Folge der Schätzer als Rekursion auffassen und Störterme abspalten, können wir zeigen, dass das mittlere Verhalten asymptotisch durch Lösungen gewöhnlicher Differentialgleichungen beschrieben werden kann. Somit folgt die Konvergenz des Schätzers aus der Asymptotik der Differentialgleichungen. Das Ergebnis sind Bedingungen an die Folge der Designpunkte, unter welchen die Folge der Schätzer konvergiert.

Ein spezielles Beispiel für die Wahl der Designpunkte ist eine adaptive Version des „Wynn-Algorithmus“. Dieser Algorithmus in seiner Grundform wird dazu verwendet optimale Designs zu bestimmen. Auf Grundlage der vorher erzielten Resultate wird dieses Verfahren untersucht und Ergebnisse zur Konvergenz des Designs sowie des Maximum-Likelihood-Schätzers präsentiert. Eine Simulationsstudie illustriert zum Abschluss diese Resultate.

Contents

1. Introduction

Planning an experiment can increase the accuracy of the results, reduces cost or saves time. So it is not astonishing, that articles on planning experiments were published already at the beginning of the 20th century. These were often inspired by agricultural questions. For a review with historical perspective on design in general see for example Atkinson and Bailey (2001).

If different plans are available to carry out an experiment, it is natural to ask the question: "Which is the best design?" While there were earlier publication by other researchers, e.g. by Elfving (1952) on linear 2-parameter models, major work was done by Kiefer (e.g. 1961, 1974). This included the original equivalence theorem by Kiefer and Wolfowitz (1960), which yields criteria to check for the optimality of a design.

Even though the theory for linear (fixed effect) models is well established, there is a crucial problem for nonlinear models, which motivated this thesis: The optimal designs depend on the value of the unknown parameter, since the information matrix usually does. For a fixed parameter value the linear theory can be generalized to what Chernoff (1953) called "locally optimal" designs. But because the actual value is not known, some prior knowledge is needed to fit a locally optimal design for the problem at hand.

If a set of possible values for the parameter is given, designs which give high information even for the worst choice of the parameter might be of interest. These are called maximin designs. In a similar way (pseudo-)Bayesian approaches introduce a weight function on the set of parameters to derive designs, which are optimal "on average". (for both approaches see e.g. Chapter 8 in Pronzato and Pázman, 2013, pp. 235)

The approach to circumvent the problem of parameter dependence chosen here is a sequential one. (see e.g. Silvey, 1980, pp. 62) After an initial phase, in which observations are obtained, the following steps are repeated until some stopping criterion is met:

- Estimate the parameter.
- Determine new design points.
- Take new observations.

The new design points added in the second step are chosen, to be locally optimal for the estimated parameter or at least would lead to a locally optimal design. The idea is now, that the estimate will be close to the actual value of the parameter after some iterations, and hence the new design points close to the locally optimal points for the actual parameter. So the two questions of interest here are:

- Does the sequence of estimators converge to the actual parameter?
- Does the sequence of designs converge to a locally optimal design?

Focusing on the sequential maximum likelihood estimation in binary response models, these questions were studied for this thesis.

Because the design depends on the estimate and vice versa, we cannot assume independent observations anymore. While we cannot use the corresponding theory, the dependence structure suggests the use of sequential methods.

For maximum likelihood estimation in location and scale families Ying and Wu (1997) formalized results of Wu (1985) concerning the estimation of the location parameter. Their paper motivated the approach of this thesis, because they apply methods from stochastic approximation, as will be done here. Started with the seminal paper of Robbins and Monro (1951) the method of stochastic approximation considers sequential problems, like root finding algorithms for stochastically perturbed functions or online estimation of parameters. Specifically, we will use the ordinary differential equation approach, where the mean behavior of the estimates is described by a differential equation. (see Ljung, 1977; Kushner and Yin, 2003; Kushner and Clark, 1978)

These models and maximum likelihood estimation therein are introduced in Chapter 2. This includes some thoughts about the existence of a global maximum. After a brief description of optimal design theory we introduce the Wynn algorithm, which is a vertex direction method to find optimal designs. It follows a review about results concerning the convergence of the estimator for adaptive designs. The chapter closes with an overview on stochastic approximation and the methods applied in the next chapter.

To ensure, that the log-likelihood eventually has a global maximum, is the goal of the first section of Chapter 3. After that the focus is on the almost sure convergence of the maximum likelihood estimator. The sequence of estimates is written recursively and split into a "mean part" and "perturbations". In a series of lemmas we find conditions, under which the "perturbations" are asymptotically negligible. Then the limit of the "mean part" is characterized.

In Chapter 4 we investigate an adaptive version of the Wynn algorithm. With conditions for almost sure convergence in place the question is, whether they hold here. We also consider asymptotic normality and if the design algorithm will be asymptotically optimal.

Chapter 5 presents the results of simulation studies for 2-parameter binary response models.

The main part of the thesis closes with some conclusions and an outlook in Chapter 6.

Supplementary material including (auxiliary) results, proofs and calculations for some examples is given in Appendix A and Appendix B.

2. The Model, Estimation and Related Topics

This section will introduce the basic notation, concepts and methods needed later. After notational conventions, the model will be described and results concerning estimation, optimal design and convergence of the estimator are summarized. These are illustrated by some examples. The chapter closes with a short review and description of stochastic approximation.

2.1. Basic Notations

Constants are usually denoted by the letters c, C or K. Column vectors or vector valued functions are denoted by bold, italic letters, e.g. $\boldsymbol{x}$, $\boldsymbol{\theta}$ or $\boldsymbol{f}$. In some cases capital letters will be used to distinguish between a random vector, e.g. $\boldsymbol{Y}$, and its realization $\boldsymbol{y}$. The j-th component of a vector $\boldsymbol{x}$ is denoted by x_j. Capital letters in a bold and upright font type are used for matrices or matrix valued functions, e.g. $\mathbf{I}$, $\mathbf{A}$ or $\mathbf{F}$. The $p \times p$ identity matrix is denoted by $\mathbf{E}_p$. As a shorthand notation for a $p \times p$ diagonal matrix with diagonal entries a_j, $j = 1, \ldots, p$, we will use $\mathrm{diag}_{j=1,\ldots,p}(a_j)$. Of special interest are nonnegative definite matrices. A matrix $\mathbf{A} \in \mathbb{R}^{p\times p}$ is called nonnegative definite, if

$$\boldsymbol{x}^\top \mathbf{A} \boldsymbol{x} \geq 0 \quad \text{for all} \quad \boldsymbol{x} \in \mathbb{R}^p \,.$$

A nonnegative definite matrix $\mathbf{A}$ is called positive definite, if

$$\boldsymbol{x}^\top \mathbf{A} \boldsymbol{x} > 0 \quad \text{for all} \quad \boldsymbol{x} \neq \mathbf{0}$$

and positive semidefinite, if additionally

$$\boldsymbol{x}^\top \mathbf{A} \boldsymbol{x} = 0 \quad \text{for some} \quad \boldsymbol{x} \neq \mathbf{0} \,.$$

For two matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{p\times p}$ we write $\mathbf{A} \leq \mathbf{B}$ if $\mathbf{B} - \mathbf{A}$ is nonnegative definite.

If not mentioned otherwise we will use the euclidean norm $\|\boldsymbol{x}\| := \sqrt{\boldsymbol{x}^\top \boldsymbol{x}}$ for a vector $\boldsymbol{x} \in \mathbb{R}^p$. The corresponding induced matrix norm for a real matrix $\mathbf{A} \in \mathbb{R}^{p\times p}$ is then defined as

$$\|\mathbf{A}\| := \sup_{\|\boldsymbol{x}\|\neq 0} \frac{\|\mathbf{A}\boldsymbol{x}\|}{\|\boldsymbol{x}\|} = \sup_{\|\boldsymbol{x}\|\neq 0} \frac{\sqrt{\boldsymbol{x}^\top \mathbf{A}^\top \mathbf{A} \boldsymbol{x}}}{\sqrt{\boldsymbol{x}^\top \boldsymbol{x}}} = \sqrt{\lambda_{\max}(\mathbf{A}^\top \mathbf{A})} \,,$$

where $\lambda_{\max}(\mathbf{A})$ denotes the largest eigenvalue of $\mathbf{A}$. Similarly $\lambda_{\min}(\mathbf{A})$ is its smallest eigenvalue. If other eigenvalues are needed, we will use the notation often used for order statistics: $\lambda_{\min} = \lambda_{(1)} \leq \lambda_{(2)} \leq \ldots \leq \lambda_{(p)} = \lambda_{\max}$. Trace and determinant of a matrix $\mathbf{A}$ are denoted by $\mathrm{tr}(\mathbf{A})$ and $\det(\mathbf{A})$, respectively.

The indicator function of a set $\mathcal{A}$ is denoted by $\mathbb{1}_{\mathcal{A}}$. A subset is denoted using the symbol $\subseteq$. A proper subset by $\subset$. The cardinality of a set $\mathcal{A}$ is denoted by $|\mathcal{A}|$.

Let $Y, X_1, X_2, \ldots$ be random variables. To denote convergence in probability we will use

$$X_n \xrightarrow{p} Y\,.$$

Similarly convergence in distribution is written as

$$X_n \xrightarrow{d} Y.$$

If the distribution is specified, e.g. the standard normal distribution $\mathcal{N}(0,1)$ we will write

$$X_n \xrightarrow{d} \mathcal{N}(0,1)\,.$$

The p-dimensional multivariate standard normal distribution is denoted by $\mathcal{N}_p(\mathbf{0}, \mathbf{E}_p)$.

Let $\mathcal{V} \subseteq \mathbb{R}^p$ and $f : \mathcal{V} \longrightarrow \mathbb{R}$, then

$$\arg\max_{\boldsymbol{v}\in\mathcal{V}} f(\boldsymbol{v}) := \{\boldsymbol{v} \in \mathcal{V} | f(\boldsymbol{v}) \geq f(\boldsymbol{w}) \text{ for all } \boldsymbol{w} \in \mathcal{V}\}$$

denotes the set of values maximizing f. In a slight abuse of notation we will use

$$\boldsymbol{w} = \arg\max_{\boldsymbol{v}\in\mathcal{V}} f(\boldsymbol{v})$$

to denote

$$\boldsymbol{w} \in \arg\max_{\boldsymbol{v}\in\mathcal{V}} f(\boldsymbol{v})\,.$$

If the maximum is not uniquely defined, we will assume, that the solution can be chosen arbitrarily or, that there is some rule, which tells us which to choose.

2.2. Binary Response Models

Binary response models are special cases of generalized linear models introduced by Nelder and Wedderburn (1972). In these models the influence of the control variable on the mean is described using a linear model and a link function. The link describes the nonlinear behavior of the mean. This means, that in our case the probability of success of a binary random variable is "linked" to the linear model. For more information on generalized linear models see McCullagh and Nelder (1997). Details concerning binary data are covered in Chapter 4 of their book.

Denote the parameter by $\boldsymbol{\theta} \in \mathbb{R}^p$, $p \in \mathbb{N}$, and the parameter space, i.e. the possible values for the parameter, by $\Theta \subseteq \mathbb{R}^p$.

A special setting of the control variable $\boldsymbol{x} \in \mathbb{R}^{p-1}$ will be called design point. The set of design points, which can be chosen in the experiment, is called design space and denoted by $\mathcal{X}$. The influence of the design points in the linear part of the model is described by the regression function $\boldsymbol{f} : \mathcal{X} \longrightarrow \mathbb{R}^p$. While in general the components of $\boldsymbol{f}$ are real valued functions in $\boldsymbol{x}$, we will consider only the following special case:

$$\boldsymbol{f}(\boldsymbol{x}) := \begin{pmatrix} 1 & x_1 & \ldots & x_{p-1} \end{pmatrix}^\top . \tag{2.1}$$

The mean function $G : \mathbb{R} \longrightarrow [0,1]$ is the essential part, which characterizes the properties of the model. It is the inverse of the link function from the generalized linear model. We will assume that it is a continuously differentiable distribution function with density G'.

An observation at $\boldsymbol{x} \in \mathbb{R}^{p-1}$ is modeled by a binary random variable Y with

$$\mathrm{P}(Y = 1) = 1 - \mathrm{P}(Y = 0) = G(\boldsymbol{f}(\boldsymbol{x})^\top \boldsymbol{\theta}) \tag{2.2}$$

or, equivalently,

$$\mathrm{E}(Y) = G(\boldsymbol{f}(\boldsymbol{x})^\top \boldsymbol{\theta})\,.$$

Example 1. The mean functions for the logit, probit, log-log and complementary log-log model are displayed in Figure 2.1. These models are also considered in McCullagh and Nelder (1997, p. 108).

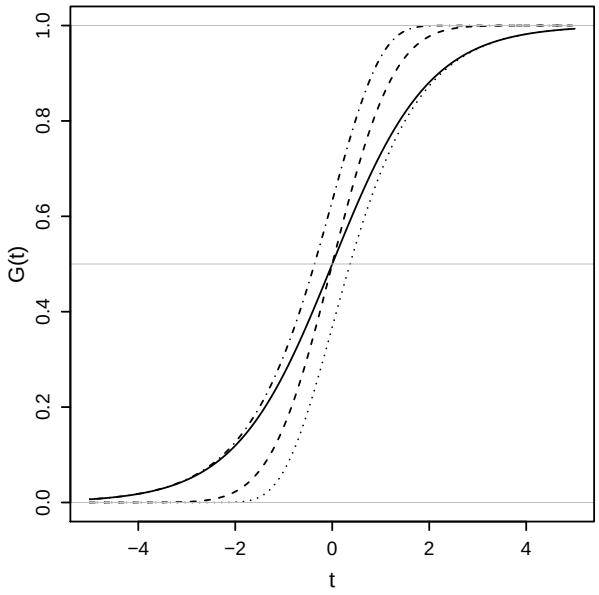


Figure 2.1.: Comparison of the mean functions for different models.
solid: Logit, dashed: Probit, dotted: Log-log, dash-dotted: Complementary log-log

Logit model:
The mean function in the logistic or logit model is

$$G(t) := \frac{1}{1 + e^{-t}}\,, \quad t \in \mathbb{R}\,, \tag{2.3}$$

which is the distribution function of the (standard) logistic distribution. It arises from modeling the logarithm of the odds by a linear model, i.e.

$$\log\left(\frac{\mathrm{P}(Y = 1)}{1 - \mathrm{P}(Y = 1)}\right) = \boldsymbol{f}(\boldsymbol{x})^\top \boldsymbol{\theta}\,. \tag{2.4}$$

The density G' is the same as the variance for one observation:

$$G'(t) = G(t)(1 - G(t))\,. \tag{2.5}$$

Figure 2.2 displays the probability of response in the logistic model with $p = 2$ for different choices of $\boldsymbol{\theta}$. The design points in this case are real numbers. The location, which is often denoted by μ, is influenced by both components: $\mu = -\theta_1/\theta_2$. It is the value of x for which $\boldsymbol{f}(x)^\top \boldsymbol{\theta} = 0$, i.e. $G(\boldsymbol{f}(x)^\top \boldsymbol{\theta}) = 0.5$. The slope is only influenced by θ_2.

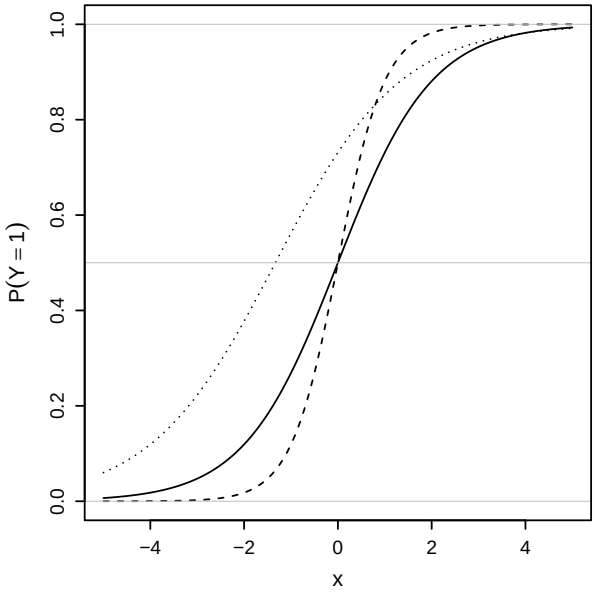


Figure 2.2.: $\mathrm{P}(Y = 1)$ as a function of x for the logit model and different values of $\boldsymbol{\theta}$.
solid: $\boldsymbol{\theta} = (0 \quad 1)^\top$, dashed: $\boldsymbol{\theta} = (0 \quad 2)^\top$, dotted: $\boldsymbol{\theta} = (1 \quad 0.75)^\top$

Probit model:
The probit model uses the distribution function of the standard normal distribution as mean function G and hence

$$G'(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}, \quad t \in \mathbb{R}. \tag{2.6}$$

It is very similar to the logit model, but with lighter tails.

Log-log model:
The link to the linear part of the model is given by

$$-\log(-\log(\mathrm{P}(Y = 1))) = \boldsymbol{f}(\boldsymbol{x})^\top \boldsymbol{\theta},$$

which explains the name of the model. The mean function

$$G(t) := e^{-e^{-t}}, \quad t \in \mathbb{R},$$

is the distribution function of the Gumbel distribution. In contrast to the logit and probit this model is asymmetric. For $t \longrightarrow \infty$ it behaves like the logit mean function.

Complementary log-log model:
This asymmetric model is closely related to the previous one. If we denote the mean function of the log-log model by G_1, then the mean function of the complementary log-log model can be written as $G(t) = 1 - G_1(-t)$. It is defined by

$$\log(-\log(1 - \mathrm{P}(Y = 1))) = \boldsymbol{f}(\boldsymbol{x})^\top \boldsymbol{\theta}\,,$$

and hence has the mean function

$$G(t) := 1 - e^{-e^t}\,, \quad t \in \mathbb{R}\,.$$

As is the case for the log-log model G approaches the logit mean function, however for $t \longrightarrow -\infty$.

Let $(\boldsymbol{x}_i)_{i\geq 1}$ be a sequence in $\mathbb{R}^{p-1}$, then $(Y_i)_{i\geq 1}$ denotes a sequence of random variables, with

$$\mathrm{P}(Y_i = 1) = 1 - \mathrm{P}(Y_i = 0) = G(\boldsymbol{f}(\boldsymbol{x}_i)^\top \boldsymbol{\theta})\,, \tag{2.7}$$

$i = 1, 2, \ldots$, i.e. where the random variable Y_i models an observation at $\boldsymbol{x}_i$. A realization of this sequence of random variables is denoted by $(y_i)_{i\geq 1}$.

A sample containing $n \in \mathbb{N}$ observations can be written as $Y_1, \ldots, Y_n$ or in vector notation as

$$\boldsymbol{Y}_n = \begin{pmatrix} Y_1 & \ldots & Y_n \end{pmatrix}^\top.$$

The corresponding $\boldsymbol{x}_i$ are stacked in an $n \times p$-matrix, which is called design matrix, and denoted by

$$\mathbf{F}_n := \begin{pmatrix} \boldsymbol{f}(\boldsymbol{x}_1) & \ldots & \boldsymbol{f}(\boldsymbol{x}_n) \end{pmatrix}^\top. \tag{2.8}$$

This matrix can be used to write the vector of mean functions corresponding to $\boldsymbol{Y}_n$ using a vector valued function $\boldsymbol{G}_n : \mathbb{R}^n \longrightarrow [0,1]^n$ defined by

$$\boldsymbol{G}_n(\boldsymbol{v}) := \begin{pmatrix} G(v_1) & \ldots & G(v_n) \end{pmatrix}^\top, \boldsymbol{v} \in \mathbb{R}^n\,. \tag{2.9}$$

As it was pointed out in the introduction we will consider procedures, where the values of the control variable are chosen depending on previous observations. We will need the following extensions of the model:

Let $\boldsymbol{x}_i$ be realizations of $\mathcal{X}$-valued random variables $\boldsymbol{X}_i$. Further, let $\mathcal{F}_n$ be the σ-field generated by $\boldsymbol{Y}_n$ and $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. Dependence on previous observations means that $\boldsymbol{X}_{n+1}$ is $\mathcal{F}_n$-measurable. Instead of (2.7) we have conditional probabilities

$$\mathrm{P}(Y_n = 1 | \boldsymbol{X}_n = \boldsymbol{x}_n) = 1 - \mathrm{P}(Y_n = 0 | \boldsymbol{X}_n = \boldsymbol{x}_n) = G(\boldsymbol{f}(\boldsymbol{x}_n)^\top \boldsymbol{\theta}) \tag{2.10}$$

and assume, that the probability of a response depends on the value of the control variable only, i.e.

$$\mathrm{P}(Y_n = y_n | \boldsymbol{Y}_{n-1} = \boldsymbol{y}_{n-1}, \boldsymbol{X}_i = \boldsymbol{x}_i, i = 1, \ldots, n) = \mathrm{P}(Y_n = y_n | \boldsymbol{X}_n = \boldsymbol{x}_n). \tag{2.11}$$

Further assume that if $m \in \mathbb{N}$ observations, $Y_{n1}, \ldots, Y_{nm}$, are taken at $\boldsymbol{X}_{n1}, \ldots, \boldsymbol{X}_{nm}$ in the n-th step of the experiment, these observations are conditionally independent in the following sense:

$$\mathrm{P}(Y_{ni} = y_{ni}, i = 1, \ldots, m | \boldsymbol{X}_{ni} = \boldsymbol{x}_{ni}, i = 1, \ldots, m) = \prod_{i=1}^{m} \mathrm{P}(Y_{ni} = y_{ni} | \boldsymbol{X}_{ni} = \boldsymbol{x}_{ni})\,.$$

2.3. Maximum Likelihood Estimation

The log-likelihood function for one observation[1] is given and denoted by

$$\begin{aligned} l(\boldsymbol{\theta}, y, \boldsymbol{f}(\boldsymbol{x})) &:= y \log(\mathrm{P}(Y = 1|\boldsymbol{X} = \boldsymbol{x})) + (1 - y) \log(\mathrm{P}(Y = 0|\boldsymbol{X} = \boldsymbol{x})) \\ &= y \log\Big(G(\boldsymbol{f}(\boldsymbol{x})^\top \boldsymbol{\theta})\Big) + (1 - y) \log\Big(1 - G(\boldsymbol{f}(\boldsymbol{x})^\top \boldsymbol{\theta})\Big) . \end{aligned}$$

For an experiment with $n \in \mathbb{N}$ observations the log-likelihood function in the above model is defined by

$$\begin{aligned} l(\boldsymbol{\theta}, \boldsymbol{y}_n, \mathbf{F}_n) &:= \sum_{i=1}^{n} l(\boldsymbol{\theta}, y_i, \boldsymbol{f}(\boldsymbol{x}_i)) \\ &= \sum_{i=1}^{n} \Big(y_i \log\Big(G(\boldsymbol{f}(\boldsymbol{x}_i)^\top \boldsymbol{\theta})\Big) + (1 - y_i) \log\Big(1 - G(\boldsymbol{f}(\boldsymbol{x}_i)^\top \boldsymbol{\theta})\Big)\Big) . \end{aligned} \tag{2.12}$$

The maximum likelihood estimate based on n observations is defined by

$$\hat{\boldsymbol{\theta}}_n := \arg\max_{\boldsymbol{\theta} \in \Theta} l(\boldsymbol{\theta}, \boldsymbol{y}_n, \mathbf{F}_n) . \tag{2.13}$$

If no maximum exists, i.e.

$$\arg\max_{\boldsymbol{\theta} \in \Theta} l(\boldsymbol{\theta}, \boldsymbol{y}_n, \mathbf{F}_n) = \emptyset$$

we will set $\hat{\boldsymbol{\theta}}_n := \boldsymbol{\theta}_0$ for some fixed $\boldsymbol{\theta}_0 \in \Theta$.

The existence of the estimate is a very important issue, as is the uniqueness. Wedderburn (1976) considers this question for generalized linear models. He takes the parameter space into account, i.e. if it is bounded or not, and gives sufficient conditions. Logit, probit and complementary log-log model are considered as examples for binary response models. These results can be applied, if there are some design points with more than one observation.

Some conditions for the existence are given in the following result by Silvapulle (1981), which is restated here using our notation. It gives conditions for the existence of a maximum of the log-likelihood over $\mathbb{R}^p$ using the separation of the design points. For the logit model see also Albert and Anderson (1984). The design points are separated, if there is a hyperplane in $\mathbb{R}^p$, such that the design points where 1's were observed are on one side and the design points with 0's on the other. More precisely, they are separated[2], if there exists $\boldsymbol{v} \in \mathbb{R}^p$, such that

$$\boldsymbol{f}(\boldsymbol{x}_i)^\top \boldsymbol{v} \leq 0 \quad \text{whenever} \quad y_i = 0 \quad \text{and} \quad \boldsymbol{f}(\boldsymbol{x}_i)^\top \boldsymbol{v} \geq 0 \quad \text{whenever} \quad y_i = 1 .$$

As in Silvapulle (1981) we define the relative interiors of the convex cones generated by the design points with observation $y_i = 1$ by

$$\mathcal{C}_n^1 := \left\{ \sum_{i=1}^{n} k_i \, y_i \boldsymbol{f}(\boldsymbol{x}_i) \,\middle|\, k_i > 0, i = 1, \ldots, n \right\}$$

[1] The notation differs slightly from the notation for more than one observation. But while using $l(\boldsymbol{\theta}, y, \boldsymbol{f}(\boldsymbol{x})^\top)$ instead would be a more consistent choice, it would also add to the length of formulas.

[2] In our definition of separation, points in the separating hyperplane are allowed. This coincides with the quasi-separation of Albert and Anderson (1984).

and for $y_i = 0$ by

$$\mathcal{C}_n^0 := \left\{ \sum_{i=1}^n k_i \, (1 - y_i) \boldsymbol{f}(\boldsymbol{x}_i) \,\middle|\, k_i > 0, i = 1, \ldots, n \right\}.$$

With this notation, separation of the design points is equivalent to $\mathcal{C}_n^0 \cap \mathcal{C}_n^1 = \emptyset$. If $\mathcal{C}_n^0 \cap \mathcal{C}_n^1 \neq \emptyset$ we will say, that there is an overlap in the design points.

Note that the design space and the regression functions are defined differently, in order to state the theorem in more general form.

Theorem 1 (Silvapulle, 1981)**.** *Let $\mathcal{X} \subseteq \mathbb{R}^p$ and $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{x}$. Let $\mathbf{F}_n$ have full column rank and $\Theta = \mathbb{R}^p$.*

(i) *If the maximum likelihood estimate $\hat{\boldsymbol{\theta}}_n$ exists and $\arg\max_{\boldsymbol{\theta} \in \mathbb{R}^p} l(\boldsymbol{\theta}, \boldsymbol{y}_n, \mathbf{F}_n)$ is bounded, then*

$$\mathcal{C}_n^0 \cap \mathcal{C}_n^1 \neq \emptyset \quad \text{or either} \quad \mathcal{C}_n^0 = \mathbb{R}^p \quad \text{or} \quad \mathcal{C}_n^1 = \mathbb{R}^p \,. \tag{2.14}$$

(ii) *Suppose that $l(\boldsymbol{\theta}, \boldsymbol{y}_n, \mathbf{F}_n)$ is a proper closed concave function with respect to $\boldsymbol{\theta}$. Then $\hat{\boldsymbol{\theta}}_n$ exists and $\arg\max_{\boldsymbol{\theta} \in \mathbb{R}^p} l(\boldsymbol{\theta}, \boldsymbol{y}_n, \mathbf{F}_n)$ is bounded if and only if* (2.14) *is satisfied.*

(iii) *Suppose that $\log G$ and $\log(1 - G)$ are concave and that $x_{i1} = 1$ for all $i = 1, \ldots, n$. Then $\hat{\boldsymbol{\theta}}_n$ exists and $\arg\max_{\boldsymbol{\theta} \in \mathbb{R}^p} l(\boldsymbol{\theta}, \boldsymbol{y}_n, \mathbf{F}_n)$ is bounded if and only if $\mathcal{C}_n^0 \cap \mathcal{C}_n^1 \neq \emptyset$.*

Further assume that G is strictly increasing at every t satisfying $0 < G(t) < 1$. Then $\hat{\boldsymbol{\theta}}_n$ is uniquely defined if and only if $\mathcal{C}_n^0 \cap \mathcal{C}_n^1 \neq \emptyset$.

Example 2. To illustrate the theorem, let us consider the case with $p = 2$ and $n = 4$. Let $x_1 = x_2 = 0$ and $x_3 = x_4 = 1$. If $y_1 = y_2 = 0$, $y_3 = y_4 = 1$ the log-likelihood has the form

$$l(\boldsymbol{\theta}, \boldsymbol{y}_4, \mathbf{F}_4) = 2 \log(G(\theta_1 + \theta_2)) + 2 \log(1 - G(\theta_1)) \,,$$

and it is strictly increasing in θ_2. Hence there exists no maximum in $\mathbb{R}^2$.

If $y_1 = y_3 = 0$, $y_2 = y_4 = 1$ instead, we get

$$l(\boldsymbol{\theta}, \boldsymbol{y}_4, \mathbf{F}_4) = \log(G(\theta_1 + \theta_2)(1 - G(\theta_1 + \theta_2))) + \log(G(\theta_1)(1 - G(\theta_1)))$$

with the maximum at $\boldsymbol{\theta} = (G^{-1}(1/2) \quad 0)^\top$.

For our model, defined in Section 2.2, some conditions can be simplified. If a constant is part of the model, $\mathcal{C}_n^0$ and $\mathcal{C}_n^1$ can be at most as large as $(0, \infty) \times \mathbb{R}^{p-1}$. As a consequence (2.14) reduces to

$$\mathcal{C}_n^0 \cap \mathcal{C}_n^1 \neq \emptyset \tag{2.15}$$

in **(i)** and **(ii)**. In fact (2.15) is sufficient for the existence and boundedness in our case.

Lemma 1. *Let $\mathbf{F}_n$ have full column rank, $\Theta = \mathbb{R}^p$ and G be a strictly increasing distribution function. Let $\mathcal{C}_n^0 \cap \mathcal{C}_n^1 \neq \emptyset$, then $\hat{\boldsymbol{\theta}}_n$ exists and $\arg\max_{\boldsymbol{\theta} \in \mathbb{R}^p} l(\boldsymbol{\theta}, \boldsymbol{y}_n, \mathbf{F}_n)$ is bounded.*

The main idea of the proof, which is given in the appendix, is to show, that the log-likelihood is bounded from below in a subset of $\mathbb{R}^p$ and tends to $-\infty$ in each direction.

In order to find the maximum likelihood estimates it is convenient to write the derivative of the log-likelihood with respect to the parameter $\boldsymbol{\theta}$ in vector notation. The derivative is

called score function. We introduce the following notations: Let the functions $\psi : \mathbb{R} \longrightarrow \mathbb{R}$ and $d : \mathbb{R} \longrightarrow \mathbb{R}$ be defined as

$$\psi(t) := \frac{G'(t)}{G(t)(1-G(t))} \quad \text{and} \quad d(t) := \frac{G'(t)^2}{G(t)(1-G(t))}$$

for $t \in \mathbb{R}$. The first function is an abbreviation for the derivative of the log-odds of $G(t)$:

$$\frac{\mathrm{d}}{\mathrm{d}t} \log\left(\frac{G(t)}{1-G(t)}\right) = \frac{G'(t)}{G(t)(1-G(t))}\,.$$

The second one will occur in the Hessian and the Fisher information matrix. Define further the corresponding matrix valued function $\boldsymbol{\Psi}_n : \mathbb{R}^n \longrightarrow \mathbb{R}^{n\times n}$ and $\mathbf{D}_n : \mathbb{R}^n \longrightarrow \mathbb{R}^{n\times n}$ for $\boldsymbol{v} \in \mathbb{R}^n$ as

$$\boldsymbol{\Psi}_n(\boldsymbol{v}) := \operatorname*{diag}_{i=1,\ldots,n} (\psi(v_i)) \qquad \mathbf{D}_n(\boldsymbol{v}) := \operatorname*{diag}_{i=1,\ldots,n} (d(v_i))\,. \tag{2.16}$$

The score function $\boldsymbol{s}_n$ based on n observations is given by

$$\boldsymbol{s}_n(\boldsymbol{\theta}) := \frac{\partial l(\boldsymbol{\theta}, \boldsymbol{y}_n, \mathbf{F}_n)}{\partial \boldsymbol{\theta}} = \mathbf{F}_n^\top \boldsymbol{\Psi}_n(\mathbf{F}_n\boldsymbol{\theta})(\boldsymbol{y}_n - \boldsymbol{G}_n(\mathbf{F}_n\boldsymbol{\theta}))\,. \tag{2.17}$$

Since

$$\boldsymbol{s}_n(\boldsymbol{\theta}) = 0 \tag{2.18}$$

is a necessary condition for a (local) maximum, the maximum likelihood estimate is often defined as the solution of (2.18). (see e.g. Fahrmeir and Kaufmann, 1985)

Assume that G is twice continuously differentiable. The Hessian matrix of the log-likelihood is

$$\begin{aligned}\mathbf{H}_n(\boldsymbol{\theta}, \mathbf{F}_n) &:= \mathbf{F}_n^\top \operatorname*{diag}_{i=1,\ldots,n} \Big(\psi'(\boldsymbol{f}(\boldsymbol{x}_i)^\top\boldsymbol{\theta})\Big(y_i - G(\boldsymbol{f}(\boldsymbol{x}_i)^\top\boldsymbol{\theta})\Big)\Big)\mathbf{F}_n - \mathbf{F}_n^\top \mathbf{D}_n(\mathbf{F}_n\boldsymbol{\theta})\mathbf{F}_n \qquad (2.19)\\ &= \sum_{i=1}^n \psi'(\boldsymbol{f}(\boldsymbol{x}_i)^\top\boldsymbol{\theta})\Big(y_i - G(\boldsymbol{f}(\boldsymbol{x}_i)^\top\boldsymbol{\theta})\Big)\boldsymbol{f}(\boldsymbol{x}_i)\boldsymbol{f}(\boldsymbol{x}_i)^\top \\ &\qquad - \sum_{i=1}^n d(\boldsymbol{f}(\boldsymbol{x}_i)^\top\boldsymbol{\theta})\, \boldsymbol{f}(\boldsymbol{x}_i)\boldsymbol{f}(\boldsymbol{x}_i)^\top\,.\end{aligned}$$

The second matrix on the right-hand side of (2.19) is the Fisher information matrix, which will be introduced in the next section. It is always nonnegative definite. The problematic part in checking the sufficient condition for a maximum is the first matrix.

Example 3. For all models introduced in Example 1 $\log G$ and $\log(1-G)$ are concave. (see Section B.1 for the details) The mean functions are also strictly increasing. Hence the second part of Theorem 1 **(iii)** can be applied and an overlap of the defined cones is equivalent to existence and uniqueness of the maximum likelihood estimate.

Logit model:
Because of (2.4) the log-likelihood simplifies to

$$l(\boldsymbol{\theta}, \boldsymbol{y}_n, \mathbf{F}_n) = \sum_{i=1}^n \Big(y_i\ \boldsymbol{f}(\boldsymbol{x}_i)^\top\boldsymbol{\theta} + \log\Big(1 - G(\boldsymbol{f}(\boldsymbol{x}_i)^\top\boldsymbol{\theta})\Big)\Big)\,.$$

The same reason yields $\psi(t) = 1$ for all $t \in \mathbb{R}$. The score function and Hessian, in which the first part vanishes, simplify considerably:

$$\boldsymbol{s}_n(\boldsymbol{\theta}) = \mathbf{F}_n^\top(\boldsymbol{y}_n - \boldsymbol{G}_n(\mathbf{F}_n\boldsymbol{\theta})) \tag{2.20}$$

$$\mathbf{H}_n(\boldsymbol{\theta}, \mathbf{F}_n) = -\sum_{i=1}^{n} G(\boldsymbol{f}(\boldsymbol{x}_i)^\top\boldsymbol{\theta})\Big(1 - G(\boldsymbol{f}(\boldsymbol{x}_i)^\top\boldsymbol{\theta})\Big)\boldsymbol{f}(\boldsymbol{x}_i)\boldsymbol{f}(\boldsymbol{x}_i)^\top \tag{2.21}$$

2.4. The Fisher Information

The Fisher information plays an important role in the theory of estimation: Its inverse yields a lower bound for the covariance matrix of an estimator (Cramér-Rao bound or information inequality) and the asymptotic covariance matrix for the maximum likelihood estimator. This holds at least, if the observations are independent identically distributed and certain regularity conditions are fulfilled. (see e.g. Lehmann and Casella, 1998) The asymptotic interpretation is of particular interest for nonlinear models. If the covariance matrix of the estimator is not known, it is substituted by the asymptotic covariance matrix, i.e. by the inverse of the Fisher information matrix.

The Fisher information matrix for a single observation[3] at a fixed design point $\boldsymbol{x}$ is defined by

$$\mathbf{I}(\boldsymbol{\theta}, \boldsymbol{f}(\boldsymbol{x})) := \operatorname{Cov}\left(\frac{\partial l(\boldsymbol{\theta}, Y, \boldsymbol{f}(\boldsymbol{x}))}{\partial\boldsymbol{\theta}}\right).$$

In the case of the binary response model this becomes

$$\mathbf{I}(\boldsymbol{\theta}, \boldsymbol{f}(\boldsymbol{x})) = d(\boldsymbol{f}(\boldsymbol{x})^\top\boldsymbol{\theta})\ \boldsymbol{f}(\boldsymbol{x})\boldsymbol{f}(\boldsymbol{x})^\top.$$

The Fisher information matrix for n independent observations $Y_1, \ldots, Y_n$ is just the sum of the information matrices of the individual observations:

$$\mathbf{I}(\boldsymbol{\theta}, \mathbf{F}_n) = \sum_{i=1}^{n} \mathbf{I}(\boldsymbol{\theta}, \boldsymbol{f}(\boldsymbol{x}_i)) = \mathbf{F}_n^\top\mathbf{D}_n(\mathbf{F}_n\boldsymbol{\theta})\mathbf{F}_n. \tag{2.22}$$

This matrix is always symmetric and nonnegative definite. If the design points are random variables, equation (2.22) is a conditional information given $\{\boldsymbol{X}_i = \boldsymbol{x}_i, i = 1, \ldots, n\}$.

2.5. Optimal Design of Experiments

This part will briefly introduce the basic notation and fundamental results from the theory of optimal design. The basis for this section was the book by Silvey (1980). We will start with a general formulation.

Denote the set of probability measures with support in $\mathcal{X}$ by Ξ. Every $\xi \in \Xi$ is called a design. Define the (weighted) information matrix (for a given $\boldsymbol{\theta}$) as

$$\mathbf{M}(\boldsymbol{\theta}, \xi) := \int_{\mathcal{X}} \mathbf{I}(\boldsymbol{\theta}, \boldsymbol{f}(\boldsymbol{x}))\ \xi(\mathrm{d}\boldsymbol{x}) \tag{2.23}$$

[3]Similar to the log-likelihood, we will suppress the transpose sign. (see footnote 1)

and denote the set of all information matrices over $\mathcal{X}$ (for a given $\boldsymbol{\theta}$) by

$$\mathcal{M}_{\boldsymbol{\theta}} := \left\{ \mathbf{M}(\boldsymbol{\theta}, \xi) \,|\, \xi \in \Xi \right\}. \tag{2.24}$$

Since the convex combination of two probability measures is again a probability measure, Ξ and $\mathcal{M}_{\boldsymbol{\theta}}$ are both convex: For every two designs $\xi_1, \xi_2 \in \Xi$ and $\lambda \in [0,1]$

$$\lambda\, \xi_1 + (1-\lambda)\, \xi_2 \in \Xi$$

and

$$\mathbf{M}(\boldsymbol{\theta}, \lambda\, \xi_1 + (1-\lambda)\, \xi_2) = \lambda\, \mathbf{M}(\boldsymbol{\theta}, \xi_1) + (1-\lambda)\, \mathbf{M}(\boldsymbol{\theta}, \xi_2) \in \mathcal{M}_{\boldsymbol{\theta}}\,.$$

These definitions are convenient from a mathematical point of view. Because a design should tell the experimenter where to observe, i.e. which values to choose for the control variable, and how many observations should be spent there, only designs which have a finite support are applicable in an experiment. Fortunately we only have to consider designs with finite support when looking for optimal designs, because we will compare designs using their information matrices. Since symmetric $p \times p$-matrices can be represented by elements of $\mathbb{R}^{p(p+1)/2}$ and $\mathcal{M}_{\boldsymbol{\theta}}$ is the closed convex hull of

$$\{\mathbf{I}(\boldsymbol{\theta}, \boldsymbol{f}(\boldsymbol{x})) \,|\, \boldsymbol{x} \in \mathcal{X}\}\,, \tag{2.25}$$

Carathéodory's theorem[4] tells us, that we need at most $p(p+1)/2 + 1$ support points to represent an element of $\mathcal{M}_{\boldsymbol{\theta}}$. A design ξ with finite support $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m \in \mathcal{X}$, $m \in \mathbb{N}$, and corresponding weights $w_i := \xi(\boldsymbol{x}_i)$, $i = 1, \ldots, m$, will be written as

$$\xi = \begin{Bmatrix} \boldsymbol{x}_1 & \ldots & \boldsymbol{x}_m \\ w_1 & \ldots & w_m \end{Bmatrix}.$$

Its information matrix is given by a sum

$$\mathbf{M}(\boldsymbol{\theta}, \xi) = \sum_{i=1}^{m} w_i\, \mathbf{I}(\boldsymbol{\theta}, \boldsymbol{f}(\boldsymbol{x}_i))\,. \tag{2.26}$$

We say that a design is singular, if its information matrix is a singular matrix. This is for example the case, if the support of a design contains less design points than p, the number of parameters. If there is only one design point $\boldsymbol{x}$ in the support of a design, we will write $\xi_{\boldsymbol{x}}$. I.e.

$$\xi_{\boldsymbol{x}} := \begin{Bmatrix} \boldsymbol{x} \\ 1 \end{Bmatrix} \quad \text{and} \quad \mathbf{M}(\boldsymbol{\theta}, \xi_{\boldsymbol{x}}) = \mathbf{I}(\boldsymbol{\theta}, \boldsymbol{f}(\boldsymbol{x}))\,.$$

Intuitively a design $\xi_1 \in \Xi$ should be better than another design $\xi_2 \in \Xi$, if its information is larger, e.g. if $\mathbf{M}(\boldsymbol{\theta}, \xi_1) - \mathbf{M}(\boldsymbol{\theta}, \xi_2)$ is positive definite. But since in general there is no design ξ, which is the "largest" in this sense (see Pukelsheim, 2006, Chapter 4, for results in linear models) one usually considers criterion functions $\phi : \mathcal{M}_{\boldsymbol{\theta}} \longrightarrow \mathbb{R}$ instead. We will consider locally optimal designs, which are optimal for a given value of the parameter. This is due to the fact, that the information matrix depends on the parameter $\boldsymbol{\theta}$. The terminology appears first in Chernoff (1953).

[4]The theorem states, that each element of the convex hull of a subset $S \subseteq \mathbb{R}^n$ can be expressed as a convex combination of $n+1$ or less elements of S. (see Silvey, 1980, Appendix 2)

A design $\xi^* \in \Xi$ is locally optimal for the parameter value $\boldsymbol{\theta}$ with respect to the criterion ϕ, if

$$\phi(\mathbf{M}(\boldsymbol{\theta}, \xi^*)) \geq \phi(\mathbf{M}(\boldsymbol{\theta}, \xi)) \quad \text{for all} \quad \xi \in \Xi\,. \tag{2.27}$$

Note, that even if the criterion function ϕ has a unique maximum in $\mathcal{M}_{\boldsymbol{\theta}}$, this does not mean, that the design ξ^* is unique, because different designs can have the same information matrix. But the set of optimal designs is convex, if ϕ is concave.

Some classical examples of characteristics, which are used to build criteria, are given in the following list:

- Average variance of the estimator (A-criterion)

$$\operatorname{tr}(\mathbf{M}(\boldsymbol{\theta}, \xi)^{-1})$$

- Largest variance component of the estimator (E-criterion)

$$\lambda_{\max}(\mathbf{M}(\boldsymbol{\theta}, \xi)^{-1})$$

- Volume of the confidence ellipsoid for the parameter (D-criterion)

$$\det(\mathbf{M}(\boldsymbol{\theta}, \xi)^{-1})$$

They all utilize the inverse of the information matrix as a substitute for the covariance matrix, i.e. they are based on asymptotic behavior. Note, that all three functions are decreasing, if the information becomes larger in the above sense. We have to take this into account for our definition of the optimal design, which maximizes a functional of the information matrix.

We will consider the D-criterion and define the criterion for $\mathbf{M} \in \mathcal{M}_{\boldsymbol{\theta}}$ by

$$\phi_D(\mathbf{M}) := \begin{cases} \log\det(\mathbf{M}) & , \det(\mathbf{M}) \neq 0 \\ -\infty & , \det(\mathbf{M}) = 0 \end{cases} \tag{2.28}$$

The advantage in using the logarithm in the definition of the criterion function is that $\log\det(\mathbf{M})$ is a concave function on $\mathcal{M}_{\boldsymbol{\theta}}$. A design ξ^* is called locally D-optimal (for the parameter value $\boldsymbol{\theta}$), if

$$\phi_D(\mathbf{M}(\boldsymbol{\theta}, \xi^*)) \geq \phi_D(\mathbf{M}(\boldsymbol{\theta}, \xi)) \quad \text{for all} \quad \xi \in \Xi\,. \tag{2.29}$$

To verify the optimality of a design one usually uses directional derivatives of the criterion: In a maximum all directional derivatives should be nonpositive. This yields necessary and sufficient conditions for optimality, if the criterion function is concave.

Let $\mathbf{M}_1, \mathbf{M}_2 \in \mathcal{M}_{\boldsymbol{\theta}}$. We define the directional derivative

$$F_\phi(\mathbf{M}_1, \mathbf{M}_2) := \lim_{\alpha \to 0+} \frac{\phi((1-\alpha)\mathbf{M}_1 + \alpha\mathbf{M}_2) - \phi(\mathbf{M_1})}{\alpha}\,. \tag{2.30}$$

This is the "usual" directional derivative in the direction $\mathbf{M}_2 - \mathbf{M}_1$, as we can see by rewriting

$$(1-\alpha)\mathbf{M}_1 + \alpha\mathbf{M}_2 = \mathbf{M}_1 + \alpha(\mathbf{M}_2 - \mathbf{M}_1)\,.$$

The benefit of this formulation is that since $(1-\alpha)\mathbf{M}_1 + \alpha\mathbf{M}_2 \in \mathcal{M}_{\boldsymbol{\theta}}$ by construction, $\phi((1-\alpha)\mathbf{M}_1 + \alpha\mathbf{M}_2)$ is always defined.

The following theorems yield the criteria to check for optimality:

Theorem 2 (Silvey, 1980,Theorem 6.1.1, p. 54)**.** *Let $\boldsymbol{\theta}$ be fixed. Let ϕ be concave on $\mathcal{M}_{\boldsymbol{\theta}}$. Then ξ^* is locally optimal with respect to ϕ if and only if $F_\phi(\mathbf{M}(\boldsymbol{\theta},\xi^*),\mathbf{M}(\boldsymbol{\theta},\xi)) \leq 0$ for all $\xi \in \Xi$.*

Theorem 3 (Silvey, 1980,Theorem 6.1.2, p. 54)**.** *Let $\boldsymbol{\theta}$ be fixed. Let ϕ be differentiable at $\mathbf{M}(\boldsymbol{\theta},\xi^*)$ and concave on $\mathcal{M}_{\boldsymbol{\theta}}$. Then ξ^* is locally optimal with respect to ϕ if and only if $F_\phi(\mathbf{M}(\boldsymbol{\theta},\xi^*),\mathbf{M}(\boldsymbol{\theta},\xi_{\boldsymbol{x}})) \leq 0$ for all $\boldsymbol{x} \in \mathcal{X}$.*

The second theorem tells us, that if ϕ is differentiable at an information matrix, it is sufficient to look in the direction of the "corners" of $\mathcal{M}_{\boldsymbol{\theta}}$ to check for optimality. It is also interesting to note, that

$$\max_{\boldsymbol{x}\in\mathcal{X}} F_\phi(\mathbf{M}(\boldsymbol{\theta},\xi^*),\mathbf{M}(\boldsymbol{\theta},\xi_{\boldsymbol{x}})) = 0\,. \tag{2.31}$$

If additionally ξ^* has finite support, then

$$F_\phi(\mathbf{M}(\boldsymbol{\theta},\xi^*),\mathbf{M}(\boldsymbol{\theta},\xi_{\boldsymbol{x}})) = 0 \tag{2.32}$$

for all support points of ξ^*.

For the special case of D-optimality in linear models, Theorem 3 and (2.31) were part of the general equivalence theorem of Kiefer and Wolfowitz (1960). It was generalized to the nonlinear case by White (1973). Our criterion is differentiable at all nonsingular matrices and we get

$$F_{\phi_D}(\mathbf{M}(\boldsymbol{\theta},\xi^*),\mathbf{M}(\boldsymbol{\theta},\xi)) = \operatorname{tr}(\mathbf{M}(\boldsymbol{\theta},\xi)\mathbf{M}(\boldsymbol{\theta},\xi^*)^{-1}) - p \tag{2.33}$$

and

$$F_{\phi_D}(\mathbf{M}(\boldsymbol{\theta},\xi^*),\mathbf{M}(\boldsymbol{\theta},\xi_{\boldsymbol{x}})) = d(\boldsymbol{f}(\boldsymbol{x})^\top\boldsymbol{\theta})\boldsymbol{f}(\boldsymbol{x})^\top\mathbf{M}(\boldsymbol{\theta},\xi^*)^{-1}\boldsymbol{f}(\boldsymbol{x}) - p \tag{2.34}$$

for nonsingular $\mathbf{M}(\boldsymbol{\theta},\xi^*)$.

To compare two competing designs, e.g. to find out which is better or how close it is to an optimal design, one uses the efficiency. For D-optimality we will define the (relative) efficiency of ξ_1 with respect to ξ_2 by

$$\operatorname{eff}(\xi_1,\xi_2,\boldsymbol{\theta}) := \left(\frac{\det(\mathbf{M}(\boldsymbol{\theta},\xi_1))}{\det(\mathbf{M}(\boldsymbol{\theta},\xi_2))}\right)^{1/p}. \tag{2.35}$$

Example 4. For the models from Example 1 the locally D-optimal designs for $\boldsymbol{\theta} = (0 \;\; 1)^\top$ and $\mathcal{X} = \mathbb{R}$ are given in Table 2.1. They can be found in Ford, Torsney and Wu (1992, Table 4, p. 579). As a feature of the D-criterion all weights are 0.5. The corresponding values of the mean function at the support points are given, too.

The designs of the symmetric distributions are symmetric, too. The relationship of the log-log and complementary log-log model and their asymmetric shape are also mirrored in the design points.

Logit model:
The locally D-optimal design for $\boldsymbol{\theta} = (0 \;\; 1)^\top$ is

$$\xi_1 := \left\{\begin{matrix} -1.5434 & 1.5434 \\ \frac{1}{2} & \frac{1}{2} \end{matrix}\right\}, \tag{2.36}$$

Table 2.1.: Locally D-optimal designs for $\boldsymbol{\theta} = (0\ 1)^\top$

	Support of ξ^*			
Model	x_1	x_2	$G(x_1)$	$G(x_2)$
Logit	-1.5434	1.5434	0.176	0.824
Probit	-1.1381	1.1381	0.128	0.872
Log-log	-0.9796	1.3377	0.070	0.769
Complementary log-log	-1.3377	0.9796	0.231	0.930

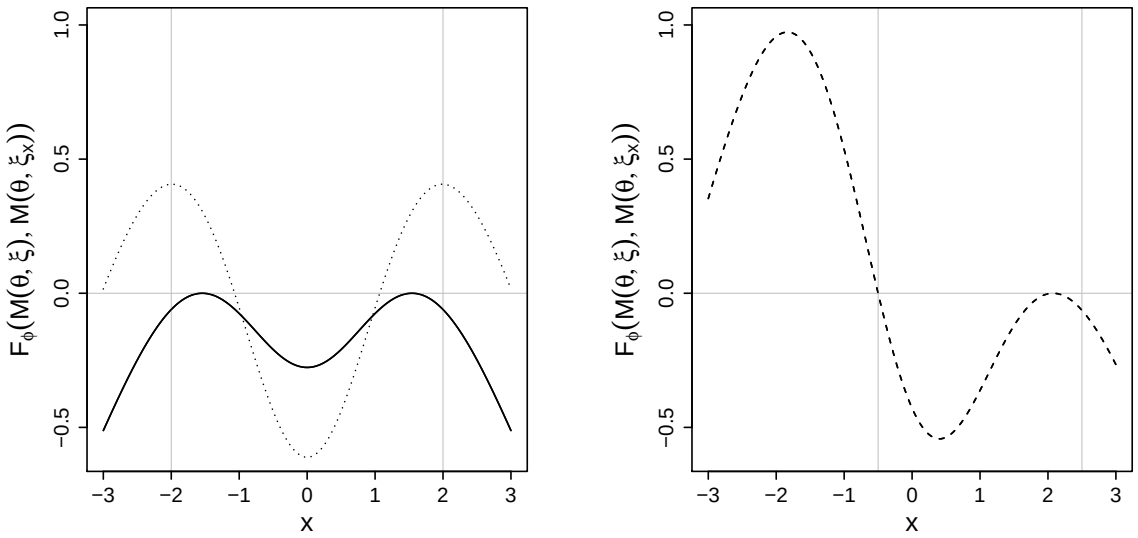
whenever $\{-1.5434, 1.5434\} \subseteq \mathcal{X}$. This is illustrated in Figure 2.3. The pictures show F_ϕ for (2.36) and the following two designs:

$$\xi_2 := \begin{Bmatrix} -0.5 & 2.0746 \\ \frac{1}{2} & \frac{1}{2} \end{Bmatrix}, \qquad \xi_3 := \begin{Bmatrix} -1.5434 & 0 & 1.5434 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{Bmatrix}.$$

The support points of ξ_1 are in $\mathcal{X} = [-2, 2]$ and the maximum of F_ϕ is attained at the support points, as it was suggested by (2.31). This is illustrated in Figure 2.3a. For the other designs F_ϕ is positive for some x. If $\mathcal{X}$ is changed to $[-0.5, 2.5]$ (Figure 2.3b) the design ξ_2 is locally optimal.

