Robust Optimal Experimental Design

for Model Discrimination of Kinetic ODE Systems

Dissertation zur Erlangung des Doktorgrades
an der Fakultät für Mathematik und Physik
der Albert-Ludwigs-Universität Freiburg im Breisgau

vorgelegt von

Dominik Skanda

April 2012

Dekan:	Prof. Dr. Kay Königsmann
1. Referent:	Prof. Dr. Dirk Lebiedz Universität Ulm Institut für Numerische Mathematik Helmholtzstraße 20 89081 Ulm
2. Referent:	Prof. Dr. Roland Herzog Fakultät für Mathematik TU Chemnitz Reichenhainer Str. 39 09126 Chemnitz
Datum der Promotion:	19. Oktober 2012

Preface

In this work, we present the development of a numerical algorithm, which calculates a design of experiments to allow for optimal discrimination of different hypothetic candidate models of a given dynamical system for the most inappropriate parameter configurations within a parameter range. The collectivity of design conditions is novel and motivated by a real biological experimental setup. The statistical discrimination criterion is worked out rigorously for these settings. The underlying problem can be classified as a semi-infinite optimization problem, which is solved in an *Outer Approximations* approach. The algorithmic framework is applied to two example problems for the calculation of optimal experimental designs. Additionally, it is applied to design a Circadian Rhythm to set its period in a robust optimal way.

Acknowledgements

First, I would like to thank Prof. Dr. D. Lebiedz for giving me the opportunity to create and realize my own project within his workgroup and for his support during the last years.
Second, I want to thank Prof. Dr. P. Beyer and Cornelia Bär, since the motivation and the inspiration for it stems from a joint project with them.
In particular I want to thank Dr. B. Bell from the Applied Physics Laboratory of the University of Washington for his great support for his software library CppAD, especially for extending the functionality of it to the needs arised during this work.
I want to thank all of my colleagues at ZBSA for any help and the nice time we had. Especially, I want to thank Marc Fein, Marcel Rehberg and Jochen Siehr for proofreading parts of this thesis.
Last but not least, I want to thank my family and friends for giving me support in any way.

This work has been partially supported by the Freiburg Initiative for Systems Biology (FRISYS), part of the BMBF FORSYS systems biology initiative, the Freiburg excellence cluster Centre for Biological Signalling Studies (BIOSS), the Helmholtz alliance Systems Biology of Cancer and the Nephage initiative (BMBF Gerontosys II - NephAge (031 5896A)).

Contents

CHAPTER 1

Introduction

Finding suitable models for dynamic systems is an important task in natural science. On the one hand a correct model helps to understand the underlying mechanisms on the other hand one can use the model to predict the behavior of a system under various circumstances. In particular in modern systems biology a related issue is to link molecular attributes to dynamic mechanisms and functional properties at the system level in order to mechanistically understand emerging functionality. For these purposes, mathematical modeling, numerical simulation and scientific computing techniques are indispensable. Quantitative modeling closely combined with experimental investigations is required if the model is supposed to be used for sound mechanistic analysis and model predictions. Typically, before an appropriate model of a system is found different hypothetical models might be reasonable and consistent with previous knowledge and available data. The main goal now is to find the best suited model out of different hypotheses. This is usually done by iterative measurements and successive fitting of the different models to the collectivity of all series of measurements. This is repeated until all inappropriate models do not fit to the collectivity of all series of measurements any more. Thus the inappropriate models are iteratively falsified. The whole process is called model discrimination. In the process of model discrimination the question arises in which way the sequential experiments have to be designed such that the goal of discarding inappropriate models is achieved best.

In this application oriented scientific computing research work we develop a numerical algorithm, which calculates in a suitable sense an optimal design of experiments, which allows the best discrimination of different hypothetic canditate models of a given dy-

namic system modeled by ordinary differential equations (ODEs). Different approaches to design experiments for model discrimination exist. Besides optimization methods (see e.g. [71, 40, 116, 68]) a model-based feedback controller see e.g. [7] and Markov chain Monte Carlo sampling methods [81] have been used to construct an appropriate design. An overview of different experimental design techniques can be found in [69]. In this work, the motivation for such an algorithm comes from the cooperation with a group of experimental biologists, working on the reconstruction of the provitamin A biosynthetic pathway in an *in vitro* biphasic system [129, 128]. Therefore, the design conditions are specifically taylored to the needs of their experimental setup. The design comprises initial values, system perturbations and the optimal placement of measurement time points. The number of measurements as well as the time points are subject to design. The parameters of the models up to an estimated confidence region are generally not known a priori. Therefore, one has to incorporate possible parameter configurations of different models into a model discrimination algorithm leading to the need for robustification. The statistical discrimination criterion is worked out rigorously for these settings. A derivation from the Kullback-Leibler divergence as optimization objective is presented for the case of discontinuous Heaviside-functions modeling the measurement decision, which are replaced by continuous approximations during the optimization procedure. The resulting problem can be classified as a semi-infinite optimization problem, which we solve in an *Outer Approximations* approach stabilized by a suggested homotopy strategy whose efficiency is demonstrated. We choose the *Outer Approximations* approach, since beside the fact that convergence can be proven, at each iteration of the *Outer Approximations* algorithm a worst case design is calculated. Therefore, although the current design might not be optimal, it can be used reliably for practical application. This behavior is especially beneficial in a biological setting, since often due to complex experimental setups and imprecise measurements the model parameters can only be calibrated with huge variances leading to a non convex and non linear robustification space.

1.1. Results and new contributions

Results and new contributions are shortly presented in this section. Parts of this work have been published in [105, 73, 104] and an electronic preprint is published in [106, 107] (the later one is a revised version).

- The algorithm for the numerical calculation of optimal experimental designs, which is developed in this work, is specifically tailored to a real life experimental situation. To our knowledge, this special scenario is not considered in existing literature.

Especially, the need to determine the best time point for a measurement and to simultaneously determine the optimum number of measurement time points does not seem to be considered satisfactorily in literature despite the fact that these demands seem natural. Therefore, a new optimization criterion for optimal experimental design in the context of model discrimination is rigerously derived utilizing discontinuous Heaviside-functions in an attempt to face this requirements.

- This new optimization criterion leads to a discontinuous semi-infinite optimization problem and thus it is not possible to solve this optimization problem by standard optimization approaches, directly. To remedy this fact a smoothing approach, motivated by a similar approach in [132], is applied. The consistency of this smoothing approach is theoretically validated. It should be mentioned that the optimization scenario, presented in [132], differs in two ways. First, in [132] only the unconstrained case is considered. Whereas in our case, inequality and equality constraints have to be tackled, as well. Second, the optimization problem, we face, is semi-infinite. To our knowledge, it is the first time that a smoothing approach is applied in this scenario and is theoretically validated.

- We solve the semi-infinite approximation utilizing the well known *Outer Approximations* scheme. Each finite subproblem within the *Outer Approximations* scheme is solved by an *Interior Point* optimization algorithm. Modern *Interior Point* algorithms are robust even under weak constraint qualifications and therefore highly suited for this type of problems. To improve robustness of the *Outer Approximations* scheme, we suggest a heuristic homotopy method, which is similar to the one presented in [93].

- A BDF-integrator has been implemented with the capability to calculate higher order sensitivities in *forward* and *reverse* mode based on the sophisticated framework of *Internal Numerical Differentiation* [25, 26].

- The whole framework for model discrimination, which is developed in this thesis, has been implemented in a software package using the third party software packages `IPOPT` [124, 127] and `CppAD` [19, 18].

1.2. Outline of this thesis

This work is organized as follows:

Chapter 2 (Theory of model discrimination).
In this chapter, we first give a short introduction to optimal experimental design in the

context of model discrimination. In particular, we introduce Kullback-Leibler-optimality as discussed by López-Fidalgo et al. [77]. The design conditions are stated. Based on Kullback-Leibler-optimality, we formally derive our optimization objective function leading to a discontinuous optimization problem.