Locally optimal designs for other values of $\boldsymbol{\theta}$ are given by

$$\frac{\pm 1.5434 - \theta_1}{\theta_2}.$$



(a) $\mathcal{X} = [-2, 2]$ (b) $\mathcal{X} = [-0.5, 2.5]$

Figure 2.3.: $F_{\phi_D}(\mathbf{M}(\boldsymbol{\theta}, \xi), \mathbf{M}(\boldsymbol{\theta}, \xi_x))$ as a function of x for ξ_1 (solid), ξ_2 (dashed) and ξ_3 (dotted). The boundaries of the design space are marked by vertical lines.

Now that we can check, if a design is optimal, there is still the question how to get candidates. With directional derivatives at hand, it seems reasonable to consider steepest

ascent algorithms, to find optimal designs. This was first done by Wynn (1970) and Fedorov (1972) for the linear case (see also Wu and Wynn, 1978), but their method works for locally optimal designs in nonlinear models, too. For D-optimality the basic algorithm is as follows: Let $(\alpha_n)_{n\geq 1}$ be a positive sequence, such that

$$\lim_{n\to\infty} \alpha_n = 0 \quad \text{and} \quad \sum_{n=1}^{\infty} \alpha_n = \infty \,.$$

While it is possible to adjust the step-length in each step by line search, we will restrict ourselves to the case of $\alpha_n = n^{-1}$, i.e. each point in the sequence $(\boldsymbol{x}_n)_{n\geq 1}$ has the same weight. Since this is the case considered by Wynn (1970), we will refer to the algorithm described below as Wynn algorithm.

Let the initial design ξ_0 have a regular information matrix.

Step 1 Calculate the next design point:

$$\boldsymbol{x}_{n+1} := \arg\max_{\boldsymbol{x}\in\mathcal{X}} F_{\phi_D}(\mathbf{M}(\boldsymbol{\theta}, \xi_n), \mathbf{M}(\boldsymbol{\theta}, \xi_{\boldsymbol{x}}))\,, \tag{2.37}$$

i.e. in the direction where the derivative is largest.

Step 2 Update the design

$$\xi_{n+1} = (1-\alpha_{n+1})\xi_n + \alpha_{n+1}\xi_{\boldsymbol{x}_{n+1}}\,. \tag{2.38}$$

Step 3 Stop, if some stopping rule is fulfilled. Otherwise return to Step 1.

The information matrix can be easily calculated in each step:

$$\mathbf{M}(\boldsymbol{\theta}, \xi_{n+1}) = (1-\alpha_{n+1})\mathbf{M}(\boldsymbol{\theta}, \xi_n) + \alpha_{n+1}\,\mathbf{I}(\boldsymbol{\theta}, \boldsymbol{f}(\boldsymbol{x}_{n+1}))\,. \tag{2.39}$$

It is interesting to note, that for D-optimality maximizing the derivative F_{ϕ_D} in (2.37) is equivalent to maximizing the efficiency by choosing the next design point. This follows from Lemma A.1 part **(ii)**:

$$\frac{\det(\mathbf{I}(\boldsymbol{\theta}, \mathbf{F}_n) + \mathbf{I}(\boldsymbol{\theta}, \boldsymbol{f}(\boldsymbol{x})))}{\det(\mathbf{I}(\boldsymbol{\theta}, \mathbf{F}_n))} = 1 + d(\boldsymbol{f}(\boldsymbol{x})^\top\boldsymbol{\theta})\boldsymbol{f}(\boldsymbol{x})^\top\mathbf{I}(\boldsymbol{\theta}, \mathbf{F}_n)^{-1}\boldsymbol{f}(\boldsymbol{x})\,.$$

A problem with locally optimal designs is that we want to use the design, which depends on the unknown parameter, to estimate this very parameter. We will consider sequential methods, where the estimation and design step are iterated. For this purpose the Wynn algorithm will be modified: In each step $\boldsymbol{\theta}$ is substituted by the last estimate $\hat{\boldsymbol{\theta}}_n$. Let $\alpha_n = n^{-1}$ and ξ_0 be as in the original algorithm.

Step 1 Calculate the next design point:

$$\boldsymbol{x}_{n+1} := \arg\max_{\boldsymbol{x}\in\mathcal{X}} F_{\phi}\Big(\mathbf{M}(\hat{\boldsymbol{\theta}}_n, \xi_n), \mathbf{M}(\hat{\boldsymbol{\theta}}_n, \xi_{\boldsymbol{x}})\Big)\,, \tag{2.40}$$

i.e. in the direction where the derivative is largest.

Step 2 Update the design

$$\xi_{n+1} = (1-\alpha_{n+1})\xi_n + \alpha_{n+1}\xi_{\boldsymbol{x}_{n+1}}\,. \tag{2.41}$$

Step 3 Stop, if some stopping rule is fulfilled. Otherwise return to Step 1.

This adaptive Wynn algorithm was recently considered by Pronzato (2010) and will be reviewed in more details in the next section.

2.6. A Review of Convergence Results for the Maximum Likelihood Estimator and Adaptive Design

From now on we will distinguish between $\boldsymbol{\theta}$, i.e. some arbitrary value chosen for the parameter, and $\bar{\boldsymbol{\theta}}$, which denotes the "actual value" of the parameter, i.e. the fixed value governing the observations in the specific experiment.

The main criterion for the convergence of the estimator is that the variance tends to 0 or equivalently, that the information tends to infinity. This translates to conditions on the eigenvalues of the Fisher information matrix, as we will see.

For independent observations Fahrmeir and Kaufmann (1985) proved results for convergence and asymptotic normality of the maximum likelihood estimator in generalized linear models, which they defined as a solution of (2.18). They also comprehensively discuss their findings in view of previous results. Their main assumption is, that the information should tend to infinity for all components of the parameter, but that the speed of divergence should not be too different. In mathematical formulation this becomes

$$\lambda_{\min}(\mathbf{I}(\bar{\boldsymbol{\theta}}, \mathbf{F}_n)) \longrightarrow \infty \tag{2.42}$$

and there is a neighborhood of $\bar{\boldsymbol{\theta}}$, as well as some constants $C > 0$, $\delta > 0$, $n_1 \in \mathbb{N}$, such that

$$\frac{\lambda_{\max}(\mathbf{I}(\bar{\boldsymbol{\theta}}, \mathbf{F}_n))^{1/2+\delta}}{\lambda_{\min}(\mathbf{I}(\boldsymbol{\theta}, \mathbf{F}_n))} \leq C \tag{2.43}$$

for all $n \geq n_1$ and all $\boldsymbol{\theta}$ in this neighborhood. In the logistic model these conditions yield $\hat{\boldsymbol{\theta}}_n \longrightarrow \bar{\boldsymbol{\theta}}$ almost surely. In binary response models with other mean functions $\mathbf{I}(\boldsymbol{\theta}, \mathbf{F}_n)$ has to be substituted by the Hessian matrix of the log-likelihood, to secure the existence of a maximum.

If the design points are realizations of random variables, as defined at the end of Section 2.2, there are similar results. Let ε_n, $n = 1, 2, \ldots$, form a martingale difference sequence with respect to $\mathcal{F}_n$, which is defined in Section 2.2, and let $\boldsymbol{X}_{n+1}$ be $\mathcal{F}_n$-measurable. Consider the linear multiple regression model

$$Y_n = \boldsymbol{f}(\boldsymbol{X}_n)^\top \boldsymbol{\theta} + \varepsilon_n \,,$$

where $\boldsymbol{\theta} \in \mathbb{R}^p$ and $\boldsymbol{f}(\boldsymbol{X}_n)$ is $\mathbb{R}^p$-valued for $n = 1, 2, \ldots$. Assume that

$$\sup_{n \geq 1} \mathrm{E}(\varepsilon_n^\alpha | \mathcal{F}_{n-1}) < \infty$$

for some $\alpha > 2$. Lai and Wei (1982) showed that the least squares estimator $\hat{\boldsymbol{\theta}}_n$ converges almost surely to the true value of the parameter $\boldsymbol{\theta}$, if the extremal eigenvalues of the matrix $\mathbf{F}_n^\top \mathbf{F}_n$, i.e. the Fisher information matrix for linear models, fulfill

$$\lambda_{\min}(\mathbf{F}_n^\top \mathbf{F}_n) \longrightarrow \infty \quad \text{and} \quad \frac{\log(\lambda_{\max}(\mathbf{F}_n^\top \mathbf{F}_n))}{\lambda_{\min}(\mathbf{F}_n^\top \mathbf{F}_n)} \longrightarrow 0 \tag{2.44}$$

almost surely. Lai (1994) later extended this to nonlinear least squares applying similar conditions as Wu (1981), who considered convergence of the nonlinear least squares estimator for independent observations. Lai assumes a compact parameter space Θ. Instead

of the eigenvalues the conditions are based on

$$D_n(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) := \sum_{i=1}^{n} \left(G(\boldsymbol{f}(\boldsymbol{x}_i)^\top \boldsymbol{\theta}_1) - G(\boldsymbol{f}(\boldsymbol{x}_i)^\top \boldsymbol{\theta}_2) \right)^2 \tag{2.45}$$

and some term $\tilde{D}_n$, which includes squares of higher order mixed partial derivatives and will not be discussed here. The conditions in (2.42) and (2.43) are replaced by the following: For each $\boldsymbol{\theta}_1 \neq \bar{\boldsymbol{\theta}}$, exist constants $1 < \delta < 2$, $K > 0$, $C > 0$ and $n_1 \in \mathbb{N}$, such that

$$\inf_{\|\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1\| \leq K} D_n(\boldsymbol{\theta}_2, \bar{\boldsymbol{\theta}}) \longrightarrow \infty \quad \text{and} \quad \frac{D_n(\boldsymbol{\theta}_1, \bar{\boldsymbol{\theta}}) + \tilde{D}_n}{\left(\inf_{\|\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1\| \leq K} D_n(\boldsymbol{\theta}_2, \bar{\boldsymbol{\theta}}) \right)^\delta} \leq C$$

almost surely.

Chen, Hu and Ying (1999) extended the result of Lai and Wei to maximum quasi-likelihood estimation in generalized linear models. This also is an extension of the results of Fahrmeir and Kaufmann (1985) for the case of the logit and similar generalized linear models.[5] Chen et al. define their estimate $\hat{\boldsymbol{\theta}}_n$ as the solution of a "quasi score function":

$$\mathbf{F}_n^\top (\boldsymbol{y}_n - \boldsymbol{G}_n(\mathbf{F}_n \boldsymbol{\theta})) = 0\,. \tag{2.46}$$

Note that their quasi-likelihood approach is a special case of the one considered in McCullagh and Nelder (1997, Chapter 9, pp. 323), with the variance function equal to G'. For the logit model the maximum quasi-likelihood approach of Chen et al. is the same as maximum likelihood, because the left hand side of (2.46) is the score function from equation (2.20) on page 11. Hence their theorem establishes the strong consistency of the maximum likelihood estimator and yields a reasonable starting point for choosing conditions needed for convergence.

Chen et al. assumed that the mean function G is continuously differentiable and strictly increasing. In addition to the conditions on the eigenvalues in (2.44) they assumed for the design points, that $\sup_{i \geq 1} \|\boldsymbol{f}(\boldsymbol{X}_i)\| < \infty$ almost surely or that G' is bounded away from 0.

Ying and Wu (1997) considered a family of similar estimation procedures for location and scale families and also specified the sequence of design points. Their estimation equation for the location parameter is given by

$$\sum_{i=1}^{n} \tilde{\psi}(\hat{\beta}_n x_i)(y_i - G(\hat{\beta}_n(x_i - \mu))) = 0\,,$$

where μ is the location and $\hat{\beta}_n$ is an estimate for the scale parameter. Assume that $\hat{\beta}_n$ converges almost surely to the true value of the parameter and that the next design point x_{n+1} is chosen as $\hat{\mu}_n$, which is the estimate of the location parameter based on n observations. In case of convergence, this yields an asymptotically optimal sequence of design points in the sense, that the points converge to the locally optimal design for estimating μ. With some additional assumptions on the weight function $\tilde{\psi}$ and the mean function G, Ying and Wu showed, that $\hat{\mu}_n$ converges almost surely to the true value of the location parameter.

[5] "Similar generalized linear model" means with natural link function, which is not discussed here.

Note that the parameters β and μ are not estimated simultaneously by maximum likelihood and that the weight function $\tilde{\psi}$ does not depend on both parameters. This helps to avoid the problem of existence and uniqueness of a maximum, as does the estimation procedure (2.46). Some of the proofs in Ying and Wu use results from the field of stochastic approximation, which motivated the approach in the next chapter.

The adaptive Wynn algorithm from the end of the preceding section is similar in spirit to the choice of design points by Ying and Wu: The points are chosen such that the design should tend to the locally optimal design for the parameter. As mentioned above this was considered by Pronzato (2010) for a finite design space $\mathcal{X}$ and compact parameter space Θ. He first proves almost sure convergence for the nonlinear least squares estimator and the maximum likelihood estimator in binary response models and then, that the adaptive Wynn algorithm yields the assumptions needed for convergence. Similar to Lai (1994) and Wu (1981), he assumes

$$D_n(\boldsymbol{\theta}, \bar{\boldsymbol{\theta}}) \longrightarrow \infty$$

for all $\boldsymbol{\theta}$, with $\|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\| > \delta$ for all $\delta > 0$. For the maximum likelihood estimator

$$\sum_{i=1}^{n} G(\boldsymbol{f}(\boldsymbol{x}_i)^\top \bar{\boldsymbol{\theta}}) \log\left(\frac{G(\boldsymbol{f}(\boldsymbol{x}_i)^\top \bar{\boldsymbol{\theta}})}{G(\boldsymbol{f}(\boldsymbol{x}_i)^\top \boldsymbol{\theta})}\right) + \left(1 - G(\boldsymbol{f}(\boldsymbol{x}_i)^\top \bar{\boldsymbol{\theta}})\right) \log\left(\frac{1 - G(\boldsymbol{f}(\boldsymbol{x}_i)^\top \bar{\boldsymbol{\theta}})}{1 - G(\boldsymbol{f}(\boldsymbol{x}_i)^\top \boldsymbol{\theta})}\right)$$

was used instead of the D_n defined above. This sum has to increase faster than $\log(\log(n))$ which is much slower than in previous results and probably due to the restrictions on $\mathcal{X}$.

For the Wynn algorithm part the first step is to show, that for any sequence $(\boldsymbol{\theta}_n)_{n\geq 1}$ in Θ, there are at least p support points with positive weights in the limiting design. Then in a second step he shows, that from convergence of $\boldsymbol{\theta}_n$ to $\bar{\boldsymbol{\theta}}$ follows that $\phi_D(\mathbf{M}(\bar{\boldsymbol{\theta}}, \xi_n))$ converges to $\phi_D(\mathbf{M}(\bar{\boldsymbol{\theta}}, \xi^*))$. To achieve this he has to assume that

$$\min_{\boldsymbol{\theta}\in\Theta} \lambda_{\min}\left(\sum_{\boldsymbol{x}\in\mathcal{X}} \mathbf{I}(\boldsymbol{\theta}, \boldsymbol{f}(\boldsymbol{x}))\right) > \delta > 0$$

and

$$\lambda_{\min}\left(\sum_{i=1}^{p} \mathbf{I}(\bar{\boldsymbol{\theta}}, \boldsymbol{f}(\boldsymbol{x}_i))\right) > \delta > 0$$

for any p distinct elements of $\mathcal{X}$.

These results on convergence for a general sequence of designs on a finite design space can also be found in Pronzato (2009). Of special interest is that asymptotic normality is considered, too, and that the information matrix acts as asymptotic covariance matrix in these results. This justifies the use of the information matrix to design experiments adaptively.

It is interesting to note, that these results yield convergence for a fully adaptive method. In order to satisfy the conditions on the eigenvalues or D_n or to secure existence of the estimate one often has to assume that the size of the initial design tends to infinity (see Chaudhuri and Mykland, 1993) or has to add deterministic design points. (see e.g. the example in Lai, 1994, pp. 1923)

2.7. Stochastic Approximation

The idea, which we will study in later chapters is to show convergence of the maximum likelihood estimator with methods from the field of stochastic approximation. It is concerned with the study of stochastic algorithms of the form

$$\hat{\boldsymbol{\theta}}_{n+1} = \hat{\boldsymbol{\theta}}_n + \alpha_{n+1} \boldsymbol{Z}_{n+1} \,,$$

where $\boldsymbol{Z}_n$ is $\mathbb{R}^p$-valued random variable depending on a parameter $\bar{\boldsymbol{\theta}}$. The goal is to estimate $\bar{\boldsymbol{\theta}}$ recursively by $\hat{\boldsymbol{\theta}}_n$. The step size $\alpha_n > 0$ will be assumed to be decreasing.

The first articles by Robbins and Monro (1951) about stochastic root finding and similarly Kiefer and Wolfowitz (1952) concerning the search for the maximum of a stochastically perturbed function considered "classical" conditions on the step size, namely

$$\sum_{i=1}^{\infty} \alpha_i = \infty \quad \text{and} \quad \sum_{i=1}^{\infty} \alpha_i^2 < \infty \,.$$

The second condition provides that sums of (conditional) variances or second moments appearing in the proofs will converge.

In the 1970th the ordinary differential equation method was introduced by Ljung (1977). It was studied and extended thereafter by several authors. (e.g. Kushner and Clark, 1978; Métivier and Priouret, 1987) For further references see for example Kushner and Yin (2003) and Benveniste, Métivier and Priouret (1990) or Benaïm (1999, for a dynamical systems point of view).

The mean behavior of a stochastic algorithm is described by the solutions of an ordinary differential equation. The intuition behind this is that the algorithm is a perturbed Euler-approximation to an ordinary differential equation. The influence of the perturbations will become negligible due to averaging with the step length. Asymptotics and convergence of the algorithm can be inferred from this by considering the limit sets of the differential equation. A benefit of this approach is that it yields a framework for a wider class of algorithms, e.g. with weaker assumptions on the step length. We will describe the approach given in Kushner and Yin (2003).

More precisely assume that the random variables $\boldsymbol{Z}_{n+1}$ from above can be split into some $\mathbb{R}^p$-valued function $\boldsymbol{z} : \mathbb{R}^p \longrightarrow \mathbb{R}^p$ acting as a "mean value", the martingale difference $\boldsymbol{\varepsilon}_{n+1} := \boldsymbol{Z}_{n+1} - \mathrm{E}(\boldsymbol{Z}_{n+1}|\mathcal{F}_n)$ with $\mathcal{F}_n$ generated by $\boldsymbol{Z}_i$, $i = 1, \ldots, n$, and a perturbation $\boldsymbol{b}_{n+1}$. Denote the initial value for the estimate by $\hat{\boldsymbol{\theta}}_0$. Then the recursion can be rewritten as

$$\hat{\boldsymbol{\theta}}_{n+1} = \hat{\boldsymbol{\theta}}_n + \alpha_{n+1}\, \boldsymbol{z}(\hat{\boldsymbol{\theta}}_n) + \alpha_{n+1}\, \boldsymbol{\varepsilon}_{n+1} + \alpha_{n+1}\, \boldsymbol{b}_{n+1}$$

After summation over n we arrive at

$$\hat{\boldsymbol{\theta}}_{n+1} = \hat{\boldsymbol{\theta}}_m + \sum_{i=m+1}^{n+1} \alpha_i\, \boldsymbol{z}(\hat{\boldsymbol{\theta}}_{i-1}) + \sum_{i=m+1}^{n+1} \alpha_i\, \boldsymbol{\varepsilon}_i + \sum_{i=m+1}^{n+1} \alpha_i \boldsymbol{b}_i \,, \quad n \geq m \,. \tag{2.47}$$

An alternative way to measure the length of the sums, or equivalently the number of steps separating $\hat{\boldsymbol{\theta}}_{n+1}$ and $\hat{\boldsymbol{\theta}}_m$, is to introduce the "natural time" defined by $t_0 := 0$ and $t_n := \sum_{i=1}^{n} \alpha_i$, $n \in \mathbb{N}$. Defining the index at time $t \in \mathbb{R}$ by

$$\nu(t) := \begin{cases} \sup\{k \in \mathbb{N} \cup \{0\} : t_k \leq t\} & , \quad t \geq 0 \\ 0 & , \quad t < 0 \end{cases} \tag{2.48}$$

i.e. $\nu(t) = n$ if and only if $t \in [t_n, t_{n+1})$, (2.47) becomes

$$\hat{\boldsymbol{\theta}}_{n+1} = \hat{\boldsymbol{\theta}}_m + \sum_{i=m+1}^{\nu(t_{n+1})} \alpha_i \, \boldsymbol{z}(\hat{\boldsymbol{\theta}}_{i-1}) + \sum_{i=m+1}^{\nu(t_{n+1})} \alpha_i \, \boldsymbol{\varepsilon}_i + \sum_{i=m+1}^{\nu(t_{n+1})} \alpha_i \, \boldsymbol{b}_i \, .$$

With the piecewise constant, continuous time interpolation

$$\hat{\boldsymbol{\theta}}(t) := \begin{cases} \hat{\boldsymbol{\theta}}_n & , \quad t_n \leq t < t_{n+1} \\ \hat{\boldsymbol{\theta}}_0 & , \quad t < 0 \end{cases} \tag{2.49}$$

i.e. $\hat{\boldsymbol{\theta}}(t) := \hat{\boldsymbol{\theta}}_{\nu(t)}$, this can be extended to arbitrary differences in time, leading to

$$\hat{\boldsymbol{\theta}}(t_n + t) = \hat{\boldsymbol{\theta}}_n + \sum_{i=n+1}^{\nu(t_n+t)} \alpha_i \, \boldsymbol{z}(\hat{\boldsymbol{\theta}}_{i-1}) + \sum_{i=n+1}^{\nu(t_n+t)} \alpha_i \, \boldsymbol{\varepsilon}_i + \sum_{i=n+1}^{\nu(t_n+t)} \alpha_i \, \boldsymbol{b}_i \tag{2.50}$$

for $t \geq 0$ and

$$\hat{\boldsymbol{\theta}}(t_n + t) = \hat{\boldsymbol{\theta}}_n - \sum_{i=\nu(t_n+t)+1}^{n} \alpha_i \, \boldsymbol{z}(\hat{\boldsymbol{\theta}}_{i-1}) - \sum_{i=\nu(t_n+t)+1}^{n} \alpha_i \, \boldsymbol{\varepsilon}_i - \sum_{i=\nu(t_n+t)+1}^{n} \alpha_i \, \boldsymbol{b}_i \tag{2.51}$$

for $t < 0$.

If we consider $\hat{\boldsymbol{\theta}}(t_n + t)$ as a function of $t \in \mathbb{R}$ then we can define a sequence of functions $(\hat{\boldsymbol{\theta}}(t_n + \cdot))_{n \geq 1}$. Assume that there is a subsequence $(\hat{\boldsymbol{\theta}}(t_{n_k} + \cdot))_{k \geq 1}$ which converges uniformly to a function $\boldsymbol{\theta}(\cdot)$, on the interval $[-1, 1]$. Then

$$\hat{\boldsymbol{\theta}}_{n_k} = \hat{\boldsymbol{\theta}}(t_{n_k} + 0) \longrightarrow \boldsymbol{\theta}(0)$$

and in general

$$\hat{\boldsymbol{\theta}}_{\nu\left(t_{n_k}+t\right)} \longrightarrow \boldsymbol{\theta}(t)$$

for all $t \in [-1, 1]$. I.e. the limit $\boldsymbol{\theta}(t)$ characterizes not only the limit of the subsequence $(\hat{\boldsymbol{\theta}}_{n_k})_{k \geq 1}$, but of all subsequences $(\hat{\boldsymbol{\theta}}_{n_{k'}})_{k' \geq 1}$ for which $n_{k'} \in [\nu(t_{n_k} - 1), \nu(t_{n_k} + 1)]$. Moreover, if there is a subsequence converging uniformly on all compact intervals, we can describe the asymptotic behavior of the whole sequence of estimates. So the next step is to characterize convergent subsequences of $(\hat{\boldsymbol{\theta}}(t_n + \cdot))_{n \geq 1}$.

If the functions $\hat{\boldsymbol{\theta}}(t_n + \cdot)$ are continuous, then the concept of equicontinuity and the Arzelà-Ascoli Theorem (see Kushner and Yin, 2003, p. 102, Theorem 4.2.1) can be used. Recall the definition of equicontinuity:

Let $f_n : \mathbb{R} \longrightarrow \mathbb{R}^p$, $n \in \mathbb{N}$, be bounded functions. We say, that the sequence $(f_n)_{n \geq 1}$ is equicontinuous if, for any $\epsilon > 0$, there is a $\delta > 0$ such that $|t - s| \leq \delta$, $t, s \in \mathbb{R}$, implies

$$\|f_n(t) - f_n(s)\| \leq \epsilon$$

for all $n \geq 1$. I.e. all functions are continuous and the bounds δ and ϵ do not depend on n.

Theorem 4 (Arzelà-Ascoli)**.** *Let $J \subset \mathbb{R}$ be a bounded interval. Let $(f_n)_{n \geq 1}$ be equicontinuous and assume $\|f_n(x)\| \leq C$ for all $x \in J$, $n \geq 1$, then $(f_n)_{n \geq 1}$ has a subsequence converging uniformly to a continuous limit on J.*

In our case the functions will have jumps, which should decrease with increasing n. Hence we have to interpolate continuously or need some equicontinuity in the extended sense (see Kushner and Yin, 2003, p. 102):

Let $f_n : \mathbb{R} \longrightarrow \mathbb{R}^p$, $n \in \mathbb{N}$, be bounded, measurable functions. We say, that the sequence $(f_n)_{n\geq 1}$ is equicontinuous in the extended sense if, for any $\epsilon > 0$, there is a $\delta > 0$ such that $|t - s| \leq \delta$, $t, s \in \mathbb{R}$, implies

$$\limsup_{n\to\infty} \|f_n(t) - f_n(s)\| \leq \epsilon \,. \tag{2.52}$$

I.e. there exists a nonnegative null sequence $(a_n)_{n\geq 1}$ such that

$$\|f_n(t) - f_n(s)\| \leq \epsilon + a_n \,.$$

With this concept the Arzelà-Ascoli Theorem can be extended (Theorem 4.2.2 in Kushner and Yin, 2003, p. 102; for a proof see page 79 of this thesis):

Theorem 5 (Extended Arzelà-Ascoli)**.** *Let $J \subset \mathbb{R}$ be a bounded interval. Let $(f_n)_{n\geq 1}$ be equicontinuous in the extended sense and assume $\|f_n(x)\| \leq C$ for all $x \in J$, $n \geq 1$. Then $(f_n)_{n\geq 1}$ has a subsequence converging uniformly to a continuous limit on J.*

Hence showing, that $(\hat{\boldsymbol{\theta}}(t_n + \cdot))_{n\geq 1}$ is equicontinuous in the extended sense would yield the desired convergence.

The only sum in (2.50), which should have an influence on the limit, is the first one involving $\boldsymbol{z}$. So this is the sum for which a limit is of interest. The following lemma shows equicontinuity in the extended sense for the sequence of sums, considered as functions in t. It follows immediately from the extended Arzelà-Ascoli Theorem (Theorem 5), that a subsequence with a continuous limit exists.

Lemma 2. *Let $\boldsymbol{z} : \mathbb{R}^p \longrightarrow \mathbb{R}^p$ be bounded for all $(\hat{\boldsymbol{\theta}}_i)_{i\geq 1}$ uniformly in i. Then*

$$f_n(t) := \sum_{i=n+1}^{\nu(t_n+t)} \alpha_i \, \boldsymbol{z}(\hat{\boldsymbol{\theta}}_{i-1})$$

is equicontinuous in the extended sense and the limit f of a convergent subsequence is Lipschitz continuous.

Proof. Without loss of generality assume that $s < t$. Since $\boldsymbol{z}$ is bounded

$$\|f_n(t) - f_n(s)\| = \left\| \sum_{i=\nu(t_n+s)+1}^{\nu(t_n+t)} \alpha_i \, \boldsymbol{z}(\hat{\boldsymbol{\theta}}_{i-1}) \right\| \leq K \sum_{i=\nu(t_n+s)+1}^{\nu(t_n+t)} \alpha_i$$

for some constant $K > 0$. The sum on the right-hand side can be written as

$$\sum_{i=\nu(t_n+s)+1}^{\nu(t_n+t)} \alpha_i = \sum_{i=1}^{\nu(t_n+t)} \alpha_i - \sum_{i=1}^{\nu(t_n+s)} \alpha_i \,.$$

By definition of t_n and $\nu(t)$ we have

$$t_n + t - \alpha_{\nu(t_n+t)} \leq \sum_{i=1}^{\nu(t_n+t)} \alpha_i \leq t_n + t \,,$$

and hence

$$\|f_n(t) - f_n(s)\| \leq K\Big(t_n + t - (t_n + s - \alpha_{\nu(t_n+s)})\Big) = K\Big(t - s + \alpha_{\nu(t_n+s)}\Big)\,. \tag{2.53}$$

Since the step size tends to 0, the equicontinuity in the extended sense follows.

For the Lipschitz continuity note, that for the convergent subsequence f_{n_k} holds

$$\|f(t) - f(s)\| \leq \|f(t) - f_{n_k}(t)\| + \|f_{n_k}(s) - f(s)\| + \|f_{n_k}(t) - f_{n_k}(s)\|\,.$$

Taking the limit for $k \longrightarrow \infty$ and using (2.53) yields that

$$\|f(t) - f(s)\| \leq K|t - s|\,.$$

□

The other two sums from (2.50), which include the martingale difference $\boldsymbol{\varepsilon}_i$ and additional perturbations $\boldsymbol{b}_i$, respectively, should be negligible for large n and all $t > 0$. We will use what in Kushner and Yin (2003, p. 137) is called the "asymptotic rate of change condition" for $\sum_{i=1}^{n} \alpha_i\, \boldsymbol{\varepsilon}_i$ and $\sum_{i=1}^{n} \alpha_i\, \boldsymbol{b}_i$. (see also Benaïm, 1999, p.12, condition **A1**). It is defined as

$$\lim_{n\to\infty} \sup\left\{\left\|\sum_{i=n+1}^{k} \alpha_i\, \boldsymbol{\varepsilon}_i\right\| : k = n+1, \ldots, \nu(t_n + t)\right\} = 0 \tag{2.54}$$

$$\lim_{n\to\infty} \sup\left\{\left\|\sum_{i=n+1}^{k} \alpha_i\, \boldsymbol{b}_i\right\| : k = n+1, \ldots, \nu(t_n + t)\right\} = 0 \tag{2.55}$$

almost surely. Roughly speaking this means that the moving average of the errors tends to 0. In this definition, we have to be aware, that it is not the limes superior $\limsup_{k\to\infty}$ which is used, but $\lim_{k\to\infty}\sup$.

If the two sums converge, then (2.54) and (2.55) are naturally fulfilled. Sufficient conditions for (2.54) are for example boundedness of the martingale differences $\boldsymbol{\varepsilon}_i$ and $\alpha_n \log(n) \longrightarrow 0$ for $n \longrightarrow \infty$. (see Theorem 5.3.3 in Kushner and Yin, 2003, p. 139)

So if Lemma 2, (2.54) and (2.55) hold, the limiting function $\boldsymbol{\theta}(t)$ would be a solution to

$$\frac{\mathrm{d}\boldsymbol{\theta}(t)}{\mathrm{d}t} = \boldsymbol{z}(\boldsymbol{\theta}(t))\,.$$

If it is not possible to find a function $\boldsymbol{z}$, which is independent of the index i, the approach can be extended to differential inclusions. (see e.g. Kushner and Yin, 2003 or Benaïm, Hofbauer and Sorin, 2005) Introduce $\boldsymbol{z}_i$ instead of $\boldsymbol{z}$ and assume, that there exist compact and convex sets $\mathcal{Z}(\boldsymbol{\theta}(t)) \subseteq \mathbb{R}^p$, depending on $\boldsymbol{\theta}(t)$, such that

$$\lim_{n\to\infty} \inf_{\boldsymbol{w}\in\mathcal{Z}(\boldsymbol{\theta}(t))} \left\|\sum_{i=n+1}^{\nu(t_n+t)} \alpha_i\, \boldsymbol{z}_i(\boldsymbol{\theta}(t)) - \boldsymbol{w}\right\| = 0\,.$$

Then the limiting function can be characterized as an absolutely continuous function which fulfills

$$\frac{\mathrm{d}\boldsymbol{\theta}(t)}{\mathrm{d}t} \in \mathcal{Z}(\boldsymbol{\theta}(t))$$

for almost every $t \in \mathbb{R}$.

3. Asymptotic Behavior of the Maximum Likelihood Estimator

The main result of this section is the almost sure convergence of the sequence of maximum likelihood estimators. To achieve this, we will rewrite the sequence as a recursion, such that the methods described in Section 2.7 can be applied. This will be done in Section 3.2. Then, since the convergence and asymptotic behavior depends on the accumulated effects in this recursion, they will be considered. The "asymptotic rate of change condition" for the perturbations, which was introduced on page 23, will be verified.

Before considering the convergence of the estimator we will assure, that it exists asymptotically.

3.1. Asymptotic Existence of the MLE

As we have seen in Lemma 1 and Theorem 1, existence of the estimate is equivalent to the overlap of the design points, i.e. $\mathcal{C}_n^0 \cap \mathcal{C}_n^1 \neq \emptyset$. We will show in Lemma 5, that there will be an overlap eventually. A prerequisite is that there are 0's and 1's in every sequence of observations. Lemma 3 shows, that in fact there are infinitely many 0's and 1's in each sequence. We will denote the smallest and the largest probability of success by $c_{\bar{\boldsymbol{\theta}}} := \min_{\boldsymbol{x}\in\mathcal{X}} G(\boldsymbol{f}(\boldsymbol{x})^\top \bar{\boldsymbol{\theta}})$ and $C_{\bar{\boldsymbol{\theta}}} := \max_{\boldsymbol{x}\in\mathcal{X}} G(\boldsymbol{f}(\boldsymbol{x})^\top \bar{\boldsymbol{\theta}})$, respectively. Also introduce the error $\varepsilon_n := Y_n - G(\boldsymbol{f}(\boldsymbol{X}_n)^\top \bar{\boldsymbol{\theta}})$.

Lemma 3. *Let $0 < G(t) < 1$ for all $t \in \mathbb{R}$ and let $\mathcal{X}$ be compact. Then*

$$0 < c_{\bar{\boldsymbol{\theta}}} \leq \liminf_{n\to\infty} \frac{1}{n}\sum_{i=1}^n Y_i \leq \limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^n Y_i \leq C_{\bar{\boldsymbol{\theta}}} < 1$$

almost surely.

Proof. Consider

$$\frac{1}{n}\sum_{i=1}^n Y_i = \frac{1}{n}\sum_{i=1}^n \varepsilon_i + \frac{1}{n}\sum_{i=1}^n G(\boldsymbol{f}(\boldsymbol{X}_i)^\top \bar{\boldsymbol{\theta}})\,.$$

For the first sum on the right-hand side a strong law of large numbers for martingale differences yields

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \varepsilon_i = 0$$

almost surely. (see e.g. Theorem A.1 in the appendix, which is Theorem 2.18 in Hall and Heyde, 1980, p. 25) The limit of the second sum is bounded by $c_{\bar{\boldsymbol{\theta}}}$ and $C_{\bar{\boldsymbol{\theta}}}$:

$$c_{\bar{\boldsymbol{\theta}}} \leq \liminf_{n\to\infty} \frac{1}{n}\sum_{i=1}^n G(\boldsymbol{f}(\boldsymbol{X}_i)^\top \bar{\boldsymbol{\theta}}) \leq \limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^n G(\boldsymbol{f}(\boldsymbol{X}_i)^\top \bar{\boldsymbol{\theta}}) \leq C_{\bar{\boldsymbol{\theta}}}$$

almost surely. Hence

$$c_{\bar{\boldsymbol{\theta}}} \leq \liminf_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} Y_i \leq \limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} Y_i \leq C_{\bar{\boldsymbol{\theta}}}$$

almost surely. Since $0 < G(t) < 1$ for all $t \in \mathbb{R}$ and $\mathcal{X}$ is compact, $c_{\bar{\boldsymbol{\theta}}} > 0$ and $C_{\bar{\boldsymbol{\theta}}} < 1$. □

Lemma 3 can be extended to certain subsequences of $(Y_i)_{i\geq 1}$, which will be needed in the proof for Lemma 5. Let $\boldsymbol{v} \in \mathbb{R}^p$, $\|\boldsymbol{v}\| = 1$, and define the sets

$$\mathcal{A}^+_{\boldsymbol{v},n} := \Big\{i \in \{1,\dots,n\} | \boldsymbol{f}(\boldsymbol{X}_i)^\top \boldsymbol{v} > 0\Big\} \quad \text{and} \quad \mathcal{A}^-_{\boldsymbol{v},n} := \Big\{i \in \{1,\dots,n\} | \boldsymbol{f}(\boldsymbol{X}_i)^\top \boldsymbol{v} < 0\Big\}.$$

The following lemma is formulated for $|\mathcal{A}^-_{\boldsymbol{v},n}|$. The analog for $|\mathcal{A}^+_{\boldsymbol{v},n}|$ can be proved similarly.

Lemma 4. *Let $0 < G(t) < 1$ for all $t \in \mathbb{R}$ and let $\mathcal{X}$ be compact. Then*

$$0 < c_{\bar{\boldsymbol{\theta}}} \leq \liminf_{n\to\infty} \frac{1}{|\mathcal{A}^-_{\boldsymbol{v},n}|}\sum_{i\in\mathcal{A}^-_{\boldsymbol{v},n}} Y_i \leq \limsup_{n\to\infty} \frac{1}{|\mathcal{A}^-_{\boldsymbol{v},n}|}\sum_{i\in\mathcal{A}^-_{\boldsymbol{v},n}} Y_i \leq C_{\bar{\boldsymbol{\theta}}} < 1$$

on $\{\lim_{n\to\infty} |\mathcal{A}^-_{\boldsymbol{v},n}| = \infty\}$.

Proof. Since $i \in \mathcal{A}^-_{\boldsymbol{v},n}$ is equivalent to $\boldsymbol{f}(\boldsymbol{X}_i)^\top \boldsymbol{v} < 0$, we can rewrite the sum of the Y_i as

$$\sum_{i\in\mathcal{A}^-_{\boldsymbol{v},n}} Y_i = \sum_{i=1}^{n} Y_i \mathbb{1}_{\{\boldsymbol{f}(\boldsymbol{X}_i)^\top \boldsymbol{v} < 0\}}$$

and $|\mathcal{A}^-_{\boldsymbol{v},n}| = \sum_{i=1}^{n} \mathbb{1}_{\{\boldsymbol{f}(\boldsymbol{X}_i)^\top \boldsymbol{v} < 0\}}$. It follows that $|\mathcal{A}^-_{\boldsymbol{v},n}|$ is $\mathcal{F}_{n-1}$-measurable, because $\boldsymbol{X}_i$ is $\mathcal{F}_{i-1}$-measurable for all $i \geq 1$. Since the sum

$$\sum_{i=1}^{n} \varepsilon_i \mathbb{1}_{\{\boldsymbol{f}(\boldsymbol{X}_i)^\top \boldsymbol{v} < 0\}}$$

is a martingale and

$$\begin{aligned}\sum_{i=1}^{n} \frac{1}{|\mathcal{A}^-_{\boldsymbol{v},i}|^2} \mathrm{E}(\varepsilon_i^2 \mathbb{1}_{\{\boldsymbol{f}(\boldsymbol{X}_i)^\top \boldsymbol{v} < 0\}} | \mathcal{F}_{i-1}) \\ \leq \sum_{i=1}^{n} \frac{1}{|\mathcal{A}^-_{\boldsymbol{v},i}|^2} G(\boldsymbol{f}(\boldsymbol{X}_i)^\top \bar{\boldsymbol{\theta}})\Big(1 - G(\boldsymbol{f}(\boldsymbol{X}_i)^\top \bar{\boldsymbol{\theta}})\Big) \mathbb{1}_{\{\boldsymbol{f}(\boldsymbol{X}_i)^\top \boldsymbol{v} < 0\}} \leq \frac{1}{4}\sum_{j=1}^{\infty} \frac{1}{j^2} < \infty,\end{aligned}$$

it follows from Theorem A.1 that

$$\lim_{n\to\infty} \frac{1}{|\mathcal{A}^-_{\boldsymbol{v},n}|}\sum_{i=1}^{n} \varepsilon_i \mathbb{1}_{\{\boldsymbol{f}(\boldsymbol{X}_i)^\top \boldsymbol{v} < 0\}} = 0.$$

The rest follows as in Lemma 3. □

Lemma 5. *Let $0 < G(t) < 1$ for all $t \in \mathbb{R}$ and let $\mathcal{X}$ be compact. Assume that $\lim_{n\to\infty} \lambda_{\min}(\mathbf{F}_n^\top \mathbf{F}_n) = \infty$ almost surely. Then there exists an integer-valued random variable N with $\mathrm{P}(N < \infty) = 1$, such that*

$$\mathrm{P}\Big(\mathcal{C}_n^0 \cap \mathcal{C}_n^1 \neq \emptyset \quad \textit{for all} \quad n \geq N\Big) = 1.$$

Proof. If the vector $\boldsymbol{v}$ does separate the design points for $n \in \mathbb{N}$, then $Y_i = 0$ for all $i \in \mathcal{A}^-_{\boldsymbol{v},n}$ and $Y_i = 1$ for all $i \in \mathcal{A}^+_{\boldsymbol{v},n}$ or vice versa. But if $\lim_{n\to\infty} |\mathcal{A}^-_{\boldsymbol{v},n}| = \infty$, Lemma 4 yields, that there are infinitely many 0's and 1's in the subsequence defined by $\mathcal{A}^-_{\boldsymbol{v},n}$ and hence that

$$N^-_{\boldsymbol{v}} := \inf\{n \in \mathbb{N} | \exists i, j \in \mathcal{A}^-_{\boldsymbol{v},n}, i \neq j : Y_i = 0, Y_j = 1\}$$

is finite. The same holds for $\lim_{n\to\infty} |\mathcal{A}^+_{\boldsymbol{v},n}| = \infty$ and

$$N^+_{\boldsymbol{v}} := \inf\{n \in \mathbb{N} | \exists i, j \in \mathcal{A}^+_{\boldsymbol{v},n}, i \neq j : Y_i = 0, Y_j = 1\}\,.$$

It follows that $N_{\boldsymbol{v}} := \inf\{N^-_{\boldsymbol{v}}, N^+_{\boldsymbol{v}}\}$, which is the smallest index, such that $\boldsymbol{v}$ is not separating the design points anymore, is finite on the event

$$\left\{\lim_{n\to\infty} |\mathcal{A}^-_{\boldsymbol{v},n}| = \infty\right\} \cup \left\{\lim_{n\to\infty} |\mathcal{A}^+_{\boldsymbol{v},n}| = \infty\right\}.$$

We will show next, that this happens almost surely. Since the minimal eigenvalue of $\mathbf{F}_n^\top \mathbf{F}_n$ tends to infinity,

$$\lim_{n\to\infty} \boldsymbol{v}^\top \mathbf{F}_n^\top \mathbf{F}_n \boldsymbol{v} = \lim_{n\to\infty} \sum_{i=1}^{n} \left(\boldsymbol{f}(\boldsymbol{X}_i)^\top \boldsymbol{v}\right)^2 = \infty$$

almost surely for all $\|\boldsymbol{v}\| = 1$. Hence there are infinitely many indices, such that $(\boldsymbol{f}(\boldsymbol{X}_i)^\top \boldsymbol{v})^2 > 0$, i.e.

$$\mathrm{P}\left(\left\{\lim_{n\to\infty} |\mathcal{A}^-_{\boldsymbol{v},n}| = \infty\right\} \cup \left\{\lim_{n\to\infty} |\mathcal{A}^+_{\boldsymbol{v},n}| = \infty\right\}\right) = 1$$

for all $\|\boldsymbol{v}\| = 1$. Thus $\mathrm{P}(N_{\boldsymbol{v}} < \infty) = 1$ for all $\|\boldsymbol{v}\| = 1$. Denote a countable subset of $\{\boldsymbol{v} \in \mathbb{R}^p | \|\boldsymbol{v}\| = 1\}$ by $\mathcal{V}$. Then $\sup_{\boldsymbol{v}\in\mathcal{V}} N_{\boldsymbol{v}}$ is an (extended) integer valued random variable and it follows, that

$$\begin{aligned}
\mathrm{P}\Big(\sup_{\boldsymbol{v}\in\mathcal{V}} N_{\boldsymbol{v}} = \infty\Big) &= \mathrm{P}\Big(\bigcap_{n\geq 1} \{\sup_{\boldsymbol{v}\in\mathcal{V}} N_{\boldsymbol{v}} > n\}\Big) \\
&= 1 - \mathrm{P}\Big(\bigcup_{n\geq 1} \{\sup_{\boldsymbol{v}\in\mathcal{V}} N_{\boldsymbol{v}} \leq n\}\Big) \\
&= 1 - \mathrm{P}\Big(\bigcup_{n\geq 1} \{\sup_{\boldsymbol{v}\in\mathcal{V}} N_{\boldsymbol{v}} = n\}\Big) = 1 - \mathrm{P}\Big(\bigcup_{n\geq 1} \bigcup_{v\in\mathcal{V}} \{N_{\boldsymbol{v}} = n\}\Big)\,.
\end{aligned}$$

Since for any $\boldsymbol{v}_1 \in \mathcal{V}$

$$\mathrm{P}\Big(\bigcup_{v\in\mathcal{V}} \bigcup_{n\geq 1} \{N_{\boldsymbol{v}} = n\}\Big) \geq \mathrm{P}\Big(\bigcup_{n\geq 1} \{N_{\boldsymbol{v}_1} = n\}\Big) = 1\,,$$

we have

$$\mathrm{P}\Big(\sup_{\boldsymbol{v}\in\mathcal{V}} N_{\boldsymbol{v}} = \infty\Big) = 0$$

for all $\mathcal{V}$. It follows from Lemma A.8 (Chow and Teicher, 1988, p. 194) that $N := \operatorname{ess\,sup}_{\|\boldsymbol{v}\|=1} N_{\boldsymbol{v}}$ is almost surely finite, i.e. $\mathrm{P}(N < \infty) = 1$, and consequently

$$\mathrm{P}\Big(\mathcal{C}^0_n \cap \mathcal{C}^1_n \neq \emptyset \quad \text{for all} \quad n \geq N\Big) = 1.$$

□

3.2. An "Essentially Recursive" Formulation of the MLE

Now that we know, that the estimate will exist eventually, the next step will be to find a recursive formulation for the maximum likelihood estimate. We will see, that there is at least an "essentially recursive" formulation, which is given in Lemma 6 below.

This will be done using a Taylor expansion of G, namely

$$G(t) = G(t_0) + G'(t_0)(t - t_0) + r(t, t_0)\,, \tag{3.1}$$

where $r(t, t_0)$ is the approximation error. Notations associated with this are

$$\begin{aligned}
\boldsymbol{R}_n(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) &:= \Big(r(\boldsymbol{f}(\boldsymbol{x}_1)^\top \boldsymbol{\theta}_1, \boldsymbol{f}(\boldsymbol{x}_1)^\top \boldsymbol{\theta}_2) \quad \dots \quad r(\boldsymbol{f}(\boldsymbol{x}_n)^\top \boldsymbol{\theta}_1, \boldsymbol{f}(\boldsymbol{x}_n)^\top \boldsymbol{\theta}_2)\Big)^\top \\
\boldsymbol{r}_n(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) &:= \mathbf{F}_n^\top \boldsymbol{\Psi}_n(\mathbf{F}_n \boldsymbol{\theta}_2)\, \boldsymbol{R}_n(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\,.
\end{aligned}$$

The random errors for one observation are defined as before by

$$\varepsilon_n := Y_n - G(\boldsymbol{f}(\boldsymbol{X}_n)^\top \bar{\boldsymbol{\theta}})\,.$$

While the notation does not distinguish between realization and random variable, it will be clear from the context.

Additionally we introduce the following notations

$$\boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}, \boldsymbol{\theta}) := \boldsymbol{f}(\boldsymbol{x}) \psi(\boldsymbol{f}(\boldsymbol{x})^\top \boldsymbol{\theta}) \Big(G(\boldsymbol{f}(\boldsymbol{x})^\top \bar{\boldsymbol{\theta}}) - G(\boldsymbol{f}(\boldsymbol{x})^\top \boldsymbol{\theta})\Big)\,,$$

the (conditional) expectation of a summand of the score function, and

$$\tilde{\boldsymbol{s}}_n(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) := \mathbf{F}_n^\top \boldsymbol{\Psi}_n(\mathbf{F}_n \boldsymbol{\theta}_2)(\boldsymbol{y}_n - \boldsymbol{G}_n(\mathbf{F}_n \boldsymbol{\theta}_1))\,,$$

a "pseudo" score function, which will be needed for the recursion. Finally, whenever $\mathbf{I}(\hat{\boldsymbol{\theta}}_i, \mathbf{F}_{i+1})$ is nonsingular, we define the following sums for the accumulated effects of the recursion:

$$\begin{aligned}
\mathcal{G}_{m,n} &:= \sum_{i=m}^{n} \mathbf{I}(\hat{\boldsymbol{\theta}}_i, \mathbf{F}_{i+1})^{-1} \boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}_{i+1}, \hat{\boldsymbol{\theta}}_i) \\
\mathcal{E}_{m,n} &:= \sum_{i=m}^{n} \mathbf{I}(\hat{\boldsymbol{\theta}}_i, \mathbf{F}_{i+1})^{-1} \boldsymbol{f}(\boldsymbol{x}_{i+1}) \psi(\boldsymbol{f}(\boldsymbol{x}_{i+1})^\top \hat{\boldsymbol{\theta}}_i)\, \varepsilon_{i+1} \\
\widetilde{\mathcal{S}}_{m,n} &:= \sum_{i=m}^{n} \mathbf{I}(\hat{\boldsymbol{\theta}}_i, \mathbf{F}_{i+1})^{-1} \Big(\boldsymbol{s}_{i+1}(\hat{\boldsymbol{\theta}}_{i+1}) - \tilde{\boldsymbol{s}}_{i+1}(\hat{\boldsymbol{\theta}}_{i+1}, \hat{\boldsymbol{\theta}}_i)\Big) \\
\mathcal{R}_{m,n} &:= -\sum_{i=m}^{n} \mathbf{I}(\hat{\boldsymbol{\theta}}_i, \mathbf{F}_{i+1})^{-1} \boldsymbol{r}_{i+1}(\hat{\boldsymbol{\theta}}_{i+1}, \hat{\boldsymbol{\theta}}_i)\,.
\end{aligned}$$

Denote also $\lambda_n := \lambda_{\min}(\mathbf{F}_n^\top \mathbf{F}_n)$. As the step length for the stochastic algorithm we will choose λ_n^{-1}, which occurs naturally, whenever the norm of one of the sums is taken: If $\hat{\boldsymbol{\theta}}_n$ is bounded, then

$$\begin{aligned}
\|\mathbf{I}(\hat{\boldsymbol{\theta}}_n, \mathbf{F}_{n+1})^{-1}\| &= \|\mathbf{I}(\hat{\boldsymbol{\theta}}_n, \mathbf{F}_{n+1})^{-1} (\mathbf{F}_{n+1}^\top \mathbf{F}_{n+1})(\mathbf{F}_{n+1}^\top \mathbf{F}_{n+1})^{-1}\| \\
&\le \max_{i=1,\dots,n+1} \Big(d(\boldsymbol{f}(\boldsymbol{x}_i)^\top \hat{\boldsymbol{\theta}}_n)\Big)^{-1} \|(\mathbf{F}_{n+1}^\top \mathbf{F}_{n+1})^{-1}\| \le K \lambda_{n+1}^{-1}\,.
\end{aligned}$$

Hence the "natural time" defined on page 20 will become $t_n = \sum_{i=1}^{n} \lambda_n^{-1}$.

The following lemma is formulated for a sequence of estimates.

Lemma 6. *If there exists $m \in \mathbb{N}$ such that the maximum likelihood estimate $\hat{\boldsymbol{\theta}}_n$ exists and $\mathbf{I}(\hat{\boldsymbol{\theta}}_n, \mathbf{F}_{n+1})$ is nonsingular for all $n \geq m$, then*

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{n+1} = \hat{\boldsymbol{\theta}}_n + \mathbf{I}(\hat{\boldsymbol{\theta}}_n, \mathbf{F}_{n+1})^{-1}\boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}_{n+1}, \hat{\boldsymbol{\theta}}_n) + \mathbf{I}(\hat{\boldsymbol{\theta}}_n, \mathbf{F}_{n+1})^{-1}\boldsymbol{f}(\boldsymbol{x}_{n+1})\psi(\boldsymbol{f}(\boldsymbol{x}_{n+1})^\top\hat{\boldsymbol{\theta}}_n)\varepsilon_{n+1} \\ - \mathbf{I}(\hat{\boldsymbol{\theta}}_n, \mathbf{F}_{n+1})^{-1}\tilde{\boldsymbol{s}}_{n+1}(\hat{\boldsymbol{\theta}}_{n+1}, \hat{\boldsymbol{\theta}}_n) - \mathbf{I}(\hat{\boldsymbol{\theta}}_n, \mathbf{F}_{n+1})^{-1}\boldsymbol{r}_{n+1}(\hat{\boldsymbol{\theta}}_{n+1}, \hat{\boldsymbol{\theta}}_n) \quad (3.2)\end{aligned}$$

for all $n \geq m$.

Proof. Consider the score function from equation (2.17). A Taylor expansion of G as shown in (3.1), with $t_0 = \boldsymbol{f}(\boldsymbol{x}_i)^\top\hat{\boldsymbol{\theta}}_n$ and $t = \boldsymbol{f}(\boldsymbol{x}_i)^\top\hat{\boldsymbol{\theta}}_{n+1}$, yields

$$\begin{aligned}\boldsymbol{s}_{n+1}(\hat{\boldsymbol{\theta}}_n) &= \mathbf{F}_{n+1}^\top\boldsymbol{\Psi}_{n+1}(\mathbf{F}_{n+1}\hat{\boldsymbol{\theta}}_n)(\boldsymbol{y}_{n+1} - \boldsymbol{G}_{n+1}(\mathbf{F}_{n+1}\hat{\boldsymbol{\theta}}_{n+1})) \\ &\quad + \mathbf{I}(\hat{\boldsymbol{\theta}}_n, \mathbf{F}_{n+1})(\hat{\boldsymbol{\theta}}_{n+1} - \hat{\boldsymbol{\theta}}_n) + \mathbf{F}_{n+1}^\top\boldsymbol{\Psi}_{n+1}(\mathbf{F}_{n+1}\hat{\boldsymbol{\theta}}_n)\boldsymbol{R}_{n+1}(\hat{\boldsymbol{\theta}}_{n+1}, \hat{\boldsymbol{\theta}}_n) \\ &= \tilde{\boldsymbol{s}}_{n+1}(\hat{\boldsymbol{\theta}}_{n+1}, \hat{\boldsymbol{\theta}}_n) + \mathbf{I}(\hat{\boldsymbol{\theta}}_n, \mathbf{F}_{n+1})(\hat{\boldsymbol{\theta}}_{n+1} - \hat{\boldsymbol{\theta}}_n) + \boldsymbol{r}_{n+1}(\hat{\boldsymbol{\theta}}_{n+1}, \hat{\boldsymbol{\theta}}_n)\,.\end{aligned}$$

By rearranging the terms we obtain

$$\mathbf{I}(\hat{\boldsymbol{\theta}}_n, \mathbf{F}_{n+1})(\hat{\boldsymbol{\theta}}_{n+1} - \hat{\boldsymbol{\theta}}_n) = \boldsymbol{s}_{n+1}(\hat{\boldsymbol{\theta}}_n) - \tilde{\boldsymbol{s}}_{n+1}(\hat{\boldsymbol{\theta}}_{n+1}, \hat{\boldsymbol{\theta}}_n) - \boldsymbol{r}_{n+1}(\hat{\boldsymbol{\theta}}_{n+1}, \hat{\boldsymbol{\theta}}_n)\,.$$

Since the maximum likelihood estimator based on the first $n \geq m$ observations exists, we know that $\boldsymbol{s}_n(\hat{\boldsymbol{\theta}}_n) = 0$ and hence

$$\begin{aligned}\boldsymbol{s}_{n+1}(\hat{\boldsymbol{\theta}}_n) &= \boldsymbol{f}(\boldsymbol{x}_{n+1})\psi(\boldsymbol{f}(\boldsymbol{x}_{n+1})^\top\hat{\boldsymbol{\theta}}_n)(Y_{n+1} - G(\boldsymbol{f}(\boldsymbol{x}_{n+1})^\top\hat{\boldsymbol{\theta}}_n)) \\ &= \boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}_{n+1}, \hat{\boldsymbol{\theta}}_n) + \boldsymbol{f}(\boldsymbol{x}_{n+1})\psi(\boldsymbol{f}(\boldsymbol{x}_{n+1})^\top\hat{\boldsymbol{\theta}}_n)\varepsilon_{n+1}\,.\end{aligned}$$

Because $\mathbf{I}(\hat{\boldsymbol{\theta}}_n, \mathbf{F}_{n+1})$ is nonsingular the statement of the lemma follows. □

To prove convergence of the estimates, we have to consider the accumulated effects of the terms in the recursion, i.e. the sums of terms on the right-hand side of (3.2).

Applying (3.2) recursively yields

$$\hat{\boldsymbol{\theta}}_{n+1} = \hat{\boldsymbol{\theta}}_m + \mathcal{G}_{m,n} + \mathcal{E}_{m,n} + \widetilde{\mathcal{S}}_{m,n} + \mathcal{R}_{m,n}\,.$$

The last term on the right-hand side is a perturbation corresponding to the sum of $\boldsymbol{b}_n$ in Section 2.7. Together with $\mathcal{E}_{m,n}$, the sum of the random errors, $\mathcal{R}_{m,n}$ should vanish in the sense of the asymptotic rate of change condition. The only part influencing the asymptotic behavior of the mean and hence yielding the differential equation should be $\mathcal{G}_{m,n}$.

Example 5. In the logistic model $\psi(t) = 1$ for all $t \in \mathbb{R}$. Because of that two major simplifications happen: Not only

$$\boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}, \boldsymbol{\theta}) = \boldsymbol{f}(\boldsymbol{x})\Big(G(\boldsymbol{f}(\boldsymbol{x})^\top\bar{\boldsymbol{\theta}}) - G(\boldsymbol{f}(\boldsymbol{x})^\top\boldsymbol{\theta})\Big)\,,$$

but, more importantly, $\tilde{\boldsymbol{s}}_n$ becomes the score function, such that

$$\tilde{\boldsymbol{s}}_n(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}) = \boldsymbol{s}_n(\hat{\boldsymbol{\theta}}_n) = 0$$

for all $\boldsymbol{\theta} \in \Theta$. Consequently, $\widetilde{\mathcal{S}}_{m,n} = 0$ and

$$\hat{\boldsymbol{\theta}}_{n+1} = \hat{\boldsymbol{\theta}}_m + \mathcal{G}_{m,n} + \mathcal{E}_{m,n} + \mathcal{R}_{m,n}\,.$$

While $\tilde{\boldsymbol{s}}_n$ vanishes in the logit model, it will be present in other models and determining the behavior of the "recursion". One reason for the importance of $\tilde{\boldsymbol{s}}_n$ is its connection with the maximum of the log-likelihood. Assume that G is twice continuously differentiable and denote the first derivative of ψ by ψ'. The asymptotic behavior of $\tilde{\boldsymbol{s}}$ is strongly related to the asymptotic behavior of the Hessian matrix of the log-likelihood. With the mean value theorem applied to $\Psi_{n+1}(\mathbf{F}_{n+1}\boldsymbol{\theta}_n) - \Psi_{n+1}(\mathbf{F}_{n+1}\boldsymbol{\theta}_{n+1})$ we get

$$\begin{aligned}
&\tilde{\boldsymbol{s}}_{n+1}(\hat{\boldsymbol{\theta}}_{n+1}, \hat{\boldsymbol{\theta}}_n) - \boldsymbol{s}_{n+1}(\hat{\boldsymbol{\theta}}_{n+1}) \\
&\quad = \mathbf{F}_{n+1}(\boldsymbol{\Psi}_{n+1}(\mathbf{F}_{n+1}\hat{\boldsymbol{\theta}}_n) - \boldsymbol{\Psi}_{n+1}(\mathbf{F}_{n+1}\hat{\boldsymbol{\theta}}_{n+1}))(\boldsymbol{y}_{n+1} - \boldsymbol{G}_{n+1}(\mathbf{F}_{n+1}\hat{\boldsymbol{\theta}}_{n+1})) \\
&\quad = \left(\sum_{i=1}^{n+1} \psi'(\boldsymbol{f}(\boldsymbol{x}_i)^\top \boldsymbol{\theta}_n^*)(y_i - G(\boldsymbol{f}(\boldsymbol{x}_i)^\top \hat{\boldsymbol{\theta}}_{n+1}))\boldsymbol{f}(\boldsymbol{x}_i)\boldsymbol{f}(\boldsymbol{x}_i)^\top\right)(\hat{\boldsymbol{\theta}}_n - \hat{\boldsymbol{\theta}}_{n+1})\,,
\end{aligned}$$

where $\boldsymbol{\theta}_n^*$ is a value on the line segment connecting $\hat{\boldsymbol{\theta}}_{n+1}$ and $\hat{\boldsymbol{\theta}}_n$. The matrix on the right-hand side is very similar to the first part of the Hessian matrix, which was introduced in (2.19):

$$\mathbf{H}_{n+1}(\boldsymbol{\theta}, \mathbf{F}_{n+1}) = \sum_{i=1}^{n+1} \psi'(\boldsymbol{f}(\boldsymbol{x}_i)^\top \boldsymbol{\theta})(y_i - G(\boldsymbol{f}(\boldsymbol{x}_i)^\top \boldsymbol{\theta}))\boldsymbol{f}(\boldsymbol{x}_i)\boldsymbol{f}(\boldsymbol{x}_i)^\top - \mathbf{I}(\boldsymbol{\theta}, \mathbf{F}_{n+1})\,.$$

The Hessian matrix should be negative definite to assure at least a local maximum of the log-likelihood. If

$$\|\mathbf{I}(\boldsymbol{\theta}, \mathbf{F}_{n+1})^{-1} \sum_{i=1}^{n+1} \psi'(\boldsymbol{f}(\boldsymbol{x}_i)^\top \boldsymbol{\theta})(y_i - G(\boldsymbol{f}(\boldsymbol{x}_i)^\top \boldsymbol{\theta}))\boldsymbol{f}(\boldsymbol{x}_i)\boldsymbol{f}(\boldsymbol{x}_i)^\top\| \longrightarrow 0$$

for $n \longrightarrow \infty$ the Hessian matrix would be at least asymptotically negative definite. (compare Fahrmeir and Kaufmann, 1985). It would also imply that

$$\|\mathbf{I}(\boldsymbol{\theta}, \mathbf{F}_{n+1})^{-1}\tilde{\boldsymbol{s}}_{n+1}(\hat{\boldsymbol{\theta}}_{n+1}, \hat{\boldsymbol{\theta}}_n)\| = \|\mathbf{I}(\boldsymbol{\theta}, \mathbf{F}_{n+1})^{-1}(\tilde{\boldsymbol{s}}_{n+1}(\hat{\boldsymbol{\theta}}_{n+1}, \hat{\boldsymbol{\theta}}_n) - \boldsymbol{s}_{n+1}(\hat{\boldsymbol{\theta}}_{n+1}))\| \longrightarrow 0\,,$$

if the difference of consecutive estimates is bounded. In this sense $\tilde{\boldsymbol{s}}_n$ is a "measure" for the existence of a maximum.

Note that the Hessian arises directly in the recursion, if a slightly different approach is considered. From the mean value theorem

$$\boldsymbol{s}_{n+1}(\hat{\boldsymbol{\theta}}_n) - \boldsymbol{s}_{n+1}(\hat{\boldsymbol{\theta}}_{n+1}) = \mathbf{H}_{n+1}(\boldsymbol{\theta}_n^*, \mathbf{F}_{n+1})(\hat{\boldsymbol{\theta}}_n - \hat{\boldsymbol{\theta}}_{n+1})\,,$$

with $\boldsymbol{\theta}_n^*$ on the line segment connecting $\hat{\boldsymbol{\theta}}_n$ and $\hat{\boldsymbol{\theta}}_{n+1}$, it follows that

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_{n+1} - \hat{\boldsymbol{\theta}}_n &= -\mathbf{H}_{n+1}(\boldsymbol{\theta}_n^*, \mathbf{F}_{n+1})^{-1}\Big(\boldsymbol{s}_{n+1}(\hat{\boldsymbol{\theta}}_n) - \boldsymbol{s}_{n+1}(\hat{\boldsymbol{\theta}}_{n+1})\Big) \\
&= -\mathbf{H}_{n+1}(\boldsymbol{\theta}_n^*, \mathbf{F}_{n+1})^{-1}\boldsymbol{f}(\boldsymbol{x}_{n+1})\psi(\boldsymbol{f}(\boldsymbol{x}_{n+1})^\top \hat{\boldsymbol{\theta}}_n)(y_{n+1} - G(\boldsymbol{f}(\boldsymbol{x}_{n+1})^\top \hat{\boldsymbol{\theta}}_n))\,, \quad (3.3)
\end{aligned}$$

if the Hessian is invertible. Connecting this to our recursion, $\tilde{\boldsymbol{s}}_n$ can be considered as an error term from the expansion of $\mathbf{H}_{n+1}(\boldsymbol{\theta}_n^*, \mathbf{F}_{n+1})^{-1}$ around $\mathbf{I}(\hat{\boldsymbol{\theta}}_n, \mathbf{F}_{n+1})^{-1}$.

Equation (3.3) yields bounds for $\|\hat{\boldsymbol{\theta}}_{n+1} - \hat{\boldsymbol{\theta}}_n\|$, which can be used to show, that this difference tends to 0. The following lemma uses a condition similar to the condition (C^*) in Fahrmeir and Kaufmann (1985, p. 360), to ensure, that the Hessian matrix is negative definite.

Lemma 7. *Let $\Theta_0 \subseteq \Theta$ be compact. Assume that there exists an $m \in \mathbb{N}$ and a constant $C > 0$, such that*

$$-\mathbf{H}_{n+1}(\boldsymbol{\theta}, \mathbf{F}_{n+1}) \geq C\, \mathbf{I}(\boldsymbol{\theta}, \mathbf{F}_{n+1}) \tag{3.4}$$

for all $n \geq m$ and all $\boldsymbol{\theta} \in \Theta_0$. If $\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\theta}}_{n+1} \in \Theta_0$, then there exists $K > 0$, not depending on n, such that

$$\|\hat{\boldsymbol{\theta}}_{n+1} - \hat{\boldsymbol{\theta}}_n\| \leq K \frac{1}{\lambda_{n+1}}\,.$$

Proof. From the condition on the Hessian matrix in (3.4) we get, that

$$\lambda_{\min}(-\mathbf{H}_{n+1}(\boldsymbol{\theta}, \mathbf{F}_{n+1})) \geq C\, \lambda_{\min}(\mathbf{I}(\boldsymbol{\theta}, \mathbf{F}_{n+1}))$$

for $\boldsymbol{\theta} \in \Theta_0$. Together with (3.3) this yields

$$\|\hat{\boldsymbol{\theta}}_{n+1} - \hat{\boldsymbol{\theta}}_n\| \leq \frac{\|\boldsymbol{s}_{n+1}(\hat{\boldsymbol{\theta}}_n) - \boldsymbol{s}_{n+1}(\hat{\boldsymbol{\theta}}_{n+1})\|}{\lambda_{\min}(-\mathbf{H}_{n+1}(\boldsymbol{\theta}_n^*, \mathbf{F}_{n+1}))} \leq K_1 C^{-1} \frac{1}{\lambda_{\min}(\mathbf{I}(\boldsymbol{\theta}_n^*, \mathbf{F}_{n+1}))}\,,$$

where

$$K_1 := \max_{\boldsymbol{x} \in \mathcal{X}} \max_{\boldsymbol{\theta} \in \Theta_0} \|\boldsymbol{f}(\boldsymbol{x}) \psi(\boldsymbol{f}(\boldsymbol{x})^\top \boldsymbol{\theta})\|\,.$$

Using Lemma A.3 and the boundedness of the function d we obtain

$$\|\mathbf{I}(\boldsymbol{\theta}_n^*, \mathbf{F}_{n+1})^{-1} \mathbf{I}(\hat{\boldsymbol{\theta}}_n, \mathbf{F}_{n+1})\| \leq \max_{\boldsymbol{x} \in \mathcal{X}} \max_{\boldsymbol{\theta} \in \Theta_0} \frac{d(\boldsymbol{f}(\boldsymbol{x})^\top \hat{\boldsymbol{\theta}}_n)}{d(\boldsymbol{f}(\boldsymbol{x})^\top \boldsymbol{\theta})} \leq K_2\,,$$

where $K_2 > 0$ is independent of n. The statement of the lemma follows from

$$\begin{aligned}
\frac{1}{\lambda_{\min}(\mathbf{I}(\boldsymbol{\theta}_n^*, \mathbf{F}_{n+1}))} &= \|\mathbf{I}(\boldsymbol{\theta}_n^*, \mathbf{F}_{n+1})^{-1}\| \\
&= \|\mathbf{I}(\boldsymbol{\theta}_n^*, \mathbf{F}_{n+1})^{-1} \mathbf{I}(\hat{\boldsymbol{\theta}}_n, \mathbf{F}_{n+1}) \mathbf{I}(\hat{\boldsymbol{\theta}}_n, \mathbf{F}_{n+1})^{-1}\| \\
&\leq \|\mathbf{I}(\boldsymbol{\theta}_n^*, \mathbf{F}_{n+1})^{-1} \mathbf{I}(\hat{\boldsymbol{\theta}}_n, \mathbf{F}_{n+1})\| \|\mathbf{I}(\hat{\boldsymbol{\theta}}_n, \mathbf{F}_{n+1})^{-1}\| \leq \frac{K_2}{\lambda_{n+1}}
\end{aligned}$$

by choosing $K = K_1 K_2 C^{-1}$. □

A closer look at the condition in (3.4) shows that it is equivalent to

$$\sum_{i=1}^{n+1} (\boldsymbol{v}^\top \boldsymbol{f}(\boldsymbol{x}_i))^2 \Big(\psi'(\boldsymbol{f}(\boldsymbol{x}_i)^\top \boldsymbol{\theta})(y_i - G(\boldsymbol{f}(\boldsymbol{x}_i)^\top \boldsymbol{\theta})) - (1 - C) d(\boldsymbol{f}(\boldsymbol{x}_i)^\top \boldsymbol{\theta}) \Big) \leq 0\,. \tag{3.5}$$

A sufficient condition for (3.5) to hold is, if

$$\psi'(\boldsymbol{f}(\boldsymbol{x}_i)^\top \boldsymbol{\theta})(y_i - G(\boldsymbol{f}(\boldsymbol{x}_i)^\top \boldsymbol{\theta})) - (1 - C) d(\boldsymbol{f}(\boldsymbol{x}_i)^\top \boldsymbol{\theta}) \leq 0$$

for all $i = 1, \ldots, n+1$. A stronger condition is that

$$\psi'(t)(1 - G(t)) - (1 - C) d(t) \leq 0 \quad \text{and} \quad -\psi'(t) G(t) - (1 - C) d(t) \leq 0 \tag{3.6}$$

for all $t \in \mathbb{R}$, which can be written as $t = \boldsymbol{f}(\boldsymbol{x})^\top \boldsymbol{\theta}$ for some $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{\theta} \in \Theta_0$.

Example 6. All four models introduced in Example 1 satisfy the stronger condition (3.6). For the logit model this is obvious, because $\psi'(t) = 0$ for all $t \in \mathbb{R}$. It holds for all $0 < C < 1$. In the log-log, complementary log-log and probit model it holds for $C = 1/2$ and hence for all $0 < C \leq 1/2$. The details are given in Section B.2. Our first hint about the result comes from the log-concavity of the mean functions. The second derivative of $\log G(t)$ is given by

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} \log G(t) = \frac{G''(t)G(t) - G'(t)^2}{G(t)^2}$$

and has to be negative because of the log-concavity. Let us rewrite the first inequality in (3.6): It is equivalent to

$$\begin{aligned} 0 \geq G''(t)G(t)(1 - G(t)) - G'(t)^2(1 - 2G(t)) - G'(t)^2 G(t) \\ = \Big(G''(t)G(t) - G'(t)^2\Big)(1 - G(t)) \end{aligned}$$

and hence to the second derivative of $\log G(t)$ being negative. The same works for $\log(1 - G(t))$. Thus (3.6) is stronger then log-concavity.

3.3. The Localized Process and Behavior of the Accumulated Effects

One of the problems is to assure, that the estimates do not tend to infinity. To do that, we will first consider the "local" behavior of the sequence in some compact set.

Let $\Theta_0 \subset \Theta$ be compact and convex. The localized versions are defined by multiplying the indicator function $\mathbb{1}_{\Theta_0}(\hat{\boldsymbol{\theta}}_i)$ to each summand:

$$\begin{aligned} \mathcal{G}^{(0)}_{m,n} &:= \sum_{i=m}^{n} \mathbb{1}_{\Theta_0}(\hat{\boldsymbol{\theta}}_i) \mathbf{I}(\hat{\boldsymbol{\theta}}_i, \mathbf{F}_{i+1})^{-1} \boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}_{i+1}, \hat{\boldsymbol{\theta}}_i) \\ \mathcal{E}^{(0)}_{m,n} &:= \sum_{i=m}^{n} \mathbb{1}_{\Theta_0}(\hat{\boldsymbol{\theta}}_i) \mathbf{I}(\hat{\boldsymbol{\theta}}_i, \mathbf{F}_{i+1})^{-1} \boldsymbol{f}(\boldsymbol{x}_{i+1}) \psi(\boldsymbol{f}(\boldsymbol{x}_{i+1})^\top \hat{\boldsymbol{\theta}}_i)\, \varepsilon_{i+1} \\ \tilde{\mathcal{S}}^{(0)}_{m,n} &:= \sum_{i=m}^{n} \mathbb{1}_{\Theta_0}(\hat{\boldsymbol{\theta}}_i) \mathbf{I}(\hat{\boldsymbol{\theta}}_i, \mathbf{F}_{i+1})^{-1} \Big(\boldsymbol{s}_{i+1}(\hat{\boldsymbol{\theta}}_{i+1}) - \tilde{\boldsymbol{s}}_{i+1}(\hat{\boldsymbol{\theta}}_{i+1}, \hat{\boldsymbol{\theta}}_i)\Big) \\ \mathcal{R}^{(0)}_{m,n} &:= -\sum_{i=m}^{n} \mathbb{1}_{\Theta_0}(\hat{\boldsymbol{\theta}}_i) \mathbf{I}(\hat{\boldsymbol{\theta}}_i, \mathbf{F}_{i+1})^{-1} \boldsymbol{r}_{i+1}(\hat{\boldsymbol{\theta}}_{i+1}, \hat{\boldsymbol{\theta}}_i)\,. \end{aligned}$$

The part representing the mean function is $\mathcal{G}^{(0)}_{m,n}$. So this is the term, which should be equicontinuous in the extended sense (see page 22 for the definition) and hence yields the limiting equation. One goal is, to show, that the asymptotic rate of change condition (see Equation 2.54) holds for all other terms, i.e. for all $t > 0$

$$\begin{aligned} \lim_{n\to\infty} \sup \Big\{ \|\mathcal{E}^{(0)}_{n,k}\| : k = n, \ldots, \nu(t_n + t) - 1 \Big\} = 0 \\ \lim_{n\to\infty} \sup \Big\{ \|\tilde{\mathcal{S}}^{(0)}_{n,k}\| : k = n, \ldots, \nu(t_n + t) - 1 \Big\} = 0 \\ \lim_{n\to\infty} \sup \Big\{ \|\mathcal{R}^{(0)}_{n,k}\| : k = n, \ldots, \nu(t_n + t) - 1 \Big\} = 0\,. \end{aligned}$$

If the set Θ_0 is visited only finitely often, all of sums are 0 eventually, since the lower bound for the index increases. The asymptotic rate of change condition will be fulfilled. We will start with the observational errors in $\mathcal{E}^{(0)}_{m,n}$.

The Behavior of $\mathcal{E}^{(0)}_{m,n}$

The error terms ε_{n+1} are forming a martingale difference sequence with respect to the sequence of σ-fields $\mathcal{F}_n$ generated by $\boldsymbol{Y}_n$ and $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. The estimator $\hat{\boldsymbol{\theta}}_n$ and $\boldsymbol{X}_{n+1}$ are measurable with respect to $\mathcal{F}_n$. Thus, if $m \in \mathbb{N}$ is fixed, $\mathcal{E}^{(0)}_{m,n}$, $n \geq m+1$, is a martingale because

$$\mathrm{E}(\mathcal{E}^{(0)}_{m,n} | \mathcal{F}_n) = \mathcal{E}^{(0)}_{m,n-1} \,.$$

Fixing the lower bound is not a problem, because of the definition of the maximum likelihood estimator: If there is no maximum in Θ we choose a fixed value $\boldsymbol{\theta}_0$. (see (2.13) and the line below it) But by Lemma 5 there are only finitely many summands (almost surely), which contain this substitute and the qualitative behavior of $\mathcal{E}^{(0)}_{m,n}$, i.e. if the asymptotic rate of change condition holds, is not influenced by them.

The next lemma shows, that $\mathcal{E}^{(0)}_{m,n}$ converges almost surely, for $n \longrightarrow \infty$ and consequently the asymptotic rate of change condition is fulfilled. The weaker condition in part **(ii)** is very close to the one in Lai and Wei (1982) and Chen et al. (1999). That $\mathbf{F}_n^\top \mathbf{F}_n$ is nonsingular for all $n \geq m$, can be easily implemented by the choice of the initial design.

Lemma 8. *Let $\mathcal{X}$ and Θ_0 be compact. Let G be continuously differentiable, strictly increasing and $0 < G(t) < 1$ for all $t \in \mathbb{R}$. Let $\mathbf{F}_n^\top \mathbf{F}_n$ be nonsingular for all $n \geq m$ almost surely and assume that $\lim_{n\to\infty} \lambda_n = \infty$ almost surely.*

(i) *If there exists $K > 0$ such that*

$$\frac{\lambda_{\max}(\mathbf{F}_n^\top \mathbf{F}_n)}{\lambda_{\min}(\mathbf{F}_n^\top \mathbf{F}_n)^\delta} \leq K$$

for some $\delta \geq 1$ and all $n \geq m$, then $\mathcal{E}^{(0)}_{m,n}$ converges almost surely.

(ii) *If there exists $K > 0$ such that*

$$\frac{\Big(\log(\lambda_{\max}(\mathbf{F}_n^\top \mathbf{F}_n))\Big)^\delta}{\lambda_{\min}(\mathbf{F}_n^\top \mathbf{F}_n)} \leq K$$

for some $\delta > 1$ and all $n \geq m$, then $\mathcal{E}^{(0)}_{m,n}$ converges almost surely.

Proof. Before we start with the proof itself, we will state some facts, which will help us later to drop the dependence on the sequence of estimators $(\hat{\boldsymbol{\theta}}_n)_{n\geq m}$. Define the constants

$$K_1 := \min_{\boldsymbol{\theta}\in\Theta_0} \min_{\boldsymbol{x}\in\mathcal{X}} d(\boldsymbol{f}(\boldsymbol{x})^\top \boldsymbol{\theta}) \quad \text{and} \quad K_2 := \max_{\boldsymbol{\theta}\in\Theta_0} \max_{\boldsymbol{x}\in\mathcal{X}} d(\boldsymbol{f}(\boldsymbol{x})^\top \boldsymbol{\theta})$$

and note, that

$$K_1 \boldsymbol{v}^\top \mathbf{F}_n^\top \mathbf{F}_n \boldsymbol{v} \leq \boldsymbol{v}^\top \mathbf{I}(\boldsymbol{\theta}, \mathbf{F}_n) \boldsymbol{v} \leq K_2 \boldsymbol{v}^\top \mathbf{F}_n^\top \mathbf{F}_n \boldsymbol{v} \tag{3.7}$$

for all $\boldsymbol{\theta} \in \Theta_0$ and all $\boldsymbol{v} \in \mathbb{R}^p$. Because $\mathcal{X}$ and Θ_0 are compact it follows, that also

$$\max_{\boldsymbol{\theta}\in\Theta_0} \max_{\boldsymbol{x}\in\mathcal{X}} |\boldsymbol{f}(\boldsymbol{x})^\top \boldsymbol{\theta}| < \infty \,,$$

and consequently $0 < K_1 < K_2 < \infty$, due to the properties of G. Hence if $\mathbf{F}_n^\top \mathbf{F}_n$ is nonsingular, so is $\mathbf{I}(\boldsymbol{\theta}, \mathbf{F}_n)$.

For showing convergence in **(i)**, we notice, that a sequence of random vectors converges almost surely if and only if all its components converge almost surely. Let $\boldsymbol{v} \in \mathbb{R}^p$, $\|\boldsymbol{v}\| = 1$. Then

$$\boldsymbol{v}^\top \sum_{i=m}^{n} \mathbb{1}_{\Theta_0}(\hat{\boldsymbol{\theta}}_i)\mathbf{I}(\hat{\boldsymbol{\theta}}_i, \mathbf{F}_{i+1})^{-1}\boldsymbol{f}(\boldsymbol{X}_{i+1})\,\psi(\boldsymbol{f}(\boldsymbol{X}_{i+1})^\top\hat{\boldsymbol{\theta}}_i)\,\varepsilon_{i+1} \tag{3.8}$$

is a martingale and converges almost surely if

$$\sum_{i=m}^{n} \mathrm{E}\Big(\Big(\mathbb{1}_{\Theta_0}(\hat{\boldsymbol{\theta}}_i)\boldsymbol{v}^\top\mathbf{I}(\hat{\boldsymbol{\theta}}_i, \mathbf{F}_{i+1})^{-1}\boldsymbol{f}(\boldsymbol{X}_{i+1})\,\psi(\boldsymbol{f}(\boldsymbol{X}_{i+1})^\top\hat{\boldsymbol{\theta}}_i)\,\varepsilon_{i+1}\Big)^2\Big|\mathcal{F}_i\Big)$$
$$= \sum_{i=m}^{n} \Big(\mathbb{1}_{\Theta_0}(\hat{\boldsymbol{\theta}}_i)\boldsymbol{v}^\top\mathbf{I}(\hat{\boldsymbol{\theta}}_i, \mathbf{F}_{i+1})^{-1}\boldsymbol{f}(\boldsymbol{X}_{i+1})\,\psi(\boldsymbol{f}(\boldsymbol{X}_{i+1})^\top\hat{\boldsymbol{\theta}}_i)\Big)^2\mathrm{E}(\varepsilon_{i+1}^2|\mathcal{F}_i)$$

converges almost surely. (see e.g. Hall and Heyde, 1980, p. 35, Theorem 2.17) Since $\mathrm{E}(\varepsilon_{i+1}^2|\mathcal{F}_i)$ and, for $\hat{\boldsymbol{\theta}}_i \in \Theta_0$, $\psi(\boldsymbol{f}(\boldsymbol{x}_{i+1})^\top\hat{\boldsymbol{\theta}}_i)$ are bounded uniformly in i, this is equivalent to the convergence of

$$\sum_{i=m}^{n} \mathbb{1}_{\Theta_0}(\hat{\boldsymbol{\theta}}_i)\Big(\boldsymbol{v}^\top\mathbf{I}(\hat{\boldsymbol{\theta}}_i, \mathbf{F}_{i+1})^{-1}\boldsymbol{f}(\boldsymbol{X}_{i+1})\Big)^2. \tag{3.9}$$

With the Cauchy-Schwarz inequality follows

$$\Big(\boldsymbol{v}^\top\mathbf{I}(\hat{\boldsymbol{\theta}}_i, \mathbf{F}_{i+1})^{-1}\boldsymbol{f}(\boldsymbol{X}_{i+1})\Big)^2 \le \boldsymbol{v}^\top\mathbf{I}(\hat{\boldsymbol{\theta}}_i, \mathbf{F}_{i+1})^{-1}\boldsymbol{v}\;\boldsymbol{f}(\boldsymbol{X}_{i+1})^\top\mathbf{I}(\hat{\boldsymbol{\theta}}_i, \mathbf{F}_{i+1})^{-1}\boldsymbol{f}(\boldsymbol{X}_{i+1}). \tag{3.10}$$

The first quadratic form on the right-hand side is bounded by the largest eigenvalue of $\mathbf{I}(\hat{\boldsymbol{\theta}}_i, \mathbf{F}_{i+1})^{-1}$:

$$\boldsymbol{v}^\top\mathbf{I}(\hat{\boldsymbol{\theta}}_i, \mathbf{F}_{i+1})^{-1}\boldsymbol{v} \le \lambda_{\max}(\mathbf{I}(\hat{\boldsymbol{\theta}}_i, \mathbf{F}_{i+1})^{-1}) \le \lambda_{\min}(\mathbf{I}(\hat{\boldsymbol{\theta}}_i, \mathbf{F}_{i+1}))^{-1}.$$

Together with (3.7) and the condition on the eigenvalues of $\mathbf{F}_i^\top\mathbf{F}_i$ follows

$$\boldsymbol{v}^\top\mathbf{I}(\hat{\boldsymbol{\theta}}_i, \mathbf{F}_{i+1})^{-1}\boldsymbol{v} \le \frac{1}{K_1}\lambda_{\min}(\mathbf{F}_{i+1}^\top\mathbf{F}_{i+1})^{-1} \le \frac{K}{K_1}\lambda_{\max}(\mathbf{F}_{i+1}^\top\mathbf{F}_{i+1})^{-1/\delta}.$$

Because $\mathbf{F}_{i+1}^\top\mathbf{F}_{i+1}$ is a positive definite $p \times p$-matrix

$$\det(\mathbf{F}_{i+1}^\top\mathbf{F}_{i+1}) \le \lambda_{\max}(\mathbf{F}_{i+1}^\top\mathbf{F}_{i+1})^p$$

and hence

$$\boldsymbol{v}^\top\mathbf{I}(\hat{\boldsymbol{\theta}}_i, \mathbf{F}_{i+1})^{-1}\boldsymbol{v} \le \frac{K}{K_1}\det(\mathbf{F}_{i+1}^\top\mathbf{F}_{i+1})^{-1/(p\delta)}. \tag{3.11}$$

Consider the second quadratic form on the right-hand side of (3.10). The inequalities in (3.7) and an application of Lemma A.2 (see also Lemma 2 (ii) in Lai and Wei, 1982) yield

$$\boldsymbol{f}(\boldsymbol{X}_{i+1})^\top\mathbf{I}(\hat{\boldsymbol{\theta}}_i, \mathbf{F}_{i+1})^{-1}\boldsymbol{f}(\boldsymbol{X}_{i+1}) \le K_2\,\boldsymbol{f}(\boldsymbol{X}_{i+1})^\top(\mathbf{F}_{i+1}^\top\mathbf{F}_{i+1})^{-1}\boldsymbol{f}(\boldsymbol{X}_{i+1})$$
$$= K_2\,\frac{\det(\mathbf{F}_{i+1}^\top\mathbf{F}_{i+1}) - \det(\mathbf{F}_i^\top\mathbf{F}_i)}{\det(\mathbf{F}_{i+1}^\top\mathbf{F}_{i+1})}. \tag{3.12}$$

Also note, that

$$\det(\mathbf{F}_{i+1}^\top \mathbf{F}_{i+1}) \geq \det(\mathbf{F}_{i+1}^\top \mathbf{F}_{i+1}) - \det(\mathbf{F}_m^\top \mathbf{F}_m) = \sum_{j=m}^{i} \Big(\det(\mathbf{F}_{j+1}^\top \mathbf{F}_{j+1}) - \det(\mathbf{F}_j^\top \mathbf{F}_j)\Big)\,, \tag{3.13}$$

which in combination with (3.11) and (3.12) leads to

$$(\boldsymbol{v}^\top \mathbf{I}(\hat{\boldsymbol{\theta}}_i, \mathbf{F}_{i+1})^{-1} \boldsymbol{f}(\boldsymbol{X}_{i+1}))^2 \leq \frac{KK_2}{K_1} \frac{\det(\mathbf{F}_{i+1}^\top \mathbf{F}_{i+1}) - \det(\mathbf{F}_i^\top \mathbf{F}_i)}{\Big(\sum_{j=m}^{i} \big(\det(\mathbf{F}_{j+1}^\top \mathbf{F}_{j+1}) - \det(\mathbf{F}_j^\top \mathbf{F}_j)\big)\Big)^{1+1/(p\delta)}}\,.$$

After summation over i, the Abel-Dini theorem (see e.g. Knopp, 1996, p. 299) yields convergence of

$$\sum_{i=m}^{n} \frac{\det(\mathbf{F}_{i+1}^\top \mathbf{F}_{i+1}) - \det(\mathbf{F}_i^\top \mathbf{F}_i)}{\Big(\sum_{j=m}^{i} \big(\det(\mathbf{F}_{j+1}^\top \mathbf{F}_{j+1}) - \det(\mathbf{F}_j^\top \mathbf{F}_j)\big)\Big)^{1+1/(p\delta)}}\,. \tag{3.14}$$

Hence (3.9) and with that (3.8) converge almost surely. Since (3.8) converges for all $\|\boldsymbol{v}\| = 1$ and uniformly in $\boldsymbol{v}$, the almost sure convergence of $\mathcal{E}_{m,n}^{(0)}$ follows.

With the less restrictive condition **(ii)**, the convergence follows similarly. Note, that $\delta > 1$ in this part. The inequality (3.11) becomes

$$\boldsymbol{v}^\top \mathbf{I}(\hat{\boldsymbol{\theta}}_i, \mathbf{F}_{i+1})^{-1} \boldsymbol{v} \leq \frac{p^\delta K}{K_1} \Big(\log(\det(\mathbf{F}_{i+1}^\top \mathbf{F}_{i+1}))\Big)^{-\delta}\,.$$

and hence

$$\Big(\boldsymbol{v}^\top \mathbf{I}(\hat{\boldsymbol{\theta}}_i, \mathbf{F}_{i+1})^{-1} \boldsymbol{f}(\boldsymbol{X}_{i+1})\Big)^2 \leq \frac{p^r KK_2}{K_1} \frac{\det(\mathbf{F}_{i+1}^\top \mathbf{F}_{i+1}) - \det(\mathbf{F}_i^\top \mathbf{F}_i)}{\det(\mathbf{F}_{i+1}^\top \mathbf{F}_{i+1}) \Big(\log(\det(\mathbf{F}_{i+1}^\top \mathbf{F}_{i+1}))\Big)^{\delta}}\,.$$

An upper bound follows as before by replacing the determinants in the denominator using (3.13). The convergence of the sum follows again by the Abel-Dini Theorem: If in (3.14) the exponent in the denominator is equal to 1, the sums tend to ∞ as a consequence of the Abel-Dini Theorem. In fact they tend to ∞ like $\log\Big(\sum_{i=m}^{n} \big(\det(\mathbf{F}_{i+1}^\top \mathbf{F}_{i+1}) - \det(\mathbf{F}_i^\top \mathbf{F}_i)\big)\Big)$, in the sense that

$$\lim_{n\to\infty} \sum_{i=m}^{n} \frac{\det(\mathbf{F}_{i+1}^\top \mathbf{F}_{i+1}) - \det(\mathbf{F}_i^\top \mathbf{F}_i)}{\sum_{j=m}^{i} \big(\det(\mathbf{F}_{j+1}^\top \mathbf{F}_{j+1}) - \det(\mathbf{F}_j^\top \mathbf{F}_j)\big)} \Big/ \log\Big(\sum_{i=m}^{n} \big(\det(\mathbf{F}_{i+1}^\top \mathbf{F}_{i+1}) - \det(\mathbf{F}_i^\top \mathbf{F}_i)\big)\Big) = 1\,.$$

(see the "Satz" on page 301 in Knopp, 1996) Since the Abel-Dini-Theorem still holds, if parts are substituted with terms, which are asymptotically equivalent in this sense, the convergence follows. □

The Behavior of $\mathcal{R}_{m,n}^{(0)}$

Since $\boldsymbol{r}_{n+1}$ consists mainly of errors from the Taylor expansion, we can expect, that it is small if the estimates are sufficiently close. The lemma is formulated for one sample path.

Lemma 9. *Let $\mathcal{X}$ and Θ_0 be compact. Let $\mathbf{F}_{n+1}^\top \mathbf{F}_{n+1}$ be nonsingular and $\hat{\boldsymbol{\theta}}_n \in \Theta_0$. Let G be twice continuously differentiable and strictly increasing. Then*

$$\|\mathbf{I}(\hat{\boldsymbol{\theta}}_n, \mathbf{F}_{n+1})^{-1} \boldsymbol{r}_{n+1}(\hat{\boldsymbol{\theta}}_{n+1}, \hat{\boldsymbol{\theta}}_n)\| \leq K \|\hat{\boldsymbol{\theta}}_{n+1} - \hat{\boldsymbol{\theta}}_n\|^2 .$$

for some $K > 0$, uniformly in n.

Proof. For this proof, introduce the notation

$$\tilde{r}_{i,n} := \int_0^1 G'\Big(\boldsymbol{f}(\boldsymbol{x}_i)^\top (\hat{\boldsymbol{\theta}}_n + s\,(\hat{\boldsymbol{\theta}}_{n+1} - \hat{\boldsymbol{\theta}}_n))\Big) - G'\Big(\boldsymbol{f}(\boldsymbol{x}_i)^\top \hat{\boldsymbol{\theta}}_n\Big) \mathrm{d}s$$

and the diagonal matrix

$$\tilde{\mathbf{R}}_{n+1} := \operatorname*{diag}_{i=1,\ldots,n+1} (\tilde{r}_{i,n}) .$$

We can write the error r, which originated in the Taylor expansion of G, as

$$r(\boldsymbol{f}(\boldsymbol{x}_i)^\top \hat{\boldsymbol{\theta}}_{n+1}, \boldsymbol{f}(\boldsymbol{x}_i)^\top \hat{\boldsymbol{\theta}}_n) = \tilde{r}_{i,n}\, \boldsymbol{f}(\boldsymbol{x}_i)^\top (\hat{\boldsymbol{\theta}}_{n+1} - \hat{\boldsymbol{\theta}}_n)$$

and $\boldsymbol{r}_{n+1}$ becomes

$$\begin{aligned} \boldsymbol{r}_{n+1}(\hat{\boldsymbol{\theta}}_{n+1}, \hat{\boldsymbol{\theta}}_n) &= \sum_{i=1}^{n+1} \tilde{r}_{i,n} \psi(\boldsymbol{f}(\boldsymbol{x}_i)^\top \hat{\boldsymbol{\theta}}_n) \boldsymbol{f}(\boldsymbol{x}_i) \boldsymbol{f}(\boldsymbol{x}_i)^\top (\hat{\boldsymbol{\theta}}_{n+1} - \hat{\boldsymbol{\theta}}_n) \\ &= \mathbf{F}_{n+1}^\top \boldsymbol{\Psi}_{n+1}(\mathbf{F}_{n+1} \hat{\boldsymbol{\theta}}_n) \tilde{\mathbf{R}}_{n+1} \mathbf{F}_{n+1} (\hat{\boldsymbol{\theta}}_{n+1} - \hat{\boldsymbol{\theta}}_n) . \end{aligned}$$

It follows with Lemma A.4 that

$$\|\mathbf{I}(\hat{\boldsymbol{\theta}}_n, \mathbf{F}_{n+1})^{-1} \boldsymbol{r}_{n+1}(\hat{\boldsymbol{\theta}}_{n+1}, \hat{\boldsymbol{\theta}}_n)\| \leq 2 \|\hat{\boldsymbol{\theta}}_{n+1} - \hat{\boldsymbol{\theta}}_n\| \max_{i=1,\ldots,n+1} \left| \frac{\tilde{r}_{i,n}}{G'(\boldsymbol{f}(\boldsymbol{x}_i)^\top \hat{\boldsymbol{\theta}}_n)} \right| . \tag{3.15}$$

As a consequence of the mean value theorem the $\tilde{r}_{i,n}$ are bounded above:

$$\begin{aligned} |\tilde{r}_{i,n}| &\leq \max_{0 \leq s \leq 1} \left| G'\Big(\boldsymbol{f}(\boldsymbol{x}_i)^\top (\hat{\boldsymbol{\theta}}_n + s\,(\hat{\boldsymbol{\theta}}_{n+1} - \hat{\boldsymbol{\theta}}_n))\Big) - G'\Big(\boldsymbol{f}(\boldsymbol{x}_i)^\top \hat{\boldsymbol{\theta}}_n\Big) \right| \\ &\leq \left| \boldsymbol{f}(\boldsymbol{x}_i)^\top (\hat{\boldsymbol{\theta}}_{n+1} - \hat{\boldsymbol{\theta}}_n) \right| \max_{0 \leq s \leq 1} \left| G''\Big(\boldsymbol{f}(\boldsymbol{x}_i)^\top (\hat{\boldsymbol{\theta}}_n + s\,(\hat{\boldsymbol{\theta}}_{n+1} - \hat{\boldsymbol{\theta}}_n))\Big) \right| . \end{aligned} \tag{3.16}$$

The second derivative $G''(t)$ is continuous and tends to 0 for $|t| \longrightarrow \infty$. Consequently $|G''(t)|$ is bounded on $\mathbb{R}$. Since $\mathcal{X}$ and Θ_0 are compact, it follows from the assumptions on G that

$$K := 2 \max_{\boldsymbol{\theta} \in \Theta_0} \max_{\boldsymbol{x} \in \mathcal{X}} \max_{t \in \mathbb{R}} \frac{\|\boldsymbol{f}(\boldsymbol{x})\|\, |G''(t)|}{G'(\boldsymbol{f}(\boldsymbol{x})^\top \boldsymbol{\theta})} < \infty$$

and in combination with (3.15) and (3.16)

$$\|\mathbf{I}(\hat{\boldsymbol{\theta}}_n, \mathbf{F}_{n+1})^{-1} \boldsymbol{r}_{n+1}(\hat{\boldsymbol{\theta}}_{n+1}, \hat{\boldsymbol{\theta}}_n)\| \leq K \|\hat{\boldsymbol{\theta}}_{n+1} - \hat{\boldsymbol{\theta}}_n\|^2 .$$

□

As a direct consequence follows, that the asymptotic rate of change condition holds:

Lemma 10. *Additionally to the assumptions of Lemma 9 assume, that for* $\hat{\boldsymbol{\theta}}_n \in \Theta_0$

$$\|\hat{\boldsymbol{\theta}}_{n+1} - \hat{\boldsymbol{\theta}}_n\| \leq C\frac{1}{\lambda_{n+1}},$$

then

$$\lim_{n\to\infty} \sup\left\{\|\mathcal{R}_{n,k}^{(0)}\| : k = n, \ldots, \nu(t_n + t) - 1\right\} = 0$$

for all $t > 0$.

Proof. From Lemma 9 and the condition on $\|\hat{\boldsymbol{\theta}}_{n+1} - \hat{\boldsymbol{\theta}}_n\|$ there exists a constant $K > 0$, such that

$$\|\mathcal{R}_{n,k}^{(0)}\| \leq \sum_{i=n}^{k} \mathbb{1}_{\Theta_0}(\hat{\boldsymbol{\theta}}_i)\left\|\mathbf{I}(\hat{\boldsymbol{\theta}}_i, \mathbf{F}_{i+1})^{-1}\boldsymbol{r}_{i+1}\left(\hat{\boldsymbol{\theta}}_{i+1}, \hat{\boldsymbol{\theta}}_i\right)\right\| \leq \sum_{i=n}^{k} K\,\frac{1}{\lambda_{i+1}^2}$$

Since the λ_i are positive and not decreasing

$$\sup\left\{\left\|\mathcal{R}_{n,k}^{(0)}\right\| : k = n, \ldots, \nu(t_n + t) - 1\right\} \leq \frac{K}{\lambda_{n+1}} \sum_{i=n+1}^{\nu(t_n+t)} \frac{1}{\lambda_i}.$$

By definition of ν and t_n

$$\sum_{i=n+1}^{\nu(t_n+t)} \frac{1}{\lambda_i} \leq t_n + t - t_n = t$$

and hence

$$\sup\left\{\|\mathcal{R}_{n,k}^{(0)}\| : k = n, \ldots, \nu(t_n + t) - 1\right\} \leq \frac{K}{\lambda_{n+1}} t$$

which tends to 0 for $n \longrightarrow \infty$ and all $t > 0$. □

The Behavior of $\widetilde{\mathcal{S}}_{m,n}^{(0)}$

Lemma 11. *Let* $\mathcal{X}$ *and* Θ_0 *be compact. Let* $\mathbf{F}_{n+1}^\top \mathbf{F}_{n+1}$ *be nonsingular. Let* G *be twice continuously differentiable, strictly increasing and* $0 < G(t) < 1$ *for all* $t \in \mathbb{R}$. *Let further be* $\hat{\boldsymbol{\theta}}_n \in \Theta_0$ *and* $\|\hat{\boldsymbol{\theta}}_{n+1} - \hat{\boldsymbol{\theta}}_{n+1}\| \leq C$ *for some* $C > 0$, *then*

$$\left\|\mathbf{I}(\hat{\boldsymbol{\theta}}_n, \mathbf{F}_{n+1})^{-1}\left(\tilde{\boldsymbol{s}}_{n+1}(\hat{\boldsymbol{\theta}}_{n+1}, \hat{\boldsymbol{\theta}}_n) - \boldsymbol{s}_{n+1}(\hat{\boldsymbol{\theta}}_{n+1})\right)\right\| \leq K\|\hat{\boldsymbol{\theta}}_{n+1} - \hat{\boldsymbol{\theta}}_n\|$$

for some $K > 0$, *independent of* n.

Proof. Since G is twice continuously differentiable, ψ is continuously differentiable. It follows from the mean value theorem, that

$$\psi(\boldsymbol{f}(\boldsymbol{x}_i)^\top \hat{\boldsymbol{\theta}}_n) - \psi(\boldsymbol{f}(\boldsymbol{x}_i)^\top \hat{\boldsymbol{\theta}}_{n+1}) = \psi'(\boldsymbol{f}(\boldsymbol{x}_i)^\top \hat{\boldsymbol{\theta}}_{i,n}^*)\boldsymbol{f}(\boldsymbol{x}_i)^\top(\hat{\boldsymbol{\theta}}_n - \hat{\boldsymbol{\theta}}_{n+1})$$

for some $\hat{\boldsymbol{\theta}}_{i,n}^*$ on the line segment between $\hat{\boldsymbol{\theta}}_n$ and $\hat{\boldsymbol{\theta}}_{n+1}$. Hence

$$\begin{aligned}\tilde{\boldsymbol{s}}_{n+1}(\hat{\boldsymbol{\theta}}_{n+1}, \hat{\boldsymbol{\theta}}_n) - \boldsymbol{s}_{n+1}(\hat{\boldsymbol{\theta}}_{n+1}) \\ = \sum_{i=1}^{n+1} \left(y_i - G(\boldsymbol{f}(\boldsymbol{x}_i)^\top \hat{\boldsymbol{\theta}}_{n+1})\right)\psi'(\boldsymbol{f}(\boldsymbol{x}_i)^\top \hat{\boldsymbol{\theta}}_{i,n}^*)\boldsymbol{f}(\boldsymbol{x}_i)\boldsymbol{f}(\boldsymbol{x}_i)^\top(\hat{\boldsymbol{\theta}}_n - \hat{\boldsymbol{\theta}}_{n+1}) \quad (3.17)\end{aligned}$$

The sum on the right-hand side is a matrix of the form $\mathbf{F}_{n+1}^\top \mathbf{A}_{n+1}\mathbf{F}_{n+1}$, where $\mathbf{A}_{n+1}$ is a diagonal matrix with entries $\Big(y_i - G(\boldsymbol{f}(\boldsymbol{x}_i)^\top \hat{\boldsymbol{\theta}}_{n+1})\Big)\psi'(\boldsymbol{f}(\boldsymbol{x}_i)^\top \hat{\boldsymbol{\theta}}^*_{i,n})$. With Lemma A.4 follows

$$\begin{aligned}\Big\|\mathbf{I}(\hat{\boldsymbol{\theta}}_n, \mathbf{F}_{n+1})^{-1}\Big(\tilde{\boldsymbol{s}}_{n+1}(\hat{\boldsymbol{\theta}}_{n+1}, \hat{\boldsymbol{\theta}}_n) - \boldsymbol{s}_{n+1}(\hat{\boldsymbol{\theta}}_{n+1})\Big)\Big\| \\ \leq 2\|\hat{\boldsymbol{\theta}}_{n+1} - \hat{\boldsymbol{\theta}}_n\| \max_{i=1,\dots,n+1}\left|\frac{\Big(y_i - G(\boldsymbol{f}(\boldsymbol{x}_i)^\top \hat{\boldsymbol{\theta}}_{n+1})\Big)\psi'(\boldsymbol{f}(\boldsymbol{x}_i)^\top \hat{\boldsymbol{\theta}}^*_{i,n})}{d(\boldsymbol{f}(\boldsymbol{x}_i)^\top \hat{\boldsymbol{\theta}}_n)}\right|.\end{aligned}$$

Since ψ' is continuous and

$$\Big|y_i - G(\boldsymbol{f}(\boldsymbol{x}_i)^\top \hat{\boldsymbol{\theta}}_{n+1})\Big| \leq 1$$

the numerator in the maximum is bounded. This together with the continuity of d and the fact, that it is bounded away from 0 on a compact set, establishes the proof by choosing

$$K := \max_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{\theta}_0\in\Theta_0} \max_{\boldsymbol{\theta}_1\in\Theta_1}\left|\frac{\psi'(\boldsymbol{f}(\boldsymbol{x})^\top \boldsymbol{\theta}_1)}{d(\boldsymbol{f}(\boldsymbol{x})^\top \boldsymbol{\theta}_0)}\right|,$$

where the set $\Theta_1 := \{\boldsymbol{\theta} \in \Theta | \forall \boldsymbol{\theta}_0 \in \Theta_0 : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq C\}$ is compact. □

Lemma 12. *Assume further, that for $\hat{\boldsymbol{\theta}}_n \in \Theta_0$ there exists $C > 0$, such that*

$$\|\hat{\boldsymbol{\theta}}_{n+1} - \hat{\boldsymbol{\theta}}_n\| \leq C\frac{1}{\lambda_{n+1}},$$

then $(\widetilde{\mathcal{S}}^{(0)}_{n,\nu(t_n+t)-1})_{n\geq m}$ is equicontinuous in the extended sense.