Chapter 3 (Theory of the solution of minMax optimization problems).
The resulting optimization problem of interest for the calculation of an optimal experimental design, which is solved numerically, can be classified as a semi-infinite inequality and finite equality constrained optimization problem (**SIECP**). Therefore, in this chapter we briefly present the theoretical basis for the solution of **SIECPs**, i.e. necessary first order optimality conditions. The concept of a continuous optimality function is introduced, which gives a measure of the degree of optimality in respect to the first order optimality conditions. Prior to that, an equivalent first order optimality condition for finite inequality and equality constrained optimization problems (**IECPs**) is introduced together with a related continuous optimality function. Based on the concept of optimality functions, the mathematical derivation of a consistent discretization scheme for the numerical solution of a **SIECP**, namely the *Outer Approximations* scheme, is finally shown.

In Chapters 4, 5 and 6, the elementary numerical techniques, which are indispensable for the numerical calculation of the resulting optimization problem of interest, are presented.

Chapter 4 (Automatic Differentiation).
In this chapter, we give a brief introduction to Automatic Differentiation (AD) utilizing truncated Taylor series propagation in *forward* and *reverse* mode. Hereafter, we treat the special case of AD of solutions of parametrized nonlinear equations, i.e. implicitly defined functions.

Chapter 5 (Calculating numerical solutions of Ordinary Differential Equations and Sensitivity Generation for Ordinary Differential Equations).
In this chapter, we treat the numerical solution of ordinary differential equations (ODEs). More precisely, we present the implementation details of a *Backward Differentiation Formula* (BDF) method based on Nordsiek array interpolation. This is a numerical integration method, capable to solve stiff ODEs. Based on the sophisticated framework of *Internal Numerical Differentiation* [25, 26], we subsequently present the implementation details of algorithmic strategies for the numerical calulation of sensitivities within the implemented BDF method.

Chapter 6 (Nonlinear Programming).
This chapter treats the numerical solution of nonlinear programming (NLP) problems by use of a primal-dual *Interior Point* method. The conceptual idea, the barrier approach, is briefly sketched.

Chapter 7 (Numerical Calculation of Robust Optimal Experimental Design).
In this chapter, we first state the formal discontinuous optimization problem for the calculation of an optimal experimental design. Since this problem is not directly solvable by the *Outer Approximations* scheme as presented in Chapter 3 (due to the discontinuities in the optimization objective), the optimization problem is first approximated by a smoothed continuous version depending on smoothing parameters. This smoothing approach is presented and the theoretical aspects of this smoothing approach are discussed. Hereafter, the application of the *Outer Approximations* scheme to the smoothed continuous optimization problem is worked out. Finally, we discuss a homotopy approach to numerically stabilize the *Outer Approximations* scheme.

Chapter 8 (Numerical results).
We have applied the algorithmic framework for the calculation of optimal experimental designs to two example problems for which we present results in this chapter, namely to models describing glycolytic oscillations and to models describing signal sensing in *dictyostelium discoideum*. Additionally, we have applied the algorithmic framework to design a Circadian Rhythm in order to set its period in a robust optimal way.

1.3. General notation

The i-th element of a vector x of an n-dimensional vector space is written as x^i, i.e. with superscript indices. The exponentiation of a scalar value a by the power of b is also written as a^b. If the i-th element of a vector x is exponentiated by the power b, it is written as $\left(x^i\right)^b$ to prevent ambiguity.
Additionally, for a sequence $\{y_j\}_{j=0}^{\infty}$ we denote by

$$y_j \to^K y$$

that $\{y_j\}_{j\in K} \subset \{y_j\}_{j=0}^{\infty}$ converges to y. For a scalar sequence $\{z_j\}_{j=0}^{\infty}$ the limit superior is denoted by $\overline{\lim}\, z_j$ and the limit inferior is denoted by $\underline{\lim}\, z_j$. These notations are taken from [91].

CHAPTER 2

Theory of model discrimination

2.1. Introduction to model discrimination

Experience has shown that a modelling approach firmly based on experimental data can lead to the generation of valuable biological knowledge [115, 100, 17].
To discriminate a set of candidate models against a given set of experimental data, often likelihood ratio tests based on bootstrap methods are applied [63, 114, 117]. Ranking methods like Stewart's method [112, 66, 20] or the well known Akaike information criterion [30] are popular as well in the field of biological modeling.
In contrast to these approaches this work deals with the problem of designing experiments so that statistical methods of this type can be exploited in an optimal sense.
This differs from the approach to find an experimental design to best estimate the parameters of a model for a given experimental system in the sense of criteria characterizing the confidence regions [67, 15, 13].
The conceptual methodology presented here goes back to ideas of Hunter and Reiner [65], they state "*... choose the experimental points which,..., will most strain the incorrect model in its attempt to jointly explain the previous data and the new observation.*"
Atkinson and Fedorov summerized in 1975 [11] a statistically rooted optimal design criterium for model discrimination, the so called T-optimum design. The T-optimum design is based on the assumption that the observations $y_{ik} \in \mathbb{R}$ can either be explained by nonlinear regression model $\eta_1(x, \theta_1)$ or $\eta_2(x, \theta_2)$ of type

$$y_{ik} = \eta(x_i) + \epsilon_{ik} \quad i = 1, ..., N, \ k = 1, ..., R_i, \tag{2.1}$$

where the design points x_i, $i = 1, ..., N$, are known and the random variables ϵ_{ik} are independently normally distributed with zero mean and constant variance σ^2. The index $k = 1, ..., R_i$, enumerates R_i repeated measurements at the design point x_i.
Assuming that the first model is true, that is $\eta_t(x) = \eta_1(x, \theta_1)$, the optimization objective for a T-optimum design is given by

$$\Delta_2(\xi_N) := \sum_{i=1}^{N} p_i \{\eta_t(x_i) - \eta_2(x_i, \hat{\theta}_2)\}^2,$$

where

$$\sum_{i=1}^{N} p_i \{\eta_t(x_i) - \eta_2(x_i, \hat{\theta}_2)\}^2 = \inf_{\theta_2 \in \Theta_2} \sum_{i=1}^{N} p_i \{\eta_t(x_i) - \eta_2(x_i, \theta_2)\}^2$$

with the design ξ_N containing probability weights p_i for the observations at points x_i. i.e.

$$\xi_N = \begin{Bmatrix} x_1, ..., x_N \\ p_1, ..., p_N \end{Bmatrix},$$

$$\sum_{i=1}^{N} p_i = 1.$$

The design $\hat{\xi}_N$ for which

$$\Delta_2(\hat{\xi}_N) = \sup_{\xi_N} \Delta_2(\xi_N)$$

is called T-optimum. The T-optimum design provides the most powerful F-test for lack of fit of the second model when the first is true [10]. Instead of analyzing such Maxmin problems of finding a design for N discrete trials, it is more convenient to replace

$$\sum_{i=1}^{N} p_i \{\eta_t(x_i) - \eta_2(x_i, \theta_2)\}^2$$

by

$$\Delta_2(\xi) = \int_{\mathcal{X}} \{\eta_t(x) - \eta_2(x, \hat{\theta}_2(\xi))\}^2 \xi(dx),$$

with

$$\hat{\theta}_2(\xi) := \arg\min_{\theta_2 \in \Theta_2} \int_{\mathcal{X}} \{\eta_t(x) - \eta_2(x, \theta_2)\}^2 \xi(dx),$$

where ξ is a normed measure defined on the design region $\mathcal{X}$ and assuming that $\hat{\theta}_2(\xi)$ exists. This leads to the Maxmin optimization problem

$$\Delta_2(\hat{\xi}) = \sup_{\xi} \Delta_2(\xi). \tag{2.2}$$

With the assumptions

(a) $\mathcal{X}$ compact and $\eta_j(x, \theta_j)$, $j = 1, 2$, continuous on $\mathcal{X}$,

(b) $\eta_j(x, \theta_j)$ differentiable with respect to θ_j on Θ_j, $j = 1, 2$,

(c) The optimal design $\hat{\xi}$ satisfying (2.2) is regular, i.e.

$$\int_{\mathcal{X}} \{\eta_t(x) - \eta_2(x, \hat{\theta}_2)\xi(dx)\} = \inf_{\theta_2 \in \Theta_2} \int_{\mathcal{X}} \{\eta_t(x) - \eta_2(x, \theta_2)\xi(dx)\},$$

has a unique solution $\hat{\theta}_2$ when $\xi = \hat{\xi}$;

the following theorem states the necessary and sufficient optimality conditions of the optimization problem (2.2).