If additionally

$$\lim_{n\to\infty}\left\|\mathbf{I}(\hat{\boldsymbol{\theta}}_n, \mathbf{F}_n)^{-1}\sum_{i=1}^{n}(y_i - G(\boldsymbol{f}(\boldsymbol{x}_i)^\top \hat{\boldsymbol{\theta}}_n))\psi'(\boldsymbol{f}(\boldsymbol{x}_i)^\top \hat{\boldsymbol{\theta}}_n)\boldsymbol{f}(\boldsymbol{x}_i)\boldsymbol{f}(\boldsymbol{x}_i)^\top\right\| = 0$$

then

$$\lim_{n\to\infty}\sup\Big\{\|\widetilde{\mathcal{S}}^{(0)}_{n,k}\| : k = n, \dots, \nu(t_n + t) - 1\Big\} = 0$$

for all $t > 0$.

Proof. Since the summands of $\widetilde{\mathcal{S}}^{(0)}_{n,k}$ are bounded by a multiple of λ^{-1}_{n+1}, it follows as in Lemma 2, that $(\widetilde{\mathcal{S}}^{(0)}_{n,\nu(t_n+t)-1})_{n\geq m}$ is equicontinuous in the extended sense.

For the additional assumption, we introduce

$$\psi'(\boldsymbol{f}(\boldsymbol{x}_i)^\top \hat{\boldsymbol{\theta}}^*_{i,n}) = \psi'(\boldsymbol{f}(\boldsymbol{x}_i)^\top \hat{\boldsymbol{\theta}}^*_{i,n}) - \psi'(\boldsymbol{f}(\boldsymbol{x}_i)^\top \hat{\boldsymbol{\theta}}_{n+1}) + \psi'(\boldsymbol{f}(\boldsymbol{x}_i)^\top \hat{\boldsymbol{\theta}}_{n+1})$$

in (3.17). Similar to the proof of Lemma 11

$$\begin{aligned}&\Big\|\mathbf{I}(\hat{\boldsymbol{\theta}}_n, \mathbf{F}_{n+1})^{-1}\Big(\tilde{\boldsymbol{s}}_{n+1}(\hat{\boldsymbol{\theta}}_{n+1}, \hat{\boldsymbol{\theta}}_n) - \boldsymbol{s}_{n+1}(\hat{\boldsymbol{\theta}}_{n+1})\Big)\Big\| \\ &\quad\leq \left\|\mathbf{I}(\hat{\boldsymbol{\theta}}_n, \mathbf{F}_n)^{-1}\sum_{i=1}^{n+1}(y_i - G(\boldsymbol{f}(\boldsymbol{x}_i)^\top \hat{\boldsymbol{\theta}}_{n+1}))\psi'(\boldsymbol{f}(\boldsymbol{x}_i)^\top \hat{\boldsymbol{\theta}}_{n+1})\boldsymbol{f}(\boldsymbol{x}_i)\boldsymbol{f}(\boldsymbol{x}_i)^\top\right\|\|\hat{\boldsymbol{\theta}}_{n+1} - \hat{\boldsymbol{\theta}}_n\| \\ &\quad+ 2\|\hat{\boldsymbol{\theta}}_{n+1} - \hat{\boldsymbol{\theta}}_n\| \max_{i=1,\dots,n+1}\left|\frac{\Big(y_i - G(\boldsymbol{f}(\boldsymbol{x}_i)^\top \hat{\boldsymbol{\theta}}_{n+1})\Big)(\psi'(\boldsymbol{f}(\boldsymbol{x}_i)^\top \hat{\boldsymbol{\theta}}^*_{i,n}) - \psi'(\boldsymbol{f}(\boldsymbol{x}_i)^\top \hat{\boldsymbol{\theta}}_{n+1}))}{d(\boldsymbol{f}(\boldsymbol{x}_i)^\top \hat{\boldsymbol{\theta}}_n)}\right|.\end{aligned}$$

Because of the additional assumption the first part on the right-hand side tends to 0 faster then the difference of the estimates. In the second term $\hat{\boldsymbol{\theta}}^*_{i,n}$ is a convex combination of the estimates. Consequently the maximum tends to 0, too, and hence

$$\lim_{n\longrightarrow\infty} \sup\left\{\|\widetilde{\mathcal{S}}^{(0)}_{n,k}\| : k = n,\dots,\nu(t_n+t)-1\right\} = 0\,.$$

□

The Behavior of $\mathcal{G}^{(0)}_{m,n}$

This is the key part contributing to the asymptotic behavior of $\hat{\boldsymbol{\theta}}_n$. Since the summands are all bounded it follows almost directly, that this part is equicontinuous in the extended sense. The proof is omitted.

Lemma 13. *Let $\mathcal{X}$ and Θ_0 be compact. Let $\hat{\boldsymbol{\theta}}_n$ exist for all $n \geq m$. Let G be continuous. Then the sequence of functions $(\mathcal{G}^{(0)}_{n,\nu(t_n+t)-1})_{n\geq m}$ is equicontinuous in the extended sense.*

3.4. Characterizing the Limit

So far we have seen, that the only two parts contributing to the asymptotics of $\hat{\boldsymbol{\theta}}(t_{n_k}+t)$ are $\mathcal{G}$ and $\widetilde{\mathcal{S}}$. We will assume the additional assumptions on $\widetilde{\mathcal{S}}$, i.e. that its asymptotic rate of change tends to 0, and only consider $\mathcal{G}$.

The next question to be answered is: What is the limit? Since we know, that $\hat{\boldsymbol{\theta}}(t_{n_k}+t)$ converges to some limit $\boldsymbol{\theta}(t)$, we can use that to get rid of the direct dependence on the estimates. In a second step the same is done for the dependence on the individual design points.

The following decomposition and proofs are based on the proof for Theorem 6.1.1 in Kushner and Yin (2003, p. 168).

Let us consider $\mathcal{G}^{(0)}_{m,\nu(t_m+t)-1}$ and split it into batches corresponding to time intervals of length $\Delta > 0$. In the following denote $\nu_{m,j} := \nu(t_m + j\Delta)$ for fixed $\Delta > 0$.

$$\begin{aligned}\mathcal{G}^{(0)}_{m,\nu(t_m+t)-1} = &\sum_{j=0}^{\lfloor t/\Delta\rfloor-1}\sum_{i=\nu_{m,j}}^{\nu_{m,j+1}-1} \mathbb{1}_{\Theta_0}(\hat{\boldsymbol{\theta}}_i)\mathbf{I}(\hat{\boldsymbol{\theta}}_i,\mathbf{F}_{i+1})^{-1}\boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}_{i+1},\hat{\boldsymbol{\theta}}_i)\\ &+\sum_{i=\nu(t_m+\lfloor t/\Delta\rfloor\Delta)}^{\nu(t_m+t)-1} \mathbb{1}_{\Theta_0}(\hat{\boldsymbol{\theta}}_i)\mathbf{I}(\hat{\boldsymbol{\theta}}_i,\mathbf{F}_{i+1})^{-1}\boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}_{i+1},\hat{\boldsymbol{\theta}}_i)\,. \quad (3.18)\end{aligned}$$

Similarly define $\mathcal{G}^{(1)}_{m,\nu(t_m+t)-1}$, where the estimates are substituted by the limit $\boldsymbol{\theta}(\cdot)$ of a subsequence at times $j\Delta$ by

$$\begin{aligned}\mathcal{G}^{(1)}_{m,\nu(t_m+t)-1} := &\sum_{j=0}^{\lfloor t/\Delta\rfloor-1}\sum_{i=\nu_{m,j}}^{\nu_{m,j+1}-1} \mathbb{1}_{\Theta_0}(\hat{\boldsymbol{\theta}}_i)\mathbf{I}(\boldsymbol{\theta}(j\Delta),\mathbf{F}_{i+1})^{-1}\boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}_{i+1},\boldsymbol{\theta}(j\Delta))\\ &+\sum_{i=\nu(t_m+\lfloor t/\Delta\rfloor\Delta)}^{\nu(t_m+t)-1} \mathbb{1}_{\Theta_0}(\hat{\boldsymbol{\theta}}_i)\mathbf{I}(\boldsymbol{\theta}(\lfloor t/\Delta\rfloor\Delta),\mathbf{F}_{i+1})^{-1}\boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}_{i+1},\boldsymbol{\theta}(\lfloor t/\Delta\rfloor\Delta))\,. \quad (3.19)\end{aligned}$$

Evaluating the limiting function at $j\Delta$ means, that we compare the estimates with the beginning of a batch.

The influence of the design points is smoothed by taking batch-wise averages:

$$\tilde{\boldsymbol{g}}_{m,j} := \frac{1}{\nu_{m,j+1}-\nu_{m,j}} \sum_{i=\nu_{m,j}}^{\nu_{m,j+1}-1} \mathbb{1}_{\Theta_0}(\hat{\boldsymbol{\theta}}_i)\boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}_{i+1},\boldsymbol{\theta}(j\Delta))$$

$$\tilde{\mathbf{M}}_{m,j} := \frac{1}{\nu_{m,j+1}-\nu_{m,j}} \sum_{i=\nu_{m,j}}^{\nu_{m,j+1}-1} \lambda_{i+1}\mathbf{I}(\boldsymbol{\theta}(j\Delta),\mathbf{F}_{i+1})^{-1}\,.$$

For the "incomplete" sum ending at $\nu(t_m+t)-1$, we will write $\tilde{\boldsymbol{g}}_{m,\lfloor t/\Delta\rfloor}$ and $\tilde{\mathbf{M}}_{m,\lfloor t/\Delta\rfloor}$. The part capturing these moving averages is denoted by

$$\mathcal{G}^{(2)}_{m,\nu(t_m+t)-1} := \Delta \sum_{j=0}^{\lfloor t/\Delta\rfloor-1} \tilde{\mathbf{M}}_{m,j}\tilde{\boldsymbol{g}}_{m,j} + \Big(t-\Big\lfloor\frac{t}{\Delta}\Big\rfloor\Delta\Big)\tilde{\mathbf{M}}_{m,\lfloor t/\Delta\rfloor}\tilde{\boldsymbol{g}}_{m,\lfloor t/\Delta\rfloor}\,. \tag{3.20}$$

The following Lemmas 14 and 15 show, that the differences $\|\mathcal{G}^{(0)}_{n_k,\nu(t_{n_k}+t)-1}-\mathcal{G}^{(1)}_{n_k,\nu(t_{n_k}+t)-1}\|$ and $\|\mathcal{G}^{(1)}_{n_k,\nu(t_{n_k}+t)-1}-\mathcal{G}^{(2)}_{n_k,\nu(t_{n_k}+t)-1}\|$ tend to 0 for $k\longrightarrow\infty$ and $\Delta\longrightarrow 0$. Because

$$\begin{aligned}\|\mathcal{G}^{(0)}_{n_k,\nu(t_{n_k}+t)-1}&-\mathcal{G}^{(2)}_{n_k,\nu(t_{n_k}+t)-1}\|\\ &\le \|\mathcal{G}^{(0)}_{n_k,\nu(t_{n_k}+t)-1}-\mathcal{G}^{(1)}_{n_k,\nu(t_{n_k}+t)-1}\| + \|\mathcal{G}^{(1)}_{n_k,\nu(t_{n_k}+t)-1}-\mathcal{G}^{(2)}_{n_k,\nu(t_{n_k}+t)-1}\|\end{aligned}$$

it follows for $\|\mathcal{G}^{(0)}_{n_k,\nu(t_{n_k}+t)-1}-\mathcal{G}^{(2)}_{n_k,\nu(t_{n_k}+t)-1}\|$, too. We will consider the difference between $\mathcal{G}^{(0)}_{n_k,\nu(t_{n_k}+t)-1}$ and $\mathcal{G}^{(1)}_{n_k,\nu(t_{n_k}+t)-1}$ first.

Lemma 14. *Let $(\hat{\boldsymbol{\theta}}(t_n+\cdot))_{n\ge m}$ be equicontinuous in the extended sense. Denote by $(\hat{\boldsymbol{\theta}}(t_{n_k}+\cdot))_{k\ge 1}$ a convergent subsequence and its continuous limit by $\boldsymbol{\theta}(\cdot)$. Then for each $t>0$*

$$\lim_{\Delta\to\infty}\lim_{k\to\infty}\|\mathcal{G}^{(0)}_{n_k,\nu(t_{n_k}+t)-1}-\mathcal{G}^{(1)}_{n_k,\nu(t_{n_k}+t)-1}\| = 0\,.$$

Proof. In the first part of the proof, we will establish upper bounds for the summands of the difference. With that in place we can use the definition of the index function ν and the discrete time t_n to switch to continuous time, in order to use the uniform convergence to the limit function $\boldsymbol{\theta}(\cdot)$.

For the summands of the inner sums of $\mathcal{G}^{(0)}_{m,\nu(t_m+t)-1}-\mathcal{G}^{(1)}_{m,\nu(t_m+t)-1}$ we have

$$\begin{aligned}\|\mathbf{I}(\hat{\boldsymbol{\theta}}_i,\mathbf{F}_{i+1})^{-1}&\boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}_{i+1},\hat{\boldsymbol{\theta}}_i)-\mathbf{I}(\boldsymbol{\theta}(j\Delta),\mathbf{F}_{i+1})^{-1}\boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}_{i+1},\boldsymbol{\theta}(j\Delta))\|\\ &\le \Big\|\Big(\mathbf{I}(\hat{\boldsymbol{\theta}}_i,\mathbf{F}_{i+1})^{-1}-\mathbf{I}(\boldsymbol{\theta}(j\Delta),\mathbf{F}_{i+1})^{-1}\Big)\boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}_{i+1},\boldsymbol{\theta}(j\Delta)\Big\|\\ &\qquad + \Big\|\mathbf{I}(\hat{\boldsymbol{\theta}}_i,\mathbf{F}_{i+1})^{-1}\Big(\boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}_{i+1},\hat{\boldsymbol{\theta}}_i)-\boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}_{i+1},\boldsymbol{\theta}(j\Delta))\Big)\Big\|\,.\end{aligned} \tag{3.21}$$

For the first part on the right-hand side of the inequality follows

$$\begin{aligned}\Big\|\Big(\mathbf{I}(\hat{\boldsymbol{\theta}}_i,\mathbf{F}_{i+1})^{-1}&-\mathbf{I}(\boldsymbol{\theta}(j\Delta),\mathbf{F}_{i+1})^{-1}\Big)\boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}_{i+1},\boldsymbol{\theta}(j\Delta))\Big\|\\ &\le \Big\|\Big(\mathbf{I}(\hat{\boldsymbol{\theta}}_i,\mathbf{F}_{i+1})^{-1}-\mathbf{I}(\boldsymbol{\theta}(j\Delta),\mathbf{F}_{i+1})^{-1}\Big)\Big\|\|\boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}_{i+1},\boldsymbol{\theta}(j\Delta))\|\,.\end{aligned}$$

Since $0 \leq j\Delta \leq t$

$$\|\boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}_{i+1}, \boldsymbol{\theta}(j\Delta))\| \leq \max_{\boldsymbol{x}\in\mathcal{X}} \sup_{0\leq s\leq t} \|\boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}, \boldsymbol{\theta}(s))\| \,. \tag{3.22}$$

The right-hand side is bounded, because the function $\boldsymbol{g}_{\bar{\boldsymbol{\theta}}}$ is continuous in both components, the limit function $\boldsymbol{\theta}(\cdot)$ is continuous, too, and $\mathcal{X}$ and $[0,t]$ are compact. The difference of the inverse matrices can be rewritten as

$$\begin{aligned}&\mathbf{I}(\hat{\boldsymbol{\theta}}_i, \mathbf{F}_{i+1})^{-1} - \mathbf{I}(\boldsymbol{\theta}(j\Delta), \mathbf{F}_{i+1})^{-1} \\ &\qquad = \mathbf{I}(\boldsymbol{\theta}(j\Delta), \mathbf{F}_{i+1})^{-1}\Big(\mathbf{I}(\hat{\boldsymbol{\theta}}_i, \mathbf{F}_{i+1}) - \mathbf{I}(\boldsymbol{\theta}(j\Delta), \mathbf{F}_{i+1})\Big)\mathbf{I}(\hat{\boldsymbol{\theta}}_i, \mathbf{F}_{i+1})^{-1} \,.\end{aligned}$$

Note, that

$$\mathbf{I}(\hat{\boldsymbol{\theta}}_i, \mathbf{F}_{i+1}) - \mathbf{I}(\boldsymbol{\theta}(j\Delta), \mathbf{F}_{i+1}) = \mathbf{F}_{i+1}^\top\Big(\mathbf{D}_{i+1}(\mathbf{F}_{i+1}\hat{\boldsymbol{\theta}}_i) - \mathbf{D}_{i+1}(\mathbf{F}_{i+1}\boldsymbol{\theta}(j\Delta))\Big)\mathbf{F}_{i+1} \,.$$

An application of Lemma A.4 yields

$$\begin{aligned}&\Big\|\mathbf{I}(\boldsymbol{\theta}(j\Delta), \mathbf{F}_{i+1})^{-1}\Big(\mathbf{I}(\hat{\boldsymbol{\theta}}_i, \mathbf{F}_{i+1}) - \mathbf{I}(\boldsymbol{\theta}(j\Delta), \mathbf{F}_{i+1})\Big)\Big\| \\ &\qquad \leq 2 \max_{j=1,\ldots,i+1} \left|\frac{d(\boldsymbol{f}(\boldsymbol{x}_j)^\top\hat{\boldsymbol{\theta}}_i) - d(\boldsymbol{f}(\boldsymbol{x}_j)^\top\boldsymbol{\theta}(j\Delta))}{d(\boldsymbol{f}(\boldsymbol{x}_j)^\top\boldsymbol{\theta}(j\Delta))}\right| .\end{aligned}$$

A bound for the numerator follows from the mean value theorem:

$$|d(\boldsymbol{f}(\boldsymbol{x}_j)^\top\hat{\boldsymbol{\theta}}_i) - d(\boldsymbol{f}(\boldsymbol{x}_j)^\top\boldsymbol{\theta}(j\Delta))| \leq |\boldsymbol{f}(\boldsymbol{x}_j)^\top(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}(j\Delta))| \sup_{u\in\mathbb{R}} |d'(u)| \,.$$

Consequently

$$\Big\|\mathbf{I}(\boldsymbol{\theta}(j\Delta), \mathbf{F}_{i+1})^{-1}\Big(\mathbf{I}(\hat{\boldsymbol{\theta}}_i, \mathbf{F}_{i+1}) - \mathbf{I}(\boldsymbol{\theta}(j\Delta), \mathbf{F}_{i+1})\Big)\Big\| \leq K_1\|\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}(j\Delta)\| \,,$$

where the constant is

$$K_1 := 2\max_{\boldsymbol{x}\in\mathcal{X}} \sup_{0\leq s\leq t} \sup_{u\in\mathbb{R}} \left|\frac{d'(u)\|\boldsymbol{f}(\boldsymbol{x})\|}{d(\boldsymbol{f}(\boldsymbol{x})^\top\boldsymbol{\theta}(s))}\right| .$$

Also

$$\|\mathbf{I}(\hat{\boldsymbol{\theta}}_i, \mathbf{F}_{i+1})^{-1}\| \leq \lambda_{i+1}^{-1} \max_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{\theta}\in\Theta_0} (d(\boldsymbol{f}(\boldsymbol{x})^\top\boldsymbol{\theta}))^{-1}$$

and hence with $K_2 := \max_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{\theta}\in\Theta_0} (d(\boldsymbol{f}(\boldsymbol{x})^\top\boldsymbol{\theta}))^{-1}$

$$\begin{aligned}&\|\mathbf{I}(\hat{\boldsymbol{\theta}}_i, \mathbf{F}_{i+1})^{-1} - \mathbf{I}(\boldsymbol{\theta}(j\Delta), \mathbf{F}_{i+1})^{-1}\| \\ &\qquad \leq \Big\|\mathbf{I}(\boldsymbol{\theta}(j\Delta), \mathbf{F}_{i+1})^{-1}\Big(\mathbf{I}(\hat{\boldsymbol{\theta}}_i, \mathbf{F}_{i+1}) - \mathbf{I}(\boldsymbol{\theta}(j\Delta), \mathbf{F}_{i+1})\Big)\Big\| \, \|\mathbf{I}(\hat{\boldsymbol{\theta}}_i, \mathbf{F}_{i+1})^{-1}\| \\ &\qquad\qquad \leq \frac{K_1K_2}{\lambda_{i+1}}\|\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}(j\Delta)\| \,. \end{aligned} \tag{3.23}$$

Let us consider the second term on the right-hand side of (3.21). The function $\boldsymbol{g}_{\bar{\boldsymbol{\theta}}}$ is continuously differentiable in $\boldsymbol{\theta}$ for fixed $\boldsymbol{x} \in \mathcal{X}$. The mean value theorem yields

$$\begin{aligned}&\|\boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}, \hat{\boldsymbol{\theta}}_i) - \boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}, \boldsymbol{\theta}(j\Delta))\| \\ &\qquad \leq |\boldsymbol{f}(\boldsymbol{x})^\top(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}(j\Delta))| \sup_{0\leq s\leq 1} \Big\|\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{g}_{\bar{\boldsymbol{\theta}}}\Big(\boldsymbol{x}, \boldsymbol{\theta}(j\Delta) + s\,(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}(j\Delta))\Big)\Big\| \\ &\qquad\qquad \leq K_3\|\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}(j\Delta)\|\end{aligned}$$

for some constant $K_3 > 0$. Denoting by Θ_2 the closed convex hull of Θ_1 and the image of $[0,t]$ under $\boldsymbol{\theta}(\cdot)$, K_3 can be chosen as

$$K_3 := \max_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{\theta}\in\Theta_2} \|\boldsymbol{f}(\boldsymbol{x})\|^2 |\psi'(\boldsymbol{f}(\boldsymbol{x})^\top \boldsymbol{\theta})| + \sup_{s\in\mathbb{R}} d(s)\,.$$

Now by definition of the matrix norm

$$\Big\| \mathbf{I}(\hat{\boldsymbol{\theta}}_i, \mathbf{F}_{i+1})^{-1} \Big(\boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}_{i+1}, \hat{\boldsymbol{\theta}}_i) - \boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}_{i+1}, \boldsymbol{\theta}(j\Delta)) \Big) \Big\| \le \frac{K_2 K_3}{\lambda_{i+1}} \|\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}(j\Delta)\| \tag{3.24}$$

Combining the bounds from (3.21) to (3.24) gives

$$\|\mathbf{I}(\hat{\boldsymbol{\theta}}_i, \mathbf{F}_{i+1})^{-1} \boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}_{i+1}, \hat{\boldsymbol{\theta}}_i) - \mathbf{I}(\boldsymbol{\theta}(j\Delta), \mathbf{F}_{i+1})^{-1} \boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}_{i+1}, \boldsymbol{\theta}(j\Delta))\| \le \frac{K_4}{\lambda_{i+1}} \|\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}(j\Delta)\|\,.$$

This inequality holds also for the summands of the last sum, running from $\nu(t_m + \lfloor t/\Delta \rfloor \Delta)$ to $\nu(t_m + t) - 1$, and yields

$$\begin{aligned}\|\mathcal{G}^{(0)}_{m,\nu(t_m+t)-1} - \mathcal{G}^{(1)}_{m,\nu(t_m+t)-1}\| \le & \sum_{j=0}^{\lfloor t/\Delta \rfloor - 1} \sum_{i=\nu_{m,j}}^{\nu_{m,j+1}-1} \frac{K_4}{\lambda_{i+1}} \|\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}(j\Delta)\| \\ & + \sum_{i=\nu(t_m+\lfloor t/\Delta \rfloor \Delta)}^{\nu(t_m+t)-1} \frac{K_4}{\lambda_{i+1}} \|\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}(j\Delta)\|\,.\end{aligned}$$

By taking the maximum over i we get the bounds

$$\sum_{i=\nu_{m,j}}^{\nu_{m,j+1}-1} \frac{1}{\lambda_{i+1}} \|\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}(j\Delta)\| \le \max_{\nu_{m,j} \le i \le \nu_{m,j+1}-1} \|\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}(j\Delta)\| \sum_{i=\nu_{m,j}}^{\nu_{m,j+1}-1} \frac{1}{\lambda_{i+1}}$$

for $j = 0, \ldots, \lfloor t/\Delta \rfloor - 1$. If we choose $j = \lfloor t/\Delta \rfloor$ a bound for the last sum follows

$$\sum_{i=\nu(t_m+\lfloor t/\Delta \rfloor \Delta)}^{\nu(t_m+t)-1} \frac{1}{\lambda_{i+1}} \|\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}(j\Delta)\| \le \max_{\nu_{m,j} \le i \le \nu_{m,j+1}-1} \|\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}(j\Delta)\| \sum_{i=\nu_{m,j}}^{\nu(t_m+t)-1} \frac{1}{\lambda_{i+1}}\,,$$

since

$$t \le \left\lfloor \frac{t}{\Delta} \right\rfloor \Delta + \Delta\,.$$

Taking the maximum over $j = 0, \ldots, \lfloor t/\Delta \rfloor$ yields the bound

$$\begin{aligned}&\|\mathcal{G}^{(0)}_{m,\nu(t_m+t)-1} - \mathcal{G}^{(1)}_{m,\nu(t_m+t)-1}\| \\ &\qquad \le K_5 \max_{j=0,\ldots,\lfloor t/\Delta \rfloor} \max_{\nu_{m,j} \le i \le \nu_{m,j+1}-1} \|\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}(j\Delta)\| \\ &\qquad\qquad \times \left(\sum_{j=0}^{\lfloor t/\Delta \rfloor - 1} \sum_{i=\nu_{m,j}}^{\nu_{m,j+1}-1} \frac{1}{\lambda_{i+1}} + \sum_{i=\nu(t_m+\lfloor t/\Delta \rfloor \Delta)}^{\nu(t_m+t)-1} \frac{1}{\lambda_{i+1}} \right).\end{aligned}$$

The remaining sums on the right-hand side can be combined to

$$\sum_{j=0}^{\lfloor t/\Delta \rfloor - 1} \sum_{i=\nu_{m,j}}^{\nu_{m,j+1}-1} \frac{1}{\lambda_{i+1}} + \sum_{i=\nu(t_m+\lfloor t/\Delta \rfloor \Delta)}^{\nu(t_m+t)-1} \frac{1}{\lambda_{i+1}} = \sum_{i=m}^{\nu(t_m+t)-1} \frac{1}{\lambda_{i+1}}\,.$$

By definition of t_m and ν

$$\sum_{i=m}^{\nu(t_m+t)-1} \frac{1}{\lambda_{i+1}} \leq t$$

for all $m \in \mathbb{N}$ and hence

$$\|\mathcal{G}^{(0)}_{m,\nu(t_m+t)-1} - \mathcal{G}^{(1)}_{m,\nu(t_m+t)-1}\| \leq K_3\, t \max_{j=0,\dots,\lfloor t/\Delta \rfloor} \max_{\nu_{m,j} \leq i \leq \nu_{m,j+1}-1} \|\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}(j\Delta)\| .$$

Next we will determine upper bounds for the maxima on the right-hand side. The continuous time interpolation of the estimates, which was defined in (2.49), is the tool to give upper bounds in terms of the time instead of the index i in the inner maximum. We will use, that $\hat{\boldsymbol{\theta}}_i = \hat{\boldsymbol{\theta}}(t_i)$ and $t_i = t_m + t_i - t_m$. The difference $t_i - t_m$ will be used to switch from index to time.

Remember the definition of ν in (2.48):

$$\nu(t) := \sup\{k \in \mathbb{N} : t_k \leq t\} .$$

If $\nu_{m,j} = \nu(t_m + j\Delta) \leq i$, then by this definition

$$t_m + j\Delta \leq t_{i+1} .$$

Since $t_{i+1} = t_i + \lambda_{i+1}^{-1}$ and $\lambda_{i+1} \longrightarrow \infty$ for $i \longrightarrow \infty$

$$t_m + j\Delta - \Delta \leq t_i$$

for large enough m. Similarly, if $i \leq \nu_{m,j+1} - 1 = \nu(t_m + j\Delta + \Delta) - 1$ we get

$$t_i \leq t_m + j\Delta + \Delta - \frac{1}{\lambda_{i+1}} \leq t_m + j\Delta + \Delta$$

and combing both inequalities

$$j\Delta - \Delta \leq t_i - t_m \leq j\Delta + \Delta . \tag{3.25}$$

The set of indices i described by

$$\nu(t_m + j\Delta) \leq i \leq \nu(t_m + j\Delta + \Delta) - 1$$

is a subset of the one described by (3.25). It follows, that

$$\begin{aligned}\max_{\nu_{m,j} \leq i \leq \nu_{m,j+1}-1} \|\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}(j\Delta)\| &\leq \max_{j\Delta-\Delta \leq t_i - t_m \leq j\Delta+\Delta} \|\hat{\boldsymbol{\theta}}(t_m + t_i - t_m) - \boldsymbol{\theta}(j\Delta)\| \\ &\leq \sup_{j\Delta-\Delta \leq T \leq j\Delta+\Delta} \|\hat{\boldsymbol{\theta}}(t_m + T) - \boldsymbol{\theta}(j\Delta)\| .\end{aligned}$$

The last inequality follows from substituting $t_i - t_m$ by T. Note also, that

$$j\Delta - \Delta \leq T \leq j\Delta + \Delta \Longleftrightarrow |T - j\Delta| \leq \Delta .$$

For the outer maximum and indices j we get

$$0 \leq j\Delta \leq \left\lfloor \frac{t}{\Delta} \right\rfloor \Delta \leq t$$

and after replacing $j\Delta$ by $s > 0$

$$\max_{j=0,\dots,\lfloor t/\Delta \rfloor} \max_{\nu_{m,j} \le i \le \nu_{m,j+1}-1} \|\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}(j\Delta)\| \le \sup_{0\le s\le t} \sup_{|T-s|\le\Delta} \|\hat{\boldsymbol{\theta}}(t_m + T) - \boldsymbol{\theta}(s)\| .$$

For the last step introduce $\boldsymbol{\theta}(T)$ on the right-hand side:

$$\begin{aligned}\sup_{0\le s\le t} \sup_{|T-s|\le\Delta} \|\hat{\boldsymbol{\theta}}(t_m + T) - \boldsymbol{\theta}(s)\| &\le \sup_{0\le s\le t} \sup_{|T-s|\le\Delta} \left(\|\hat{\boldsymbol{\theta}}(t_m + T) - \boldsymbol{\theta}(T)\| + \|\boldsymbol{\theta}(T) - \boldsymbol{\theta}(s)\|\right) \\ &\le \sup_{-\Delta\le T\le t+\Delta} \|\hat{\boldsymbol{\theta}}(t_m + T) - \boldsymbol{\theta}(T)\| \\ &\qquad + \sup_{0\le s\le t} \sup_{|T-s|\le\Delta} \|\boldsymbol{\theta}(T) - \boldsymbol{\theta}(s)\|\end{aligned}$$

For the subsequence $(\hat{\boldsymbol{\theta}}(t_{n_k} + \cdot))_{k\ge1}$ follows

$$\lim_{k\to\infty} \sup_{-\Delta\le T\le t+\Delta} \|\hat{\boldsymbol{\theta}}(t_{n_k} + T) - \boldsymbol{\theta}(T)\| = 0$$

because $\hat{\boldsymbol{\theta}}(t_{n_k} + \cdot)$ converges uniformly to $\boldsymbol{\theta}(\cdot)$. The limit of the subsequence is continuous and uniformly continuous on $[0, t]$. Consequently for arbitrarily small $\Delta > 0$

$$\sup_{0\le s\le t} \sup_{|T-s|\le\Delta} \|\boldsymbol{\theta}(T) - \boldsymbol{\theta}(s)\|$$

is arbitrarily small, too. Hence

$$\lim_{\Delta\to0} \lim_{k\to\infty} \|\mathcal{G}^{(0)}_{n_k,\nu(t_{n_k}+t)-1} - \mathcal{G}^{(1)}_{n_k,\nu(t_{n_k}+t)-1}\| = 0$$

as desired. □

Lemma 15. *Let* $\left(\hat{\boldsymbol{\theta}}(t_n + \cdot)\right)_{n\ge m}$ *be equicontinuous in the extended sense. Let* $\left(\hat{\boldsymbol{\theta}}(t_{n_k} + \cdot)\right)_{k\ge1}$ *be a convergent subsequence and denote its continuous limit by* $\boldsymbol{\theta}(\cdot)$*. Then for each* $t > 0$

$$\lim_{\Delta\to0} \lim_{k\to\infty} \|\mathcal{G}^{(1)}_{n_k,\nu(t_{n_k}+t)} - \mathcal{G}^{(2)}_{n_k,\nu(t_{n_k}+t)}\| = 0$$

Proof. We will start again with the inner sums of $\mathcal{G}^{(1)}_{m,\nu(t_m+t)-1} - \mathcal{G}^{(2)}_{m,\nu(t_m+t)-1}$, i.e.

$$\sum_{i=\nu_{m,j}}^{\nu_{m,j+1}-1} \mathbb{1}_{\Theta_0}(\hat{\boldsymbol{\theta}}_i)\mathbf{I}(\boldsymbol{\theta}(j\Delta), \mathbf{F}_{i+1})^{-1}\boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}_{i+1}, \boldsymbol{\theta}(j\Delta)) - \Delta\tilde{\mathbf{M}}_{m,j}\tilde{\boldsymbol{g}}_{m,j} .$$

We will show, that its limit with respect to m is bounded by a multiple of Δ^2. This yields that $\|\mathcal{G}^{(1)}_{m,\nu(t_m+t)-1} - \mathcal{G}^{(2)}_{m,\nu(t_m+t)-1}\|$ is bounded by a multiple of Δ and tends to 0 for $\Delta \longrightarrow 0$.

Introducing

$$\begin{aligned}\sum_{i=\nu_{m,j}}^{\nu_{m,j+1}-1} \mathbf{I}(\boldsymbol{\theta}(j\Delta), \mathbf{F}_{i+1})^{-1} \frac{1}{\nu_{m,j+1} - \nu_{m,j}} \sum_{k=\nu_{m,j}}^{\nu_{m,j+1}-1} \mathbb{1}_{\Theta_0}(\hat{\boldsymbol{\theta}}_k)\boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}_{k+1}, \boldsymbol{\theta}(j\Delta)) \\ = \sum_{i=\nu_{m,j}}^{\nu_{m,j+1}-1} \mathbf{I}(\boldsymbol{\theta}(j\Delta), \mathbf{F}_{i+1})^{-1}\tilde{\boldsymbol{g}}_{m,j}\end{aligned}$$

it follows from the triangle inequality, that

$$\begin{aligned}\Big\| \sum_{i=\nu_{m,j}}^{\nu_{m,j+1}-1} \mathbb{1}_{\Theta_0}(\hat{\boldsymbol{\theta}}_i)\mathbf{I}(\boldsymbol{\theta}(j\Delta),\mathbf{F}_{i+1})^{-1}\boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}_{i+1},\boldsymbol{\theta}(j\Delta)) - \Delta\tilde{\mathbf{M}}_{m,j}\tilde{\boldsymbol{g}}_{m,j}\Big\| \\ \leq \Big\| \sum_{i=\nu_{m,j}}^{\nu_{m,j+1}-1} \mathbb{1}_{\Theta_0}(\hat{\boldsymbol{\theta}}_i)\mathbf{I}(\boldsymbol{\theta}(j\Delta),\mathbf{F}_{i+1})^{-1}\boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}_{i+1},\boldsymbol{\theta}(j\Delta)) \\ - \sum_{i=\nu_{m,j}}^{\nu_{m,j+1}-1} \mathbf{I}(\boldsymbol{\theta}(j\Delta),\mathbf{F}_{i+1})^{-1}\tilde{\boldsymbol{g}}_{m,j}\Big\| \\ + \Big\| \sum_{i=\nu_{m,j}}^{\nu_{m,j+1}-1} \mathbf{I}(\boldsymbol{\theta}(j\Delta),\mathbf{F}_{i+1})^{-1}\tilde{\boldsymbol{g}}_{m,j} - \Delta\tilde{\mathbf{M}}_{m,j}\tilde{\boldsymbol{g}}_{m,j}\Big\| . \quad (3.26)\end{aligned}$$

We will start with the first difference on the right-hand side. By partial summation

$$\begin{aligned}&\sum_{i=\nu_{m,j}}^{\nu_{m,j+1}-1} \mathbb{1}_{\Theta_0}(\hat{\boldsymbol{\theta}}_i)\mathbf{I}(\boldsymbol{\theta}(j\Delta),\mathbf{F}_{i+1})^{-1}\boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}_{i+1},\boldsymbol{\theta}(j\Delta)) - \sum_{i=\nu_{m,j}}^{\nu_{m,j+1}-1} \mathbf{I}(\boldsymbol{\theta}(j\Delta),\mathbf{F}_{i+1})^{-1}\tilde{\boldsymbol{g}}_{m,j} \\ &\quad = \sum_{i=\nu_{m,j}}^{\nu_{m,j+1}-1} \Bigg(\Big(\mathbf{I}(\boldsymbol{\theta}(j\Delta),\mathbf{F}_{i+1})^{-1} - \mathbf{I}(\boldsymbol{\theta}(j\Delta),\mathbf{F}_{i+2})^{-1} \Big) \\ &\qquad\qquad \times \sum_{l=\nu_{m,j}}^{i} \Big(\mathbb{1}_{\Theta_0}(\hat{\boldsymbol{\theta}}_l)\boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}_{l+1},\boldsymbol{\theta}(j\Delta)) - \tilde{\boldsymbol{g}}_{m,j} \Big) \Bigg) .\end{aligned}$$

Because $\boldsymbol{g}_{\bar{\boldsymbol{\theta}}}$ is bounded

$$\|\mathbb{1}_{\Theta_0}(\hat{\boldsymbol{\theta}}_l)\boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}_{l+1},\boldsymbol{\theta}(j\Delta)) - \tilde{\boldsymbol{g}}_{m,j}\| \leq K_1$$

for some $K_1 > 0$ and we get

$$\Big\| \sum_{l=\nu_{m,j}}^{i} \Big(\mathbb{1}_{\Theta_0}(\hat{\boldsymbol{\theta}}_l)\boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}_{l+1},\boldsymbol{\theta}(j\Delta)) - \tilde{\boldsymbol{g}}_{m,j} \Big) \Big\| \leq K_1(i - \nu_{m,j} + 1) \leq K_1(\nu_{m,j+1} - \nu_{m,j}) .$$

An upper bound for the difference of the information matrices is given by

$$\|\mathbf{I}(\boldsymbol{\theta}(j\Delta),\mathbf{F}_{i+1})^{-1} - \mathbf{I}(\boldsymbol{\theta}(j\Delta),\mathbf{F}_{i+2})^{-1}\| \leq \frac{K}{\lambda_{i+1}\lambda_{i+2}} ,$$

where $K > 0$. Putting both together yields

$$\begin{aligned}\Big\| \sum_{i=\nu_{m,j}}^{\nu_{m,j+1}-1} \mathbb{1}_{\Theta_0}(\hat{\boldsymbol{\theta}}_i)\mathbf{I}(\boldsymbol{\theta}(j\Delta),\mathbf{F}_{i+1})^{-1}\boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}_{i+1},\boldsymbol{\theta}(j\Delta)) - \sum_{i=\nu_{m,j}}^{\nu_{m,j+1}-1} \mathbf{I}(\boldsymbol{\theta}(j\Delta),\mathbf{F}_{i+1})^{-1}\tilde{\boldsymbol{g}}_{m,j} \Big\| \\ \leq K K_1 \sum_{i=\nu_{m,j}}^{\nu_{m,j+1}-1} \frac{1}{\lambda_{i+2}} \frac{\nu_{m,j+1} - \nu_{m,j}}{\lambda_{i+1}} .\end{aligned}$$

With the properties of ν and t_n (see Lemma A.7) follows, that $\nu_{m,j+1} - \nu_{m,j}$ can be bounded in terms of λ_i and Δ. Consequently

$$\sum_{i=\nu_{m,j}}^{\nu_{m,j+1}-1} \frac{1}{\lambda_{i+2}} \frac{\nu_{m,j+1} - \nu_{m,j}}{\lambda_{i+1}} \leq \sum_{i=\nu_{m,j}}^{\nu_{m,j+1}-1} \frac{1}{\lambda_{i+2}} \left(\Delta \frac{\lambda_{\nu_{m,j}}}{\lambda_{i+1}} + \frac{\lambda_{\nu_{m,j}}}{\lambda_{i+1}\lambda_{\nu_{m,j}+1}} \right) \leq \Delta \sum_{i=\nu_{m,j}}^{\nu_{m,j+1}-1} \frac{1}{\lambda_{i+1}} + \sum_{i=\nu_{m,j}}^{\nu_{m,j+1}-1} \frac{1}{\lambda_{i+1}^2} . \quad (3.27)$$

Again by definition of ν and t_m

$$\sum_{i=\nu_{m,j}}^{\nu_{m,j+1}-1} \frac{1}{\lambda_{i+1}} \leq t_m + j\Delta + \Delta - \left(t_m + j\Delta - \frac{1}{\lambda_m} \right) = \Delta + \frac{1}{\lambda_m}$$

and

$$\sum_{i=\nu_{m,j}}^{\nu_{m,j+1}-1} \frac{1}{\lambda_{i+1}} \geq t_m + j\Delta + \Delta - \frac{1}{\lambda_m} - (t_m + j\Delta) = \Delta - \frac{1}{\lambda_m} .$$

If we take the limit with respect to m, it follows, that

$$\lim_{m\to\infty} \sum_{i=\nu_{m,j}}^{\nu_{m,j+1}-1} \frac{1}{\lambda_{i+1}} = \Delta \quad (3.28)$$

and since λ_{i+1}^{-1} is nonincreasing

$$\lim_{m\to\infty} \sum_{i=\nu_{m,j}}^{\nu_{m,j+1}-1} \frac{1}{\lambda_{i+1}^2} \leq \lim_{m\to\infty} \frac{1}{\lambda_{\nu_{m,j}+1}} \sum_{i=\nu_{m,j}}^{\nu_{m,j+1}-1} \frac{1}{\lambda_{i+1}} = 0 .$$

In total this yields

$$\lim_{m\to\infty} \left\| \sum_{i=\nu_{m,j}}^{\nu_{m,j+1}-1} \mathbb{1}_{\Theta_0}(\hat{\boldsymbol{\theta}}_i) \mathbf{I}(\boldsymbol{\theta}(j\Delta), \mathbf{F}_{i+1})^{-1} \boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}_{i+1}, \boldsymbol{\theta}(j\Delta)) - \sum_{i=\nu_{m,j}}^{\nu_{m,j+1}-1} \mathbf{I}(\boldsymbol{\theta}(j\Delta), \mathbf{F}_{i+1})^{-1} \tilde{\boldsymbol{g}}_{m,j} \right\| \leq K K_1 \lim_{m\to\infty} \left(\Delta \sum_{i=\nu_{m,j}}^{\nu_{m,j+1}-1} \frac{1}{\lambda_{i+1}} + \sum_{i=\nu_{m,j}}^{\nu_{m,j+1}-1} \frac{1}{\lambda_{i+1}^2} \right) = K_2 \Delta^2 . \quad (3.29)$$

For the remaining term in (3.26) follows

$$\left\| \sum_{i=\nu_{m,j}}^{\nu_{m,j+1}-1} \mathbf{I}(\boldsymbol{\theta}(j\Delta), \mathbf{F}_{i+1})^{-1} \tilde{\boldsymbol{g}}_{m,j} - \Delta \tilde{\mathbf{M}}_{m,j} \tilde{\boldsymbol{g}}_{m,j} \right\| \leq \|\tilde{\boldsymbol{g}}_{m,j}\| \left\| \sum_{i=\nu_{m,j}}^{\nu_{m,j+1}-1} \mathbf{I}(\boldsymbol{\theta}(j\Delta), \mathbf{F}_{i+1})^{-1} - \Delta \tilde{\mathbf{M}}_{m,j} \right\| . \quad (3.30)$$

Partial summation for the sum on the right-hand side yields

$$\begin{aligned}
\sum_{i=\nu_{m,j}}^{\nu_{m,j+1}-1} \mathbf{I}(\boldsymbol{\theta}(j\Delta), \mathbf{F}_{i+1})^{-1} &= \sum_{i=\nu_{m,j}}^{\nu_{m,j+1}-1} \frac{\lambda_{i+2}-\lambda_{i+1}}{\lambda_{i+1}\lambda_{i+2}} \sum_{k=\nu_{m,j}}^{i} \lambda_{k+1}\mathbf{I}(\boldsymbol{\theta}(j\Delta), \mathbf{F}_{k+1})^{-1} \\
&\quad - \sum_{i=\nu_{m,j+1}}^{\nu_{m,j+1}-1} \frac{1}{\lambda_{i+1}} \sum_{k=\nu_{m,j}}^{i-1} \lambda_{k+1}\mathbf{I}(\boldsymbol{\theta}(j\Delta), \mathbf{F}_{k+1})^{-1} \\
&\quad + \sum_{i=\nu_{m,j}}^{\nu_{m,j+1}-1} \frac{1}{\lambda_{i+2}} \sum_{k=\nu_{m,j}}^{i} \lambda_{k+1}\mathbf{I}(\boldsymbol{\theta}(j\Delta), \mathbf{F}_{k+1})^{-1} \\
&= \sum_{i=\nu_{m,j}}^{\nu_{m,j+1}-1} \frac{\lambda_{i+2}-\lambda_{i+1}}{\lambda_{i+1}\lambda_{i+2}} \sum_{k=\nu_{m,j}}^{i} \lambda_{k+1}\mathbf{I}(\boldsymbol{\theta}(j\Delta), \mathbf{F}_{k+1})^{-1} \\
&\quad + \frac{1}{\lambda_{\nu_{m,j+1}+1}} \sum_{k=\nu_{m,j}}^{\nu_{m,j+1}-1} \lambda_{k+1}\mathbf{I}(\boldsymbol{\theta}(j\Delta), \mathbf{F}_{k+1})^{-1} \\
&= \sum_{i=\nu_{m,j}}^{\nu_{m,j+1}-1} \frac{\lambda_{i+2}-\lambda_{i+1}}{\lambda_{i+1}\lambda_{i+2}} \sum_{k=\nu_{m,j}}^{i} \lambda_{k+1}\mathbf{I}(\boldsymbol{\theta}(j\Delta), \mathbf{F}_{k+1})^{-1} \\
&\quad + \frac{\nu_{m,j+1}-\nu_{m,j}}{\lambda_{\nu_{m,j+1}+1}} \tilde{\mathbf{M}}_{m,j} \,.
\end{aligned}$$

Note that the sum indexed by k is a sum of nonnegative definite matrices and can be bounded by $\tilde{\mathbf{M}}_{m,j}$:

$$\begin{aligned}
&\Big\| \sum_{k=\nu_{m,j}}^{i} \lambda_{k+1}\mathbf{I}(\boldsymbol{\theta}(j\Delta), \mathbf{F}_{k+1})^{-1} \Big\| \\
&\qquad \le \Big\| \sum_{k=\nu_{m,j}}^{\nu_{m,j+1}-1} \lambda_{k+1}\mathbf{I}(\boldsymbol{\theta}(j\Delta), \mathbf{F}_{k+1})^{-1} \Big\| = (\nu_{m,j+1}-\nu_{m,j}) \|\tilde{\mathbf{M}}_{m,j}\| \,,
\end{aligned}$$

and hence

$$\begin{aligned}
&\Big\| \sum_{i=\nu_{m,j}}^{\nu_{m,j+1}-1} \mathbf{I}(\boldsymbol{\theta}(j\Delta), \mathbf{F}_{i+1})^{-1} - \Delta\tilde{\mathbf{M}}_{m,j} \Big\| \\
&\le \Big\| \sum_{i=\nu_{m,j}}^{\nu_{m,j+1}-1} \frac{\lambda_{i+2}-\lambda_{i+1}}{\lambda_{i+1}\lambda_{i+2}} \sum_{k=\nu_{m,j}}^{i} \lambda_{k+1}\mathbf{I}(\boldsymbol{\theta}(j\Delta), \mathbf{F}_{k+1})^{-1} \Big\| + \Big\| \frac{\nu_{m,j+1}-\nu_{m,j}}{\lambda_{\nu_{m,j+1}+1}} \tilde{\mathbf{M}}_{m,j} - \Delta\tilde{\mathbf{M}}_{m,j} \Big\| \\
&\le \|\tilde{\mathbf{M}}_{\nu_{m,j},\nu_{m,j}-1}\| \left((\nu_{m,j+1}-\nu_{m,j}) \sum_{i=\nu_{m,j}}^{\nu_{m,j+1}-1} \frac{\lambda_{i+2}-\lambda_{i+1}}{\lambda_{i+1}\lambda_{i+2}} + \Big| \frac{\nu_{m,j+1}-\nu_{m,j}}{\lambda_{\nu_{m,j+1}+1}} - \Delta \Big| \right) . \quad (3.31)
\end{aligned}$$

The difference $\lambda_{i+2}-\lambda_{i+1}$ is bounded above, so we can use the same bound as in (3.27), i.e.

$$(\nu_{m,j+1}-\nu_{m,j}) \sum_{i=\nu_{m,j}}^{\nu_{m,j+1}-1} \frac{\lambda_{i+2}-\lambda_{i+1}}{\lambda_{i+1}\lambda_{i+2}} \le \Delta \sum_{i=\nu_{m,j}}^{\nu_{m,j+1}-1} \frac{1}{\lambda_{i+1}} + \sum_{i=\nu_{m,j}}^{\nu_{m,j+1}-1} \frac{1}{\lambda_{i+1}^2} \,.$$

A bound for the second part in (3.31) is

$$\begin{aligned}
&\left|\frac{\nu_{m,j+1}-\nu_{m,j}}{\lambda_{\nu_{m,j+1}+1}}-\Delta\right| \\
&\quad\le \max\left\{\left|\Delta\frac{\lambda_{\nu_{m,j}}}{\lambda_{\nu_{m,j+1}+1}}+\frac{\lambda_{\nu_{m,j}}}{\lambda_{\nu_{m,j+1}+1}\lambda_{\nu_{m,j}+1}}-\Delta\right|,\left|\Delta\frac{\lambda_{\nu_{m,j+1}}}{\lambda_{\nu_{m,j+1}+1}}+\frac{\lambda_{\nu_{m,j+1}}}{\lambda^2_{\nu_{m,j+1}+1}}-\Delta\right|\right\} \\
&\qquad\le \Delta\max\left\{\left|\frac{\lambda_{\nu_{m,j}}}{\lambda_{\nu_{m,j+1}+1}}-1\right|,\left|\frac{\lambda_{\nu_{m,j+1}}}{\lambda_{\nu_{m,j+1}+1}}-1\right|\right\}+\frac{1}{\lambda_{\nu_{m,j}}} \\
&\qquad\le \Delta\left|\frac{\lambda_{\nu_{m,j+1}+1}-\lambda_{\nu_{m,j}}}{\lambda_{\nu_{m,j+1}+1}}\right|+\frac{1}{\lambda_{\nu_{m,j}}}\,.
\end{aligned}$$

With Lemma A.6 follows that the $\lambda_n=\lambda_{\min}(\mathbf{F}_n^\top\mathbf{F}_n)$ increases at most linearly in n:

$$\lambda_{\nu_{m,j+1}+1}-\lambda_{\nu_{m,j}}\le\lambda_{\max}\left(\sum_{i=\nu_{m,j}+1}^{\nu_{m,j+1}+1}\boldsymbol{f}(\boldsymbol{x}_i)\boldsymbol{f}(\boldsymbol{x}_i)^\top\right)\le\nu_{m,j+1}-\nu_{m,j}+1\,.$$

The leads to

$$\frac{\lambda_{\nu_{m,j+1}+1}-\lambda_{\nu_{m,j}}}{\lambda_{\nu_{m,j+1}+1}}\le\Delta\frac{\lambda_{\nu_{m,j}}}{\lambda_{\nu_{m,j+1}+1}}+\frac{\lambda_{\nu_{m,j}}}{\lambda_{\nu_{m,j+1}+1}\lambda_{\nu_{m,j}+1}}+\frac{1}{\lambda_{\nu_{m,j+1}+1}}\le\Delta+\frac{2}{\lambda_{\nu_{m,j+1}}}$$

and

$$\left|\frac{\nu_{m,j+1}-\nu_{m,j}}{\lambda_{\nu_{m,j+1}+1}}-\Delta\right|\le\Delta^2+\frac{\Delta+3}{\lambda_{\nu_{m,j}}}\,.$$

Combining (3.30) and (3.31) with their bounds and taking the limit with respect to m yields

$$\begin{aligned}
&\lim_{m\to\infty}\left\|\sum_{i=\nu_{m,j}}^{\nu_{m,j+1}-1}\mathbf{I}(\boldsymbol{\theta}(j\Delta),\mathbf{F}_{i+1})^{-1}\tilde{\boldsymbol{g}}_{m,j}-\Delta\tilde{\mathbf{M}}_{m,j}\tilde{\boldsymbol{g}}_{m,j}\right\| \\
&\qquad\le K_3\lim_{m\to\infty}\left(\Delta\sum_{i=\nu_{m,j}}^{\nu_{m,j+1}-1}\frac{1}{\lambda_{i+1}}+\sum_{i=\nu_{m,j}}^{\nu_{m,j+1}-1}\frac{1}{\lambda^2_{i+1}}+\Delta^2+\frac{\Delta+3}{\lambda_{\nu_{m,j}}}\right)=2K_3\Delta^2
\end{aligned}$$

and finally with (3.29)

$$\begin{aligned}
&\lim_{m\to\infty}\left\|\sum_{i=\nu_{m,j}}^{\nu_{m,j+1}-1}\mathbb{1}_{\Theta_0}(\hat{\boldsymbol{\theta}}_i)\mathbf{I}(\boldsymbol{\theta}(j\Delta),\mathbf{F}_{i+1})^{-1}\boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}_{i+1},\boldsymbol{\theta}(j\Delta))-\Delta\tilde{\mathbf{M}}_{m,j}\tilde{\boldsymbol{g}}_{m,j}\right\| \\
&\qquad\le K_2\Delta^2+2K_3\Delta^2=K_4\Delta^2\,.
\end{aligned}$$

The same follows for the remaining sum from $\nu(t_m+\lfloor t/\Delta\rfloor\Delta)$ to $\nu(t_m+t)-1$. The result follows since

$$\begin{aligned}
\lim_{k\to\infty}\|\mathcal{G}^{(1)}_{n_k,\nu(t_{n_k}+t)-1}-\mathcal{G}^{(2)}_{n_k,\nu(t_{n_k}+t)-1}\|&\le\sum_{j=0}^{\lfloor t/\Delta\rfloor-1}K_4\Delta^2+K_4\Big(t-\Big\lfloor\frac{t}{\Delta}\Big\rfloor\Delta\Big)^2 \\
&\le K_4\Delta\left(\sum_{j=0}^{\lfloor t/\Delta\rfloor-1}\Delta+\Big(t-\Big\lfloor\frac{t}{\Delta}\Big\rfloor\Delta\Big)\right)=K_4\Delta t\,,
\end{aligned}$$

which tends to 0, if $\Delta\longrightarrow 0$. □

The last step in the characterization of the limit is to investigate

$$\mathcal{G}^{(2)}_{n,\nu(t_n+t)-1} = \sum_{j=0}^{\lfloor t/\Delta \rfloor -1} \Delta\, \tilde{\mathbf{M}}_{n_k,j} \tilde{\boldsymbol{g}}_{n_k,j} + \Big(t - \Big\lfloor \frac{t}{\Delta} \Big\rfloor \Delta\Big) \tilde{\mathbf{M}}_{n_k,\lfloor t/\Delta \rfloor} \tilde{\boldsymbol{g}}_{n_k,\lfloor t\Delta \rfloor}\,.$$

We know, that $\mathcal{G}^{(0)}_{n,\nu(t_n+t)-1}$ is equicontinuous in the extended sense and by the same argument as in Lemma 2 the limit of the convergent subsequence is Lipschitz and absolutely continuous. Hence it can be written as an integral using some measurable function $\boldsymbol{\gamma} : \mathbb{R} \longrightarrow \mathbb{R}^p$. From Lemma 14 and Lemma 15 follows

$$\lim_{\Delta \to 0} \lim_{k \to \infty} \Big\| \sum_{j=0}^{\lfloor t/\Delta \rfloor -1} \Delta\, \tilde{\mathbf{M}}_{n_k,j} \tilde{\boldsymbol{g}}_{n_k,j} + \Big(t - \Big\lfloor \frac{t}{\Delta} \Big\rfloor \Delta\Big) \tilde{\mathbf{M}}_{n_k,\lfloor t/\Delta \rfloor} \tilde{\boldsymbol{g}}_{n_k,\lfloor t/\Delta \rfloor} - \int_0^t \boldsymbol{\gamma}(s) \mathrm{d}s \Big\| = 0\,. \quad (3.32)$$

What remains is to characterize the limits of $\tilde{\mathbf{M}}_{n_k,j}$ and $\tilde{\boldsymbol{g}}_{n_k,j}$ and with that $\boldsymbol{\gamma}$.