Theorem 1 (Theorem 1 in [11])**.** *Given the preceeding assumptions:*

(i) A necessary and sufficient condition for a design $\hat{\xi}$ to be T-optimum is the inequality

$$\psi_2(x, \hat{\xi}) \leq \Delta_2(\hat{\xi}) \ \forall x \in \mathcal{X},$$

where

$$\psi_2(x, \hat{\xi}) := \{\eta_t(x) - \eta_2(x, \hat{\theta}_2)\};$$

(ii) at the points of the optimum design $\psi_2(x, \hat{\xi})$ achieves its upper bound;

(iii) for any non-optimal design ξ', i.e. a design for which $\Delta_2(\xi') < \Delta_2(\hat{\xi})$, it holds

$$\sup_{x \in \mathcal{X}} \psi_2(x, \xi') > \Delta_2(\hat{\xi});$$

(iv) the set of T-optimum designs is convex.

A proof of this theorem can be found in [43].

In the literature [10, 43, 11, 12] one can find two common classes of algorithms to compute such T-optimum designs.
First, the following iterative algorithms for the computation of a T-optimum design:

Algorithm 1 (The algorithm is taken from [11])**.**

(i) Let ξ_s be the design at iteration s. Find x_{s+1} according to

$$\psi_2(x_{s+1}, \xi_s) = \sup_{x \in \mathcal{X}} \psi_2(x, \xi_s).$$

(ii) Compute the next design as

$$\xi_{s+1} = (1 - \alpha_s)\xi_s + \alpha_s \xi(x_{s+1}),$$

where $\xi(x_{s+1})$ is a design with respect to a single point measure in x_{s+1}.

α_s has to be a sequence of one of following forms:

(a) any sequence which satisfies

$$\alpha_s \to 0, \quad \sum_{s=0}^{\infty} \alpha_s = \infty, \quad \sum_{s=0}^{\infty} \alpha_s^2 < \infty;$$

(b) α_s maximizes $\Delta_2((1 - \alpha)\xi_s + \alpha\xi(x_{s+1}))$;

(c) if $\Delta_2(\xi_s) \geq \Delta_2(\xi_{s+1})$, α_s is taken as $\min(\bar{\alpha}_s, \alpha_{s-1}/\beta)$, for $\beta > 0$ fixed, with

$$\sum \bar{\alpha}_s = \infty \quad \text{and} \quad \lim \bar{\alpha}_s = 0, \quad \text{as} \quad s \to \infty.$$

Details can be found in [11].

The second common class of important algorithms to construct discriminating designs are the sequential algorithms leading to designs which are asymptotically T-optimum and which give at each trial the largest increase in the expected value of the sum of squares of differences between the responses of the two models.

Algorithm 2 (The algorithm is taken from [11])**.**

---

1. *Given an initial nonsingular design ξ_{N_0} (a design is called singular if its information matrix is singular [43]), where N_0 is the number of observations. Find the estimates $\hat{\theta}_{1N_0}$ and $\hat{\theta}_{2N_0}$ satisfying*

$$\sum_{i=1}^{N_0} \{y_i - \eta_j(x_i, \hat{\theta}_{jN_0})\}^2 = \inf_{\theta_j \in \theta_j} \sum_{i=1}^{N_0} \{y_i - \eta_j(x_i, \theta_j)\}^2 \quad j = 1, 2.$$

2. *The point x_{N_0+1} is found for which*

$$\{\eta_1(x_{N_0+1}, \hat{\theta}_{1N_0}) - \eta_2(x_{N_0+1}, \hat{\theta}_{2N_0})\}^2 = \max_{x \in \mathcal{X}} \{\eta_1(x, \hat{\theta}_{1N_0}) - \eta_2(x, \hat{\theta}_{2N_0})\}^2$$

holds.

3. *The* (N_0+1)*st observation is taken at* x_{N_0+1}.

4. *Repeat steps 1 to 3.*

For details we refer to [10, 43, 11, 12].

A generalization of the T-optimum design to the case of multiresponse heteroskedastic regression models of type (2.1) was given by Uciński and Bogacka [119, 120] in 2004 called generalized T-optimality.

2.2. KL-optimal design

In this section a model discrimination criterion based on the Kullback-Leibler (KL) divergence called KL-optimality as discussed by López-Fidalgo et al. [77] is introduced. López-Fidalgo et al. [77] demonstrate that KL-optimality is consistent with T-optimality [11] and generalized T-optimality [119].

We introduce the concept of a probability space and formally define the KL-divergence.

Definition 1. *A probability space is a triple* $\left(\Omega, \mathcal{F}, P\right)$ *consisting of*

- *a non-empty set* Ω *(sample space),*
- *a* σ*-algebra* $\mathcal{F} \subseteq \mathcal{P}(\Omega)$, $E \in \mathcal{F}$ *is called an event,*
- *a probability measure* $P : \mathcal{F} \to [0, 1]$.

Definition 2. *Two probability spaces* $\left(\Omega, \mathcal{F}, P_i\right)$, $i = 1, 2$, *are called absolutely continuous with respect to each other, in symbols* $P_1 \equiv P_2$, *if* $\nexists\ E \in \mathcal{F} : (P_1(E) = 0\ AND\ P_2(E) \neq 0)\ OR\ (P_1(E) \neq 0\ AND\ P_2(E) = 0)$.

The Radon-Nikodym Theorem allows a representation of a probability measure via a measurable probability density function.

Theorem 2. *(Radon-Nikodym)*
Let λ *be a probability measure such that* $\lambda \equiv P_1$, $\lambda \equiv P_2$. *Then* λ*-measurable functions* $f_i : \Omega \to \mathbb{R}, i = 1, 2$, *called generalized probability densities, exist which are unique up to sets of measure zero and non-negative, such that*

$$P_i(E) = \int_E f_i(x) \mathrm{d}\lambda(x), \quad i = 1, 2,$$

for all $E \in \mathcal{F}$.

A proof of this theorem can be found e.g. in [22].

In the following, we use X for the generic variable and x for a specific value of X. If H_i, $i = 1, 2$ is the hypothesis that X is from the statistical population with probability measure P_i, the mean information for discrimination in favor of H_1 against H_2 given $x \in E \in \mathcal{F}$, for P_1 is given by the Kullback–Leibler divergence.

Definition 3. *Kullback–Leibler (KL) divergence*

$$\begin{aligned}\mathcal{I}(1:2;E) :&= \frac{1}{P_1(E)} \int_E \log \frac{f_1(x)}{f_2(x)} \mathrm{d}P_1(x) \\ &= \begin{cases} \dfrac{1}{P_1(E)} \displaystyle\int_E f_1(x) \log \frac{f_1(x)}{f_2(x)} \mathrm{d}\lambda(x) & \text{if} \quad P_1(E) > 0, \\ 0 & \text{if} \quad P_1(E) = 0, \end{cases}\end{aligned}$$

with

$$\mathrm{d}P_1(x) = f_1(x)\mathrm{d}\lambda(x).$$

When E is the entire sample space Ω, we shorten the notation to $\mathcal{I}(1:2)$ and omit the region of integration. For discrete sets E the integral is substituted by a sum.

For details we refer to [70].

Now assume that the sample space Ω is split into two disjoint sets E_1 and E_2, $\Omega = E_1 \cup E_2$. We define a statistical test procedure to choose between hypotheses H_1 and H_2 by accepting H_1 if $x \in E_1$ and accepting H_2 if $x \in E_2$. Assuming that one of the hypotheses has to be true we treat H_2 as the null hypothesis and call E_1 the critical region. The following wrong test decisions can occur.