If the sequence of designs converges, then we can directly characterize the function $\boldsymbol{\gamma}$.

Lemma 16. *Let $\lambda_n \geq cn$ for some $c > 0$ and all $n \geq m$. Assume further, that for all $\boldsymbol{\theta} \in \Theta_0$*

$$\mathbf{M}(\boldsymbol{\theta}, \xi_n)^{-1} \longrightarrow \mathbf{M}(\boldsymbol{\theta}, \xi)^{-1}$$

and

$$\int_{\mathcal{X}} \boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}, \boldsymbol{\theta})\, \xi_n(\mathrm{d}\boldsymbol{x}) \longrightarrow \int_{\mathcal{X}} \boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}, \boldsymbol{\theta})\, \xi(\mathrm{d}\boldsymbol{x})$$

for $n \longrightarrow \infty$. Then

$$\lim_{k \to \infty} \Big\| \mathcal{G}^{(2)}_{n_k,\nu(t_{n_k}+t)} - \int_0^t \mathbf{M}(\boldsymbol{\theta}(s), \xi)^{-1} \int_{\mathcal{X}} \boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}, \boldsymbol{\theta}(s))\, \xi(\mathrm{d}\boldsymbol{x})\, \mathrm{d}s \Big\| = 0\,.$$

Proof. Because of the convergence in the prerequisites

$$\lim_{k \to \infty} \Big\| \tilde{\boldsymbol{g}}_{n_k,j} - \int_{\mathcal{X}} \boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}, \boldsymbol{\theta}(j\Delta))\, \xi(\mathrm{d}\boldsymbol{x}) \Big\| = 0$$

and

$$\lim_{k \to \infty} \| \tilde{\mathbf{M}}_{n_k,j} - \mathbf{M}(\boldsymbol{\theta}(j\Delta), \xi)^{-1} \| = 0\,.$$

Consequently also

$$\lim_{k \to \infty} \Big\| \tilde{\mathbf{M}}_{n_k,j} \tilde{\boldsymbol{g}}_{n_k,j} - \mathbf{M}(\boldsymbol{\theta}(j\Delta), \xi)^{-1} \int_{\mathcal{X}} \boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}, \boldsymbol{\theta}(j\Delta))\, \xi(\mathrm{d}\boldsymbol{x}) \Big\| = 0$$

and

$$\lim_{k \to \infty} \Big\| \mathcal{G}^{(2)}_{n_k,\nu(n_k+t)-1} - \Big(\Delta \sum_{j=1}^{\lfloor t/\Delta \rfloor \Delta} \mathbf{M}(\boldsymbol{\theta}(j\Delta), \xi)^{-1} \int_{\mathcal{X}} \boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}, \boldsymbol{\theta}(j\Delta))\, \xi(\mathrm{d}\boldsymbol{x}) - \Big(t - \Big\lfloor \frac{t}{\Delta} \Big\rfloor \Delta\Big) \mathbf{M}(\boldsymbol{\theta}(j\Delta), \xi)^{-1} \int_{\mathcal{X}} \boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}, \boldsymbol{\theta}(j\Delta))\, \xi(\mathrm{d}\boldsymbol{x}) \Big) \Big\| = 0\,.$$

Taking the limit for $\Delta \longrightarrow 0$ yields the result. □

If the design does not converge, we can still do some analysis. We will briefly present some ideas into the direction of set-valued analysis and differential inclusions. Even though further development in this direction is beyond the scope of this thesis the ideas look promising.

While it is not possible to characterize the limit as a single valued function, we can say something about the sets, which contain the values of $\tilde{\mathbf{M}}_{n_k,j}$ and $\tilde{\boldsymbol{g}}_{n_k,j}$.

For the matrix term $\tilde{\mathbf{M}}_{n_k,j}$ we have that for all $\|\boldsymbol{v}\| = 1$ and $\boldsymbol{\theta}(j\Delta) \in \Theta_0$

$$\begin{aligned} 0 < \min_{i=\nu_{m,j},\dots,\nu_{m,j+1}-1} \lambda_{i+1} \boldsymbol{v}^\top \mathbf{I}(\boldsymbol{\theta}(j\Delta), \mathbf{F}_{i+1})^{-1} \boldsymbol{v} \\ \leq \boldsymbol{v}^\top \tilde{\mathbf{M}}_{n_k,j} \boldsymbol{v} \leq \max_{i=\nu_{m,j},\dots,\nu_{m,j+1}-1} \lambda_{i+1} \boldsymbol{v}^\top \mathbf{I}(\boldsymbol{\theta}(j\Delta), \mathbf{F}_{i+1})^{-1} \boldsymbol{v} \leq K \end{aligned} \tag{3.33}$$

for some $K > 0$, depending only on Θ_0 and $\mathcal{X}$. I.e. $\tilde{\mathbf{M}}_{n_k,j}$ lies in a convex and compact subset of the nonnegative definite matrices. Denote this subset by $\widetilde{\mathcal{M}}_{\boldsymbol{\theta}(j\Delta)}$. Note that

$$\lambda_{i+1} \boldsymbol{v}^\top \mathbf{I}(\boldsymbol{\theta}(j\Delta), \mathbf{F}_{i+1})^{-1} \boldsymbol{v} = \frac{\lambda_{i+1}}{i+1} \boldsymbol{v}^\top \mathbf{M}(\boldsymbol{\theta}(j\Delta), \xi_{i+1})^{-1} \boldsymbol{v}\,,$$

and

$$\frac{\lambda_{i+1}}{i+1} = \frac{\lambda_{\min}(\mathbf{F}_{i+1}^\top \mathbf{F}_{i+1})}{i+1} = \lambda_{\min}\Big(\int_{\mathcal{X}} \boldsymbol{f}(\boldsymbol{x}) \boldsymbol{f}(\boldsymbol{x})^\top \xi_{i+1}(\mathrm{d}\boldsymbol{x}) \Big)\,.$$

Denote $\lambda_{\min}(\xi) := \lambda_{\min}\Big(\int_{\mathcal{X}} \boldsymbol{f}(\boldsymbol{x}) \boldsymbol{f}(\boldsymbol{x})^\top \xi(\mathrm{d}\boldsymbol{x}) \Big)$. With this notation follows

$$\widetilde{\mathcal{M}}_{\boldsymbol{\theta}} \subseteq \overline{\mathrm{conv}} \left\{ \mathbf{A} \in \mathbb{R}^{p\times p} \,\Big|\, \exists (\tilde{\xi}_i)_{i\geq 1} \subset \Xi : \mathbf{A} = \lim_{i\to\infty} \lambda_{\min}\big(\tilde{\xi}_i\big) \mathbf{M}(\boldsymbol{\theta}, \tilde{\xi}_i)^{-1}, \sup_{\|\boldsymbol{v}\|=1} \boldsymbol{v}^\top \mathbf{A} \boldsymbol{v} \leq K \right\}.$$

The second part of interest $\tilde{\boldsymbol{g}}_{n_k,j}$, can be written as

$$\tilde{\boldsymbol{g}}_{n_k,j} = \frac{1}{\nu_{n_k,j+1} - \nu_{n_k,j}} \sum_{i=\nu_{n_k,j}}^{\nu_{n_k,j+1}-1} \boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}_{i+1}, \boldsymbol{\theta}(j\Delta)) = \int_{\mathcal{X}} \boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}, \boldsymbol{\theta}(j\Delta)) \xi(\mathrm{d}\boldsymbol{x})$$

for some design $\xi \in \Xi$. It follows, that

$$\tilde{\boldsymbol{g}}_{n_k,j} \in \overline{\mathrm{conv}} \Big\{ \boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}, \boldsymbol{\theta}(j\Delta)) \,\Big|\, \boldsymbol{x} \in \mathcal{X} \Big\}\,.$$

Hence the product $\tilde{\mathbf{M}}_{n_k,j} \tilde{\boldsymbol{g}}_{n_k,j}$ is an element of the set

$$\left\{ \mathbf{A}\boldsymbol{w} \,\Big|\, \mathbf{A} \in \widetilde{\mathcal{M}}_{\boldsymbol{\theta}}, \exists \xi \in \Xi : \boldsymbol{w} = \int_{\mathcal{X}} \boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}, \boldsymbol{\theta}) \xi(\mathrm{d}\boldsymbol{x}) \right\}.$$

Also there exists a set-valued function $\boldsymbol{\Gamma}(\boldsymbol{\theta})$, with

$$\boldsymbol{\Gamma}(\boldsymbol{\theta}) \subseteq \left\{ \mathbf{A}\boldsymbol{w} \,\Big|\, \mathbf{A} \in \widetilde{\mathcal{M}}_{\boldsymbol{\theta}}, \exists \xi \in \Xi : \boldsymbol{w} = \int_{\mathcal{X}} \boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}, \boldsymbol{\theta}) \xi(\mathrm{d}\boldsymbol{x}) \right\},$$

such that

$$\lim_{k\to\infty} \inf_{\boldsymbol{w} \in \boldsymbol{\Gamma}(\boldsymbol{\theta}(j\Delta))} \|\tilde{\mathbf{M}}_{n_k,j} \tilde{\boldsymbol{g}}_{n_k,j} - \boldsymbol{w}\| = 0\,.$$

Because of (3.32) follows $\boldsymbol{\gamma}(t) \in \boldsymbol{\Gamma}(\boldsymbol{\theta}(t))$. So $\boldsymbol{\Gamma}$ might be used to study the asymptotic behavior further.

Especially if $\lambda_n \geq cn$ for some $c > 0$, this seems promising. In this case exists a lower bound for (3.33), which is larger than 0, and the set $\widetilde{\mathcal{M}}_{\boldsymbol{\theta}}$ becomes more tractable:

$$\widetilde{\mathcal{M}}_{\boldsymbol{\theta}} \subseteq \overline{\operatorname{conv}}\Big\{\lambda_{\min}(\xi)\mathbf{M}(\boldsymbol{\theta},\xi)^{-1} \,\Big|\, \xi \in \Xi, \lambda_{\min}(\mathbf{M}(\boldsymbol{\theta},\xi)) \geq \tilde{K}\Big\},$$

for some $\tilde{K} > 0$. All matrices in $\widetilde{\mathcal{M}}_{\boldsymbol{\theta}}$ are nonsingular.

$$\boldsymbol{\Gamma}(\boldsymbol{\theta}) \subseteq \left\{\mathbf{A}\boldsymbol{w} \,\middle|\, \mathbf{A} \in \widetilde{\mathcal{M}}_{\boldsymbol{\theta}}, \exists \xi \in \Xi : \lambda_{\min}(\mathbf{M}(\boldsymbol{\theta},\xi)) > 0 \quad \text{and} \quad \boldsymbol{w} = \int_{\mathcal{X}} \boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x},\boldsymbol{\theta})\xi(\mathrm{d}\boldsymbol{x})\right\}.$$

It follows, that $\boldsymbol{\Gamma}(\bar{\boldsymbol{\theta}}) = \{\mathbf{0}\}$ and that this is the only "equilibrium" in the sense, that $\mathbf{0} \in \boldsymbol{\Gamma}(\boldsymbol{\theta})$: Since all matrices in $\mathbf{A} \in \widetilde{\mathcal{M}}_{\boldsymbol{\theta}}$ are nonsingular, only the vector $\mathbf{0}$ is mapped onto $\mathbf{0}$. Hence $\mathbf{0} \in \boldsymbol{\Gamma}(\boldsymbol{\theta})$ if and only if $\int_{\mathcal{X}} \boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x},\boldsymbol{\theta})\xi(\mathrm{d}\boldsymbol{x}) = \mathbf{0}$ for some ξ. The mean value theorem yields

$$\int_{\mathcal{X}} \boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x},\boldsymbol{\theta})\xi(\mathrm{d}\boldsymbol{x}) = \int_{\mathcal{X}} \boldsymbol{f}(\boldsymbol{x})\boldsymbol{f}(\boldsymbol{x})^\top \psi(\boldsymbol{f}(\boldsymbol{x})^\top\boldsymbol{\theta})G'(\boldsymbol{f}(\boldsymbol{x})^\top\boldsymbol{\theta}^*)\xi(\mathrm{d}\boldsymbol{x})(\boldsymbol{\theta}-\bar{\boldsymbol{\theta}}),$$

where $\boldsymbol{\theta}^*$ is on the line segment connecting $\boldsymbol{\theta}$ and $\bar{\boldsymbol{\theta}}$. Since the design ξ is nonsingular, the matrix

$$\int_{\mathcal{X}} \boldsymbol{f}(\boldsymbol{x})\boldsymbol{f}(\boldsymbol{x})^\top \psi(\boldsymbol{f}(\boldsymbol{x})^\top\boldsymbol{\theta})G'(\boldsymbol{f}(\boldsymbol{x})^\top\boldsymbol{\theta}^*)\xi(\mathrm{d}\boldsymbol{x})$$

is nonsingular, too. Hence $\int_{\mathcal{X}} \boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x},\boldsymbol{\theta})\xi(\mathrm{d}\boldsymbol{x}) = \mathbf{0}$ if and only if $\boldsymbol{\theta} = \bar{\boldsymbol{\theta}}$.

3.5. A Convergence Result

Using Lemma 16, we will show that the maximum likelihood estimator converges under relatively strong conditions. We state some of the assumptions beforehand:

(I) Let $\mathcal{X}$ be compact.

(II) Let G be twice continuously differentiable and strictly increasing.

(III) Let $0 < G(t) < 1$ for all $t \in \mathbb{R}$.

(IV) Assume that there exists $m \in \mathbb{N}$ such that $\mathbf{F}_n^\top\mathbf{F}_n$ is nonsingular almost surely for all $n \geq m$.

(V) Let $\lim_{n\to\infty} \lambda_{\min}(\mathbf{F}_n^\top\mathbf{F}_n) = \infty$ almost surely.

(VI) Assume that there exist $c > 0$ and $m \in \mathbb{N}$, such that for all $n \geq m$

$$\lambda_{\min}(\mathbf{F}_n^\top\mathbf{F}_n) \geq cn$$

almost surely.

Theorem 6. *Assume (I) to (VI) and let*

$$\lim_{n\to\infty} \sup\left\{\|\widetilde{\mathcal{S}}_{n,k}^{(0)}\| : k = n, \ldots, \nu(t_n + t) - 1\right\} = 0$$

almost surely for all $t > 0$. Assume that the sequence of designs converges almost surely as in the prerequisites of Lemma 16.

Assume that there exists $C > 0$ such that $\sup_{n\geq m} \|\hat{\boldsymbol{\theta}}_n - \bar{\boldsymbol{\theta}}\| \leq C$ almost surely. then $\lim_{n\to\infty} \|\hat{\boldsymbol{\theta}}_n - \bar{\boldsymbol{\theta}}\| = 0$ almost surely.

Proof. The assumptions ensure, that Lemma 8 and Lemma 10 hold with Θ_0 being the closed ball with center $\bar{\boldsymbol{\theta}}$ and radius C.

From Lemma 16 follows, that the limiting behavior of $\hat{\boldsymbol{\theta}}_n$ is described by solutions of the differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\theta}(t) = \mathbf{M}(\boldsymbol{\theta}(t), \xi)^{-1} \int_{\mathcal{X}} \boldsymbol{g}_{\bar{\theta}}(\boldsymbol{x}, \boldsymbol{\theta}(t))\, \xi(\mathrm{d}\boldsymbol{x})\,.$$

This differential equation has one asymptotically stable point, which is $\bar{\boldsymbol{\theta}}$. This is a consequence of the fact that the design ξ is nonsingular. A Lyapunov function, to show this is

$$L(\boldsymbol{\theta}) := \int_{\mathcal{X}} l(\bar{\boldsymbol{\theta}}, \boldsymbol{f}(\boldsymbol{x}), G(\boldsymbol{f}(\boldsymbol{x})^\top \bar{\boldsymbol{\theta}})) - l(\boldsymbol{\theta}, \boldsymbol{f}(\boldsymbol{x}), G(\boldsymbol{f}(\boldsymbol{x})^\top \bar{\boldsymbol{\theta}}))\xi(\mathrm{d}\boldsymbol{x})\,.$$

Since all solutions of the differential equation converge to $\bar{\boldsymbol{\theta}}$ follows the almost sure convergence of the estimator. □

4. Adaptive Wynn Algorithm

A special choice for the design sequence is if the design points are generated by an adaptive Wynn algorithm. Its basic steps were introduced in Section 2.5, more precisely in (2.40) and (2.41). In this chapter we show and discuss results concerning the asymptotic behavior of this algorithm.

We will assume that Θ is compact. Hence the estimator $\hat{\boldsymbol{\theta}}_n$ always exists as a maximum in Θ, but it may happen, that the estimate is on the boundary of the parameter space. In order to use the recursion from Lemma 6 we have to ensure that there are finitely many estimates on the boundary. On the other hand e.g. Lemma 8 can be applied directly.

4.1. Information Tends to Infinity

The adaptive Wynn algorithm chooses the next design point as

$$\boldsymbol{x}_{n+1} := \arg\max_{\boldsymbol{x}\in\mathcal{X}} \left(d(\boldsymbol{f}(\boldsymbol{x})^\top \hat{\boldsymbol{\theta}}_n) \boldsymbol{f}(\boldsymbol{x})^\top \mathbf{M}(\hat{\boldsymbol{\theta}}_n, \xi_n)^{-1} \boldsymbol{f}(\boldsymbol{x}) - p \right).$$

A first problem to solve is that the weighted information matrix is bounded below by a positive definite matrix and consequently

$$\lambda_{\min}(\mathbf{I}(\boldsymbol{\theta}, \mathbf{F}_n)) > c\, n$$

for some $c > 0$. Otherwise $\mathbf{M}(\hat{\boldsymbol{\theta}}_n, \xi_n)$ would become singular. If the design space $\mathcal{X}$ consists of a finite number of design points and the Θ is bounded Pronzato (2010) showed that there are at least p design points, such that the weights at these points are bounded below by some constant.

In the following lemma we show that the sequence of weighted information matrices $(\mathbf{M}(\boldsymbol{\theta}, \xi_n))_{n\geq m}$ has a subsequence of nonsingular matrices for all $\boldsymbol{\theta} \in \Theta$. As in the previous chapter we use the notation $\lambda_n := \lambda_{\min}(\mathbf{F}_n^\top \mathbf{F}_n)$.

Lemma 17. *Let $\mathcal{X}$ and Θ be compact. Let $(\boldsymbol{\theta}_n)_{n\geq 1}$ be an arbitrary sequence in Θ. Let d be continuous and assume $d(t) > 0$ for all $t \in \mathbb{R}$. Let $\mathbf{F}_m^\top \mathbf{F}_m$ be nonsingular for some $m \in \mathbb{N}$. For $n \geq m$ let the design points be chosen using*

$$\boldsymbol{x}_{n+1} := \arg\max_{\boldsymbol{x}\in\mathcal{X}} \left(d(\boldsymbol{f}(\boldsymbol{x})^\top \boldsymbol{\theta}_n) \boldsymbol{f}(\boldsymbol{x})^\top \mathbf{M}(\boldsymbol{\theta}_n, \xi_n)^{-1} \boldsymbol{f}(\boldsymbol{x}) - p \right).$$

Then $\lim_{n\to\infty} \lambda_{\min}(\mathbf{F}_n^\top \mathbf{F}_n) = \infty$ and there exist constants $0 < c \leq C < \infty$, such that

$$c \leq \limsup_{n\to\infty} \frac{\lambda_{\min}(\mathbf{F}_n^\top \mathbf{F}_n)}{n} \leq C\,.$$

Proof. Because $\mathbf{F}_n^\top \mathbf{F}_n$ is nonsingular for $n = m \in \mathbb{N}$, it is nonsingular for all $n \geq m$. The upper bound follows, because $\mathcal{X}$ is compact and

$$\lambda_n \leq \lambda_{\max}(\mathbf{F}_n^\top \mathbf{F}_n) \leq \operatorname{tr}(\mathbf{F}_n^\top \mathbf{F}_n) \leq n \max_{\boldsymbol{x} \in \mathcal{X}} \|\boldsymbol{f}(\boldsymbol{x})\|^2 .$$

For the lower bound note that

$$d(\boldsymbol{f}(\boldsymbol{x}_{n+1})^\top \boldsymbol{\theta}) \boldsymbol{f}(\boldsymbol{x}_{n+1})^\top \mathbf{M}(\boldsymbol{\theta}, \xi_n)^{-1} \boldsymbol{f}(\boldsymbol{x}_{n+1}) = \max_{\boldsymbol{v} \in \mathbb{R}^p} \left(2\sqrt{d(\boldsymbol{f}(\boldsymbol{x}_{n+1})^\top \boldsymbol{\theta})} \boldsymbol{f}(\boldsymbol{x}_{n+1})^\top \boldsymbol{v} - \boldsymbol{v}^\top \mathbf{M}(\boldsymbol{\theta}, \xi_n) \boldsymbol{v} \right).$$

From this follows, with (3.7) that for all $n \geq m$

$$\max_{\boldsymbol{v} \in \mathbb{R}^p} \left(2\sqrt{d(\boldsymbol{f}(\boldsymbol{x}_{n+1})^\top \boldsymbol{\theta})} \boldsymbol{f}(\boldsymbol{x}_{n+1})^\top \boldsymbol{v} - \boldsymbol{v}^\top \mathbf{M}(\boldsymbol{\theta}, \xi_n) \boldsymbol{v} \right) \geq \max_{\boldsymbol{v} \in \mathbb{R}^p} \left(2\sqrt{d(\boldsymbol{f}(\boldsymbol{x}_{n+1})^\top \boldsymbol{\theta})} \boldsymbol{f}(\boldsymbol{x}_{n+1})^\top \boldsymbol{v} - K_{\boldsymbol{\theta},n} \boldsymbol{v}^\top \mathbf{M}(\boldsymbol{\theta}_n, \xi_n) \boldsymbol{v} \right)$$

where

$$K_{\boldsymbol{\theta},n} := \max_{\boldsymbol{x} \in \mathcal{X}} \frac{d(\boldsymbol{f}(\boldsymbol{x})^\top \boldsymbol{\theta})}{d(\boldsymbol{f}(\boldsymbol{x})^\top \boldsymbol{\theta}_n)} .$$

Combining both inequalities yields

$$d(\boldsymbol{f}(\boldsymbol{x}_{n+1})^\top \boldsymbol{\theta}) \boldsymbol{f}(\boldsymbol{x}_{n+1})^\top \mathbf{M}(\boldsymbol{\theta}, \xi_n)^{-1} \boldsymbol{f}(\boldsymbol{x}_{n+1}) \geq K_1 d(\boldsymbol{f}(\boldsymbol{x}_{n+1})^\top \boldsymbol{\theta}_n) \boldsymbol{f}(\boldsymbol{x}_{n+1})^\top \mathbf{M}(\boldsymbol{\theta}_n, \xi_n)^{-1} \boldsymbol{f}(\boldsymbol{x}_{n+1}) \quad (4.1)$$

for all $\boldsymbol{\theta} \in \Theta$. The constant is given by

$$K_1 := \min_{\boldsymbol{\theta} \in \Theta} \min_{\boldsymbol{\theta}_1 \in \Theta} \left(\left(\max_{\boldsymbol{x} \in \mathcal{X}} \frac{d(\boldsymbol{f}(\boldsymbol{x})^\top \boldsymbol{\theta})}{d(\boldsymbol{f}(\boldsymbol{x})^\top \boldsymbol{\theta}_1)} \right)^{-1} \min_{\boldsymbol{x} \in \mathcal{X}} \frac{d(\boldsymbol{f}(\boldsymbol{x})^\top \boldsymbol{\theta})}{d(\boldsymbol{f}(\boldsymbol{x})^\top \boldsymbol{\theta}_1)} \right).$$

As a consequence of Lemma A.5 and (3.7) we obtain a lower bound for the right-hand side of (4.1):

$$\begin{aligned} K_1 d(\boldsymbol{f}(\boldsymbol{x}_{n+1})^\top \boldsymbol{\theta}_n) \boldsymbol{f}(\boldsymbol{x}_{n+1})^\top \mathbf{M}(\boldsymbol{\theta}_n, \xi_n)^{-1} \boldsymbol{f}(\boldsymbol{x}_{n+1}) &\geq K_1 K \lambda_{\max}(\mathbf{M}(\boldsymbol{\theta}_n, \xi_n)^{-1}) = K_1 K n \lambda_{\max}(\mathbf{I}(\boldsymbol{\theta}_n, \mathbf{F}_n)^{-1}) \\ &= \frac{K_1 K n}{\lambda_{\min}(\mathbf{I}(\boldsymbol{\theta}_n, \mathbf{F}_n))} \geq \frac{K_2 n}{\lambda_n} , \quad (4.2) \end{aligned}$$

for some $K_2 > 0$. By Lemma A.1 part **(ii)**

$$\frac{\det(\mathbf{M}(\boldsymbol{\theta}, \xi_{n+1}))}{\det(\mathbf{M}(\boldsymbol{\theta}, \xi_n))} = \left(\frac{n}{n+1} \right)^p \left(1 + n^{-1} d(\boldsymbol{f}(\boldsymbol{x}_{n+1})^\top \boldsymbol{\theta}) \boldsymbol{f}(\boldsymbol{x}_{n+1})^\top \mathbf{M}(\boldsymbol{\theta}, \xi_n)^{-1} \boldsymbol{f}(\boldsymbol{x}_{n+1}) \right)$$

for all $\boldsymbol{\theta} \in \Theta$. Together with (4.1) and (4.2) this yields

$$\begin{aligned} \frac{\det(\mathbf{M}(\boldsymbol{\theta}, \xi_{n+1}))}{\det(\mathbf{M}(\boldsymbol{\theta}, \xi_n))} &\geq \left(\frac{n}{n+1} \right)^p (1 + K_2 \lambda_n^{-1}) \\ &= 1 + K_2 \lambda_n^{-1} + \left(\left(\frac{n}{n+1} \right)^p - 1 \right) (1 + K_2 \lambda_n^{-1}) \\ &\geq 1 + K_2 \lambda_n^{-1} + p \left(\frac{n}{n+1} - 1 \right) \left(1 + K_2 \lambda_n^{-1} \right) \\ &= 1 + K_2 \lambda_n^{-1} - p \lambda_n^{-1} \frac{\lambda_n + K_2}{n+1} = 1 + \lambda_n^{-1} \left(K_2 - p \frac{\lambda_n + K_2}{n+1} \right). \end{aligned}$$

Assume that there is no lower bound, i.e. $\limsup_{n\to\infty} \lambda_n n^{-1} = \lim_{n\to\infty} \lambda_n n^{-1} = 0$. Then there exists $K_3 > 0$, such that for a sufficiently large $m_1 \geq m$

$$\left(K_2 - p\frac{\lambda_n + K_2}{n+1}\right) \geq K_3 > 0$$

for all $n \geq m_1$ and consequently

$$\frac{\det(\mathbf{M}(\boldsymbol{\theta}, \xi_{n+1}))}{\det(\mathbf{M}(\boldsymbol{\theta}, \xi_n))} \geq 1 + \lambda_n^{-1} K_3 \,.$$

If we apply this recursively, we obtain

$$\frac{\det(\mathbf{M}(\boldsymbol{\theta}, \xi_{m_1+n+1}))}{\det(\mathbf{M}(\boldsymbol{\theta}, \xi_{m_1}))} \geq \prod_{i=m_1}^{m_1+n} \left(1 + \lambda_i^{-1} K_3\right) \geq 1 + K_3 \sum_{i=m_1}^{m_1+n} \lambda_i^{-1} \,.$$

The sum on the right-hand side tends to ∞ for $n \longrightarrow \infty$, which is a contradiction, because of the boundedness of the left-hand side. Hence the lower bound follows. Since λ_n is monotonously increasing, $\lim_{n\to\infty} \lambda_n = \infty$. □

The question is: Can it happen that $\liminf_{n\to} \lambda_n n^{-1} = 0$? Note that since λ_n tends to infinity, $\lambda_n n^{-1}$ can never be equal to 0. It can only be arbitrarily close.

We will consider the following subsequence of the λ_n: Let $c > 0$ be the lower bound from Lemma 17 and let $\delta \in (0, 1)$. Let n_1 be an index, such that $\lambda_{n_1} n_1^{-1} > c$, and define

$$n_{2k} := \inf\left\{n > n_{2k-1} \middle| \frac{\lambda_n}{n} < \delta c\right\} \quad \text{and} \quad n_{2k+1} := \inf\left\{n > n_{2k} \middle| \frac{\lambda_n}{n} > c\right\}.$$

Because λ_n is nondecreasing in n it follows that

$$\delta c > \frac{\lambda_{n_{2k}}}{n_{2k}} \geq \frac{\lambda_{n_{2k-1}}}{n_{2k}} \geq \frac{n_{2k-1}}{n_{2k}} \frac{\lambda_{n_{2k-1}}}{n_{2k-1}} > \frac{n_{2k-1}}{n_{2k}} c \,.$$

As a direct consequence we obtain

$$n_{2k-1} < \delta n_{2k} \tag{4.3}$$

and that

$$\frac{n_2}{\delta^{k-1}} < n_{2k} \,.$$

Thus n_{2k} grows exponentially in k.

The gap between n_{2k-1} and n_{2k} also has to grow. In fact it increases at least as fast as a multiple of n_{2k}:

$$(1 - \delta) n_{2k} < n_{2k} - n_{2k-1} \,.$$

Next we consider the other direction: How many steps does it take to reach the threshold c again? The speed of this is limited by the fact that

$$\lambda_{n_{2k+1}} - \lambda_{n_{2k}} \leq \lambda_{\max}(\mathbf{F}_{n_{2k+1}}^\top \mathbf{F}_{n_{2k+1}} - \mathbf{F}_{n_{2k}}^\top \mathbf{F}_{n_{2k}}) \leq (n_{2k+1} - n_{2k}) \max_{\boldsymbol{x}\in\mathcal{X}} \|\boldsymbol{f}(\boldsymbol{x})\| \,,$$

which follows from Lemma A.6. This leads to a similar bound as above:

$$c < \frac{\lambda_{n_{2k+1}}}{n_{2k+1}} = \frac{\lambda_{n_{2k}}}{n_{2k+1}} + \frac{\lambda_{n_{2k+1}} - \lambda_{n_{2k}}}{n_{2k+1}} \leq \frac{\lambda_{n_{2k}}}{n_{2k}} + \frac{(n_{2k+1} - n_{2k})}{n_{2k+1}} \max_{\boldsymbol{x} \in \mathcal{X}} \|\boldsymbol{f}(\boldsymbol{x})\|$$
$$< \delta c + \frac{(n_{2k+1} - n_{2k})}{n_{2k+1}} \max_{\boldsymbol{x} \in \mathcal{X}} \|\boldsymbol{f}(\boldsymbol{x})\|$$

$$\iff \qquad n_{2k} < \left(1 - \frac{(1-\delta)\, c}{\max_{\boldsymbol{x} \in \mathcal{X}} \|\boldsymbol{f}(\boldsymbol{x})\|}\right) n_{2k+1}\,. \tag{4.4}$$

This means the periods in which $\lambda_n n^{-1} > \delta c$ can be arbitrarily long. In an actual trajectory these will be even larger than suggested by (4.3) and (4.4), because $\lambda_n n^{-1}$ is not necessarily strictly decreasing or increasing, respectively. While this is no proof for the desired inequality $\lambda_n \geq cn$, it is a strong hint, that the inequality holds. A scenario in which for example $\lim_{k\to\infty} \lambda_{n_k} n_k^{-1/2} = 0$ is even less likely.

With regard to Lemma 5 in Chapter 3, the result yields that the log-likelihood function has a global maximum in $\mathbb{R}^p$. With the additional conditions of Lemma 7, follows that the difference between neighboring estimates tends to zero as λ_n^{-1}.

4.2. Convergence of the Design and Asymptotic Normality

These two properties follow, if the sequence of estimators converges. While we will state the results using almost sure convergence, they hold for convergence in probability. The proofs are mostly identical to those in (Pronzato, 2009) and (Pronzato, 2010). The main difference is, that the design space is not assumed to be finite.

We will start with the convergence of the design. It converges in the sense, that the determinant of the information matrix, i.e. the value of the D-criterion, converges almost surely to the value of the locally optimal design. Remember, that the D-criterion was defined as $\phi_D(\mathbf{M}(\boldsymbol{\theta}, \xi)) := \log\det(\mathbf{M}(\boldsymbol{\theta}, \xi))$.

Lemma 18. *Let $\mathcal{X}$ and Θ be compact. Assume that there exists $m \in \mathbb{N}$, such that $\mathbf{M}(\boldsymbol{\theta}, \xi_n)$ is nonsingular almost surely for all $n \geq m$ and all $\boldsymbol{\theta} \in \Theta$. Let d be continuously differentiable and $0 < d(t) < 1$ for all $t \in \mathbb{R}$. Let $\lim_{n\to\infty} \|\hat{\boldsymbol{\theta}}_n - \bar{\boldsymbol{\theta}}\| = 0$ almost surely, then*

$$\lim_{n\to\infty} \log\det(\mathbf{M}(\hat{\boldsymbol{\theta}}_n, \xi_n)) = \log\det(\mathbf{M}(\bar{\boldsymbol{\theta}}, \xi^*_{\bar{\boldsymbol{\theta}}}))$$

almost surely.

Proof. The only difference to Lemma 3 in Pronzato (2010, p. 225, for the proof see pp. 235) is that the design space $\mathcal{X}$ is not finite. The first step in the proof is to show, that for all $\epsilon > 0$ exists a $\delta > 0$, such that for all $\|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\| \leq \delta$ follows

$$\max_{\boldsymbol{x} \in \mathcal{X}} \left\| \sqrt{d(\boldsymbol{f}(\boldsymbol{x})^\top \boldsymbol{\theta})}\boldsymbol{f}(\boldsymbol{x}) - \sqrt{d(\boldsymbol{f}(\boldsymbol{x})^\top \bar{\boldsymbol{\theta}})}\boldsymbol{f}(\boldsymbol{x}) \right\| \leq \epsilon\,.$$

This continuity property follows in our case from

$$\left\| \sqrt{d(\boldsymbol{f}(\boldsymbol{x})^\top \boldsymbol{\theta})}\boldsymbol{f}(\boldsymbol{x}) - \sqrt{d(\boldsymbol{f}(\boldsymbol{x})^\top \bar{\boldsymbol{\theta}})}\boldsymbol{f}(\boldsymbol{x}) \right\| \leq \frac{1}{2} \|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\| \|\boldsymbol{f}(\boldsymbol{x})\|^2 \max_{\|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\| \leq \delta} \left| \frac{d'(\boldsymbol{f}(\boldsymbol{x})^\top \boldsymbol{\theta})}{\sqrt{d(\boldsymbol{f}(\boldsymbol{x})^\top \boldsymbol{\theta})}} \right|, \tag{4.5}$$

and is a consequence of the differentiability of d and the mean value theorem.

If there exists an $m \in \mathbb{N}$, such that $\mathbf{M}(\boldsymbol{\theta}, \xi_n)$ is nonsingular for all $n \geq m$, then

$$\max_{\boldsymbol{x} \in \mathcal{X}} \max_{\|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\| \leq \delta} \left| d(\boldsymbol{f}(\boldsymbol{x})^\top \boldsymbol{\theta}) \boldsymbol{f}(\boldsymbol{x})^\top \mathbf{M}(\boldsymbol{\theta}, \xi_n)^{-1} \boldsymbol{f}(\boldsymbol{x}) - d(\boldsymbol{f}(\boldsymbol{x})^\top \bar{\boldsymbol{\theta}}) \boldsymbol{f}(\boldsymbol{x})^\top \mathbf{M}(\bar{\boldsymbol{\theta}}, \xi_n)^{-1} \boldsymbol{f}(\boldsymbol{x}) \right| \leq C\epsilon$$

for all $n \geq m$ and some $C > 0$. This yields that for all $\epsilon > 0$ exists a $\delta > 0$, such that for all $n \geq m$

$$d(\boldsymbol{f}(\boldsymbol{x}_{n+1})^\top \bar{\boldsymbol{\theta}}) \boldsymbol{f}(\boldsymbol{x}_{n+1})^\top \mathbf{M}(\bar{\boldsymbol{\theta}}, \xi_n)^{-1} \boldsymbol{f}(\boldsymbol{x}_n) > \max_{\boldsymbol{x} \in \mathcal{X}} \Big(d(\boldsymbol{f}(\boldsymbol{x})^\top \bar{\boldsymbol{\theta}}) \boldsymbol{f}(\boldsymbol{x})^\top \mathbf{M}(\bar{\boldsymbol{\theta}}, \xi_n)^{-1} \boldsymbol{f}(\boldsymbol{x}) \Big) - \epsilon \,,$$

whenever $\|\hat{\boldsymbol{\theta}}_n - \bar{\boldsymbol{\theta}}\| < \delta$.

From this point on the proof is the same as in said article of Pronzato. As in other proofs for adaptive or nonadaptive versions of the Wynn algorithm, a contradiction is used to prove, that the sequence of designs will be close to the optimal value infinitely often. In a second step it is shown, that we will be arbitrarily close if n is large enough. □

The asymptotic normality yields, that the asymptotic variance is in fact given by the inverse of the Fisher information matrix, since $\mathbf{M}(\hat{\boldsymbol{\theta}}_n, \xi_n)^{-1}$ appears as the normalizing sequence. This justifies, that it is used in the D-criterion and adaptive Wynn algorithm to find the optimal design.

Theorem 7. *Let $\mathcal{X}$ and Θ be compact. Assume that there exists $m \in \mathbb{N}$, such that $\mathbf{M}(\boldsymbol{\theta}, \xi_n)$ is nonsingular almost surely for all $n \geq m$ and all $\boldsymbol{\theta} \in \Theta$. Let G be twice continuously differentiable and strictly increasing. Assume further, that $0 < G(t) < 1$ for all $t \in \mathbb{R}$. Let $\lim_{n \to \infty} \|\hat{\boldsymbol{\theta}}_n - \bar{\boldsymbol{\theta}}\| = 0$ almost surely, then*

$$\sqrt{n} \mathbf{M}(\hat{\boldsymbol{\theta}}_n, \xi_n)^{1/2} (\hat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}) \xrightarrow{d} \mathcal{N}_p(\mathbf{0}, \mathbf{E}_p) \,.$$

Proof. With a Taylor expansion of the score function around the actual parameter follows

$$0 = \boldsymbol{s}_n(\hat{\boldsymbol{\theta}}_n) = \boldsymbol{s}_n(\bar{\boldsymbol{\theta}}) + \mathbf{H}_n(\boldsymbol{\theta}_n^*, \mathbf{F}_n)(\hat{\boldsymbol{\theta}}_n - \bar{\boldsymbol{\theta}})$$

for $\boldsymbol{\theta}_n^*$ on the line segment connecting $\hat{\boldsymbol{\theta}}_n$ and $\bar{\boldsymbol{\theta}}$. Bringing the second term onto the other side yields

$$\boldsymbol{s}_n(\bar{\boldsymbol{\theta}}) = -\mathbf{H}_n(\boldsymbol{\theta}_n^*, \mathbf{F}_n)(\hat{\boldsymbol{\theta}}_n - \bar{\boldsymbol{\theta}}) \,. \tag{4.6}$$

Now we will show, that the normalized $\boldsymbol{s}_n(\bar{\boldsymbol{\theta}})$ is asymptotically normal and hence is the right-hand side of (4.6). Let $\boldsymbol{v} \in \mathbb{R}^p$ with $\|\boldsymbol{v}\| = 1$ and consider

$$\frac{1}{\sqrt{n}} \boldsymbol{v}^\top \mathbf{M}(\bar{\boldsymbol{\theta}}, \xi_{\bar{\boldsymbol{\theta}}}^*)^{-1/2} \boldsymbol{s}_n(\bar{\boldsymbol{\theta}}) \,.$$

By a central limit theorem for martingales (see Theorem A.2 in the appendix, compare with Corollary 3.1 in Hall and Heyde, 1980, p. 58) this is asymptotically standard normal, if

$$\mathrm{E}\Big(\frac{1}{\sqrt{n}} \boldsymbol{v}^\top \mathbf{M}(\bar{\boldsymbol{\theta}}, \xi_{\bar{\boldsymbol{\theta}}}^*)^{-1/2} \boldsymbol{s}_n(\bar{\boldsymbol{\theta}}) \Big) = 0 \tag{4.7}$$

for all $n \geq 1$ and

$$\frac{1}{n} \sum_{i=1}^{n} \mathrm{E}\Big((\boldsymbol{v}^\top \mathbf{M}(\bar{\boldsymbol{\theta}}, \xi_{\bar{\boldsymbol{\theta}}}^*)^{-1/2} \boldsymbol{f}(\boldsymbol{X}_i) \psi(\boldsymbol{f}(\boldsymbol{X}_i)^\top \bar{\boldsymbol{\theta}}) \varepsilon_i)^2 | \mathcal{F}_i \Big) \xrightarrow{p} 1 \,. \tag{4.8}$$

The conditional Lindeberg condition (A.5) in Theorem A.2 is fulfilled automatically, since the summands in (4.8) are almost surely bounded, i.e.

$$\left|\boldsymbol{v}^\top \mathbf{M}(\bar{\boldsymbol{\theta}}, \xi_{\bar{\boldsymbol{\theta}}}^*)^{-1/2} \boldsymbol{f}(\boldsymbol{X}_i) \psi(\boldsymbol{f}(\boldsymbol{X}_i)^\top \bar{\boldsymbol{\theta}})\, \varepsilon_i\right| \leq K_1$$

for some $K_1 > 0$. The conditional expectations of this summands of the normalized score function are

$$\mathrm{E}\Big(\boldsymbol{v}^\top \mathbf{M}(\bar{\boldsymbol{\theta}}, \xi_{\bar{\boldsymbol{\theta}}}^*)^{-1/2} \boldsymbol{f}(\boldsymbol{X}_i) \psi(\boldsymbol{f}(\boldsymbol{X}_i)^\top \bar{\boldsymbol{\theta}}) \varepsilon_i | \mathcal{F}_i\Big) = 0$$

and thus (4.7) holds. The second condition holds, since

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^{n} \mathrm{E}\Big((\boldsymbol{v}^\top \mathbf{M}(\bar{\boldsymbol{\theta}}, \xi_{\bar{\boldsymbol{\theta}}}^*)^{-1/2} \boldsymbol{f}(\boldsymbol{X}_i) \psi(\boldsymbol{f}(\boldsymbol{X}_i)^\top \bar{\boldsymbol{\theta}}) \varepsilon_i)^2 | \mathcal{F}_i\Big) \\
&= \frac{1}{n} \boldsymbol{v}^\top \mathbf{M}(\bar{\boldsymbol{\theta}}, \xi_{\bar{\boldsymbol{\theta}}}^*)^{-1/2} \mathbf{I}(\bar{\boldsymbol{\theta}}, \mathbf{F}_n) \mathbf{M}(\bar{\boldsymbol{\theta}}, \xi_{\bar{\boldsymbol{\theta}}}^*)^{-1/2} \boldsymbol{v} \\
&= \boldsymbol{v}^\top \mathbf{M}(\bar{\boldsymbol{\theta}}, \xi_{\bar{\boldsymbol{\theta}}}^*)^{-1/2} \mathbf{M}(\bar{\boldsymbol{\theta}}, \xi_n) \mathbf{M}(\bar{\boldsymbol{\theta}}, \xi_{\bar{\boldsymbol{\theta}}}^*)^{-1/2} \boldsymbol{v}
\end{aligned}$$

and because of the convergence of the design sequence

$$\lim_{n \to \infty} \boldsymbol{v}^\top \mathbf{M}(\bar{\boldsymbol{\theta}}, \xi_{\bar{\boldsymbol{\theta}}}^*)^{-1/2} \mathbf{M}(\bar{\boldsymbol{\theta}}, \xi_n) \mathbf{M}(\bar{\boldsymbol{\theta}}, \xi_{\bar{\boldsymbol{\theta}}}^*)^{-1/2} \boldsymbol{v} = 1 \tag{4.9}$$

almost surely. This establishes

$$\frac{1}{\sqrt{n}} \boldsymbol{v}^\top \mathbf{M}(\bar{\boldsymbol{\theta}}, \xi_{\bar{\boldsymbol{\theta}}}^*)^{-1/2} \boldsymbol{s}_n(\bar{\boldsymbol{\theta}}) \xrightarrow{d} \mathcal{N}(0, 1)$$

and since it holds for all $\|\boldsymbol{v}\| = 1$

$$\frac{1}{\sqrt{n}} \mathbf{M}(\bar{\boldsymbol{\theta}}, \xi_{\bar{\boldsymbol{\theta}}}^*)^{-1/2} \boldsymbol{s}_n(\bar{\boldsymbol{\theta}}) \xrightarrow{d} \mathcal{N}_p(\mathbf{0}, \mathbf{E}_p)\,.$$

For the right-hand side of (4.6) we have to show, that

$$-\frac{1}{\sqrt{n}} \mathbf{M}(\bar{\boldsymbol{\theta}}, \xi_{\bar{\boldsymbol{\theta}}}^*)^{-1/2} \mathbf{H}_n(\boldsymbol{\theta}_n^*, \mathbf{F}_n)(\hat{\boldsymbol{\theta}}_n - \bar{\boldsymbol{\theta}}) \tag{4.10}$$

behaves asymptotically like

$$\sqrt{n} \mathbf{M}(\hat{\boldsymbol{\theta}}_n, \xi_n)^{1/2} (\hat{\boldsymbol{\theta}}_n - \bar{\boldsymbol{\theta}})\,.$$

If we can show that it is asymptotically equivalent to

$$\sqrt{n} \mathbf{M}(\bar{\boldsymbol{\theta}}, \xi_n)^{1/2} (\hat{\boldsymbol{\theta}}_n - \bar{\boldsymbol{\theta}})\,, \tag{4.11}$$

then the last step follows from the convergence of the estimator.

To do this we will expand (4.10) with $\sqrt{n}\mathbf{M}(\bar{\boldsymbol{\theta}}, \xi_{\bar{\boldsymbol{\theta}}}^*)^{1/2}$:

$$-\frac{1}{\sqrt{n}} \mathbf{M}(\bar{\boldsymbol{\theta}}, \xi_{\bar{\boldsymbol{\theta}}}^*)^{-1/2} \mathbf{H}_n(\boldsymbol{\theta}_n^*, \mathbf{F}_n) \frac{1}{\sqrt{n}} \mathbf{M}(\bar{\boldsymbol{\theta}}, \xi_{\bar{\boldsymbol{\theta}}}^*)^{-1/2} \sqrt{n} \mathbf{M}(\bar{\boldsymbol{\theta}}, \xi_{\bar{\boldsymbol{\theta}}}^*)^{1/2} (\hat{\boldsymbol{\theta}}_n - \bar{\boldsymbol{\theta}})\,.$$

If we rewrite the Hessian matrix

$$\begin{aligned}-\mathbf{H}_n(\boldsymbol{\theta}_n^*, \mathbf{F}_n) &= \mathbf{I}(\boldsymbol{\theta}_n^*, \mathbf{F}_n) - \sum_{i=1}^n \psi'(\boldsymbol{f}(\boldsymbol{X}_i)^\top \boldsymbol{\theta}_n^*)\Big(Y_i - G(\boldsymbol{f}(\boldsymbol{X}_i)^\top \boldsymbol{\theta}_n^*)\Big)\boldsymbol{f}(\boldsymbol{X}_i)\boldsymbol{f}(\boldsymbol{X}_i)^\top \\ &= \mathbf{I}(\boldsymbol{\theta}_n^*, \mathbf{F}_n) - \mathbf{I}(\bar{\boldsymbol{\theta}}, \mathbf{F}_n) + \mathbf{I}(\bar{\boldsymbol{\theta}}, \mathbf{F}_n) - \sum_{i=1}^n \psi'(\boldsymbol{f}(\boldsymbol{X}_i)^\top \boldsymbol{\theta}_n^*)\boldsymbol{f}(\boldsymbol{X}_i)\boldsymbol{f}(\boldsymbol{X}_i)^\top \varepsilon_i \\ &\quad - \sum_{i=1}^n \psi'(\boldsymbol{f}(\boldsymbol{X}_i)^\top \boldsymbol{\theta}_n^*)\Big(G(\boldsymbol{f}(\boldsymbol{X}_i)^\top \bar{\boldsymbol{\theta}}) - G(\boldsymbol{f}(\boldsymbol{X}_i)^\top \boldsymbol{\theta}_n^*)\Big)\boldsymbol{f}(\boldsymbol{X}_i)\boldsymbol{f}(\boldsymbol{X}_i)^\top ,\end{aligned} \tag{4.12}$$

we see that $\mathbf{I}(\boldsymbol{\theta}_n^*, \mathbf{F}_n) - \mathbf{I}(\bar{\boldsymbol{\theta}}, \mathbf{F}_n)$ and the last sum on the right-hand side of (4.12) can be bounded by $nK_j\|\boldsymbol{\theta}_n^* - \bar{\boldsymbol{\theta}}\|$, $K_j > 0$, $j = 2, 3$. This follows from the properties of the function G, especially its differentiability, the compactness of $\mathcal{X}$ and the fact, that $\|\boldsymbol{\theta}_n^* - \bar{\boldsymbol{\theta}}\|$ is also bounded, because of the convergence of the estimates. The first bound is

$$\|\mathbf{I}(\boldsymbol{\theta}_n^*, \mathbf{F}_n) - \mathbf{I}(\bar{\boldsymbol{\theta}}, \mathbf{F}_n)\| = n\|\mathbf{M}(\boldsymbol{\theta}_n^*, \xi_n) - \mathbf{M}(\bar{\boldsymbol{\theta}}, \xi_n)\| \le nK_2\|\boldsymbol{\theta}_n^* - \bar{\boldsymbol{\theta}}\|$$

and hence

$$\|\frac{1}{n}\mathbf{M}(\bar{\boldsymbol{\theta}}, \xi_{\bar{\boldsymbol{\theta}}}^*)^{-1/2}\Big(\mathbf{I}(\boldsymbol{\theta}_n^*, \mathbf{F}_n) - \mathbf{I}(\bar{\boldsymbol{\theta}}, \mathbf{F}_n)\Big)\mathbf{M}(\bar{\boldsymbol{\theta}}, \xi_{\bar{\boldsymbol{\theta}}}^*)^{-1/2}\| \le K_2\|\mathbf{M}(\bar{\boldsymbol{\theta}}, \xi_{\bar{\boldsymbol{\theta}}}^*)^{-1}\|\|\boldsymbol{\theta}_n^* - \bar{\boldsymbol{\theta}}\| .$$

For the second term mentioned above we get

$$\begin{aligned}&\Big\|\sum_{i=1}^n \psi'(\boldsymbol{f}(\boldsymbol{X}_i)^\top \boldsymbol{\theta}_n^*)\Big(G(\boldsymbol{f}(\boldsymbol{X}_i)^\top \bar{\boldsymbol{\theta}}) - G(\boldsymbol{f}(\boldsymbol{X}_i)^\top \boldsymbol{\theta}_n^*)\Big)\boldsymbol{f}(\boldsymbol{X}_i)\boldsymbol{f}(\boldsymbol{X}_i)^\top\Big\| \\ &\quad \le n \max_{i=1,\dots,n}\Big\|\psi'(\boldsymbol{f}(\boldsymbol{X}_i)^\top \boldsymbol{\theta}_n^*)\Big(G(\boldsymbol{f}(\boldsymbol{X}_i)^\top \bar{\boldsymbol{\theta}}) - G(\boldsymbol{f}(\boldsymbol{X}_i)^\top \boldsymbol{\theta}_n^*)\Big)\boldsymbol{f}(\boldsymbol{X}_i)\boldsymbol{f}(\boldsymbol{X}_i)^\top\Big\| \\ &\quad \le n \max_{\boldsymbol{x}\in\mathcal{X}}\Big(\Big|\psi'(\boldsymbol{f}(\boldsymbol{x})^\top \boldsymbol{\theta}_n^*)G'(\boldsymbol{f}(\boldsymbol{x})^\top \boldsymbol{\theta}_n^{**})\Big|\,\|\boldsymbol{f}(\boldsymbol{x})\|^2\|\boldsymbol{\theta}_n^* - \bar{\boldsymbol{\theta}}\|\Big) \\ &\quad \le nK_3\|\boldsymbol{\theta}_n^* - \bar{\boldsymbol{\theta}}\|\end{aligned}$$

and $K_3\|\mathbf{M}(\bar{\boldsymbol{\theta}}, \xi_{\bar{\boldsymbol{\theta}}}^*)^{-1}\|\|\boldsymbol{\theta}_n^* - \bar{\boldsymbol{\theta}}\|$ as a bound for the normalized version.

For the remaining sum on the right-hand side of (4.12) follows, that it converges to 0 almost surely, since for all $\|\boldsymbol{v}\| = 1$

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \psi'(\boldsymbol{f}(\boldsymbol{X}_i)^\top \boldsymbol{\theta}_n^*)\Big(\boldsymbol{v}^\top \mathbf{M}(\bar{\boldsymbol{\theta}}, \xi_{\bar{\boldsymbol{\theta}}}^*)^{-1/2}\boldsymbol{f}(\boldsymbol{X}_i)\Big)^2 \varepsilon_i = 0$$

almost surely. This is a consequence of Theorem A.1, because

$$\begin{aligned}&\frac{1}{n}\boldsymbol{v}^\top \mathbf{M}(\bar{\boldsymbol{\theta}}, \xi_{\bar{\boldsymbol{\theta}}}^*)^{-1/2}\left(\sum_{i=1}^n \psi'(\boldsymbol{f}(\boldsymbol{X}_i)^\top \boldsymbol{\theta}_n^*)\boldsymbol{f}(\boldsymbol{X}_i)\boldsymbol{f}(\boldsymbol{X}_i)^\top \varepsilon_i\right)\mathbf{M}(\bar{\boldsymbol{\theta}}, \xi_{\bar{\boldsymbol{\theta}}}^*)^{-1/2}\boldsymbol{v} \\ &\qquad = \frac{1}{n}\sum_{i=1}^n \psi'(\boldsymbol{f}(\boldsymbol{X}_i)^\top \boldsymbol{\theta}_n^*)\Big(\boldsymbol{v}^\top \mathbf{M}(\bar{\boldsymbol{\theta}}, \xi_{\bar{\boldsymbol{\theta}}}^*)^{-1/2}\boldsymbol{f}(\boldsymbol{X}_i)\Big)^2 \varepsilon_i\end{aligned}$$

and

$$\begin{aligned}&\sum_{i=1}^n \mathrm{E}\left(\left(\frac{1}{i}\psi'(\boldsymbol{f}(\boldsymbol{X}_i)^\top \boldsymbol{\theta}_n^*)\Big(\boldsymbol{v}^\top \mathbf{M}(\bar{\boldsymbol{\theta}}, \xi_{\bar{\boldsymbol{\theta}}}^*)^{-1/2}\boldsymbol{f}(\boldsymbol{X}_i)\Big)^2 \varepsilon_i\right)^2 \Big| \mathcal{F}_i\right) \\ &\quad \le \sup_{\|\boldsymbol{\theta}_n^* - \bar{\boldsymbol{\theta}}\|} \max_{\boldsymbol{x}\in\mathcal{X}}\left|\Big(\boldsymbol{v}^\top \mathbf{M}(\bar{\boldsymbol{\theta}}, \xi_{\bar{\boldsymbol{\theta}}}^*)^{-1/2}\boldsymbol{f}(\boldsymbol{X}_i)\Big)^4 \psi'(\boldsymbol{f}(\boldsymbol{X}_i)^\top \boldsymbol{\theta}_n^*)\right| \sum_{i=1}^n \frac{1}{i^2}\mathrm{E}\Big(\varepsilon_i^2 | \mathcal{F}_i\Big) \le K_4 \sum_{i=1}^\infty \frac{1}{i^2} ,\end{aligned}$$

for some $K_4 > 0$. As in (4.9) the remaining term

$$\frac{1}{n}\mathbf{M}(\bar{\boldsymbol{\theta}}, \xi^*_{\bar{\boldsymbol{\theta}}})^{-1/2}\mathbf{I}(\bar{\boldsymbol{\theta}}, \mathbf{F}_n)\mathbf{M}(\bar{\boldsymbol{\theta}}, \xi^*_{\bar{\boldsymbol{\theta}}})^{-1/2}$$

converges to the identity matrix. It follows that (4.10) and (4.11) are asymptotically equivalent. This in combination with the convergence of the estimator establishes the proof. □

5. Simulations

This chapter will present the results of the simulation studies. While the description of results will be contained in this chapter, only a representative selection of graphs are included here. All other figures, which might also be referenced in the text, are in Appendix C.

5.1. Setup

The software package **R** (version 2.14.1, 64bit; see R Core Team, 2014) on a desktop computer with an AMD Phenom II x4 955 (3.2 GHz) as processor and the operating system Ubuntu 12.04 LTS was used to run the main simulations.

Starting with a fixed initial design, the observations from a binary response model were simulated. The design points were generated using the adaptive Wynn algorithm for D-optimal design. Each simulation had 5000 replications, with 500 steps each.

The models under consideration were the logit, probit, log-log and complementary log-log model, which were described in Example 1.

The choices for $\mathcal{X}$, Θ and the true parameter $\bar{\boldsymbol{\theta}}$ are presented in Table 5.1. The initial designs was $0, \pm 0.5, \pm 1.5$, with one observation per design point.

For all optimizations, i.e. calculation of the estimates and finding the design points, the build in "general-purpose" method of **R** was used.[1] Some problems in the optimization procedures arose because of numerical instabilities, if the value of the mean function G was very close to 0 or 1. This occurred especially for the probit model. Choosing the design region appropriately, solved this problem.

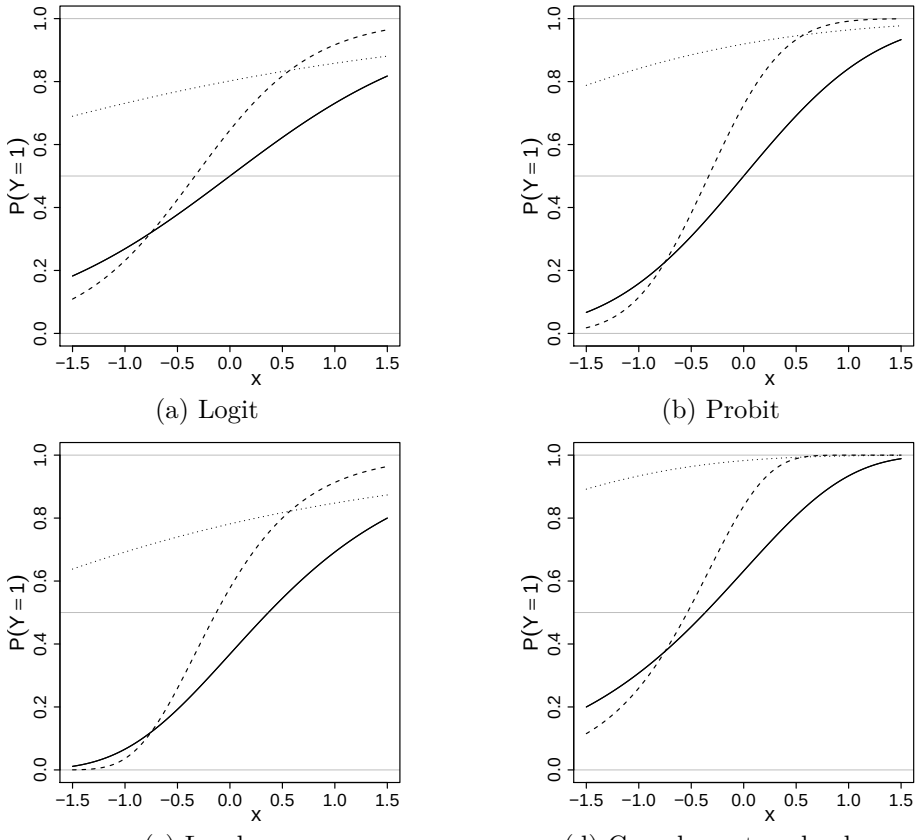
The parameter setting $(1.4 \quad 0.4)$ is a very special one, as can be seen in Figure 5.1. Especially in the complementary log-log case. The design region is not chosen appropriately. But it was of interest to see, how the procedure works in these situations.

The locally optimal designs for the simulation settings are given in Table 5.2. As a feature of the D-criterion, the weights are always 0.5.

Table 5.1.: $\mathcal{X}$, Θ and $\bar{\boldsymbol{\theta}}$ for the simulations

Θ	$\mathcal{X}$	$\bar{\boldsymbol{\theta}}^\top$
$[-2 \quad 2] \times [0 \quad 2]$	$[-1.5 \quad 1.5]$	$(0 \quad 1)$ $(0.6 \quad 1.8)$ $(1.4 \quad 0.4)$

[1] Because of the constraints it was "L-BFGS-B", which according to the built-in documentation is due to Byrd, Lu, Nocedal and Zhu (1995).



(a) Logit (b) Probit

(c) Log-log (d) Complementary log-log

Figure 5.1.: Probabilities of the models used in the simulation.
solid: $\bar{\boldsymbol{\theta}} = (0 \quad 1)^\top$, dashed: $\bar{\boldsymbol{\theta}} = (0.6 \quad 1.8)^\top$, dotted: $\bar{\boldsymbol{\theta}} = (1.4 \quad 0.4)^\top$

Table 5.2.: Locally D-optimal designs for the simulations

Model	$\bar{\boldsymbol{\theta}}^\top$	x_1	x_2
Logit	(0 1)	-1.5000	1.5000
	(0.6 1.8)	-1.1908	0.5241
	(1.4 0.4)	-1.5000	1.5000
Probit	(0 1)	-1.1381	1.1381
	(0.6 1.8)	-0.9656	0.2989
	(1.4 0.4)	-1.5000	1.5000
Log-log	(0 1)	-0.9796	1.3377
	(0.6 1.8)	-0.8776	0.4098
	(1.4 0.4)	-1.5000	1.5000
Complementary log-log	(0 1)	-1.3377	0.9796
	(0.6 1.8)	-1.0765	0.2109
	(1.4 0.4)	-1.5000	0.3481

5.2. Results

5.2.1. Separation of the Design Points

The separation of the design points in the sense of Section 2.3 was determined, too. Separation meant, that the relative interiors of the convex cones corresponding to observed 0's and 1's are disjoint: $\mathcal{C}_n^0 \cap \mathcal{C}_n^1 = \emptyset$. Otherwise there is an overlap in the design points. While in the setting of the simulations this was not necessary to ensure the existence of the estimate, it is of interest to verify the asymptotic existence from Lemma 5 and to find out how long it takes, until the estimates in $\mathbb{R}^2$ exist. Table 5.3 summarizes the results.

For the parameter settings $(0 \quad 1)$ and $(0.6 \quad 1.8)$, the number of steps, until overlap is reasonably small: To achieve 99%, between 20 and 30 observations are needed. If $\bar{\boldsymbol{\theta}} = (1.4 \quad 0.4)$ it takes more than two or even three times the number of observations for the complementary log-log and the probit model, if compared to the other two parameter settings. This can be explained with the corresponding probabilities shown in Figure 5.1: In the complementary log-log model for example, the probability to observe a 0 is approximately 0.1 on the left boundary of the design space and around 0.001 on the right. So it naturally needs more steps, to achieve an overlap in the design points corresponding to 0's and 1's. For the probit model the situation is a bit better, which is reflected in the values of the quantiles.

Table 5.3.: Overlap in the design points

			Sample Quantiles[2]				
Model	$\bar{\boldsymbol{\theta}}^\top$	Initial[1]	75%	90%	95%	99%	100%
Logit	(0 1)	0.3678	8	10	12	16	30
	(0.6 1.8)	0.1844	10	14	17	23	43
	(1.4 0.4)	0.1884	14	20	26	42	95
Probit	(0 1)	0.2564	8	10	12	16	26
	(0.6 1.8)	0.0312	10	14	16	22	34
	(1.4 0.4)	0.0348	24	35	46	73	175
Log-log	(0 1)	0.2938	8	10	12	16	25
	(0.6 1.8)	0.1412	10	14	17	29	48
	(1.4 0.4)	0.2170	12	17	22	39	78
complementary Log-log	(0 1)	0.1434	8	11	12	16	32
	(0.6 1.8)	0.0058	10	13	16	23	45
	(1.4 0.4)	0.0008	50	74	93	138	306

[1] Initial: Proportion of replications with overlap in the initial design.
[2] Sample Quantiles: These are for the number of steps, until there is overlap.

5.2.2. Distribution of the Estimates, Mean Squared Error and Bias

Distribution of the Estimates

The Figures 5.2 to 5.5 show histograms of the components $\hat{\theta}_1$ and $\hat{\theta}_2$ of the estimates and different sample sizes. From left to right these are 50, 250 and 500. The general shape of the histograms are very similar for all models, hence only the logit model (for all parameter values) and the complementary log-log model for one value is displayed.

For small sample sizes, the distribution of the components has a higher variability than that for large n, as was to be expected. The estimates accumulate around the actual value, which is represented by the dashed vertical line, and the shape of the histogram becomes more symmetric. The proportion of estimates taking values on the boundary is also decreasing as the sample sizes increases. This is visible especially in the cases where $\bar{\boldsymbol{\theta}} = (0.6 \quad 1.8)$ (e.g. Figure 5.3) and $\bar{\boldsymbol{\theta}} = (1.4 \quad 0.4)$ (e.g. Figure 5.4 and Figure 5.5).

For $\bar{\boldsymbol{\theta}} = (1.4 \quad 0.4)$ the distribution becomes visibly smoother for larger sample sizes. While there is a "baseline distribution" in Figure 5.5 (a), the spikes can be attributed to the replications, where no overlap occurred. (see Table 5.3)

Eigenvalues of the Mean Squared Error Matrix

In Figure 5.6 and Figure 5.7 the square root of the eigenvalues of the estimated mean squared error matrix $\mathbf{MSE}_n$ are displayed. They are all decreasing and, as indicated by the gray solid curves, are of order $\sqrt{n}$. Again the graph is qualitatively very similar between the models. The main difference is, the distance between the curves shown in the figures.

The standard setting $\bar{\boldsymbol{\theta}} = (0 \quad 1)$ shows the best performance. Also the eigenvalues are closer to each other. The bad choice of the design space for $\bar{\boldsymbol{\theta}} = (1.4 \quad 0.4)$ in the complementary log-log model is especially seen in the beginning steps. This is also visible for the probit model (Figure C.9), but not as distinctive as it is here.

Bias

The estimated bias for different sample sizes is shown in Table 5.4 and Table 5.5. In the standard case and if $\bar{\boldsymbol{\theta}} = (1.4 \quad 0.4)^\top$, the bias for the slope component θ_2 is decreasing. The bias for θ_1 is either decreasing or already small. For the bias we will write it is "small" or "decreasing", if its absolute value is.

For the remaining parameter setting the bias seems to increase. It is closer to 0 for sample sizes of 50 or 100, than it is for 500 observations. This can be explained by the fact, that the actual parameter is relatively close to the boundary of the parameter space and a considerable amount of estimates is still on the boundary. This can be seen in Figure 5.3 (c). Comparing the adaptive versions to one with a fixed sequence of design points, which alternated between two design points, showed a bias of the same order.

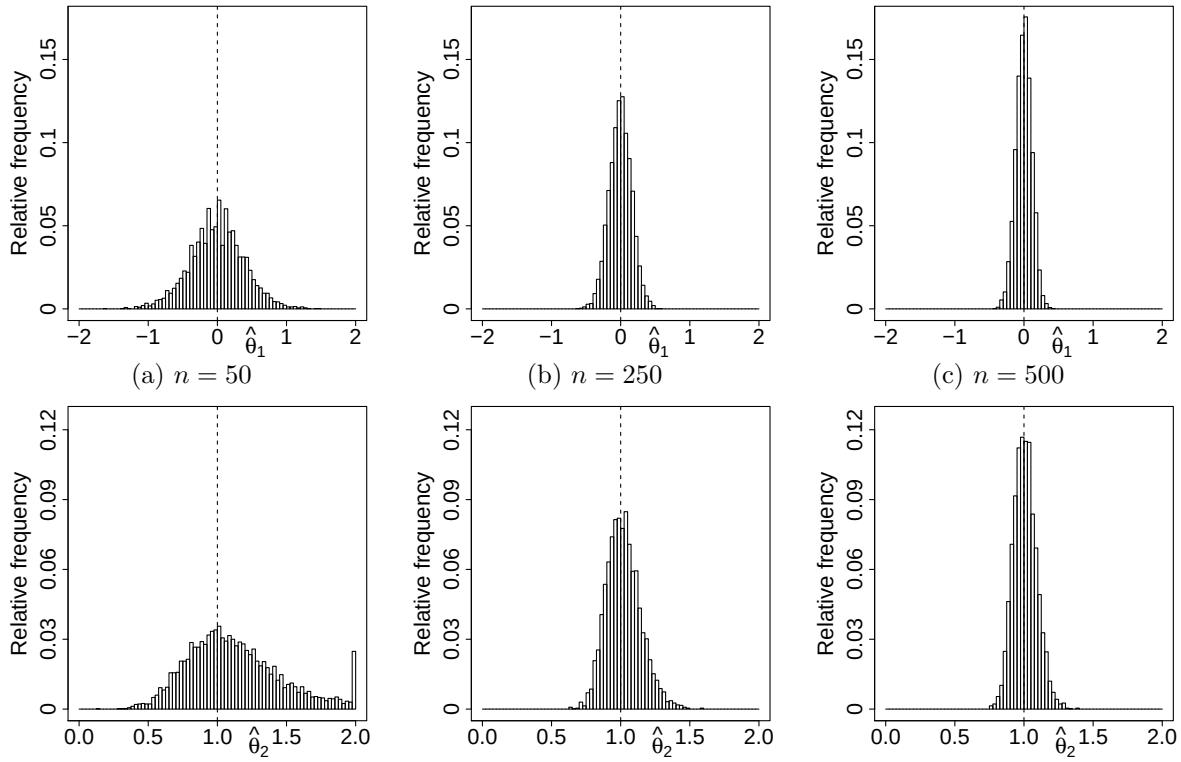


Figure 5.2.: Histograms of the components of the estimates and different sample sizes for the logit model with $\bar{\boldsymbol{\theta}} = (0 \quad 1)^\top$

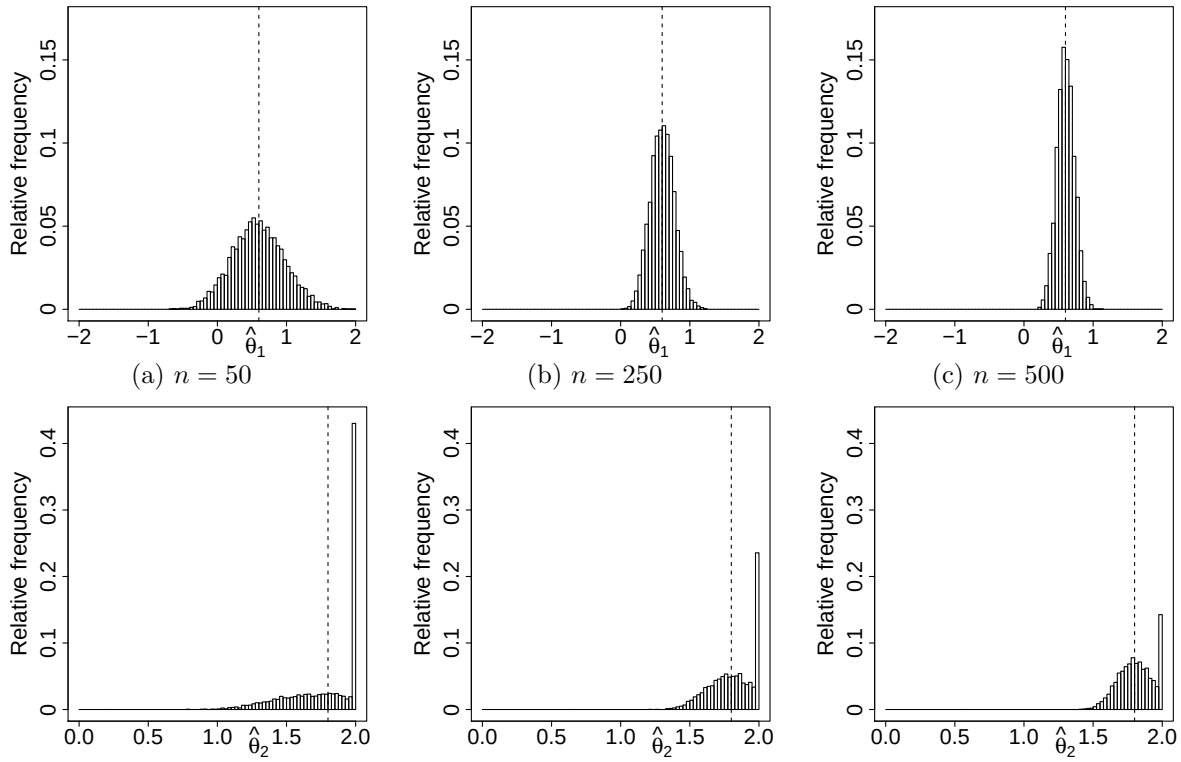
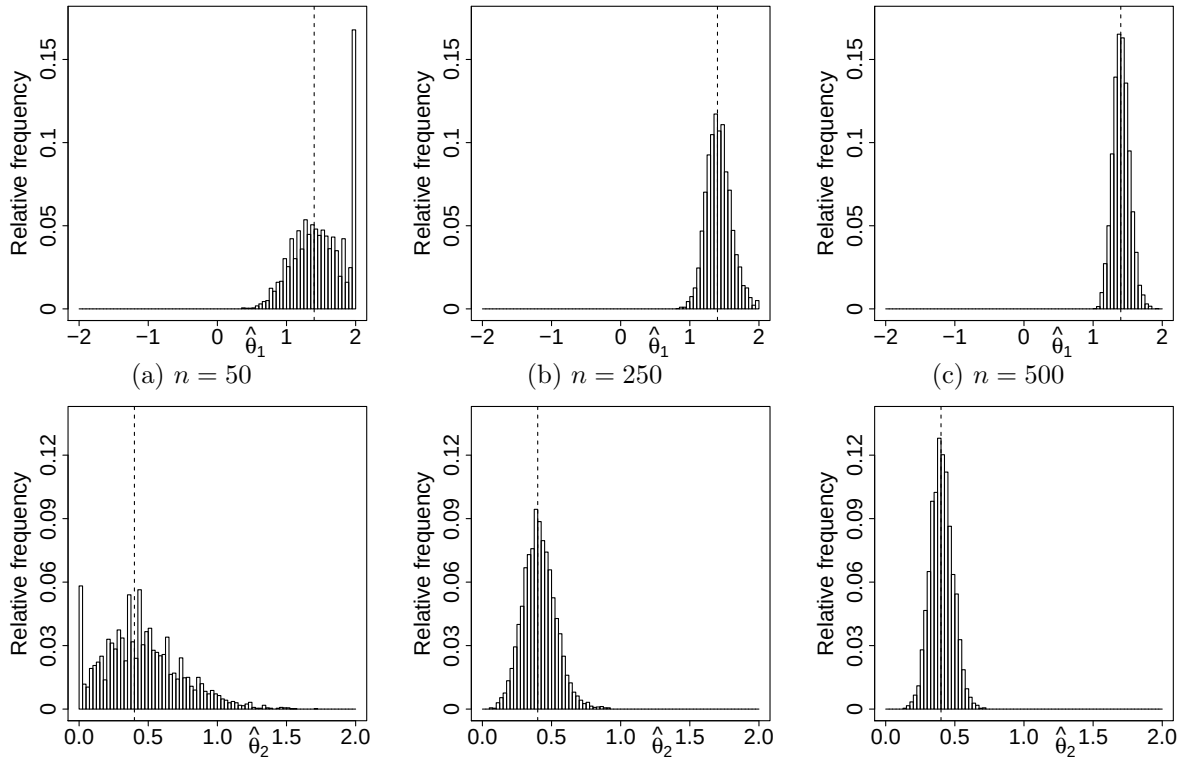
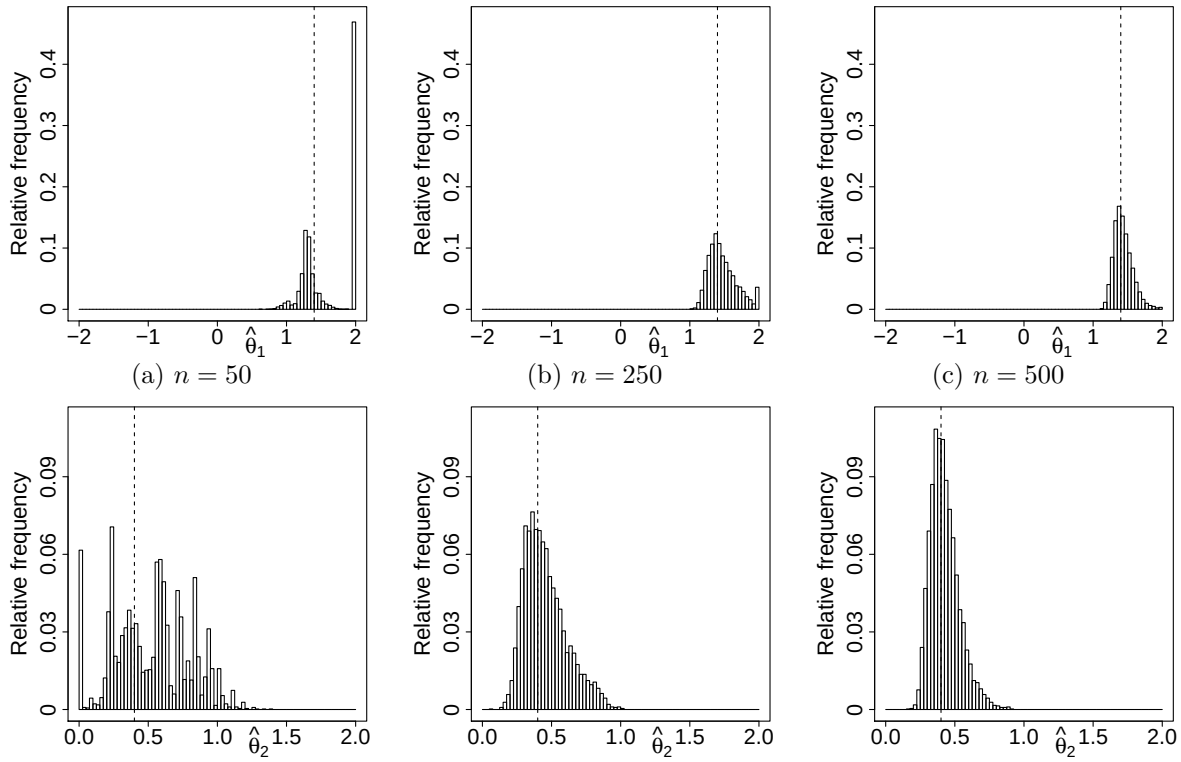


Figure 5.3.: Histograms of the components of the estimates and different sample sizes for the logit model with $\bar{\boldsymbol{\theta}} = (0.6 \quad 1.8)^\top$



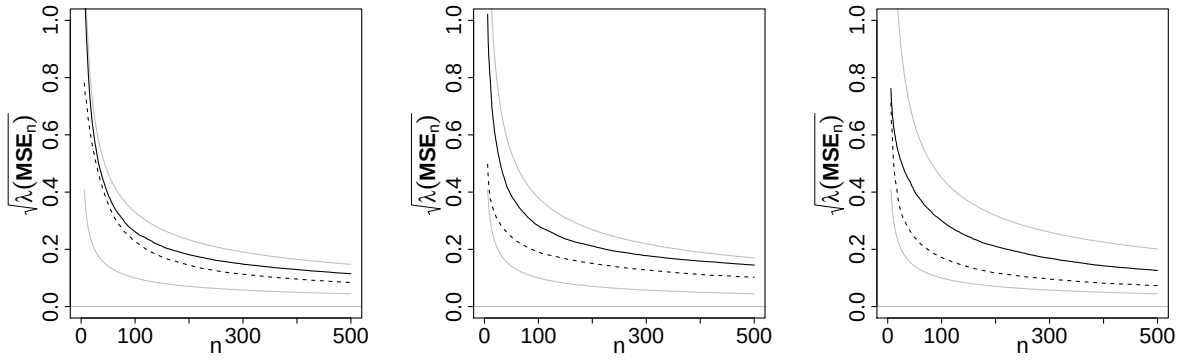
(a) $n = 50$ (b) $n = 250$ (c) $n = 500$

Figure 5.4.: Histograms of the components of the estimates and different sample sizes for the logit model with $\bar{\boldsymbol{\theta}} = (1.4 \quad 0.4)^\top$



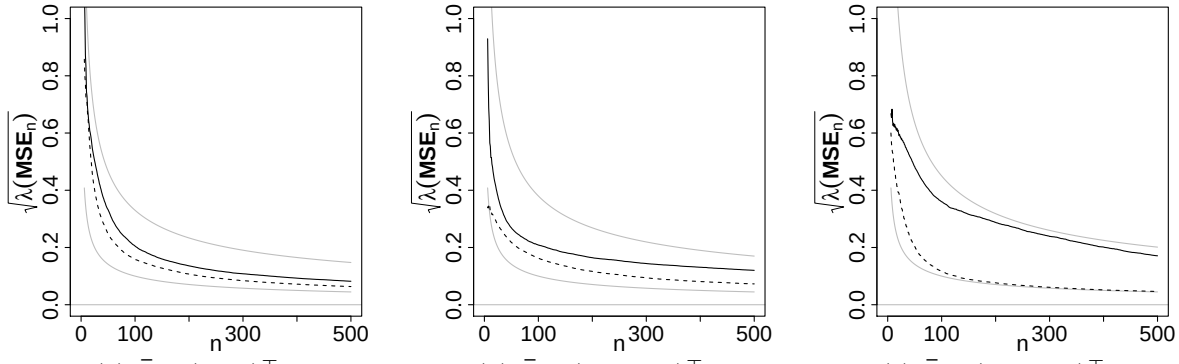
(a) $n = 50$ (b) $n = 250$ (c) $n = 500$

Figure 5.5.: Histograms of the components of the estimates and different sample sizes for the complementary log-log model with $\bar{\boldsymbol{\theta}} = (1.4 \quad 0.4)^\top$



(a) $\bar{\boldsymbol{\theta}} = (0 \quad 1)^\top$ (b) $\bar{\boldsymbol{\theta}} = (0.6 \quad 1.8)^\top$ (c) $\bar{\boldsymbol{\theta}} = (1.4 \quad 0.4)^\top$

Figure 5.6.: Square root of the eigenvalues of $\mathbf{MSE}_n$ for the logit model.
solid: $\lambda_{\max}(\mathbf{MSE}_n)$, dashed: $\lambda_{\min}(\mathbf{MSE}_n)$, gray solid: curves of order $n^{-1/2}$



(a) $\bar{\boldsymbol{\theta}} = (0 \quad 1)^\top$ (b) $\bar{\boldsymbol{\theta}} = (0.6 \quad 1.8)^\top$ (c) $\bar{\boldsymbol{\theta}} = (1.4 \quad 0.4)^\top$

Figure 5.7.: Square root of the eigenvalues of $\mathbf{MSE}_n$ for the complementary log-log model.
solid: $\lambda_{\max}(\mathbf{MSE}_n)$, dashed: $\lambda_{\min}(\mathbf{MSE}_n)$, gray solid: curves of order $n^{-1/2}$

Table 5.4.: Bias for the estimates of θ_1 after n steps

Model	$\bar{\boldsymbol{\theta}}^\top$	n = 50	100	250	400	500
Logit	(0 1)	-0.0048	-0.0010	-0.0044	-0.0017	-0.0016
	(0.6 1.8)	-0.0068	0.0041	0.0032	0.0045	0.0046
	(1.4 0.4)	0.0798	0.0438	0.0176	0.0092	0.0079
Probit	(0 1)	0.0044	0.0003	0.0004	-0.0008	-0.0005
	(0.6 1.8)	0.0021	0.0004	0.0017	0.0043	0.0047
	(1.4 0.4)	0.1366	0.0866	0.0378	0.0233	0.0193
Log-log	(0 1)	0.0015	0.0023	0.0005	0.0001	-0.0002
	(0.6 1.8)	-0.0109	0.0010	0.0045	0.0053	0.0055
	(1.4 0.4)	0.0827	0.0525	0.0221	0.0138	0.0115
Complementary log-log	(0 1)	-0.0027	-0.0028	-0.0028	-0.0010	-0.0010
	(0.6 1.8)	0.0197	0.0128	0.0070	0.0053	0.0050
	(1.4 0.4)	0.2266	0.1281	0.0737	0.0487	0.0357

Table 5.5.: Bias for the estimates of θ_2 after n steps

Model	$\bar{\boldsymbol{\theta}}^\top$	n = 50	100	250	400	500
Logit	(0 1)	0.1235	0.0565	0.0201	0.0121	0.0080
	(0.6 1.8)	-0.0170	0.0006	0.0084	0.0102	0.0118
	(1.4 0.4)	0.0583	0.0282	0.0115	0.0061	0.0042
Probit	(0 1)	0.1131	0.0515	0.0187	0.0101	0.0082
	(0.6 1.8)	0.0052	0.0130	0.0150	0.0155	0.0134
	(1.4 0.4)	0.0864	0.0586	0.0285	0.0173	0.0139
Log-log	(0 1)	0.1381	0.0615	0.0211	0.0126	0.0099
	(0.6 1.8)	0.0036	0.0128	0.0141	0.0144	0.0134
	(1.4 0.4)	0.0453	0.0244	0.0100	0.0054	0.0051
Complementary log-log	(0 1)	0.1332	0.0586	0.0231	0.0136	0.0104
	(0.6 1.8)	-0.0002	0.0069	0.0112	0.0097	0.0104
	(1.4 0.4)	0.1257	0.0872	0.0547	0.0364	0.0274

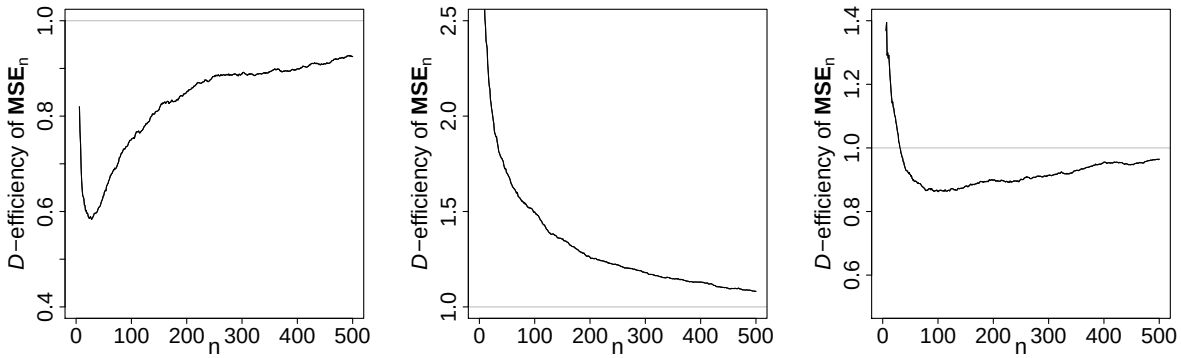
Comparison of Mean Squared Error and Fisher Information

Figure 5.8 shows the efficiency of $\mathbf{MSE}_n$ for the logit model. Here the efficiency is calculated as

$$\left(\frac{\det(\mathbf{M}(\bar{\boldsymbol{\theta}}, \xi^*_{\bar{\boldsymbol{\theta}}})^{-1})}{\det(n\,\mathbf{MSE}_n)}\right)^{1/p}.$$

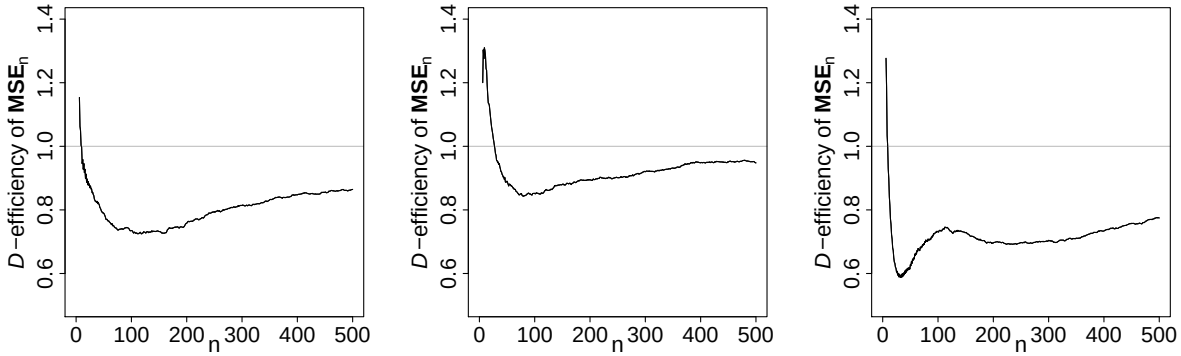
It shows how close the mean squared error and the asymptotic covariance matrix are. For large sample sizes the efficiency becomes closer to 1, which is consistent with the asymptotic behavior mentioned in Section 4.2. Even in the case $\bar{\boldsymbol{\theta}} = (1.4 \quad 0.4)^\top$ the efficiency is increasing, but the effects described above are clearly visible in the probit and complementary log-log model as illustrated in Figure 5.9.

For $\bar{\boldsymbol{\theta}} = (0.6 \quad 1.8)^\top$ the efficiency is larger than 1. This is again due to the fact, that a lot of estimates for the slope parameter θ_2 are on the boundary of the parameter space, i.e. equal to 2, and hence the variance component is comparably small. For the other parameter values the efficiency is it at least smaller than 1 for sample sizes larger than 50.



(a) $\bar{\boldsymbol{\theta}} = (0 \quad 1)^\top$ (b) $\bar{\boldsymbol{\theta}} = (0.6 \quad 1.8)^\top$ (c) $\bar{\boldsymbol{\theta}} = (1.4 \quad 0.4)^\top$

Figure 5.8.: D-efficiency of the estimated MSE for the logit model



(a) Probit (b) Log-log (c) Complementary log-log

Figure 5.9.: D-efficiency of the estimated MSE for different models and $\bar{\boldsymbol{\theta}} = (1.4 \quad 0.4)^\top$

5.2.3. Efficiency of the Adaptive Designs and Behavior of the Design Points

Figure 5.10 and Figure 5.11 summarize the efficiencies $\text{eff}(\xi_n, \xi^*, \bar{\boldsymbol{\theta}})$, i.e. the determinant of the information matrix of the simulated design compared to the one of locally optimal design. (see equation (2.35) for the definition) Displayed are sample quantiles, derived from the replications of a specific setting, in dependence of the sample size. The distribution of the efficiencies for a fixed sample size is skewed with more weight close to 1. In most cases the maximum is very close to the 75%-quantile. The 25%-quantile usually reaches an efficiency of 0.8 after around 100 steps.

The logit model behaves very well in all given settings. This is also supported by Table 5.6, which shows how many replications reached an efficiency of at least 0.9 after a given number of steps. For the logit model at least 75% reached this efficiency after 50 steps. For the other models it took 100.

An exception is as before the problematic value of $\bar{\boldsymbol{\theta}} = (1.4 \quad 0.4)^\top$. In the probit and the complementary log-log model, the smallest efficiency does not seem to increase at all. One part of the explanation for this could be, that it took a lot of steps, until the design points overlapped. As Table 5.3 illustrated, it took 175 steps for the probit

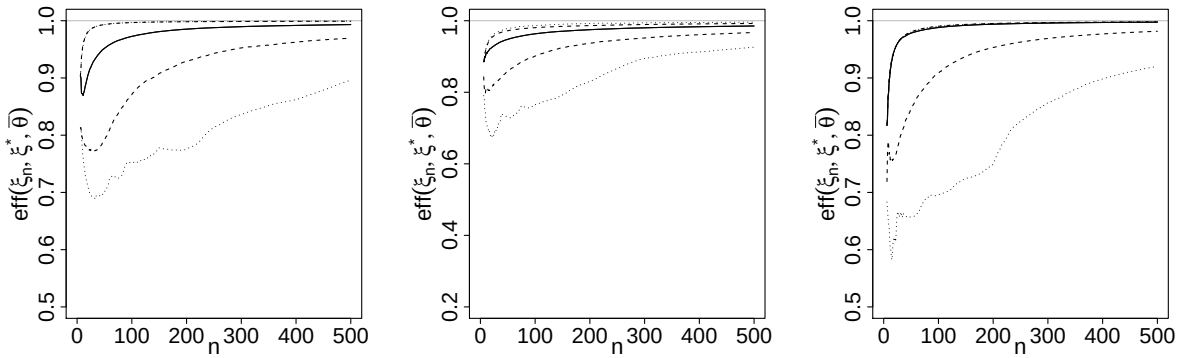
model and 306 for the complementary log-log until this happened for the last replication. So all the estimates before that were possibly "far away" from the actual value and the corresponding design points probably not optimal. One drawback of the Wynn algorithm is, that it takes a relatively long time to get these "bad" values out of the system, and hence the low efficiency for the minimum.

The choice of the new design point is also sensitive to the value of the estimates. For the complementary log-log model one of the eigenvalues of the MSE-matrix is still relatively large, resulting in large variability in the design points. The histograms of design points at steps 50, 250 and 500, shown in Figure 5.14, illustrate that. While the left design point is fixed on the boundary, design points chosen on the right vary a lot. But one can see, that they are accumulating around the optimal points, which are marked by dashed lines.

Some other examples, where the change in the distribution of the design points is more visible, are displayed in Figure 5.12 and Figure 5.13.

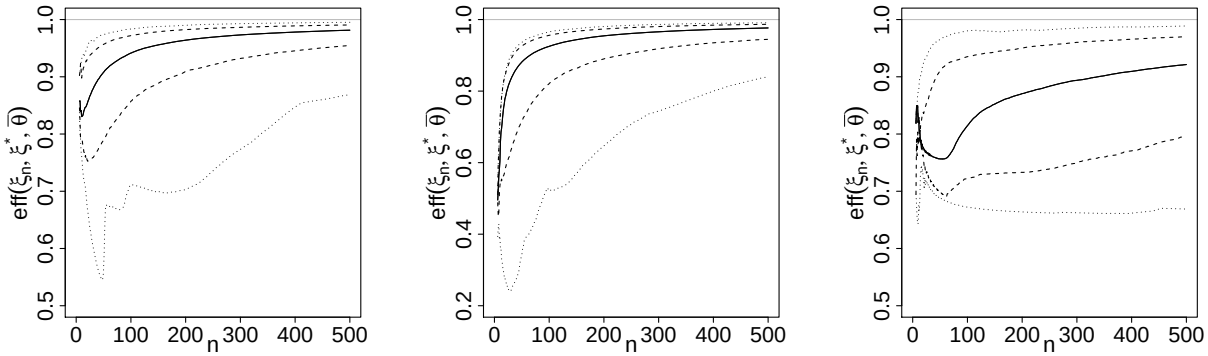
Table 5.6.: Proportion of replications with efficiency higher than 0.9 after n steps

Model	$\bar{\boldsymbol{\theta}}^\top$	n = 25	50	100	250	500
Logit	(0 1)	0.5754	0.7528	0.9056	0.9928	0.9998
	(0.6 1.8)	0.7434	0.8670	0.9508	0.9974	1.0000
	(1.4 0.4)	0.7968	0.8882	0.9586	0.9962	1.0000
Probit	(0 1)	0.5288	0.7374	0.9194	0.9954	1.0000
	(0.6 1.8)	0.0000	0.5060	0.8254	0.9860	0.9996
	(1.4 0.4)	0.5748	0.6650	0.7596	0.8872	0.9598
Log-log	(0 1)	0.2534	0.5370	0.8114	0.9822	0.9990
	(0.6 1.8)	0.0672	0.4098	0.7642	0.9688	0.9966
	(1.4 0.4)	0.8056	0.8972	0.9654	0.9968	1.0000
Complementary log-log	(0 1)	0.2828	0.5632	0.8284	0.9808	0.9986
	(0.6 1.8)	0.0000	0.3138	0.7420	0.9574	0.9972
	(1.4 0.4)	0.0110	0.0670	0.1658	0.3894	0.6616



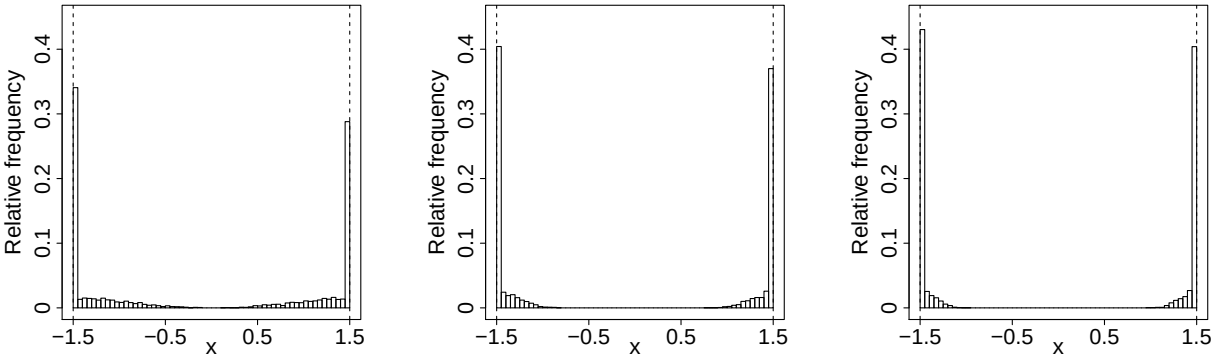
(a) $\bar{\boldsymbol{\theta}} = (0 \quad 1)^\top$ (b) $\bar{\boldsymbol{\theta}} = (0.6 \quad 1.8)^\top$ (c) $\bar{\boldsymbol{\theta}} = (1.4 \quad 0.4)^\top$

Figure 5.10.: Efficiency of the adaptive design for the logit model.
solid: median of the efficiencies, dashed: 5%- and 95%-quantile, dotted: minimum and maximum



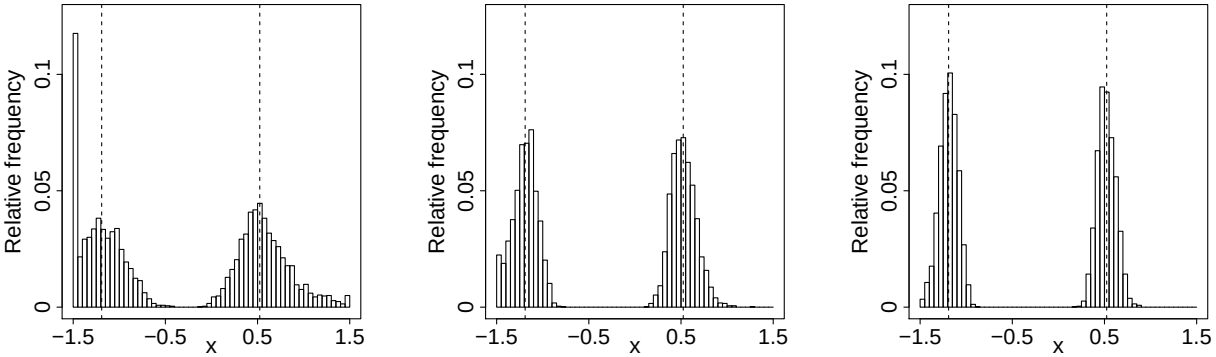
(a) $\bar{\boldsymbol{\theta}} = (0 \quad 1)^\top$ (b) $\bar{\boldsymbol{\theta}} = (0.6 \quad 1.8)^\top$ (c) $\bar{\boldsymbol{\theta}} = (1.4 \quad 0.4)^\top$

Figure 5.11.: Efficiency of the adaptive design for the complementary log-log model.
solid: median of the efficiencies, dashed: 5%- and 95%-quantile, dotted: minimum and maximum



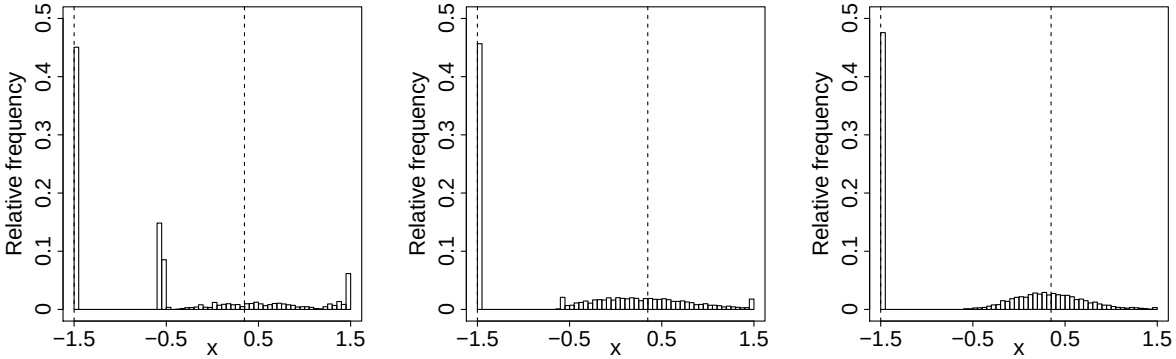
(a) $n = 50$ (b) $n = 250$ (c) $n = 500$

Figure 5.12.: Histograms of the design points at different steps calculated over all replications for the logit model with $\bar{\boldsymbol{\theta}} = (0 \quad 1)^\top$.
dashed lines: locally D-optimal design points



(a) $n = 50$ (b) $n = 250$ (c) $n = 500$

Figure 5.13.: Histograms of the design points at different steps calculated over all replications for the logit model with $\bar{\boldsymbol{\theta}} = (0.6 \quad 1.8)^\top$.
dashed lines: locally D-optimal design points



(a) $n = 50$ (b) $n = 250$ (c) $n = 500$

Figure 5.14.: Histograms of the design points at different steps calculated over all replications for the complementary log-log model with $\bar{\boldsymbol{\theta}} = (1.4 \quad 0.4)^\top$.
dashed lines: locally D-optimal design points

6. Concluding Remarks

As written in the introduction, the two main themes of this thesis were:

- Does the sequence of estimators converge to the actual parameter?
- Does the sequence of designs converge to a locally optimal design?

The results in Chapter 3 mark only the first steps for studying the convergence of the estimator using the method of ordinary differential equations. This approach has not been tried before.

While it is probably too early to think in this direction, it seems to be possible to extend the approach to other generalized linear models, too. The score function, Hessian matrix and Fisher information matrix have similar structures in these models, so it should be possible to extend the method. Especially, since in most of the proofs in Chapter 3 only boundedness of the model was used, and not specifically its binary nature.

As illustrated by Theorem 6 overly restrictive conditions on the design sequence, namely its convergence, were required. This probably can be mitigated by the characterization of the limit by the sets, which contain the trajectories. Consequently this would lead to the study of differential inclusions.

Another open point is the behavior of $\widetilde{\mathcal{S}}_{n,m}$, $\tilde{\boldsymbol{s}}_n$ and, closely related, the behavior of the Hessian matrix. The simulations and calculations for other examples suggested all that

$$\sum_{i=1}^{n} \psi'(\boldsymbol{f}(\boldsymbol{x}_i)^\top \hat{\boldsymbol{\theta}}_n)\Big(y_i - G(\boldsymbol{f}(\boldsymbol{x}_i)^\top \hat{\boldsymbol{\theta}}_n)\Big)\boldsymbol{f}(\boldsymbol{x}_i)\boldsymbol{f}(\boldsymbol{x}_i)^\top ,$$

which is the first term in the Hessian matrix, is bounded. This would imply the asymptotic rate of change condition for $\widetilde{\mathcal{S}}_{n,m}$.

However it is interesting and assuring to note that the eigenvalue conditions of Lemma 8 are close to those used in the convergence theorems in the literature.

That the investigation of the convergence is not in vain is shown by the simulations. The results of the simulations support that the estimator converges. The mean squared error matrix seems to tend to 0 and is approaching the optimal asymptotic variance. The problems which might hint for non-convergence, like a (for the author surprisingly high) bias after 500 observations can be explained within the model. Similar things can be said about the convergence of the adaptive design. The design approaches the locally optimal design even though it is slower than using the Wynn algorithm with the actual parameter.

A problem to investigate further is, which initial design to choose. The equidistant design chosen for the simulations in Chapter 5 is quite conservative and minimal in the sense, that there was only one observation per design point. If one would spend more observations in the beginning, the estimates would be more stable at the start. Karvanen (2008) proposes an interesting approach for the 2-parameter model using binary search. The goal is to shorten the time until there is an overlap in the design points.