Definition 4. *Incorrectly accepting H_1 although H_2 is true is called type I error. The probability that this error occurs is given by*

$$\alpha = \mathrm{Prob}(x \in E_1 | H_2) = P_2(E_1).$$

Definition 5. *Incorrectly accepting H_2 although H_1 is true is called type II error. The probability that this error occurs is given by*

$$\beta = \mathrm{Prob}(x \in E_2 | H_1) = P_1(E_2).$$

We assume that the test is repeated n-times and denote by $\mathcal{O}_n$ a sample of n independent observations. $\mathcal{O}_1$ represents a sample of a single observation. β_n is defined as the corresponding probability of an error of type II which depends on the number of

independent observations and the splitting of the corresponding probability space Ω into disjoint sets E_1 and E_2.

The following theorem demonstrates an asymptotic relation between the KL-divergence and the minimum possible probability β_n^* of an error of type II with respect to all possible splittings $E_1 \cup E_2 = \Omega$ with given $\alpha = \text{Prob}(x \in E_1 | H_2) = P_2(E_1)$ [37].

Theorem 3 (Theorem 3.3 in [70]). *For any value of α, say α_0, $0 < \alpha_0 < 1$,*

$$\lim_{n\to\infty} \left(\beta_n^*\right)^{1/n} = e^{-\mathcal{I}(2:1,\mathcal{O}_1)}.$$

A proof of this theorem is given in [37, 70].
Assuming probability models for the outcome of a data measurement experiment depending on experimental design parameter $\xi \in \Xi \subset \mathbb{R}^d$, this theorem justifies the KL-divergence to be an appropriate objective functional for model-based computation of an optimal experimental design for discrimination between model hypotheses. For a design with the largest possible value of $\mathcal{I}$ the asymptotical probability of encountering an error of type II β_n^* becomes minimal with respect to all possible splittings $E_1 \cup E_2 = \Omega$ with given α_0. We indicate the dependency of the KL divergence on the design by $\mathcal{I}(2:1,\mathcal{O}_1;\xi)$. Our aim is to derive an algorithm to calculate the optimal design $\hat{\xi} \in \Xi$ such that

$$\hat{\xi} = \arg\max_{\xi\in\Xi} \mathcal{I}(2:1,\mathcal{O}_1;\xi).$$

An extension of the case to test a simple null hypothesis against a simple alternative hypothesis to the more general case of both hypotheses being composite is generally of interest. This includes the situation to test if given measurement data can be explained best by the parametrized probability measure P_1, parametrized by parameters $\theta_1 \in \Theta_1$ where $\Theta_1 \subset \mathbb{R}^{p_1}$ is the set of all possible parameter values to parametrize P_1, against the hypothesis that the measurement can best be explained by P_2, parametrized by parameters $\theta_2 \in \Theta_2$ where $\Theta_2 \subset \mathbb{R}^{p_2}$ is the set of all possible parameter values to parametrize P_2.
In the following, we assume that the parameters $\theta_2 \in \Theta_2$ of P_2 are known but not the parameters $\theta_1 \in \Theta_1$ of P_1 of the alternative hypotheses. We denote the dependency of the KL divergence on the parameter vector $\theta_1 \in \Theta_1$ by $\mathcal{I}(2:1,\mathcal{O}_1;\xi,\theta_1)$. By calculating

$$\hat{\xi} = \arg\max_{\xi\in\Xi} \min_{\theta_1\in\Theta_1} \mathcal{I}(2:1,\mathcal{O}_1;\xi,\theta_1), \tag{2.3}$$

we can get a robust worst case estimate of an optimally discriminating design for the case of a composite alternative hypothesis.

2.3. Derivation of the optimal experimental design criterion

In this section we derive a numerically computable optimization objective functional based on the framework of KL divergence. The derivation is motivated by the requirements of biological *in vitro* time series experiments modeled by kinetic ODE systems. In most situations such experiments are time and cost consuming. Therefore a central issue is to get the most information out of a single time series data measurement experiment taking place within a given fixed time span $[0, T^{\text{end}}]$. This means that in an optimal experimental design the most informative measurement time points for one measurement run have to be calculated in such a way that only one measurement at one time point can be performed. Often, an experiment cannot produce measurements in a time continuous way. Therefore we assume that there has to be a minimal time span ΔT for the separation of measurement time points. Additionally, the initial species concentrations of the participating species should be chosen in a most discriminating way.
A commonly used practice is to combine kinetic time series measurements with perturbation stimuli like external adding of species quantities. From the model discrimination point of view the optimal time point of perturbation and the optimal species quantities to be added should be determined. We further assume that a measurement cannot be done at the same time as a perturbation.
In the following, we translate these experimental conditions into a statistical model. Given the measurement time-vector $t \in \mathbb{R}^n_+$ with entries t^i for the n measurement time points $t^i, i = 1, ..., n$ such that $t^{i+1} \geq t^i$, the model response vectors at measurement time t^i for hypotheses H_j, $j = 1, 2$ are given by

$$y_{j,i} := y_j(t^{i-1}, t^i, y_{j,i-1} + c_{i-1}, \theta_j), \quad i = 1, ..., n, \ j = 1, 2$$

where $y_j(t^{i-1}, t^i, y_{j,i-1} + c_{i-1}, \theta_j) \in \mathbb{R}^m$, $i = 1, ..., n$, $j = 1, 2$ is the solution of the initial value problem

$$\frac{\mathrm{d}y_j}{\mathrm{d}t} = f_j^{\text{rhs}}(y_j, \theta_j), \quad t \in [t^{\text{init}}, t^{\text{end}}], \quad j = 1, 2 \tag{2.4}$$

with initial state $y_j(t^{\text{init}}) = y_{j,i-1} + c_{i-1}$ at initial time $t^{\text{init}} = t^{i-1}$ and end time $t^{\text{end}} = t^i$, $i = 1, ..., n$, $j = 1, 2$, $t_0 := 0$ and $c_0 := 0$. The vectors $c_i \in \mathbb{R}^m$, $i = 1, ..., n-1$, denote species quantities the experimental system can be perturbed with at time points t^i, $i = 1, ..., n-1$. $f_j^{\text{rhs}}(\cdot, \cdot)$, $j = 1, 2$ are the right hand side functions of the two ODE models. We assume autonomous ODE models [58]. $y_{\text{I}} =: y_{j,0}$, $j = 1, 2$, denotes the initial species concentration of the entire experiment, which is the same for both models.

Let y_{t^i} denote the vector of species concentrations of an observation at measurement time point t^i. By assuming that the measurements at successive time points t^i, $i = 1, ..., n$,

are independent with normally distributed error vectors $\epsilon_{j,i} \in \mathbb{R}^m$, $i = 1, ..., n$, $j = 1, 2$, zero mean and variance functions $\left(v_j(y_{j,i}, t^i, \theta_j)\right)^2$, $i = 1, ..., n$, $j = 1, 2$, we get for the regression models

$$y_{t^i} = y_{j,i} + \epsilon_{j,i}, \quad i = 1, ..., n, \ j = 1, 2,$$

the two model probability densities $f_1(\cdot;\cdot)$ for hypothesis H_1 and $f_2(\cdot;\cdot)$ for hypothesis H_2 at measurement time point t^i, $i = 1, ..., n$, given by

$$f_j(y_{t^i}; y_{j,i}) = \frac{1}{\sqrt{2\pi}|v_j^i|} e^{-\frac{1}{2}(y_{j,i}-y_{t^i})^T V_{j,i}(y_{j,i}-y_{t^i})}, \quad i = 1, ..., n, \ j = 1, 2, \tag{2.5}$$

with $|v_j^i| := \prod_{k=1}^m v_j^k(y_{j,i}, t^i, \theta_j)$, $i = 1, ..., n$, $j = 1, 2$, where $v_j^k(y_{j,i}, t^i, \theta_j)$ denotes the k-th entry of the square root of the variance functions $\left(v_j(y_{j,i}, t^i, \theta_j)\right)^2$, and diagonal matrices $V_{j,i} \in \mathbb{R}^{m\times m}$, $i = 1, ..., n$, $j = 1, 2$ with diagonal entries $V_{j,i}^{kk} := (1/v_j^k(y_{j,i}, t^i, \theta_j))^2$, $i = 1, ..., n$, $j = 1, 2$, $k = 1, ..., m$.