A. Appendix

A.1. Proof of Lemma 1

Lemma 1. *Let $\mathbf{F}_n$ have full column rank, $\Theta = \mathbb{R}^p$ and G be a strictly increasing distribution function. Let $\mathcal{C}_n^0 \cap \mathcal{C}_n^1 \neq \emptyset$, then $\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ exists and $\arg\max_{\boldsymbol{\theta}\in\mathbb{R}^p} l(\boldsymbol{\theta}, \boldsymbol{y}_n, \mathbf{F}_n)$ is bounded.*

Proof. In this proof, let $\boldsymbol{v} \in \mathbb{R}^p$ be a unit vector, i.e. $\|\boldsymbol{v}\| = 1$.

All summands of the log-likelihood and hence the log-likelihood itself are strictly smaller than 0, because G takes only values in the interval $(0, 1)$. Since $\mathcal{X}$ is compact we can choose a $\boldsymbol{\theta} \in \mathbb{R}^p$, such that

$$|\boldsymbol{f}(\boldsymbol{x})^\top \boldsymbol{\theta}| \leq C$$

for some $C > 0$ and all $\boldsymbol{x} \in \mathcal{X}$ and hence $|l(\boldsymbol{\theta}, \boldsymbol{y}_n, \mathbf{F}_n)|$ is bounded. If we can show now, that the log-likelihood tends to $-\infty$ in all directions, i.e.

$$\lim_{t\to\infty} l(\boldsymbol{\theta} + t\boldsymbol{v}, \boldsymbol{y}_n, \mathbf{F}_n) = -\infty$$

for all unit vectors $\boldsymbol{v}$, the result would follow. But because l and all its summands are bounded above, it is in fact sufficient, that at least one of the summands in the log-likelihood tends to $-\infty$. Let us consider the behavior of the summands.

Since G is a distribution function we have,

$$\lim_{t\to\infty} \log G(t) = 0 \qquad \lim_{t\to\infty} \log(1 - G(t)) = -\infty$$
$$\lim_{t\to-\infty} \log G(t) = -\infty \qquad \lim_{t\to-\infty} \log(1 - G(t)) = 0$$

and the summands of the log-likelihood function

$$y_i \log G\Big(\boldsymbol{f}(\boldsymbol{x}_i)^\top(\boldsymbol{\theta} + t\boldsymbol{v})\Big) + (1 - y_i) \log\Big(1 - G\Big(\boldsymbol{f}(\boldsymbol{x}_i)^\top(\boldsymbol{\theta} + t\boldsymbol{v})\Big)\Big)\,,$$

depending on the sign of $\boldsymbol{f}(\boldsymbol{x}_i)^\top \boldsymbol{v}$ and the value of y_i, will tend to 0 or $-\infty$, too.

Let $i \in \{1, \ldots, n\}$ and $\boldsymbol{v}$ be fixed, such that $\boldsymbol{f}(\boldsymbol{x}_i)^\top \boldsymbol{v} > 0$. Assume for now, that $y_i = 1$. Because $\mathcal{C}_n^0 \cap \mathcal{C}_n^1 \neq \emptyset$ there exists $i' \in \{1, \ldots, n\}$, $i' \neq i$, such that

$$y_{i'} = 0 \qquad \text{and} \qquad \boldsymbol{f}(\boldsymbol{x}_{i'})^\top \boldsymbol{v} > 0$$

or

$$y_{i'} = 1 \qquad \text{and} \qquad \boldsymbol{f}(\boldsymbol{x}_{i'})^\top \boldsymbol{v} < 0\,.$$

For $t \longrightarrow \infty$ the i-th summand will tend to 0, but

$$\lim_{t\to\infty} \left(y_{i'} \log G\Big(\boldsymbol{f}(\boldsymbol{x}_{i'})^\top(\boldsymbol{\theta} + t\boldsymbol{v})\Big) + (1 - y_{i'}) \log\Big(1 - G\Big(\boldsymbol{f}(\boldsymbol{x}_{i'})^\top(\boldsymbol{\theta} + t\boldsymbol{v})\Big)\Big)\right) = -\infty\,.$$

For $t \longrightarrow -\infty$ the i-th summand of the log-likelihood will tend to $-\infty$ itself. It follows, that the log-likelihood tends to $-\infty$ for the directions given by $\boldsymbol{v}$ and $-\boldsymbol{v}$. The same argument holds, if $y_i = 0$.

What remains to be shown is, that for each unit vector $\boldsymbol{v}$ exists an index $i \in \{1, \ldots, n\}$, i.e. a design point $\boldsymbol{x}_i$, such that

$$\boldsymbol{f}(\boldsymbol{x}_i)^\top \boldsymbol{v} \neq 0\,.$$

But this is a consequence of the fact, that $\mathbf{F}_n$ has full column rank, and hence its rows span $\mathbb{R}^p$. Consequently the log-likelihood tends to $-\infty$ in all directions and the set of maxima, $\arg\max_{\boldsymbol{\theta}\in\mathbb{R}^p} l(\boldsymbol{\theta}, \boldsymbol{y}_n, \mathbf{F}_n)$, is bounded and not empty. □

A.2. Some Results on Matrices

Let $\boldsymbol{f}_i \in \mathbb{R}^p$, $i = 1, 2, \ldots$, be a sequence of vectors, $\mathbf{F}_n := (\boldsymbol{f}_1, \ldots, \boldsymbol{f}_n)^\top$ the $n \times p$ matrix with i-th row $\boldsymbol{f}_i^\top$. Define $\mathbf{A}_n := \mathbf{F}_n^\top \mathbf{F}_n = \sum_{i=1}^n \boldsymbol{f}_i \boldsymbol{f}_i^\top$. Let $(v_i)_{i\geq 1}$ be a sequence in $\mathbb{R}$ and $\boldsymbol{v_n} = (v_1, \ldots, v_n)^\top$ the vector of its first n components.

The following lemma shows two consequences of the fact, that

$$\det(\mathbf{E}_p + \boldsymbol{f}_1 \boldsymbol{f}_2^\top) = 1 + \boldsymbol{f}_1^\top \boldsymbol{f}_2\,.$$

(Harville, 1997, Corollary 18.1.3, p. 420) The first part can be found in Lai and Wei (1982, Lemma 2 (i), p. 156), the second for example in Wynn (1970, p. 1658):

Lemma A.1. **(i)** *If $\mathbf{A}_{n+1}$ is nonsingular, then*

$$\boldsymbol{f}_{n+1}^\top \mathbf{A}_{n+1}^{-1} \boldsymbol{f}_{n+1} = \frac{\det \mathbf{A}_{n+1} - \det \mathbf{A}_n}{\det \mathbf{A}_{n+1}}$$

(ii) *If $\mathbf{A}_n$ is nonsingular, then*

$$\boldsymbol{f}_{n+1}^\top \mathbf{A}_{n}^{-1} \boldsymbol{f}_{n+1} = \frac{\det \mathbf{A}_{n+1} - \det \mathbf{A}_n}{\det \mathbf{A}_{n}}$$

Lemma A.2 corresponds to Lemma 2 (ii) in Lai and Wei (1982). It gives a bound for sums of quadratic forms.

Lemma A.2 (Lai and Wei, 1982, Lemma 2 (ii))**.** *Assume that $\mathbf{A}_m$ is nonsingular for some $m \in \mathbb{N}$. Then $\lambda_{\max}(\mathbf{A}_n)$ is nondecreasing and $\mathbf{A}_n$ is nonsingular for all $n \geq m$.*
Moreover if $\lim_{n\to\infty} \lambda_{\max}(\mathbf{A}_n) < \infty$, then $\sum_{i=m}^\infty \boldsymbol{f}_i^\top \mathbf{A}_i^{-1} \boldsymbol{f}_i < \infty$.
On the other hand, if $\lim_{n\to\infty} \lambda_{\max}(\mathbf{A}_n) = \infty$, then there exist $n_0 \geq m$ and $K > 0$, such that

$$\frac{\sum_{i=m}^n \boldsymbol{f}_i^\top \mathbf{A}_i^{-1} \boldsymbol{f}_i}{\log\Big(\lambda_{\max}(\mathbf{A}_n)\Big)} \leq K$$

for all $n \geq m$.

The next lemma is used in equation (2.12) of Chen et al. (1999).

Lemma A.3. *Let $\mathbf{F}$ be a $n \times p$ matrix with full column rank, i.e. $\mathbf{F}^\top\mathbf{F}$ is invertible, let $\mathbf{A}$ be a $n \times n$ diagonal matrix with positive diagonal elements and let $\boldsymbol{v} \in \mathbb{R}^p$.*

Then

$$\lambda_{\min}(\mathbf{A})\|\boldsymbol{v}\| \leq \|(\mathbf{F}^\top\mathbf{F})^{-1}\mathbf{F}^\top\mathbf{A}\mathbf{F}\boldsymbol{v}\| \leq \lambda_{\max}(\mathbf{A})\|\boldsymbol{v}\|$$

Proof. Since $\lambda_{\min}(\mathbf{A})$ is the smallest diagonal element of $\mathbf{A}$

$$\frac{1}{\lambda_{\min}(\mathbf{A})}\mathbf{F}^\top\mathbf{A}\mathbf{F} - \mathbf{F}^\top\mathbf{F}$$

is positive semidefinite and hence is

$$(\mathbf{F}^\top\mathbf{F})^{-1} - \lambda_{\min}(\mathbf{A})\,(\mathbf{F}^\top\mathbf{A}\mathbf{F})^{-1}\,.$$

By definition of the norm

$$\begin{aligned}\|(\mathbf{F}^\top\mathbf{F})^{-1}\mathbf{F}^\top\mathbf{A}\mathbf{F}\boldsymbol{v}\|^2 &= \boldsymbol{v}^\top\mathbf{F}^\top\mathbf{A}\mathbf{F}(\mathbf{F}^\top\mathbf{F})^{-2}\mathbf{F}^\top\mathbf{A}\mathbf{F}\boldsymbol{v}\\ &\geq \lambda_{\min}(\mathbf{A})^2\ \boldsymbol{v}^\top\mathbf{F}^\top\mathbf{A}\mathbf{F}(\mathbf{F}^\top\mathbf{A}\mathbf{F})^{-2}\mathbf{F}^\top\mathbf{A}\mathbf{F}\boldsymbol{v} = \lambda_{\min}(\mathbf{A})^2\,\|\boldsymbol{v}\|^2\,.\end{aligned}$$

The second inequality follows similarly, since

$$\mathbf{F}^\top\mathbf{F} - \frac{1}{\lambda_{\max}(\mathbf{A})}\mathbf{F}^\top\mathbf{A}\mathbf{F}$$

is positive semidefinite and hence

$$\boldsymbol{v}^\top\mathbf{F}^\top\mathbf{A}\mathbf{F}(\mathbf{F}^\top\mathbf{F})^{-2}\mathbf{F}^\top\mathbf{A}\mathbf{F}\boldsymbol{v} \leq \lambda_{\max}(\mathbf{A})^2\ \boldsymbol{v}^\top\mathbf{F}^\top\mathbf{A}\mathbf{F}(\mathbf{F}^\top\mathbf{A}\mathbf{F})^{-2}\mathbf{F}^\top\mathbf{A}\mathbf{F}\boldsymbol{v}\,.$$

□

The proof for the upper bound in Lemma A.3 is valid for diagonal matrices with non-negative entries, as long as $\lambda_{\max}(\mathbf{A}) > 0$. For arbitrary diagonal matrices $\mathbf{A}$ we get the following result.

Lemma A.4. *Let $\mathbf{F}$ be a $n \times p$ matrix with full column rank, i.e. $\mathbf{F}^\top\mathbf{F}$ is invertible, let $\mathbf{A}$ be a $n \times n$ diagonal matrix and let $\boldsymbol{v} \in \mathbb{R}^p$.*

Then

$$\|(\mathbf{F}^\top\mathbf{F})^{-1}\mathbf{F}^\top\mathbf{A}\mathbf{F}\boldsymbol{v}\| \leq 2\max\{|\lambda_{\min}(\mathbf{A})|, |\lambda_{\max}(\mathbf{A})|\}\|\boldsymbol{v}\|$$

Proof. Denote the entries of the diagonal matrix $\mathbf{A}$ by a_i, $i = 1, \ldots, n$. This matrix can be written as the difference of its positive and negative parts $\mathbf{A}_+$ and $\mathbf{A}_-$, which are defined by

$$\mathbf{A}_+ = \operatorname*{diag}_{i=1,\ldots,n}(\max\{a_i, 0\}) \qquad \text{and} \qquad \mathbf{A}_- = \operatorname*{diag}_{i=1,\ldots,n}(\max\{-a_i, 0\})\,.$$

With the triangle inequality follows

$$\|(\mathbf{F}^\top\mathbf{F})^{-1}\mathbf{F}^\top\mathbf{A}\mathbf{F}\boldsymbol{v}\| \leq \|(\mathbf{F}^\top\mathbf{F})^{-1}\mathbf{F}^\top\mathbf{A}_+\mathbf{F}\boldsymbol{v}\| + \|(\mathbf{F}^\top\mathbf{F})^{-1}\mathbf{F}^\top\mathbf{A}_-\mathbf{F}\boldsymbol{v}\|\,.$$

Now we can apply the upper bound from Lemma A.3 to both summands on the right-hand side and get

$$\|(\mathbf{F}^\top\mathbf{F})^{-1}\mathbf{F}^\top\mathbf{A}\mathbf{F}\boldsymbol{v}\| \leq \lambda_{\max}(\mathbf{A}_+)\|\boldsymbol{v}\| + \lambda_{\max}(\mathbf{A}_-)\|\boldsymbol{v}\|\,.$$

By definition

$$\lambda_{\max}(\mathbf{A}_+) = \max\{\lambda_{\max}(\mathbf{A}), 0\} \qquad \text{and} \qquad \lambda_{\max}(\mathbf{A}_-) = \max\{-\lambda_{\min}(\mathbf{A}), 0\}$$

which finishes the proof. □

Corollary A.1. *Let* $\mathbf{F}$ *be a* $n \times p$ *matrix with full column rank, i.e.* $\mathbf{F}^\top\mathbf{F}$ *is invertible, let* $\mathbf{A}_1$ *and* $\mathbf{A}_2$ *be a* $n \times n$ *diagonal matrices with positive diagonals and let* $\boldsymbol{v} \in \mathbb{R}^p$.
Then

$$\begin{aligned}\|(\mathbf{F}^\top\mathbf{A}_1\mathbf{F})^{-1}\boldsymbol{v} - (\mathbf{F}^\top\mathbf{A}_2\mathbf{F})^{-1}\boldsymbol{v}\| \\ \leq 2\max\Big\{\Big|\lambda_{\min}(\mathbf{A}_1^{-1}(\mathbf{A}_2-\mathbf{A}_1))\Big|, \Big|\lambda_{\max}(\mathbf{A}_1^{-1}(\mathbf{A}_2-\mathbf{A}_1))\Big|\Big\}\|(\mathbf{F}^\top\mathbf{A}_2\mathbf{F})^{-1}\boldsymbol{v}\|\end{aligned}$$

Proof. With the properties of the norm follows

$$\|(\mathbf{F}^\top\mathbf{A}_1\mathbf{F})^{-1}\boldsymbol{v} - (\mathbf{F}^\top\mathbf{A}_2\mathbf{F})^{-1}\boldsymbol{v}\| \leq \|(\mathbf{F}^\top\mathbf{A}_1\mathbf{F})^{-1}(\mathbf{F}^\top\mathbf{A}_2\mathbf{F} - \mathbf{F}^\top\mathbf{A}_1\mathbf{F})\|\|(\mathbf{F}^\top\mathbf{A}_2\mathbf{F})^{-1}\boldsymbol{v}\|\,.$$

Application of Lemma A.4 yields the result:

$$\begin{aligned}\|(\mathbf{F}^\top\mathbf{A}_1\mathbf{F})^{-1}(\mathbf{F}^\top\mathbf{A}_2\mathbf{F} - \mathbf{F}^\top\mathbf{A}_1\mathbf{F})\| \\ \leq 2\max\Big\{\Big|\lambda_{\min}(\mathbf{A}_1^{-1}(\mathbf{A}_2-\mathbf{A}_1))\Big|, \Big|\lambda_{\max}(\mathbf{A}_1^{-1}(\mathbf{A}_2-\mathbf{A}_1))\Big|\Big\}\,.\end{aligned}$$

□

The following lemma is also mentioned in Wu and Wynn (1978, p. 1280)

Lemma A.5. *Let* $\mathcal{X}$ *be compact and* $\xi \in \Xi$ *a nonsingular design. Let* d *be continuous and* $d(t) > 0$ *for all* $t \in \mathbb{R}$. *Then there exist constants* $0 < c \leq C < \infty$ *such that*

$$c\lambda_{\max}(\mathbf{M}(\boldsymbol{\theta},\xi)^{-1}) \leq \max_{\boldsymbol{x}\in\mathcal{X}}\Big(d(\boldsymbol{f}(\boldsymbol{x})^\top\boldsymbol{\theta})\boldsymbol{f}(\boldsymbol{x})^\top\mathbf{M}(\boldsymbol{\theta},\xi)^{-1}\boldsymbol{f}(\boldsymbol{x})\Big) \leq C\lambda_{\max}(\mathbf{M}(\boldsymbol{\theta},\xi)^{-1})\,.$$

Proof. The upper bound follows from

$$\begin{aligned}&\max_{\boldsymbol{x}\in\mathcal{X}}\Big(d(\boldsymbol{f}(\boldsymbol{x})^\top\boldsymbol{\theta})\boldsymbol{f}(\boldsymbol{x})^\top\mathbf{M}(\boldsymbol{\theta},\xi)^{-1}\boldsymbol{f}(\boldsymbol{x})\Big) \\ &\quad\leq \max_{\boldsymbol{x}\in\mathcal{X}}\frac{d(\boldsymbol{f}(\boldsymbol{x})^\top\boldsymbol{\theta})\boldsymbol{f}(\boldsymbol{x})^\top\mathbf{M}(\boldsymbol{\theta},\xi)^{-1}\boldsymbol{f}(\boldsymbol{x})}{d(\boldsymbol{f}(\boldsymbol{x})^\top\boldsymbol{\theta})\boldsymbol{f}(\boldsymbol{x})^\top\boldsymbol{f}(\boldsymbol{x})}\max_{\boldsymbol{x}\in\mathcal{X}}\Big(d(\boldsymbol{f}(\boldsymbol{x})^\top\boldsymbol{\theta})\boldsymbol{f}(\boldsymbol{x})^\top\boldsymbol{f}(\boldsymbol{x})\Big) \\ &\quad\leq \lambda_{\max}(\mathbf{M}(\boldsymbol{\theta},\xi)^{-1})\max_{\boldsymbol{x}\in\mathcal{X}}\Big(d(\boldsymbol{f}(\boldsymbol{x})^\top\boldsymbol{\theta})\boldsymbol{f}(\boldsymbol{x})^\top\boldsymbol{f}(\boldsymbol{x})\Big)\end{aligned}$$

since $C := \max_{\boldsymbol{x}\in\mathcal{X}}\Big(d(\boldsymbol{f}(\boldsymbol{x})^\top\boldsymbol{\theta})\boldsymbol{f}(\boldsymbol{x})^\top\boldsymbol{f}(\boldsymbol{x})\Big)$ is bounded. For the lower bound consider the spectral decomposition of $\mathbf{M}(\boldsymbol{\theta},\xi)^{-1}$. Denote its eigenvalues by $\tilde{\lambda}_{(1)} \leq \ldots \leq \tilde{\lambda}_{(p)}$ and the corresponding normed eigenvectors by $\boldsymbol{z}_i$, $i = 1,\ldots,p$, then

$$\begin{aligned}\max_{\boldsymbol{x}\in\mathcal{X}}\Big(d(\boldsymbol{f}(\boldsymbol{x})^\top\boldsymbol{\theta})\boldsymbol{f}(\boldsymbol{x})^\top\mathbf{M}(\boldsymbol{\theta},\xi)^{-1}\boldsymbol{f}(\boldsymbol{x})\Big) &= \max_{\boldsymbol{x}\in\mathcal{X}}\sum_{i=1}^{p} d(\boldsymbol{f}(\boldsymbol{x})^\top\boldsymbol{\theta})(\boldsymbol{f}(\boldsymbol{x})^\top\boldsymbol{z}_i)^2\tilde{\lambda}_{(i)} \\ &\geq \max_{\boldsymbol{x}\in\mathcal{X}} d(\boldsymbol{f}(\boldsymbol{x})^\top\boldsymbol{\theta})(\boldsymbol{f}(\boldsymbol{x})^\top\boldsymbol{z}_p)^2\tilde{\lambda}_{(p)}\,.\end{aligned}$$

For any nonsingular design $\eta \in \Xi$ follows, that

$$\begin{aligned}\max_{\boldsymbol{x}\in\mathcal{X}} d(\boldsymbol{f}(\boldsymbol{x})^\top\boldsymbol{\theta})(\boldsymbol{f}(\boldsymbol{x})^\top\boldsymbol{z}_p)^2\tilde{\lambda}_{(p)} &\geq \int_{\mathcal{X}} d(\boldsymbol{f}(\boldsymbol{x})^\top\boldsymbol{\theta})(\boldsymbol{f}(\boldsymbol{x})^\top\boldsymbol{z}_p)^2\eta(\mathrm{d}\boldsymbol{x})\tilde{\lambda}_{(p)} \\ &\geq \boldsymbol{z}_p^\top\mathbf{M}(\boldsymbol{\theta},\eta)\boldsymbol{z}_p\tilde{\lambda}_{(p)} \geq \lambda_{\min}(\mathbf{M}(\boldsymbol{\theta},\eta))\lambda_{\max}(\mathbf{M}(\boldsymbol{\theta},\xi)^{-1})\end{aligned}$$

and $\lambda_{\min}(\mathbf{M}(\boldsymbol{\theta},\eta)) > 0$. Choosing a nonsingular design η_0, which is not ξ, but otherwise arbitrary, closes the proof with $c := \lambda_{\min}(\mathbf{M}(\boldsymbol{\theta},\eta_0))$. □

The following lemma gives a bound for the difference of the eigenvalues of two matrices. It is in fact valid for Hermitian matrices. (see Serre, 2010, Proposition 6.2, p. 112)

Lemma A.6. *Let the matrices* $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{p\times p}$ *be symmetric. Denote by* $\lambda_{(k)}$ *the k-th smallest eigenvalue, then*

$$|\lambda_{(k)}(A) - \lambda_{(k)}(B)| \leq \max\{|\lambda_{\max}(A-B)|, |\lambda_{\min}(A-B)|\}\,.$$

A.3. Results Concerning Stochastic Approximation

Next the extended Arzelà-Ascoli theorem from page 22 is proved. The theorem itself appears as Theorem 4.2.2 in Kushner and Yin (2003, p. 102). The functions $f_n : \mathbb{R} \longrightarrow \mathbb{R}^p$ in the sequence $(f_n)_{n\geq 1}$ are assumed to be bounded and measurable.

Theorem 5 (Extended Arzelà-Ascoli)**.** *Let* $J \subset \mathbb{R}$ *be a bounded interval. Let* $(f_n)_{n\geq 1}$ *be equicontinuous in the extended sense and assume* $\|f_n(x)\| \leq C$ *for all* $x \in J$, $n \geq 1$, *then* $(f_n)_{n\geq 1}$ *has a subsequence converging uniformly to a continuous limit on* J.

Proof. Since the f_n are bounded, we can find a convergent subsequence on $\mathbb{Q}\cap J$ using a diagonal argument: I.e. there is a subsequence $(f_{n_i})_{i\geq 1}$ such that $f_{n_i}(x_1)$ converges for some $x_1 \in \mathbb{Q}\cap J$. Now $(f_{n_i})_{i\geq 1}$ includes a further subsequence, which converges at some point $x_2 \in \mathbb{Q}\cap J$, $x_1 \neq x_2$. By induction we get a subsequence converging pointwise on $\mathbb{Q}\cap J$.

Denote this subsequence by $(\tilde{f}_n)_{n\geq 1}$ and its limit by $\tilde{f}_\infty$. Next we will extend the convergence to J. Let $t \in J$ and $\epsilon > 0$. Since $\mathbb{Q}$ is dense in $\mathbb{R}$, there exist $s \in \mathbb{Q}\cap J$ such that $|t-s| \leq \epsilon$.

With the equicontinuity in the extended sense there exists an $n_1 \in \mathbb{N}$, such that

$$\|\tilde{f}_n(t) - \tilde{f}_n(s)\| \leq \epsilon \tag{A.1}$$

for all $n \geq n_1$. Since $\tilde{f}_n(s)$ converges for all $s \in \mathbb{Q}\cap J$ there exists an $n_2 \in \mathbb{N}$, such that

$$\|\tilde{f}_n(s) - \tilde{f}_m(s)\| \leq \epsilon \tag{A.2}$$

for all $n, m \geq n_2$. Combining both we get

$$\|\tilde{f}_n(t) - \tilde{f}_m(t)\| \leq \|\tilde{f}_n(t) - \tilde{f}_n(s)\| + \|\tilde{f}_n(s) - \tilde{f}_m(s)\| + \|\tilde{f}_n(s) - \tilde{f}_m(t)\| \leq 3\epsilon$$

for $n, m \geq \max\{n_1, n_2\}$. But this means, that $(\tilde{f}_n)_{n\geq 1}$ converges uniformly to $\tilde{f}_\infty$ on J.

The continuity of the limit follows from

$$\|\tilde{f}_\infty(t) - \tilde{f}_\infty(s)\| \leq \|\tilde{f}_\infty(t) - \tilde{f}_n(t)\| + \|\tilde{f}_n(t) - \tilde{f}_n(s)\| + \|\tilde{f}_\infty(s) - \tilde{f}_n(s)\|$$

and equations (A.1), (A.2). □

The following lemma summarizes some facts about t_n, ν and related sums.

Lemma A.7. *Let* $s < t$.

(i) $t - \alpha_{\nu(t_n+t)+1} \leq \sum_{i=n+1}^{\nu(t_n+t)} \alpha_i \leq t$

(ii) $t - s - \alpha_{\nu(t_n+t)+1} \leq \sum_{i=\nu(t_n+s)+1}^{\nu(t_n+t)} \alpha_i \leq t - s + \alpha_{\nu(t_n+s)+1}$

(iii)

$$\frac{t - s - \alpha_{\nu(t_n+t)+1}}{\alpha_{\nu(t_n+t)}} \leq \nu(t_n + t) - \nu(t_n + s) \leq \frac{t - s + \alpha_{\nu(t_n+s)+1}}{\alpha_{\nu(t_n+s)}}$$

Proof. By definition of the time t_n

$$\sum_{i=n+1}^{\nu(t_n+t)} \alpha_i = t_{\nu(t_n+t)} - t_n \leq t_n + t - t_n = t\,.$$

The lower bound follows similarly:

$$t_{\nu(t_n+t)+1} \geq t_n + t \Longrightarrow t_{\nu(t_n+t)} \geq t_n + t - \alpha_{\nu(t_n+t)+1}\,.$$

Combining both inequalities from before yields

$$\sum_{i=\nu(t_n+s)+1}^{\nu(t_n+t)} \alpha_i = \sum_{i=n+1}^{\nu(t_n+t)} \alpha_i - \sum_{i=n+1}^{\nu(t_n+s)} \alpha_i \leq t - s + \alpha_{\nu(t_n+s)+1}$$

and

$$\sum_{i=\nu(t_n+s)+1}^{\nu(t_n+t)} \alpha_i \geq t - s - \alpha_{\nu(t_n+t)+1}\,.$$

Since

$$(\nu(t_n + t) - \nu(t_n + s))\alpha_{\nu(t_n+t)} \leq \sum_{i=\nu(t_n+s)+1}^{\nu(t_n+t)} \alpha_i \leq (\nu(t_n + t) - \nu(t_n + s))\alpha_{\nu(t_n+s)+1}$$

this follows from item **(ii)**. □

A.4. Limit Theorems for Martingales

Theorem A.1 (Theorem 2.18, Hall and Heyde, 1980, p. 35)**.** *Let $S_n = \sum_{i=1}^n \varepsilon_i$ be a martingale with respect to $\mathcal{F}_n$ and $(U_n)_{n\geq 1}$ a nondecreasing sequence of positive random variables such that U_n is $\mathcal{F}_{n-1}$-measurable for each n.*

If $1 \leq p \leq 2$ then

$$\sum_{i=1}^{\infty} U_i^{-1}\varepsilon_i \tag{A.3}$$

converges almost surely on the set $\{\sum_{i=1}^{\infty} U_i^{-p}\mathrm{E}(|\varepsilon_i|^p|\mathcal{F}_{i-1}) < \infty\}$, and

$$\lim_{n\to\infty} U_n^{-1} \sum_{i=1}^{n} \varepsilon_i = 0 \tag{A.4}$$

almost surely on the set $\{\lim_{n\to\infty} U_n = \infty, \sum_{i=1}^{\infty} U_i^{-p}\mathrm{E}(|\varepsilon_i|^p|\mathcal{F}_{i-1}) < \infty\}$.

If $1 \leq p \leq 2$, then (A.3) *and* (A.4) *both hold on the set*

$$\Big\{\sum_{i=1}^{\infty} U_i^{-1} < \infty, \sum_{i=1}^{\infty} U_i^{-1-p/2}\mathrm{E}(|\varepsilon_i|^p|\mathcal{F}_{i-1}) < \infty\Big\}\,.$$

The following central limit theorem follows from Corollary 3.1 in Hall and Heyde (1980, p. 58). Similar results can be found in Dvoretzky (1972).

Theorem A.2. *Let $S_n := \sum_{i=1}^n \varepsilon_i$ be a zero mean, square integrable martingale with respect to $\mathcal{F}_n$. Let*

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\Big(\varepsilon_i^2 \Big| \mathcal{F}_i\Big) \xrightarrow{p} 1$$

and for all $\delta > 0$

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\Big(\varepsilon_i^2 \mathbb{1}_{\{|n^{-1/2}\varepsilon_i|>\delta\}} \Big| \mathcal{F}_i\Big) \xrightarrow{p} 0\,. \tag{A.5}$$

Then $n^{-1/2} S_n \xrightarrow{d} N(0,1)$.

A.5. The Essential Supremum

This can be found in Chow and Teicher (1988, p. 194). While they assume a general measure space with a σ-finite measure and measurable function, we restrict the definition to a probability space $(\Omega, \mathcal{F}, \mathrm{P})$. Let T be an arbitrary nonempty set and $X_t : \Omega \longrightarrow \mathbb{R} \cup \{\pm\infty\}$, $t \in T$, a family of random variables. Then $\operatorname{ess\,sup}_{t\in T} X_t$ is defined by

(i) $\operatorname{ess\,sup}_{t\in T} X_t$ is a random variable

(ii) For all $t \in T$: $\mathrm{P}(X_t \leq \operatorname{ess\,sup}_{t\in T} X_t) = 1$

(iii) Let Z be a random variable satisfying **(i)** and **(ii)** then $\mathrm{P}(Z \geq \operatorname{ess\,sup}_{t\in T} X_t) = 1$

Lemma A.8 (Lemma 6.5.1, Chow and Teicher, 1988, p. 194)**.** *Under the previous assumptions, there exists a countable subset $T_0 \subseteq T$, such that*

$$\sup_{t\in T_0} X_t = \operatorname*{ess\,sup}_{t\in T} X_t\,.$$

B. Calculations for the Examples

This section summarizes some facts about models introduced in Example 1, i.e. logit, probit, log-log and complementary log-log. They differ in the choice of the mean function as can be seen in Figure 2.1. The Table B.1 shows mean functions, derivatives and ψ for the models. For the probit model ψ does not simplify. The corresponding entry is marked by an asterisk. The distribution function of the standard normal distribution is denoted by Φ.

Table B.1.: Mean functions, derivatives and ψ for different models

Model	$G(t)$	$G'(t)$	$G''(t)$	$\psi(t)$
Logit	$(1+e^{-t})^{-1}$	$G(t)(1-G(t))$	$G'(t)(1-2G(t))$	1
Probit	$\Phi(t)$	$(2\pi)^{-1/2}\, e^{-t^2/2}$	$-t\, G'(t)$	$*$
Log-log	$e^{-e^{-t}}$	$e^{-t}\, G(t)$	$(e^{-t}-1)\, G'(t)$	$e^{-t}\,(1-G(t))^{-1}$
complementary log-log	$1-e^{-e^{t}}$	$e^{t}\,(1-G(t))$	$(1-e^{t})\, G'(t)$	$e^{t}\, G(t)^{-1}$

B.1. Log-concavity of G and $1-G$

Logit model
Log-concavity of $G(t)$ follows since

$$\frac{\mathrm{d}}{\mathrm{d}t}\log G(t) = \frac{G(t)(1-G(t))}{G(t)} = 1-G(t)$$

and G is strictly increasing. By symmetry $\log(1-G(t)) = \log G(-t)$, which yields the concavity of $\log(1-G(t))$.

Probit model
The derivative of the standard normal density is

$$G''(t) = -t\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}t^2} = -tG'(t)$$

and hence

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\log G(t) = \frac{G''(t)G(t)-G'(t)^2}{G(t)^2} = -(t\,G(t)+G'(t))\,\frac{G'(t)}{G(t)^2}\,.$$

This is negative, if $t\,G(t) + G'(t) > 0$. For $t \geq 0$ it obviously holds. If t is negative, this follows from an inequality for Mill's ratio (see e.g. Gut, 2013, p. 558): For $s > 0$

$$1 - G(s) < \frac{G'(s)}{s}\,. \tag{B.1}$$

Substitute s by $-t > 0$, then this is equivalent to

$$-t\,(1 - G(-t)) < G'(-t) \iff -t\,(1 - G(-t)) - G'(-t) < 0\,,$$

which by symmetry of G and G' is equivalent to $t\,G(t) + G'(t) > 0$. Concavity of $\log(1 - G(t))$ follows again by symmetry.

Log-log model
The mean function is log-concave, since

$$\log G(t) = -e^{-t}$$

is concave. For $\log(1 - G(t))$ consider its derivatives:

$$\frac{\mathrm{d}}{\mathrm{d}t} \log(1 - G(t)) = e^{-t}\Big(1 - e^{e^{-t}}\Big)^{-1}$$
$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} \log(1 - G(t)) = \frac{-e^{-t}}{\left(1 - e^{e^{-t}}\right)^2}\Big(1 - e^{e^{-t}} + e^{e^{-t}} e^{-t}\Big)$$

It is negative if and only if

$$1 - e^{e^{-t}} + e^{e^{-t}} e^{-t} > 0$$

or equivalently if

$$e^{-e^{-t}} > 1 - e^{-t}\,.$$

This is true for all $t \in \mathbb{R}$ as can be checked using the Taylor approximation of the exponential function, i.e. by

$$e^{-x} > 1 - x$$

and choosing $x = \exp(-t)$. Hence the log-likelihood is concave, since G and $1 - G$ are both log-concave.

Complementary log-log model
Showing concavity here is virtually the same as in the log-log case. The reason is the relationship of the two models: Denote the mean function of the log-log model by G_1, then $G(t) = 1 - G_1(-t)$. Consequently the log-concavity follows immediately from the results for the log-log model.

B.2. Calculations for Example 6

The conditions to check for the models is, that

$$\psi'(t)(1 - G(t)) - (1 - C)d(t) \leq 0 \quad \text{and} \quad -\psi'(t)G(t) - (1 - C)d(t) \leq 0\,.$$

We will see, that these hold for $C = 1/2$. Consider the first inequality:

$$\begin{aligned} & 0 \geq \psi'(t)(1-G(t)) - \frac{1}{2}d(t) \\ & \quad = \frac{G''(t)G(t)(1-G(t)) - G'(t)^2(1-2G(t))}{G(t)^2(1-G(t))} - \frac{1}{2}\frac{G'(t)^2}{G(t)(1-G(t))} \\ \iff \quad & 0 \geq G''(t)G(t)(1-G(t)) - G'(t)^2(1-2G(t)) - \frac{1}{2}G'(t)^2G(t) \\ & \quad = G''(t)G(t)(1-G(t)) - G'(t)^2\Big(1-\frac{3}{2}G(t)\Big)\,. \end{aligned} \tag{B.2}$$

Similarly, for the second inequality:

$$\begin{aligned} & 0 \geq -\psi'(t)G(t) - \frac{1}{2}d(t) \\ \iff \quad & 0 \leq G''(t)G(t)(1-G(t)) - G'(t)^2(1-2G(t)) + \frac{1}{2}G'(t)^2(1-G(t)) \\ & \quad = G''(t)G(t)(1-G(t)) - \frac{1}{2}G'(t)^2(1-3G(t))\,. \end{aligned} \tag{B.3}$$

Log-log model
For the first inequality, the one in (B.2), we get:

$$\begin{aligned} & 0 \geq (e^{-t}-1)G'(t)G(t)(1-G(t)) - G'(t)e^{-t}G(t)\Big(1-\frac{3}{2}G(t)\Big) \\ \iff \quad & 0 \geq (e^{-t}-1)(1-G(t)) - e^{-t}\Big(1-\frac{3}{2}G(t)\Big) \\ & \quad = -1 + G(t) + \frac{1}{2}e^{-t}G(t) \\ \iff \quad & G(t)^{-1} = e^{e^{-t}} \geq 1 + \frac{1}{2}e^{-t}\,. \end{aligned}$$

That the last inequality holds follows, as in the previous section, with a Taylor expansion of the exponential function:

$$e^{e^{-t}} \geq 1 + e^{-t} > 1 + \frac{1}{2}e^{-t}\,. \tag{B.4}$$

The second inequality we have to check is (B.3):

$$\begin{aligned} & 0 \leq (e^{-t}-1)G'(t)G(t)(1-G(t)) - \frac{1}{2}G'(t)e^{-t}G(t)(1-3G(t)) \\ \iff \quad & 0 \leq (e^{-t}-1)(1-G(t)) - \frac{1}{2}e^{-t}(1-3G(t)) \\ & \quad = \Big(e^{-t}-1-\frac{1}{2}e^{-t}\Big)(1-G(t)) + e^{-t}G(t) \\ & \quad = \Big(\frac{1}{2}e^{-t}-1\Big)(1-G(t)) - e^{-t}(1-G(t)) + e^{-t} \\ & \quad = -\Big(1+e^{-t}\frac{1}{2}\Big)(1-G(t)) + e^{-t} \end{aligned} \tag{B.5}$$

We will show, that the right-hand side is a decreasing function in t, which is positive at $t = 0$ and tends to 0 for $t \longrightarrow \infty$. From this follows, that it has to be positive on the whole real line.

The right-hand side of (B.5) is a decreasing function in t: Its derivative

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\Big(-\Big(1+\frac{1}{2}e^{-t}\Big)(1-G(t))+e^{-t}\Big) &= e^{-t}\frac{1}{2}(1-G(t))+\Big(1+\frac{1}{2}e^{-t}\Big)G'(t)-e^{-t}\\ &= e^{-t}\frac{1}{2}\Big(-1+G(t)(1+e^{-t})\Big)\end{aligned}$$

is negative, if and only if

$$0 \geq -1+G(t)(1+e^{-t}) \Longleftrightarrow G(t)^{-1} \geq 1+e^{-t}\,,$$

which, as stated in (B.4), holds for all $t \in \mathbb{R}$. The limit of (B.5) for $t \longrightarrow \infty$ is

$$\lim_{t\to\infty}\Big(-\Big(1+\frac{1}{2}e^{-t}\Big)(1-G(t))+e^{-t}\Big) = 0\,.$$

Since for $t = 0$

$$-\Big(1+\frac{1}{2}e^{-t}\Big)(1-G(t))+e^{-t} = \frac{3-e}{2e} > 0\,.$$

the right-hand side of (B.5) has to be nonnegative.

Complementary log-log model

As in the previous section denote the mean function of the log-log model by G_1, then $G(t) = 1 - G_1(-t)$. With this convention

$$\begin{aligned}& 0 \geq G''(t)G(t)(1-G(t)) - G'(t)^2\Big(1-\frac{3}{2}G(t)\Big)\\ \Longleftrightarrow\quad & 0 \geq -G_1''(-t)G_1(-t)(1-G_1(-t)) - G_1'(-t)^2\Big(1-\frac{3}{2}(1-G_1(-t))\Big)\\ \Longleftrightarrow\quad & 0 \leq G_1''(-t)G_1(-t)(1-G_1(-t)) - \frac{1}{2}G_1'(-t)^2\Big(1-3G_1(-t))\Big)\end{aligned}$$

and

$$\begin{aligned}& 0 \leq G''(t)G(t)(1-G(t)) - \frac{1}{2}G'(t)^2\Big(1-3G(t))\Big)\\ \Longleftrightarrow\quad & 0 \leq -G_1''(-t)G_1(-t)(1-G_1(-t)) - \frac{1}{2}G_1'(-t)^2\Big(1-3(1-G_1(-t))\Big)\\ \Longleftrightarrow\quad & 0 \geq G_1''(-t)G_1(-t)(1-G_1(-t)) + G_1'(-t)^2\Big(1-\frac{3}{2}G_1(-t)\Big)\,,\end{aligned}$$

which are the inequalities for the log-log model with $-t$ instead of t. Since (B.2) and (B.3) hold for all $t \in \mathbb{R}$ in the log-log case, the proof is finished.

Probit model

Because $G''(t) = -tG'(t)$ the inequalities (B.2) and (B.3) reduce to

$$0 \leq tG(t)(1-G(t)) + G'(t)\Big(1-\frac{3}{2}G(t)\Big) \tag{B.6}$$

and

$$0 \geq tG(t)(1-G(t)) + \frac{1}{2}G'(t)(1-3G(t))\,.$$

The symmetry of the normal distribution yields, that they are equivalent. We will consider (B.6). Let $t < 0$, then $G(t) < 1/2$ and by direct calculations

$$1 - \frac{3}{2}G(t) > (1-G(t))^2\,.$$

It follows, that

$$\begin{aligned} tG(t)(1-G(t)) + G'(t)\Big(1-\frac{3}{2}G(t)\Big) &\geq tG(t)(1-G(t)) + G'(t)(1-G(t))^2 \\ &= \Big(tG(t) + G'(t)(1-G(t))\Big)(1-G(t))\,. \end{aligned}$$

The right-hand side is positive, if

$$tG(t) + G'(t)(1-G(t)) > 0\,. \tag{B.7}$$

We will use a similar argument as for the log-log model to show, that this holds: An increasing function, which is positive at some point and tends to 0 for $t \longrightarrow -\infty$ has to be nonnegative.

The left-hand side of (B.7) is positive at $t = 0$ and its limit for $t \longrightarrow -\infty$ is

$$\lim_{t\to-\infty}(tG(t) + G'(t)(1-G(t))) = 0\,.$$

It remains, to show, that it is increasing. The first derivative of the left-hand side of (B.7) is

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}(tG(t) + G'(t)(1-G(t))) &= G(t) + tG'(t) + G''(t)(1-G(t)) - G'(t)^2 \\ &= G(t) + tG'(t)G(t) - G'(t)^2 \\ &= G(t) - G'(t)(-tG(t) + G'(t)) \\ &= G(t) - G'(t)\Big(-t(1-G(-t)) + G'(t)\Big). \end{aligned}$$

The last equality holds because of the symmetry of the normal distribution. With (B.1), i.e. from the properties of Mill's ratio, follows

$$G(t) - G'(t)\Big(-t(1-G(-t)) + G'(t)\Big) \geq G(t) - G'(t)(G'(-t) + G'(t)) = G(t) - 2G'(t)^2\,,$$

where the last equality is again because of the symmetry. Now $G(t) - 2G'(t)^2$ tends to 0 for $t \longrightarrow -\infty$ and is positive for $t = 0$, too. Its derivative is given by

$$\frac{\mathrm{d}}{\mathrm{d}t}\Big(G(t) - 2G'(t)^2\Big) = G'(t)(1 + 4tG'(t))\,,$$

which is positive if and only if $1 + 4tG'(t)$ is.

The function $-x\exp(-x^2)$ has a local maximum at $x = -1/\sqrt{2}$ and hence, with the substitution $t = \sqrt{2}x$,

$$\frac{\sqrt{\pi}}{4} > 0.44 > \frac{1}{\sqrt{2}e^{1/2}} \geq \frac{-t}{\sqrt{2}e^{t^2/2}}$$

for $t \leq 0$. But because

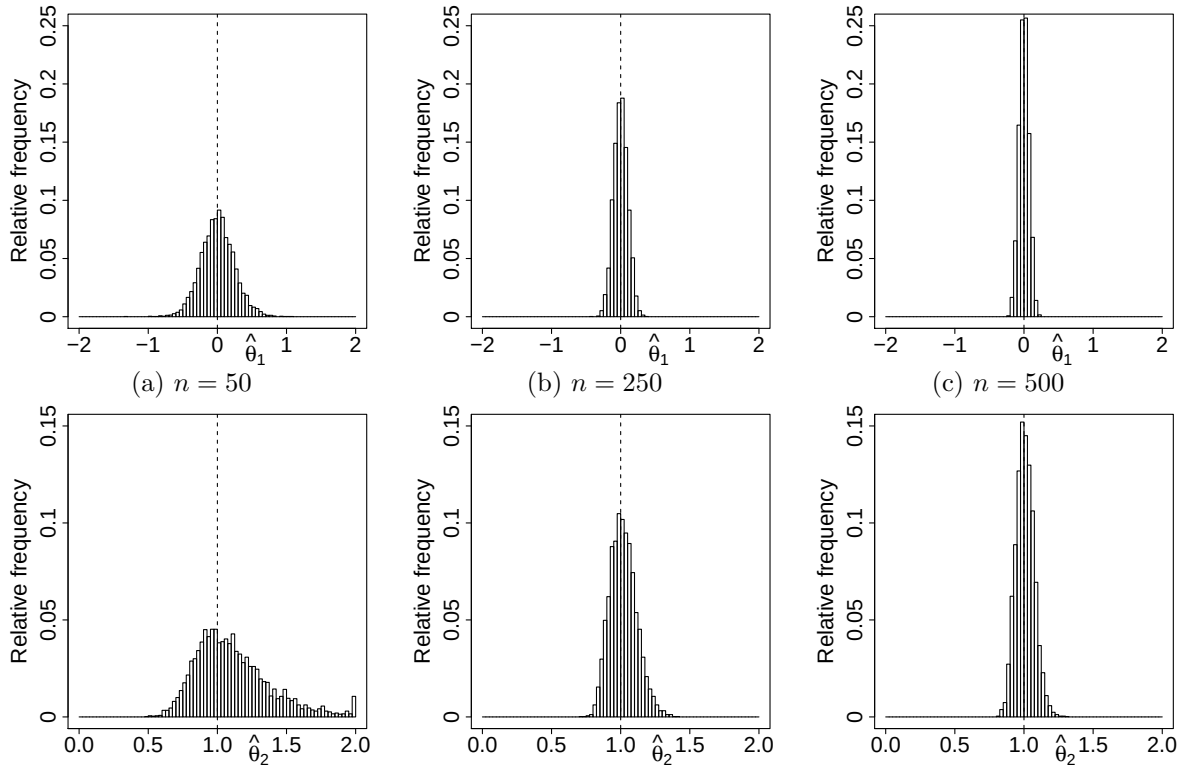
$$\frac{\sqrt{\pi}}{4} > \frac{-t}{\sqrt{2}e^{t^2/2}} \Longleftrightarrow 1 + 4tG'(t) > 0\,,$$

follows, that $G(t) - 2G'(t)^2$ is increasing. Combining this with its limit yields, that the left-hand side of (B.7) is increasing and itself is positive. Hence the inequality (B.7) holds for $t < 0$ and consequently (B.6).

The result for the positive half-axis follows similarly.