We generally allow for different error models for both hypotheses. The error model might dependent on the species concentrations $y_{j,i}$, $i = 1, ..., n$, $j = 1, 2$, the time t^i, $i = 1, ..., n$, and possibly on parameters θ_j, $j = 1, 2$.

For the sake of notational simplicity we define

$$f_j(y_{t^i}) := f_j(y_{t^i}; y_{j,i}), \quad i = 1, ..., n, \ j = 1, 2.$$

For the full measurement run containing n measurement time points we get the probability density models

$$f_j(y) := \prod_{i=1}^n f_j(y_{t^i}), \quad j = 1, 2.$$

However, by assuming such a model probability distribution we still allow that two measurements are separated by a time span less than ΔT. To overcome this problem we extend the probability spaces $\Omega_i = \mathbb{R}^m$, $i = 1, ..., n$ of a measurement at one measurement time point by one-element-containing sets $\mathcal{N}_i$, $i = 1, ..., n$ to

$$\widetilde{\Omega}_i = \Omega_i \dot{\cup} \mathcal{N}_i, \quad i = 1, ..., n,$$

where $\widetilde{\Omega}_i$ is the disjoint union of Ω_i and $\mathcal{N}_i$, $i = 1, ..., n$. The element of the set $\mathcal{N}_i$ with measure $P(\mathcal{N}_i) \in [0, 1]$, $i = 1, ..., n$ represents the event "no measurement", i.e. $\tilde{y}_i \in \mathcal{N}_i \Leftrightarrow$ "no measurement performed at time point t^i".

In order to derive measures on $\widetilde{\Omega}_i,\ i = 1, ..., n$ that allow a density function representation according to the Radon-Nikodym theorem (Theorem 2), we introduce the Heaviside-functions

$$\mathcal{H}, \mathcal{H}^* : \mathbb{R} \longrightarrow [0, 1]$$

with

$$\mathcal{H}(t^i) = \begin{cases} 1 & \text{if } t^i - t^{i-1} \geq \Delta T \\ 0 & \text{if } t^i - t^{i-1} < \Delta T \end{cases} \quad i = 1, ..., n$$

and

$$\mathcal{H}^*(t^i) := \begin{cases} 0 & \text{if } t^i - t^{i-1} \geq \Delta T \\ 1 & \text{if } t^i - t^{i-1} < \Delta T. \end{cases} \quad i = 1, ..., n$$

By use of these Heaviside-functions and σ-algebras $\mathcal{F}_i$, where $\mathcal{F}_i$ contains the Lebesgue measurable sets on Ω_i and additionally the union of them with the set $\mathcal{N}_i$, we define probability spaces $(\widetilde{\Omega}_i, \mathcal{F}_i, \widetilde{P}_{i,j})$ with measures

$$\widetilde{P}_{i,j} : E_i \in \mathcal{F}_i \mapsto \widetilde{P}_{i,j}(E_i) \in [0, 1], \quad i = 1, ..., n,\ j = 1, 2,$$

with respect to H_1 and H_2. Three cases have to be distinguished:

1. $E_i \subset \Omega_i$,
2. $E_i \subset \mathcal{N}_i$,
3. $E_i \cap \Omega_i \neq \emptyset$ and $E_i \cap \mathcal{N}_i \neq \emptyset$, $i = 1, ..., n$.

For case one with $E_i \subset \Omega_i$ we set

$$\widetilde{P}_{i,j}(E_i) := \mathcal{H}(t^i) \int_{E_i} f_j(y_{t^i}) \mathrm{d}y_{t^i}, \quad i = 1, ..., n,\ j = 1, 2.$$

For case two with $E_i \subset \mathcal{N}_i$ we set

$$\widetilde{P}_{i,j}(E_i) := \mathcal{H}^*(t^i), \quad i = 1, ..., n,\ j = 1, 2.$$

For case three with $E_i \cap \Omega_i \neq \emptyset$ and $E_i \cap \mathcal{N}_i \neq \emptyset$ we set

$$\widetilde{P}_{i,j}(E_i) := \mathcal{H}(t^i) \int_{E_i \cap \Omega_i} f_j(y_{t^i}) \mathrm{d}y_{t^i} + \mathcal{H}^*(t^i), \quad i = 1, ..., n,\ j = 1, 2.$$

By introducing these modifications the probability measures $\widetilde{P}_{i,j},\ i = 1, ..., n,\ j = 1, 2$, do not depend on measurements which are performed in less than ΔT after the previous measurement any more.

To take into account that a species concentration perturbation to the system can only be applied if no measurement is done at the same time, the same procedure is repeated with the Heaviside-functions

$$\begin{aligned}\mathcal{H}(t^i) &= \begin{cases} 1 & \text{if } t^i - t^{i-1} \geq \Delta T \\ 0 & \text{if } t^i - t^{i-1} < \Delta T \end{cases} \quad i = 1, ..., n, \\ \mathcal{H}^*(t^i) &= \begin{cases} 0 & \text{if } t^i - t^{i-1} \geq \Delta T \\ 1 & \text{if } t^i - t^{i-1} < \Delta T \end{cases} \quad i = 1, ..., n \end{aligned} \tag{2.6}$$

and

$$\begin{aligned}\widetilde{\mathcal{H}}(c_i) &= \begin{cases} 0 & \text{if } c_i > 0 \\ 1 & \text{if } c_i = 0 \end{cases} \quad i = 1, ..., n-1, \\ \widetilde{\mathcal{H}}^*(c_i) &= \begin{cases} 1 & \text{if } c_i > 0 \\ 0 & \text{if } c_i = 0 \end{cases} \quad i = 1, ..., n-1. \end{aligned} \tag{2.7}$$

The measures $\widetilde{P}_{i,j}$, $i = 1, ..., n$, $j = 1, 2$ are defined in the same way as above by replacing $\mathcal{H}(t^i)$ with $\mathcal{H}(t^i)\widetilde{\mathcal{H}}(c_i)$. One further has to exchange $\mathcal{H}^*(t^i)$:

$$\mathcal{H}^*(t^i) \to \left(\mathcal{H}(t^i)\widetilde{\mathcal{H}}^*(c_i) + \mathcal{H}^*(t^i)\widetilde{\mathcal{H}}(c_i) + \mathcal{H}^*(t^i)\widetilde{\mathcal{H}}^*(c_i)\right).$$

Inserting the two probability models into the KL divergence (Definition 3), i.e. using $\lambda_i := \widetilde{P}_{i,1}$, $i = 1, ..., n$ and the additivity of the KL divergence for independent events one gets the following expression

$$\begin{aligned}\mathcal{I}(2:1, \mathcal{O}_1; \xi, \theta_1) = \sum_{i=1}^{n} \Bigg[\int \mathcal{H}(t^i)\widetilde{\mathcal{H}}(c_i) f_2(y_{t^i}) \log \left\{ \frac{\mathcal{H}(t^i)\widetilde{\mathcal{H}}(c_i) f_2(y_{t^i})}{\mathcal{H}(t^i)\widetilde{\mathcal{H}}(c_i) f_1(y_{t^i})} \right\} \mathrm{d}y_{t^i} + \\ \left(\mathcal{H}(t^i)\widetilde{\mathcal{H}}^*(c_i) + \mathcal{H}^*(t^i)\widetilde{\mathcal{H}}(c_i) + \mathcal{H}^*(t^i)\widetilde{\mathcal{H}}^*(c_i)\right) \cdot \\ \log \left\{ \frac{\mathcal{H}(t^i)\widetilde{\mathcal{H}}^*(c_i) + \mathcal{H}^*(t^i)\widetilde{\mathcal{H}}(c_i) + \mathcal{H}^*(t^i)\widetilde{\mathcal{H}}^*(c_i)}{\mathcal{H}(t^i)\widetilde{\mathcal{H}}^*(c_i) + \mathcal{H}^*(t^i)\widetilde{\mathcal{H}}(c_i) + \mathcal{H}^*(t^i)\widetilde{\mathcal{H}}^*(c_i)} \right\} \Bigg],\end{aligned}$$

where $c_n := 0$. With $\log(1) = 0$ this simplifies to

$$\mathcal{I}(2:1, \mathcal{O}_1; \xi, \theta_1) = \sum_{i=1}^{n} \mathcal{H}(t^i)\widetilde{\mathcal{H}}(c_i) \int f_2(y_{t^i}) \cdot \log \left\{ \frac{f_2(y_{t^i})}{f_1(y_{t^i})} \right\} \mathrm{d}y_{t^i}. \tag{2.8}$$