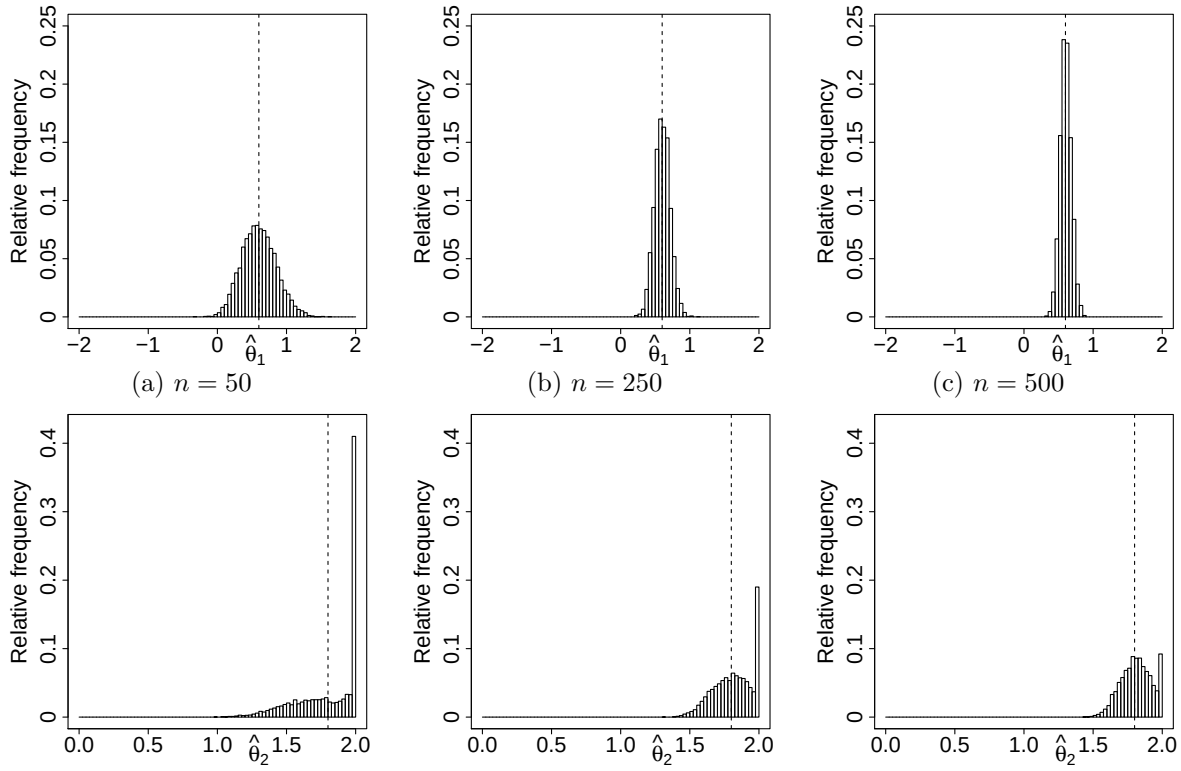
C. Additional Figures from the Simulations

Distribution of the Estimates



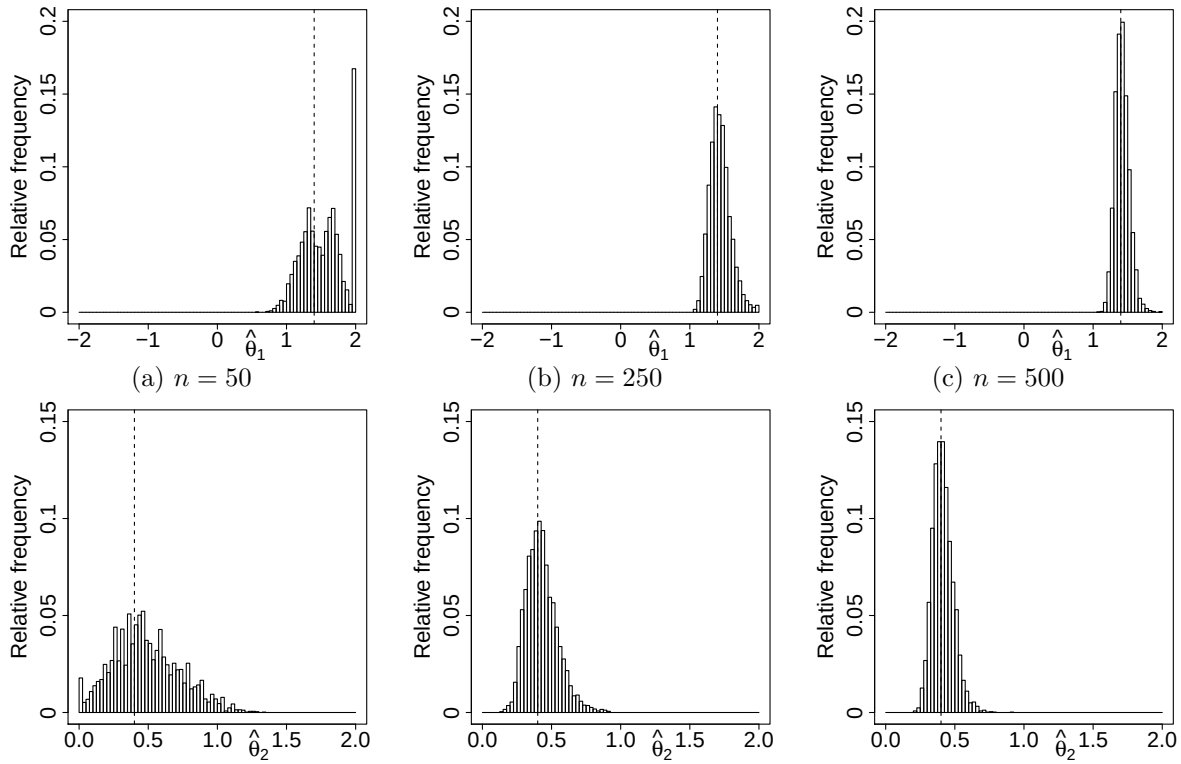
(a) $n = 50$ (b) $n = 250$ (c) $n = 500$

Figure C.1.: Histograms of the components of the estimate and different sample sizes for the probit model with $\bar{\boldsymbol{\theta}} = (0 \quad 1)^\top$



(a) $n = 50$ (b) $n = 250$ (c) $n = 500$

Figure C.2.: Histograms of the components of the estimate and different sample sizes for the probit model with $\bar{\boldsymbol{\theta}} = (0.6 \quad 1.8)^\top$



(a) $n = 50$ (b) $n = 250$ (c) $n = 500$

Figure C.3.: Histograms of the components of the estimate and different sample sizes for the probit model with $\bar{\boldsymbol{\theta}} = (1.4 \quad 0.4)^\top$

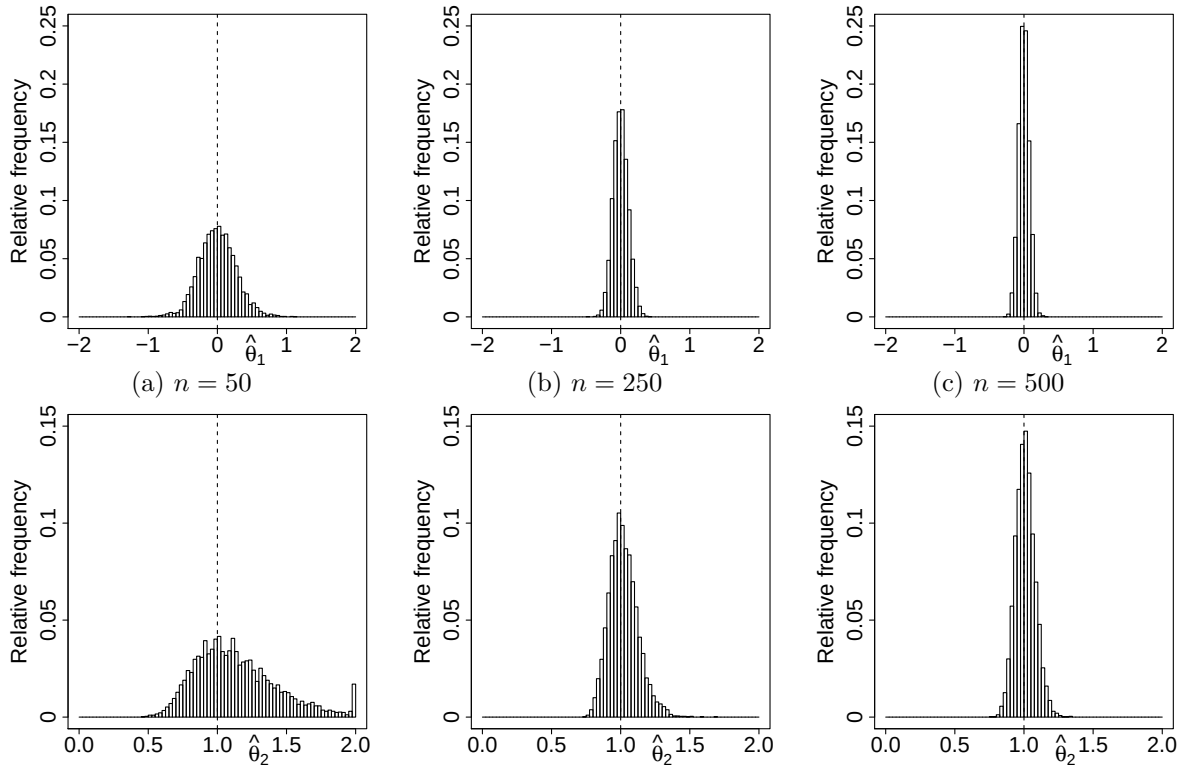


Figure C.4.: Histograms of the components of the estimate and different sample sizes for the log-log model with $\bar{\boldsymbol{\theta}} = (0 \quad 1)^\top$

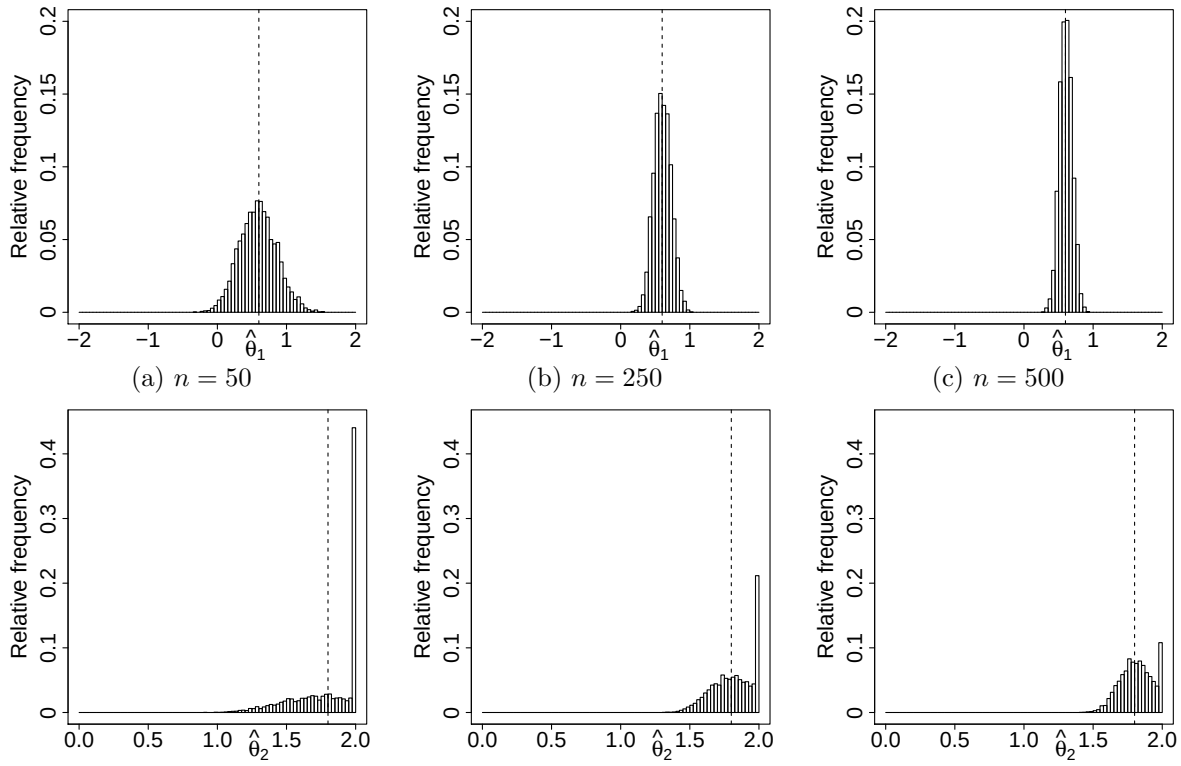
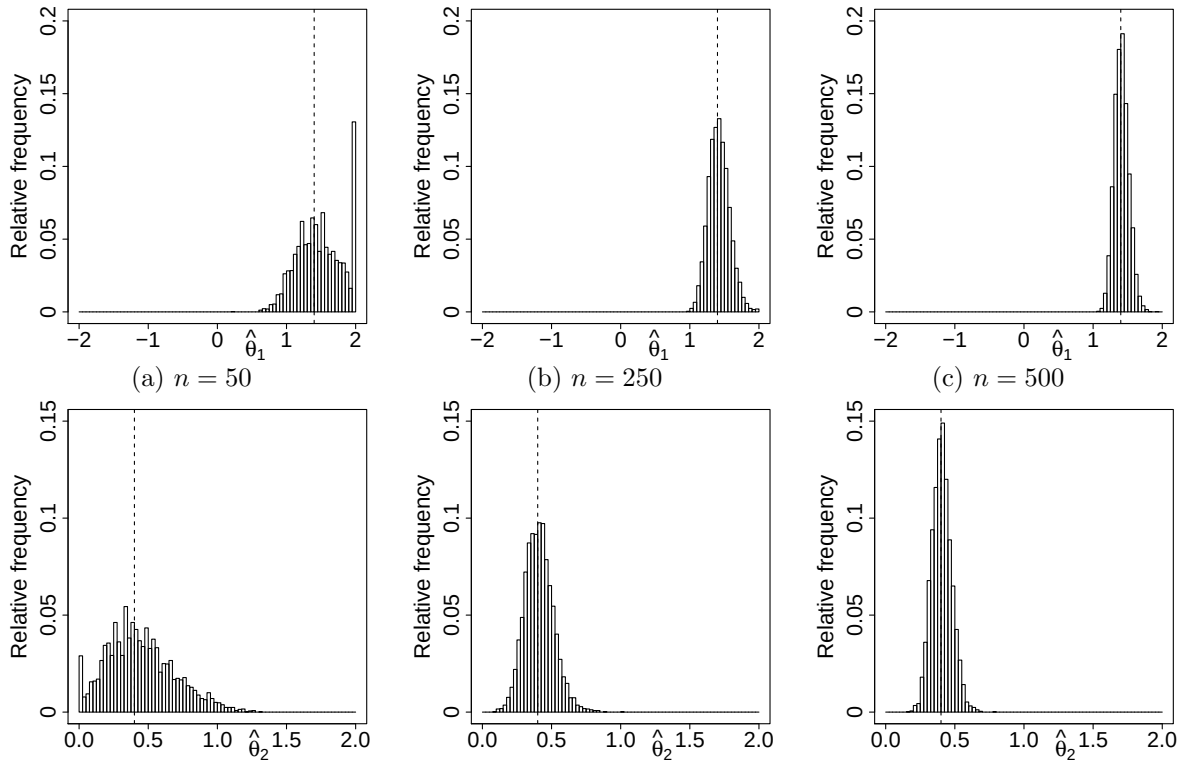
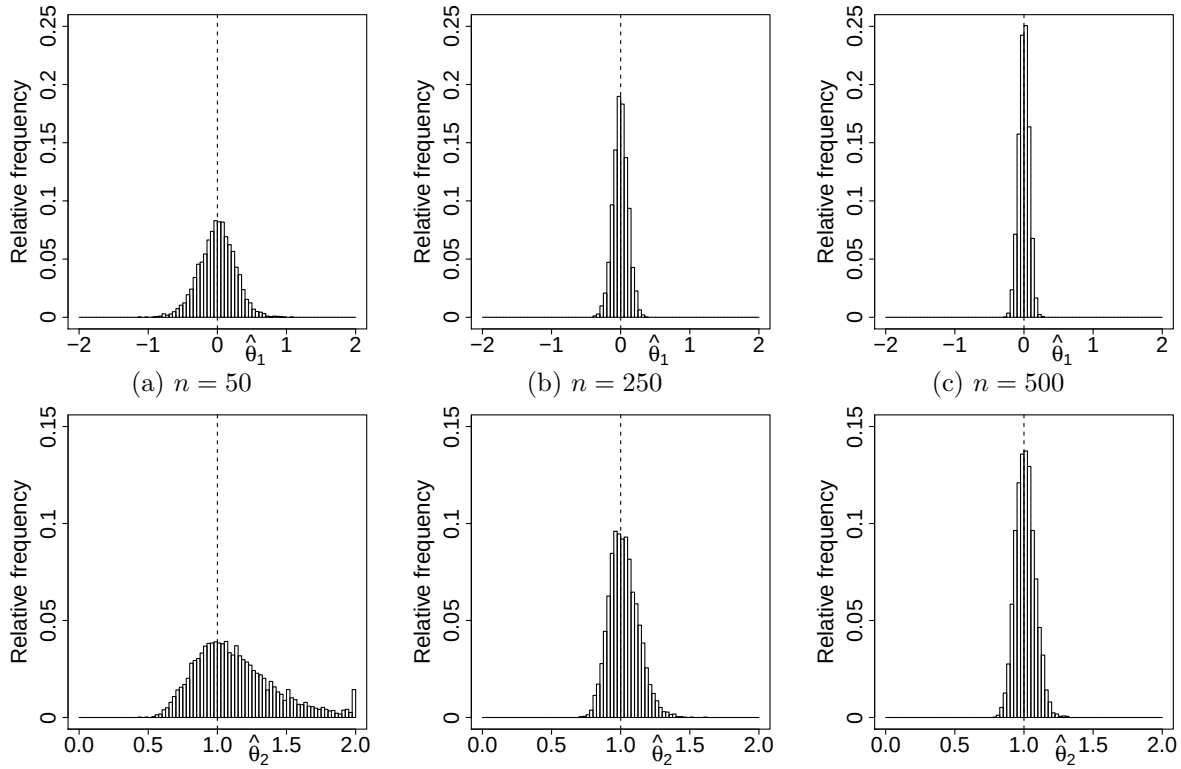


Figure C.5.: Histograms of the components of the estimate and different sample sizes for the log-log model with $\bar{\boldsymbol{\theta}} = (0.6 \quad 1.8)^\top$



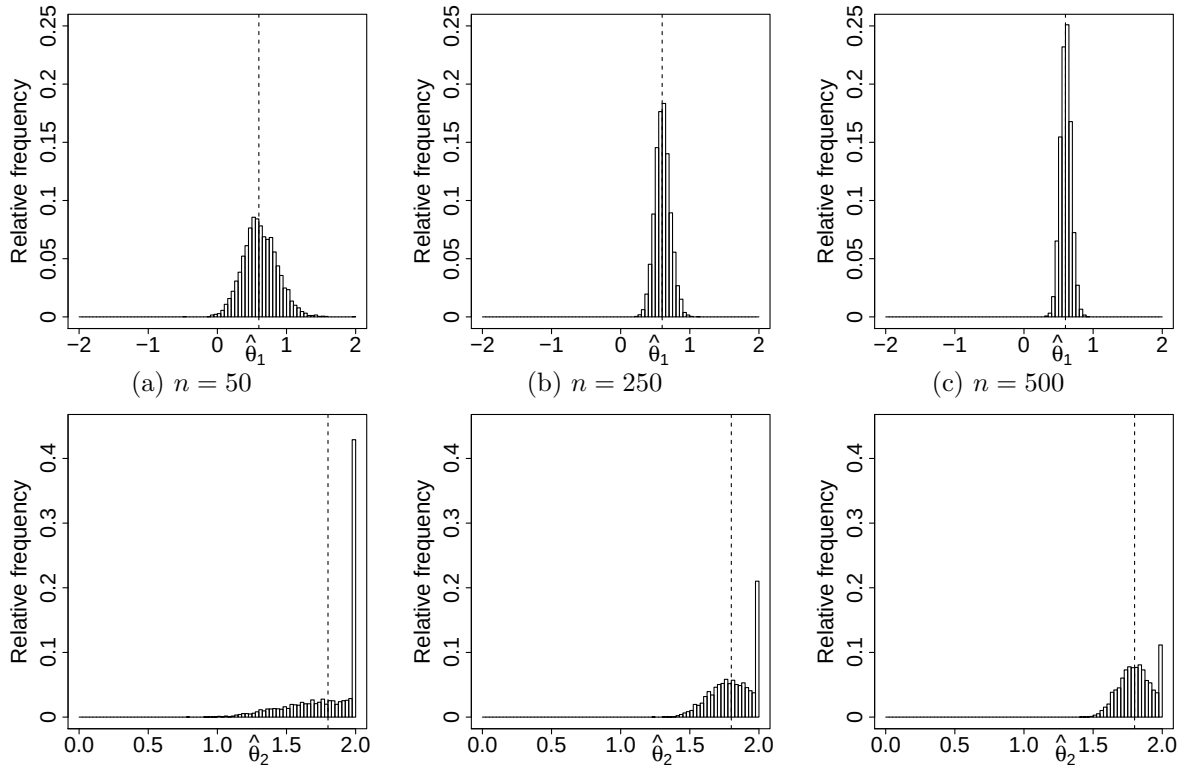
(a) $n = 50$ (b) $n = 250$ (c) $n = 500$

Figure C.6.: Histograms of the components of the estimate and different sample sizes for the log-log model with $\bar{\boldsymbol{\theta}} = (1.4 \quad 0.4)^\top$



(a) $n = 50$ (b) $n = 250$ (c) $n = 500$

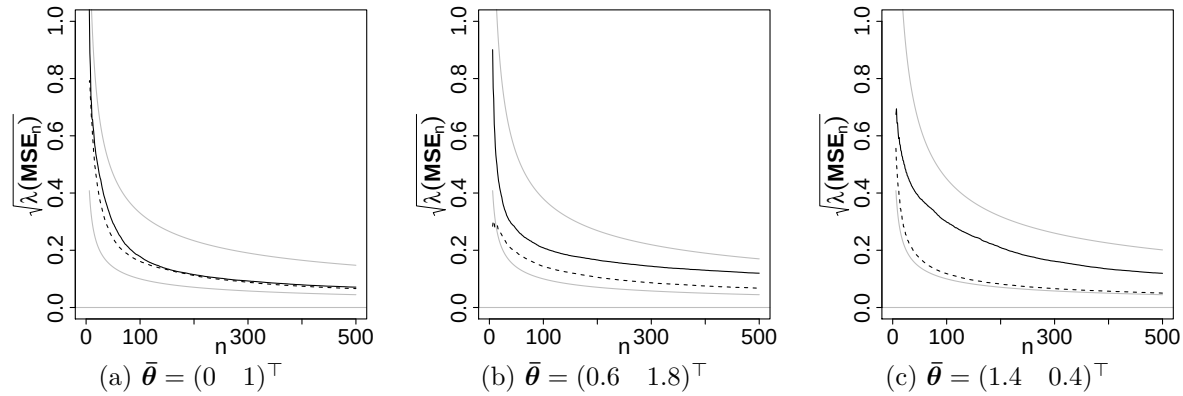
Figure C.7.: Histograms of the components of the estimate and different sample sizes for the complementary log-log model with $\bar{\boldsymbol{\theta}} = (0 \quad 1)^\top$



(a) $n = 50$ (b) $n = 250$ (c) $n = 500$

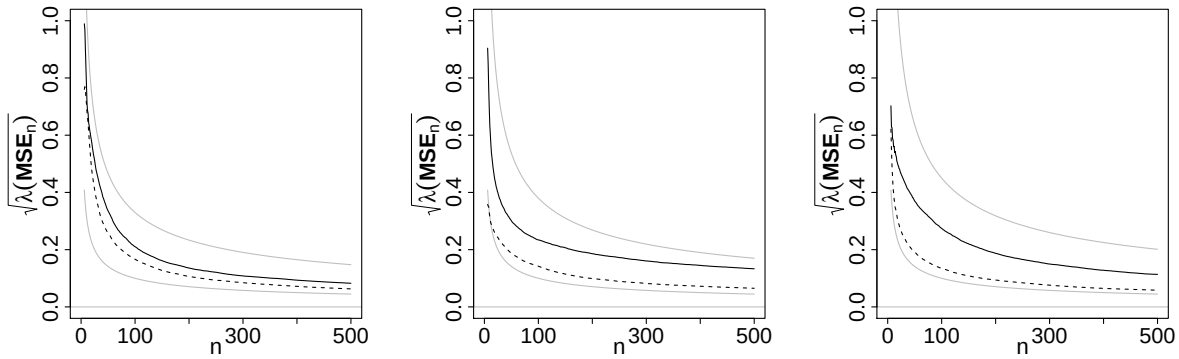
Figure C.8.: Histograms of the components of the estimate and different sample sizes for the complementary log-log model with $\bar{\boldsymbol{\theta}} = (0.6 \quad 1.8)^\top$

Eigenvalues of the Mean Squared Error Matrix



(a) $\bar{\boldsymbol{\theta}} = (0 \quad 1)^\top$ (b) $\bar{\boldsymbol{\theta}} = (0.6 \quad 1.8)^\top$ (c) $\bar{\boldsymbol{\theta}} = (1.4 \quad 0.4)^\top$

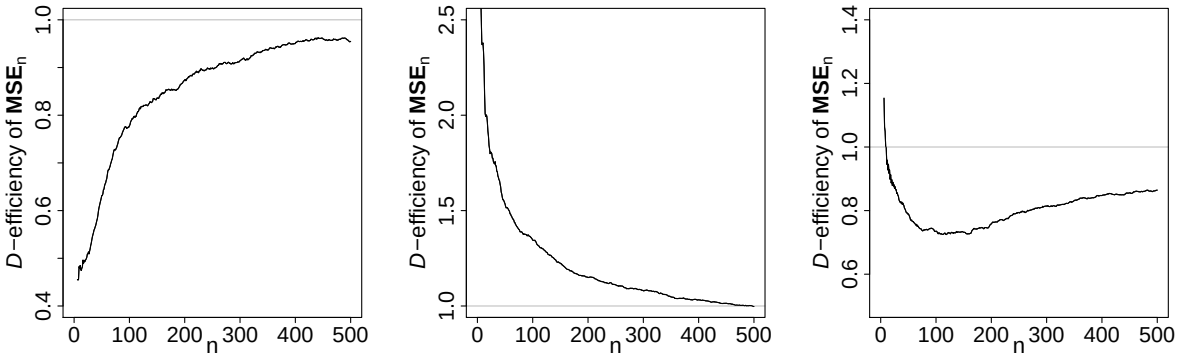
Figure C.9.: Square root of the eigenvalues of $\mathbf{MSE}_n$ for the probit model.
solid: $\lambda_{\max}(\mathbf{MSE}_n)$, dashed: $\lambda_{\min}(\mathbf{MSE}_n)$, gray solid: curves of order $n^{-1/2}$



(a) $\bar{\boldsymbol{\theta}} = (0 \quad 1)^\top$ (b) $\bar{\boldsymbol{\theta}} = (0.6 \quad 1.8)^\top$ (c) $\bar{\boldsymbol{\theta}} = (1.4 \quad 0.4)^\top$

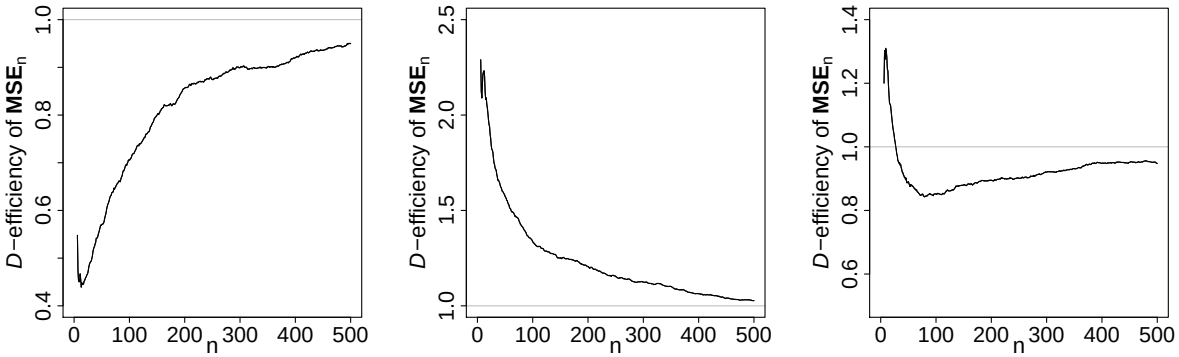
Figure C.10.: Square root of the eigenvalues of $\mathbf{MSE}_n$ for the log-log model.
solid: $\lambda_{\max}(\mathbf{MSE}_n)$, dashed: $\lambda_{\min}(\mathbf{MSE}_n)$, gray solid: curves of order $n^{-1/2}$

Comparison of Mean Squared Error and Fisher Information



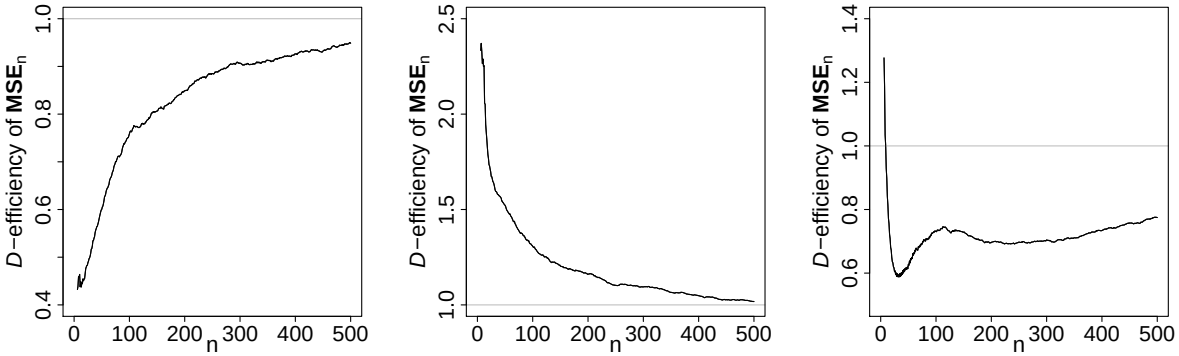
(a) $\bar{\boldsymbol{\theta}} = (0 \quad 1)^\top$ (b) $\bar{\boldsymbol{\theta}} = (0.6 \quad 1.8)^\top$ (c) $\bar{\boldsymbol{\theta}} = (1.4 \quad 0.4)^\top$

Figure C.11.: D-efficiency of the estimated MSE for the probit model



(a) $\bar{\boldsymbol{\theta}} = (0 \quad 1)^\top$ (b) $\bar{\boldsymbol{\theta}} = (0.6 \quad 1.8)^\top$ (c) $\bar{\boldsymbol{\theta}} = (1.4 \quad 0.4)^\top$

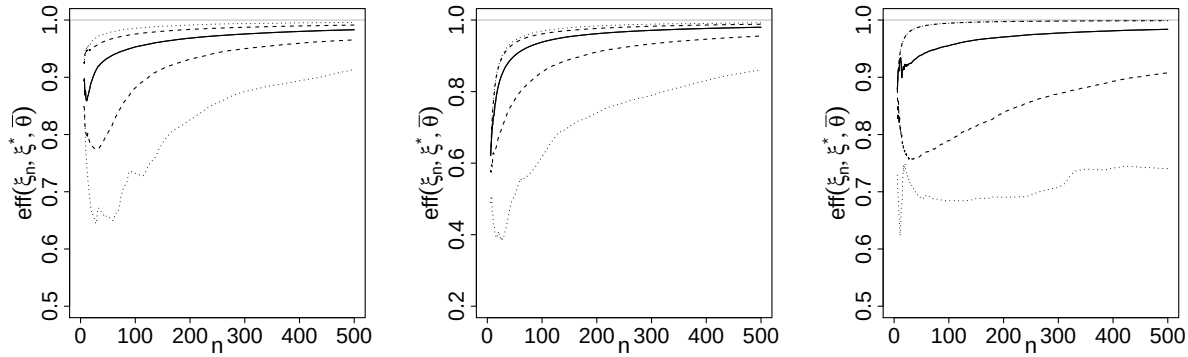
Figure C.12.: D-efficiency of the estimated MSE for the log-log model



(a) $\bar{\boldsymbol{\theta}} = (0 \quad 1)^\top$ (b) $\bar{\boldsymbol{\theta}} = (0.6 \quad 1.8)^\top$ (c) $\bar{\boldsymbol{\theta}} = (1.4 \quad 0.4)^\top$

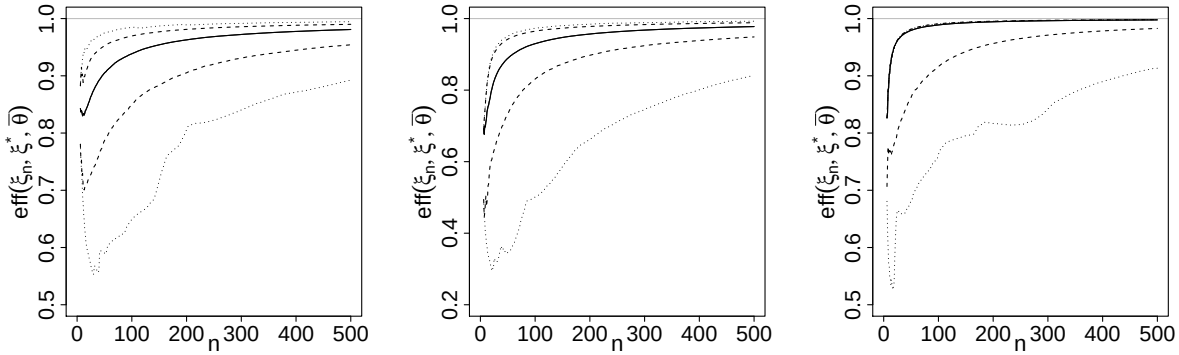
Figure C.13.: D-efficiency of the estimated MSE for the complementary log-log model

Efficiency of the Adaptive Designs



(a) $\bar{\boldsymbol{\theta}} = (0 \quad 1)^\top$ (b) $\bar{\boldsymbol{\theta}} = (0.6 \quad 1.8)^\top$ (c) $\bar{\boldsymbol{\theta}} = (1.4 \quad 0.4)^\top$

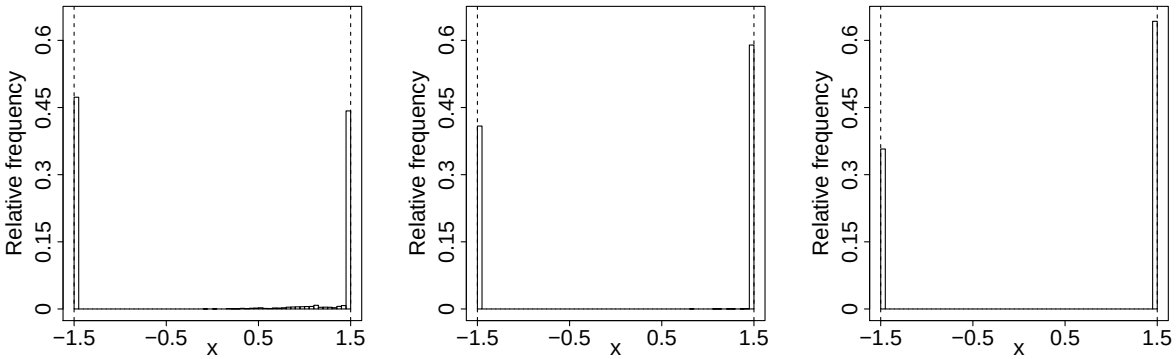
Figure C.14.: Efficiency of the adaptive design for the probit model.
solid: median of the efficiencies, dashed: 5%- and 95%-quantile, dotted: minimum and maximum



(a) $\bar{\boldsymbol{\theta}} = (0 \quad 1)^\top$ (b) $\bar{\boldsymbol{\theta}} = (0.6 \quad 1.8)^\top$ (c) $\bar{\boldsymbol{\theta}} = (1.4 \quad 0.4)^\top$

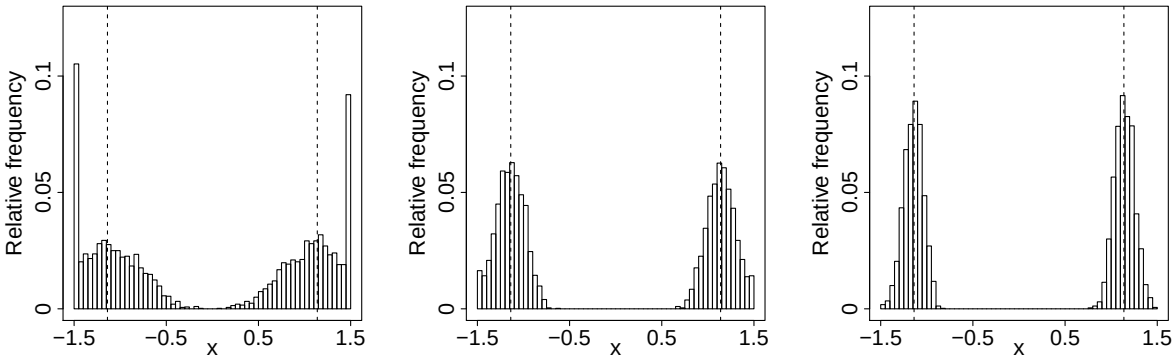
Figure C.15.: Efficiency of the adaptive design for the log-log model.
solid: median of the efficiencies, dashed: 5%- and 95%-quantile, dotted: minimum and maximum

Histograms of the Design Points



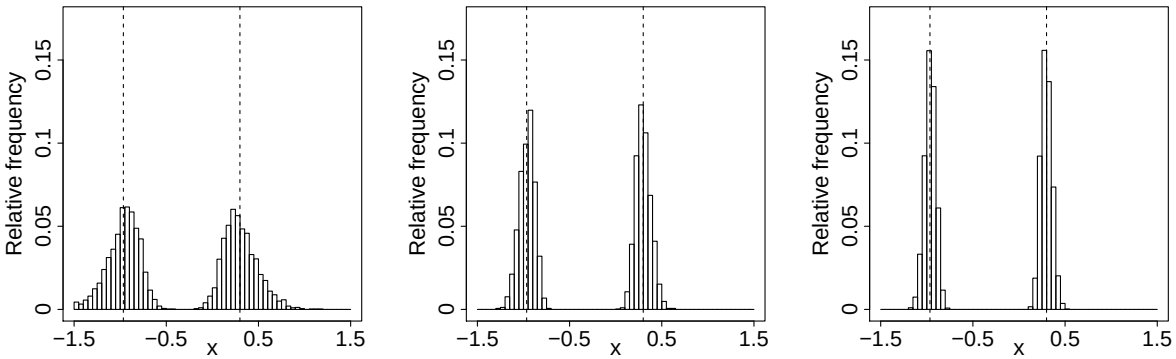
(a) $n = 50$ (b) $n = 250$ (c) $n = 500$

Figure C.16.: Histograms of the design points at different steps calculated over all replications for the logit model with $\bar{\boldsymbol{\theta}} = (1.4 \quad 0.4)^\top$.
dashed lines: locally D-optimal design points



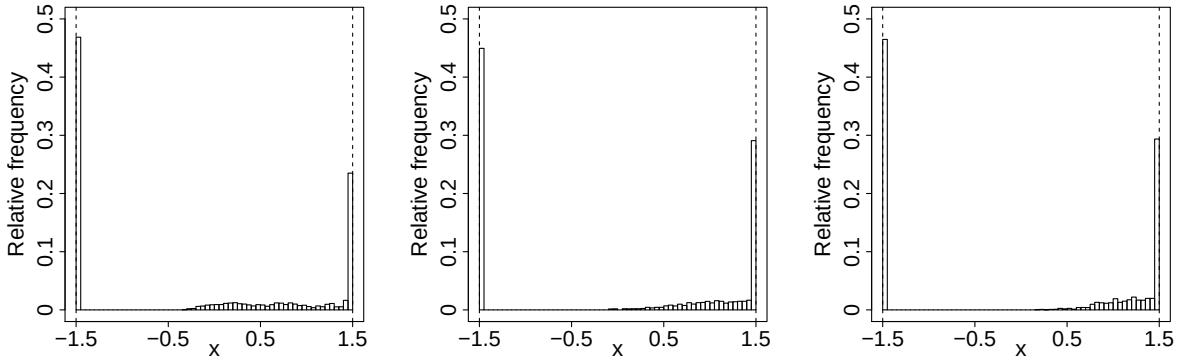
(a) $n = 50$ (b) $n = 250$ (c) $n = 500$

Figure C.17.: Histograms of the design points at different steps calculated over all replications for the probit model with $\bar{\boldsymbol{\theta}} = (0 \quad 1)^\top$.
dashed lines: locally D-optimal design points



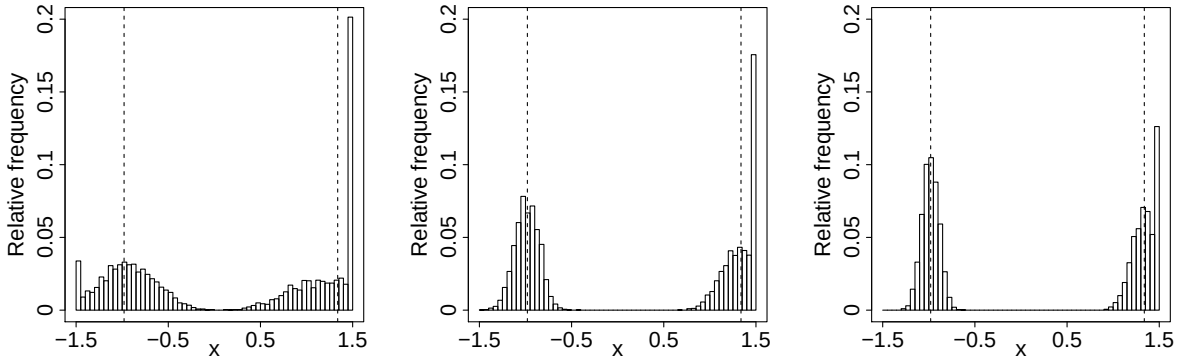
(a) $n = 50$ (b) $n = 250$ (c) $n = 500$

Figure C.18.: Histograms of the design points at different steps calculated over all replications for the probit model with $\bar{\boldsymbol{\theta}} = (0.6 \quad 1.8)^\top$.
dashed lines: locally D-optimal design points



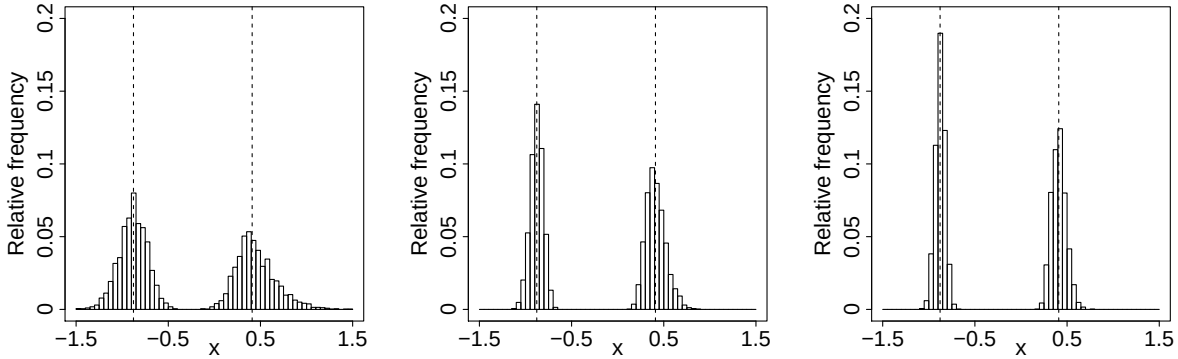
(a) $n = 50$ (b) $n = 250$ (c) $n = 500$

Figure C.19.: Histograms of the design points at different steps calculated over all replications for the probit model with $\bar{\boldsymbol{\theta}} = (1.4 \quad 0.4)^\top$.
dashed lines: locally D-optimal design points



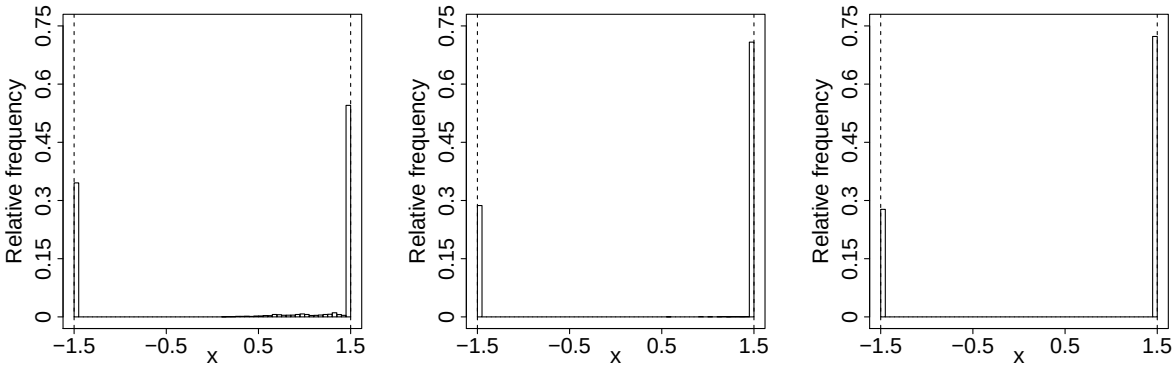
(a) $n = 50$ (b) $n = 250$ (c) $n = 500$

Figure C.20.: Histograms of the design points at different steps calculated over all replications for the log-log model with $\bar{\boldsymbol{\theta}} = (0 \quad 1)^\top$.
dashed lines: locally D-optimal design points



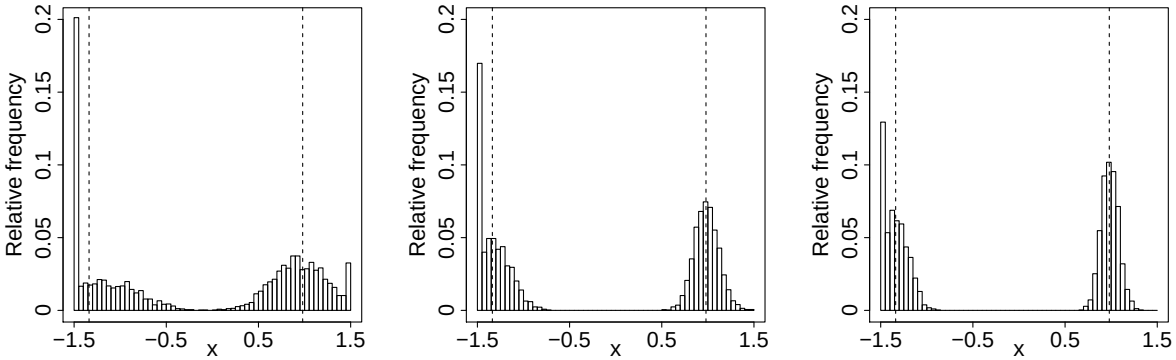
(a) $n = 50$ (b) $n = 250$ (c) $n = 500$

Figure C.21.: Histograms of the design points at different steps calculated over all replications for the log-log model with $\bar{\boldsymbol{\theta}} = (0.6 \quad 1.8)^\top$.
dashed lines: locally D-optimal design points



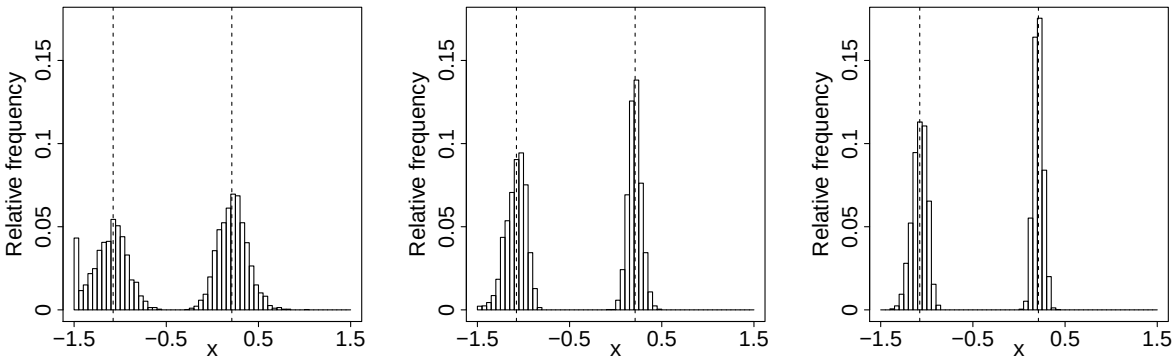
(a) $n = 50$ (b) $n = 250$ (c) $n = 500$

Figure C.22.: Histograms of the design points at different steps calculated over all replications for the log-log model with $\bar{\boldsymbol{\theta}} = (1.4 \quad 0.4)^\top$.
dashed lines: locally D-optimal design points



(a) $n = 50$ (b) $n = 250$ (c) $n = 500$

Figure C.23.: Histograms of the design points at different steps calculated over all replications for the complementary log-log model with $\bar{\boldsymbol{\theta}} = (0 \quad 1)^\top$.
dashed lines: locally D-optimal design points



(a) $n = 50$ (b) $n = 250$ (c) $n = 500$

Figure C.24.: Histograms of the design points at different steps calculated over all replications for the complementary log-log model with $\bar{\boldsymbol{\theta}} = (0.6 \quad 1.8)^\top$.
dashed lines: locally D-optimal design points

Bibliography

A. Albert and J. A. Anderson. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71:1–10, 1984.

A. C. Atkinson and R. A. Bailey. One hundred years of the design of experiments on and off the pages of Biometrika. *Biometrika*, 88:53–97, 2001.

M. Benaïm. Dynamics of stochastic approximation algorithms. In J. Azéma, M. Émery, M. Ledoux and M. Yor, editors, *Séminaire de Probabilités XXXIII*, pages 1–68. Springer, Berlin, 1999.

M. Benaïm, J. Hofbauer and S. Sorin. Stochastic approximations and differential inclusions. *SIAM Journal on Control and Optimization*, 44:328–348, 2005.

A. Benveniste, M. Métivier and P. Priouret. *Adaptive algorithms and stochastic approximations*. Springer, Berlin, 1990.

R. H. Byrd, P. Lu, J. Nocedal and C. Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16:1190–1208, 1995.

P. Chaudhuri and P. A. Mykland. Nonlinear experiments: Optimal design and inference based on likelihood. *Journal of the American Statistical Association*, 88:538–546, 1993.

K. Chen, I. Hu and Z. Ying. Strong consistency of maximum quasi-likelihood estimators in generalized linear models with fixed and adaptive designs. *The Annals of Statistics*, 27:1155–1163, 1999.

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

Y. S. Chow and H. Teicher. *Probability theory. Independence, interchangeability, martingales*. Springer-Verlag, New York, 2nd edition, 1988.

A. Dvoretzky. Asymptotic normality for sums of dependent random variables. In L. M. L. Cam, J. Neyman and E. L. Scott, editors, *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, pages 513–535, Berkeley (California), 1972. University of California Press.

G. Elfving. Optimum allocation in linear regression theory. *The Annals of Mathematical Statistics*, 23:255–262, 1952.

L. Fahrmeir and H. Kaufmann. Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics*, 13:342–346, 1985.

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

I. Ford, B. Torsney and C. F. J. Wu. The use of a canonical form in the construction of locally optimal designs for non-linear problems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 54:569–583, 1992.

A. Gut. *Probability: A Graduate Course*. Springer, New York, 2nd edition, 2013.

P. Hall and C. C. Heyde. *Martingale Limit Theory and Its Application*. Academic Press, New York, 1980.

D. A. Harville. *Matrix Algebra From a Statistician's Perspective*. Springer-Verlag, New York, 1997.

J. Karvanen. Efficient initial designs for binary response data. *Statistical Methodology*, 5: 462–473, 2008.

J. Kiefer. Optimum designs in regression problems. II. *Annals of Mathematical Statistics*, 32:298–325, 1961.

J. Kiefer. General equivalence theory for optimum designs (approximate theory). *The Annals of Statistics*, 2:849–879, 1974.

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

J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23:462–466, 1952.

K. Knopp. *Theorie und Anwendung der unendlichen Reihen*. Springer-Verlag, Berlin, 6th edition, 1996.

H. J. Kushner and D. S. Clark. *Stochastic approximation methods for constrained and unconstrained systems*. Springer-Verlag, New York, 1978.

H. J. Kushner and G. G. Yin. *Stochastic approximation and recursive algorithms and applications*. Springer-Verlag, New York, 2nd edition, 2003.

T. L. Lai. Asymptotic properties of nonlinear least squares estimates in stochastic regression models. *Ann. Statist.*, 22:1917–1930, 1994.

T. L. Lai and C. Z. Wei. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals of Statistics*, 10:154–166, 1982.

E. L. Lehmann and G. Casella. *Theory of point estimation*. Springer-Verlag, New York, 2nd edition, 1998.

L. Ljung. Analysis of recursive stochastic algorithms. *IEEE Transactions on Automatic Control*, 22:551–575, 1977.

P. McCullagh and J. A. Nelder. *Generalized linear models*. Chapman & Hall/CRC, London, 2nd edition, 1997.

M. Métivier and P. Priouret. Théorèmes de convergence presque sure pour une classe d'algorithmes stochastiques à pas décroissant. *Probability Theory and Related Fields*, 74:402–428, 1987.

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

L. Pronzato. Asymptotic properties of nonlinear estimates in stochastic models with finite design space. *Statistics and Probability Letters*, 79:2307–2313, 2009.

L. Pronzato. One-step ahead adaptive D-optimal designs on a finite design space is asymptotically optimal. *Metrika*, 71:219–238, 2010.

L. Pronzato and A. Pázman. *Design of experiments in nonlinear models*. Springer, New York, 2013.

F. Pukelsheim. *Optimal Design of Experiments*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, reprint of the 1993 original edition, 2006.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL `http://www.r-project.org`.

H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22:400–407, 1951.

D. Serre. *Matrices*. Springer, New York, 2nd edition, 2010.

M. J. Silvapulle. On the existence of maximum likelihood estimators for binomial response models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 43:310–313, 1981.

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

R. W. M. Wedderburn. On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika*, 63:27–32, 1976.

L. V. White. An extension of the general equivalence theorem to nonlinear models. *Biometrika*, 2:345–348, 1973.

C. F. J. Wu. Asymptotic theory of nonlinear least squares estimation. *The Annals of Statistics*, 9:501–513, 1981.

C. F. J. Wu. Efficient sequential designs with binary data. *Journal of the American Statistical Association*, 80:974–984, 1985.

C. F. J. Wu and H. P. Wynn. The convergence of general step-length algorithms for regular optimum design criteria. *The Annals of Statistics*, 6:1273–1285, 1978.

H. P. Wynn. The sequential generation of D-optimum experimental designs. *The Annals of Mathematical Statistics*, 41:1655–1664, 1970.

Z. Ying and C. F. J. Wu. An asymptotic theory of sequential designs based on maximum likelihood recursion. *Statistica Sinica*, 7:75–91, 1997.

List of Symbols

$\mathbb{1}_{\mathcal{A}}$	indicator function of the set $\mathcal{A}$.
$\mathcal{C}_n^j$	relative interior of the convex cone generated by design points where $y_i = j$, $j = 0, 1$.
$d(t)$	weight function in the Fisher information matrix.
$\mathbf{D}_n(\mathbf{F}_n\boldsymbol{\theta})$	diagonal matrix with entries $d(\boldsymbol{f}(\boldsymbol{x}_i)^\top\boldsymbol{\theta})$.
$D_n(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$	sum of squares for nonlinear models.
$\text{eff}(\xi_1, \xi_2, \boldsymbol{\theta})$	efficiency of the design ξ_1 with respect to ξ_2.
$\mathcal{E}_{m,n}$	accumulated effects of the observational error ε_n.
$\mathbf{E}_p$	indentity matrix.
ε_n	random / observational error of the n-th observation.
$\boldsymbol{f}(\boldsymbol{x})$	regression function.
$\mathbf{F}_n$	design matrix.
$\mathcal{F}_n$	σ-field generated by $\boldsymbol{Y}_n$ and $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$.
$F_\phi(\mathbf{M}_1, \mathbf{M_2})$	directional derivative of the criterion function ϕ at $\mathbf{M}_1$ in direction of $\mathbf{M}_2 - \mathbf{M}_1$.
$G(t)$	mean function.
$\mathcal{G}_{m,n}$	accumulated effects of the mean function.
$\boldsymbol{G}_n(\mathbf{F}_n\boldsymbol{\theta})$	vector of mean functions corresponding to $\boldsymbol{Y}_n$.
$\mathbf{H}_n(\boldsymbol{\theta}, \mathbf{F}_n)$	Hessian matrix of the log-likelihood.
$\mathbf{I}(\boldsymbol{\theta}, \mathbf{F})$	Fisher information matrix.
$l(\boldsymbol{\theta}, \boldsymbol{y}_n, \mathbf{F}_n)$	log-likelihood function for n observations.
λ_n	minimal eigenvalue of $\mathbf{F}_n^\top\mathbf{F}_n$.
$\lambda_{\min}(\mathbf{A})$, $\lambda_{\max}(\mathbf{A})$	smallest/largest eigenvalue of a matrix $\mathbf{A}$.
$\mathcal{M}_{\boldsymbol{\theta}}$	set of all weighted information matrices over $\mathcal{X}$.
$\mathbf{M}(\boldsymbol{\theta}, \xi)$	weighted (Fisher) information matrix for a design ξ.
$\mathbf{MSE}_n$	estimated mean squared error matrix.
$\nu(t)$	index at time t.
ϕ_D	D-criterion.

$\psi(t)$	first derivative of $\log\Big(G(t)/(1-G(t))\Big)$.
$\boldsymbol{\Psi}_n(\mathbf{F}_n\boldsymbol{\theta})$	diagonal matrix with entries $\psi(\boldsymbol{f}(\boldsymbol{x}_i)^\top\boldsymbol{\theta})$.
$r(t,t_0)$	error of the Taylor expansion of G.
$\boldsymbol{R}_n(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2), \boldsymbol{R}_n$	vector of errors of the Taylor expansion of G.
$\mathcal{R}_{m,n}$	accumulated effects of $\boldsymbol{R}_n$.
$\boldsymbol{s}_n(\boldsymbol{\theta})$	score function.
$\tilde{\mathcal{S}}_{m,n}$	accumulated effects of $\tilde{\boldsymbol{s}}$.
$\tilde{\boldsymbol{s}}_n(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$	"pseudo" score function in the recursion of $\hat{\boldsymbol{\theta}}_n$.
Θ	parameter space.
$\boldsymbol{\theta}$	parameter.
$\bar{\boldsymbol{\theta}}$	"actual value" of the parameter in the experiment.
$\hat{\boldsymbol{\theta}}_n$	maximum likelihood estimator/estimate based on the first n observations.
$\hat{\boldsymbol{\theta}}(t)$	piecewise constant time interpolation of the sequence of estimates.
$\boldsymbol{\theta}(t)$	limit function of the interpolated process $\hat{\boldsymbol{\theta}}(t_n+t)$.
t_n	"natural" time after n steps.
$\mathcal{X}$	design space.
$\boldsymbol{X}$, $\boldsymbol{X}_n$	$\mathcal{X}$-valued random variables.
$\boldsymbol{x}$, $\boldsymbol{x}_n$	design point, n-th design point.
ξ	design.
$\xi_{\boldsymbol{x}}$	one-point design concentrated at $\boldsymbol{x}$.
Ξ	set of all designs over $\mathcal{X}$.
Y, Y_n	observation, n-th observation.
$\boldsymbol{Y}_n$	vector of the first n observations.

List of Figures

List of Tables