By inserting the normal distribution (2.5) in (2.8) one gets

$$\mathcal{I}(2:1, \mathcal{O}_1; \xi, \theta_1) = \sum_{i=1}^{n} \mathcal{H}(t^i)\widetilde{\mathcal{H}}(c_i) \int f_2(y_{t^i}) \log \left\{ \frac{\frac{1}{\sqrt{2\pi}|v_2^i|} e^{-\frac{1}{2}(y_{2,i} - y_{t^i})^T V_{2,i} (y_{2,i} - y_{t^i})}}{\frac{1}{\sqrt{2\pi}|v_1^i|} e^{-\frac{1}{2}(y_{1,i} - y_{t^i})^T V_{1,i} (y_{1,i} - y_{t^i})}} \right\} \mathrm{d}y_{t^i}.$$

In the next step we obtain

$$\mathcal{I}(2:1,\mathcal{O}_1;\xi,\theta_1) = \sum_{i=1}^{n} \mathcal{H}(t^i)\widetilde{\mathcal{H}}(c_i) \left(\sum_{k=1}^{m} \log\left(\frac{v_1^k(y_{1,i},t^i,\theta_1)}{v_2^k(y_{2,i},t^i,\theta_2)} \right) + A^k \right) \tag{2.9}$$

with

$$A^k = \frac{1}{2} \int f_2^k(y_{t^i}) \left[-\frac{1}{\left(v_2^k(y_{2,i},t^i,\theta_2)\right)^2} \left(\left(y_{2,i}^k\right)^2 - 2y_{2,i}^k y_{t^i}^k + \left(y_{t^i}^k\right)^2 \right) + \frac{1}{\left(v_1^k(y_{1,i},t^i,\theta_1)\right)^2} \left(\left(y_{1,i}^k\right)^2 - 2y_{1,i}^k y_{t^i}^k + \left(y_{t^i}^k\right)^2 \right) \right] \mathrm{d}y_{t^i}^k.$$

A^k reduces using the well known moments of the normal distribution to

$$A^k = \frac{1}{2} \left[-\frac{1}{\left(v_2^k(y_{2,i},t^i,\theta_2)\right)^2} \left(\left(y_{2,i}^k\right)^2 - 2\left(y_{2,i}^k\right)^2 + \left(y_{2,i}^k\right)^2 + \left(v_2^k(y_{2,i},t^i,\theta_2)\right)^2 \right) + \frac{1}{\left(v_1^k(y_{1,i},t^i,\theta_1)\right)^2} \left(\left(y_{1,i}^k\right)^2 - 2y_{1,i}^k y_{2,i}^k + \left(y_{2,i}^k\right)^2 + \left(v_2^k(y_{2,i},t^i,\theta_2)\right)^2 \right) \right].$$

This further simplifies to

$$A^k = \frac{1}{2} \left[-1 + \frac{\left(\left(y_{1,i}^k\right)^2 - 2y_{1,i}^k y_{2,i}^k + \left(y_{2,i}^k\right)^2 + \left(v_2^k(y_{2,i},t^i,\theta_2)\right)^2 \right)}{\left(v_1^k(y_{1,i},t^i,\theta_1)\right)^2} \right].$$

Substituting A^k back into (2.9) we get

$$\mathcal{I}(2:1,\mathcal{O}_1;\xi,\theta_1) = \sum_{i=1}^{n} \mathcal{H}(t^i)\widetilde{\mathcal{H}}(c_i) \left(\sum_{k=1}^{m} \log\left(\frac{v_1^k(y_{1,i},t^i,\theta_2)}{v_2^k(y_{2,i},t^i,\theta_1)} \right) + \frac{1}{2} \left[-1 + \frac{\left(y_{1,i}^k - y_{2,i}^k\right)^2 + \left(v_2^k(y_{2,i},t^i,\theta_2)\right)^2}{\left(v_1^k(y_{1,i},t^i,\theta_1)\right)^2} \right] \right).$$

This reduces to

$$\mathcal{I}(2:1,\mathcal{O}_1;\xi,\theta_1) = \frac{1}{2} \sum_{i=1}^{n} \mathcal{H}(t^i)\widetilde{\mathcal{H}}(c_i) \cdot \left(\sum_{k=1}^{m} \left[\frac{\left(v_2^k(y_{2,i},t^i,\theta_2)\right)^2 + \left(y_{2,i}^k - y_{1,i}^k\right)^2}{\left(v_1^k(y_{1,i},t^i,\theta_1)\right)^2} - 2\log\left(\frac{v_2^k(y_{2,i},t^i,\theta_2)}{v_1^k(y_{1,i},t^i,\theta_1)} \right) \right] - m \right). \tag{2.10}$$

This criterion has to be maximized with respect to the initial concentrations $y_{\mathrm{I}} := y(t_0)$, the measurement time points t and the system perturbations c, thus $\xi := (y_{\mathrm{I}}, t, c) \in \Xi \subset \mathbb{R}^d$.

For our optimal experimental design we generally start with a large number of measurement time points. By use of the Heaviside-functions the number of measurement time points gets reduced in the sense that for $t^i - t^{i-1} < \Delta T$ the corresponding measurement time point is "turned off".

Remark. *The Heaviside-functions can be replaced by any appropriate switching functions*

$$\mathcal{H}'(t) + \mathcal{H}'^*(t) \equiv 1 \quad \text{with} \quad \mathcal{H}'(t), \mathcal{H}'^*(t) \in [0,1]$$

and

$$\widetilde{\mathcal{H}}'(c) + \widetilde{\mathcal{H}}'^*(c) \equiv 1 \quad \text{with} \quad \widetilde{\mathcal{H}}'(c), \widetilde{\mathcal{H}}'^*(c) \in [0,1],$$

e.g. continuously differentiable functions.

CHAPTER 3

Theory of the solution of minMax optimization problems

In this chapter we present first order optimality conditions for semi-infinite inequality and finite equality constrained optimization problems (**SIECP**) and an algorithmic scheme to find solution points which satisfy first order optimality conditions. Several methods to solve such **SIECP** are available, an overview can be found in [60, 91]. We choose the method of *Outer Approximations* [95, 103, 91], whose origin can be traced back to cutting plane methods for convex problems [91]. This approach is beneficial in the presence of a complex inner problem. The *Outer Approximations* algorithm solves iteratively discretized finite counterparts of the semi-infinite problem in each step refining the discretization until a sufficient approximation of the original problem is reached. The relation between the semi-infinite problem and an infinite sequence of finite problems can be formalized in the theory of consistent approximations and epi-convergence [88, 89, 90, 91].

In the following, we first present a first order optimality condition for finite inequality and equality constrained optimization problems (**IECP**). Thereafter, we present the concept of a continuous optimality function, which gives a measure of the degree of optimality. At a point where the first order optimality condition is fullfilled the value of the optimality function is zero. Hereafter we present first order optimalty conditions for **SIECP** and a related optimality function. At the end of the chapter we derive the *Outer Approximations* scheme utilizing the concept of optimality functions for **IECP** and **SIECP**.

We stay close to the presentation in [91], but it should be noted that the presentation is partially extended, i.e. in Section 3.1 Corollaries 1 and 2, as well as the proofs of

Proposition 2 and Theorem 8 are not contained in the presentation in [91]. In Section 3.2 Corollary 4 and the proofs of the presented results are not contained in [91]. The proof of Theorem 14 in Section 3.3, the main result of this chapter is not included in [91], as well. To be consistent with [91], we drop the meaning of the indices as used in Chapter 2 and use the same notation as in [91]. For preliminary definitions and results refer to Appendix A.

3.1. Finite inequality and equality constrained optimization problem

Definition 6 (Definition (0c), page 167 in [91]). *The problem*

$$\min \{f^0(x) | f^j(x) \leq 0,\ j \in \mathbf{q},\ g^k(x) = 0,\ k \in \mathbf{r}\}, \tag{3.1}$$

where the constraint functions $f^j : \mathbb{R}^n \to \mathbb{R},\ j \in \mathbf{q} := \{1, \ldots, q\}$, *and* $g^k : \mathbb{R}^n \to \mathbb{R},\ k \in \mathbf{r} := \{1, \ldots, r\}$ *are continuously differentiable, while the cost function* $f^0 : \mathbb{R}^n \to \mathbb{R}$ *can be either continuously differentiable on* $\mathbb{R}^n$ *or a max function of the form*

$$f^0(x) := \max_{k \in \mathbf{p}} c^k(x),$$

with the $c^k : \mathbb{R}^n \to \mathbb{R},\ k \in \mathbf{p} := \{1, \ldots, p\}$ *continuously differentiable is called a finite inequality and equality constrained optimization problem (***IECP***).*

Definition 7 (Definition 2.2.15 in [91]). *Let* $X_{IE} := \{x \in \mathbb{R}^n | \psi(x) \leq 0,\ g(x) = 0\}$. *We will say that* $\hat{x} \in X_{IE}$ *is a local minimizer for* **IECP** *if there exists a* $\rho > 0$ *such that* $f^0(x) \geq f^0(\hat{x})$ *for all* $x \in X_{IE} \cap B(\hat{x}, \rho)$ *with*

$$\psi(x) := \max_{j \in \mathbf{q}} f^j(x)$$

and

$$g(x) := (g^1(x), g^2(x), \ldots, g^r(x)).$$

If $f^0(x) > f^0(\hat{x})$ *for all* $x \in X_{IE} \cap B(\hat{x}, \rho)$, $x \neq \hat{x}$, $\hat{x}$ *is called a strict local minimizer.*

3.1.1. First order optimality conditions for IECP

Theorem 4 (Theorem 2.2.16 in [91]). *(a) Suppose* $\hat{x}$ *is a local minimizer for the problem* **IECP**, *then* $\hat{x}$ *is a local minimizer for the problem*

$$\min \left\{ \widehat{F}(x) | g(x) = 0 \right\}, \tag{3.2}$$

where

$$\widehat{F}(x) := \max\left\{f^0(x) - f^0(\hat{x}), \psi(x)\right\} = \max\left\{\max_{k\in\mathbf{p}}[c^k(x) - f^0(\hat{x})], \max_{j\in\mathbf{q}} f^j(x)\right\}.$$

(b) suppose that, for any $x \in \mathbb{R}^n$ such that both $\psi(x) = 0$ and $g(x) = 0$, there exists a sequence of vectors $\{x_i\}_{i=0}^{\infty}$ converging to x, such that $\psi(x_i) < 0$ and $g(x_i) = 0$ for all $i \in \mathbb{N}$. If $\hat{x}$ is a local minimizer for (3.2) such that $\psi(\hat{x}) \leq 0$, then $\hat{x}$ is a local minimizer for (3.1).

Proof (Modification of the proofs of Theorem 2.2.2 and Theorem 2.2.3 in [91]).
(a) First, since $\psi(\hat{x}) \leq 0$, by assumption, it holds that $\widehat{F}(\hat{x}) = 0$. Next, let $\hat{\rho} > 0$ be the radius associated with $\hat{x}$ (see Definition 7). Then, for all $x \in B(\hat{x}, \hat{\rho}) \cap X_E$ with

$$X_E := \{x \in \mathbb{R}^n | g(x) = 0\},$$

$\widehat{F}(x) \geq 0$, if $\psi(x) \geq 0$, and $f^0(x) - f^0(\hat{x}) \geq 0$, if $\psi(x) \leq 0$, the latter also implies that $\widehat{F}(x) \geq 0$. Hence $\hat{x}$ is a local minimizer for (3.2).
(b) First suppose that $\psi(\hat{x}) < 0$. Since $\psi(\cdot)$ is continuous, and $\widehat{F}(\hat{x}) = 0$, and $\hat{x}$ is a local minimizer of (3.2), there exists a $\hat{\rho} > 0$ such that, for all $x \in B(\hat{x}, \hat{\rho}) \cap X_E$, $\widehat{F}(x) \geq 0$ and $\psi(x) < 0$. It now follows by inspection that, for all $x \in B(\hat{x}, \hat{\rho}) \cap X_E$, $f^0(x) - f^0(\hat{x}) \geq 0$, which shows that $\hat{x}$ is a local minimizer for **IECP**.
Next suppose that $\psi(\hat{x}) = 0$. Since $\hat{x}$ is a local minimizer of (3.2), it follows that there exists a $\hat{\rho} > 0$ such that for all $x \in B(\hat{x}, \hat{\rho}) \cap X_E$, $\widehat{F}(x) \geq 0$. Now, if $x \in B(\hat{x}, \hat{\rho}) \cap X_E$ is such that $\psi(x) < 0$, then $\widehat{F}(x) \geq 0$ implies that $f^0(x) - f^0(\hat{x}) \geq 0$. If $x \in B(\hat{x}, \hat{\rho}) \cap X_E$ is such that $\psi(x) = 0$ then by assumption there exists a sequence of vectors $\{x_i\}_{i=0}^{\infty}$ converging to x, such that $\psi(x_i) < 0$ and $g(x_i) = 0$ for all $i \in \mathbb{N}$. Consequently there exist $i_0 \in \mathbb{N}$ so that for all $i > i_0$, $x_i \in B(\hat{x}, \hat{\rho}) \cap X_E$ with $\psi(x_i) < 0$, which implies that $\widehat{F}(x_i) = f^0(x_i) - f^0(\hat{x}) \geq 0$ for all $i > i_0$. It now follows from the continuity of $f^0(\cdot)$ that $f^0(x) - f^0(\hat{x}) \geq 0$. Hence we conclude that $\hat{x}$ is a local minimizer for **IECP**. □

Proposition 1 (Proposition 2.2.18 in [91])**.** *Suppose that C is a convex, compact set in $\mathbb{R}^n$ and that H is a subspace of $\mathbb{R}^n$. Then*

$$\max_{\xi\in C}\langle \xi, h\rangle \geq 0, \forall h \in H \tag{3.3}$$

if and only if

$$C \cap H^{\perp} \neq \emptyset, \tag{3.4}$$

where $H^{\perp}$ is the orthogonal complement of H in $\mathbb{R}^n$.

Proof (The proof is taken from [91]). "⇒" We give a proof by contraposition. Suppose

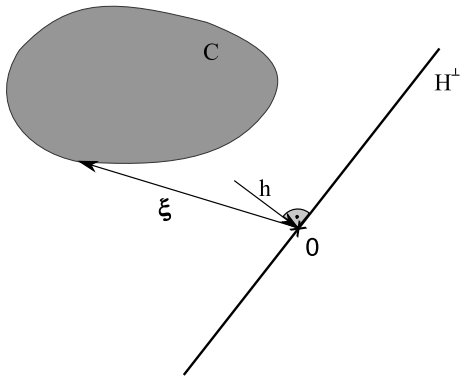


Figure 3.1.: Visualization for the proof of Proposition 1.

$C \cap H^{\perp} = \emptyset$. Then C can be separated strictly from $H^{\perp}$ (see Theorem 19), i.e., there exists a $h \neq 0$ such that $\langle \nu, h \rangle = 0$ for all $\nu \in H^{\perp}$, and $\langle \xi, h \rangle < 0$ for all $\xi \in C$, which shows that (3.3) does not hold. See Figure 3.1.
"$\Leftarrow$" Suppose that there is a vector $g \in C \cap H^{\perp}$. Then for any $h \in H$,

$$\max_{\xi \in C} \langle \xi, h \rangle \geq \langle g, h \rangle = 0,$$

i.e., (3.3) holds. □

Theorem 5 (Theorem 2.2.19 in [91])**.** *Consider problem* **IECP**. *Suppose that the functions* $c^k : \mathbb{R}^n \to \mathbb{R}, k \in \mathbf{p}$ *and the functions* $f^j, g^l : \mathbb{R}^n \to \mathbb{R}, j \in \mathbf{q}, l \in \mathbf{r}$ *are at least once continuously differentiable. If* $\hat{x}$ *is a local minimizer for* **IECP** *and the vectors* $\nabla g^l(\hat{x}), l \in \mathbf{r}$, *are linearly independent, then*
(a)

$$\mathrm{d}\widehat{F}(\hat{x}; h) \geq 0, \forall h \in \mathcal{H}_E(\hat{x}), \tag{3.5}$$

where

$$\mathcal{H}_E(\hat{x}) := \{h \in \mathbb{R}^n | g_x(\hat{x})h = 0\}, \tag{3.6}$$

and $\mathrm{d}\widehat{F}(\hat{x}; h)$ *is the directional derivative (see Theorem 24) of function* $\widehat{F}(\cdot)$ *at* $\hat{x} \in \mathbb{R}^n$ *in direction* $h \in \mathbb{R}^n$ *as given in Definition 24;*
(b) there exist multipliers $\hat{\mu} \in \Sigma_q^0 := \{(\mu^0, \mu) | \mu^0 \in \mathbb{R}_+, \mu \in \mathbb{R}_+^q, \sum_{j=0}^q \mu^j = 1\}$, $\hat{\nu} \in \Sigma_p := \{\nu | \nu \in \mathbb{R}_+^p, \sum_{j=0}^p \nu^j = 1\}$, *and* $\hat{\zeta} \in \mathbb{R}^r$ *such that*

$$\hat{\mu}^0 \left[\sum_{k=1}^{p} \hat{\nu}^k \nabla c^k(\hat{x}) \right] + \sum_{j=1}^{q} \hat{\mu}^j \nabla f^j(\hat{x}) + \sum_{l=1}^{r} \hat{\zeta}^l \nabla g^l(\hat{x}) = 0, \tag{3.7}$$

and

$$\sum_{k=1}^{p} \hat{\nu}^k \left[c^k(\hat{x}) - f^0(\hat{x}) \right] + \sum_{j=1}^{q} \hat{\mu}^j f^j(\hat{x}) = 0. \tag{3.8}$$

Proof (The proof is taken from [91]). (a) Since $\hat{x}$ is a local minimizer for **IECP**, it must also be a local minimizer for the problem (3.2). Hence, for the sake of contradiction, suppose that there is a vector $h \in \mathcal{H}_E(\hat{x})$ such that $\mathrm{d}\widehat{F}(\hat{x}; h) < 0$. Since the matrix $g_x(\hat{x})$ has maximum row rank, it follows from Corollary 6 that there exists a $t_h > 0$ and continuously differentiable function $s : [0, t_h] \to \mathbb{R}^n$ such that $s(0) = \hat{x}$, $\dot{s}(0) = h$, and $g(s(t)) = 0$ for all $t \in [0, t_h]$. Let $\sigma : [0, t_h] \to \mathbb{R}$ be defined by $\sigma(t) = \widehat{F}(s(t))$. Then by the Chain Rule Theorem (Theorem 27), the directional derivative $\mathrm{d}\sigma(0; 1) = \mathrm{d}\widehat{F}(\hat{x}; h) < 0$, and hence there exists a $t' \in (0, t_h]$ such that $\sigma(t) < \sigma(0)$ for all $t \in (0, t')$. Consequently, for any $t \in (0, t'), g(s(t)) = 0$ and $\widehat{F}(s(t)) < \widehat{F}(\hat{x}) = 0$, which contradicts the fact that $\hat{x}$ is a local minimizer for the problem (3.2) and hence for the problem **IECP**.
(b) Now, by Theorem 24, (a) for any $h \in \mathbb{R}^n$,

$$\mathrm{d}\widehat{F}(\hat{x}; h) = \max_{\xi \in \partial \widehat{F}(\hat{x})} \langle \xi, h \rangle,$$

and (c),

$$\partial \widehat{F}(\hat{x}) = \mathrm{conv} \left(\bigcup_{\substack{j \in \mathbf{q_A}(\hat{x}) \\ k \in \hat{\mathbf{p}}(\hat{x})}} \left\{ \nabla f^j(\hat{x}), \nabla c^k(\hat{x}) \right\} \right), \tag{3.9}$$

with

$$\hat{\mathbf{p}}(x) := \left\{ k \in \mathbf{p} | c^k(x) = f^0(x) \right\}$$

and

$$\mathbf{q_A}(x) := \left\{ j \in \mathbf{q} | f^j(x) \geq 0 \right\}.$$

It now follows from (3.5), (3.6) and Proposition 1 that

$$\partial \widehat{F}(\hat{x}) \cap \mathcal{H}_E^{\perp}(\hat{x}) \neq \emptyset.$$

Since the vectors $\nabla g^l(\hat{x}), l \in \mathbf{r}$, form a basis for $\mathcal{H}_E^{\perp}(\hat{x})$, it follows that there exists $\hat{\mu} \in \Sigma_q^0$, $\hat{\nu} \in \Sigma_p$ and $\hat{\zeta} \in \mathbb{R}^r$ such that (3.7) and (3.8) hold with (3.8) ensuring that only the vectors appearing in the expression (3.9) for $\partial \widehat{F}(\hat{x})$ have nonzero coefficients in (3.7) since $[c^k(\hat{x}) - f^0(\hat{x})] \leq 0$ and $f^j(\hat{x}) \leq 0$. □

For later purpose, we now introduce the so called Mangasarian-Fromowitz constraint qualification ($MFCQ$).

Definition 8. *Consider problem* **IECP**. *We say that the Mangasarian-Fromowitz constraint qualification holds at $x \in \mathbb{R}^n$, if the vectors $\nabla g^l(x), l \in \mathbf{r}$, are linearly independent, and there exists $\tilde{h} \in \mathbb{R}^n$ such that (strictly feasible direction),*

$$\nabla g^l(x)^T \tilde{h} = 0, \quad l \in \mathbf{r},$$

and

$$\nabla f^j(x)^T \tilde{h} < 0, \quad j \in \mathbf{q_A}(x).$$

Corollary 1. *If $\hat{x}$ is a local minimizer for problem (3.2) with $\psi(\hat{x}) \leq 0$ and $MFCQ$ is satisfied at all points $x \in \mathbb{R}^n$ with $\psi(x) = 0$ and $g(x) = 0$, then $\hat{x}$ is a local minimizer for problem* **IECP**.

Proof. It is sufficient to show (Theorem 4(b)), that if $x \in \mathbb{R}^n$ with $\psi(x) = 0$ and $g(x) = 0$ there exists a sequence of vectors $\{x_i\}_{i=0}^{\infty}$ converging to x, such that $\psi(x_i) < 0$ and $g(x_i) = 0$ for all $i \in \mathbb{N}$. Be $x \in \mathbb{R}^n$ with $\psi(x) = 0$ and $g(x) = 0$. By assumption x satisfies $MFCQ$ and thus there is a direction $\tilde{h}$ such that $g_x(x)^T \tilde{h} = 0$ and $\mathrm{d}\psi(x; \tilde{h}) < 0$. It follows from Corollary 6 that there exists a $t_{\tilde{h}} > 0$ and continuously differentiable function $s : [0, t_{\tilde{h}}] \to \mathbb{R}^n$ such that $s(0) = x$, $\dot{s}(0) = \tilde{h}$, and $g(s(t)) = 0$ for all $t \in [0, t_{\tilde{h}}]$. Let $\sigma : [0, t_{\tilde{h}}] \to \mathbb{R}$ be defined by $\sigma(t) = \psi(s(t))$. Then by the Chain Rule Theorem (Theorem 27), the directional derivative $\mathrm{d}\sigma(0; 1) = \mathrm{d}\psi(x; \tilde{h}) < 0$, and hence there exists a $t' \in (0, t_{\tilde{h}}]$ such that $\sigma(t) < \sigma(0)$ for all $t \in (0, t')$. Consequently, for any $t \in (0, t'), g(s(t)) = 0$ and $\psi(s(t)) < \psi(x) = 0$ and thus there exists a sequence of vectors $\{x_i\}_{i=0}^{\infty}$ converging to x, such that $\psi(x_i) < 0$ and $g(x_i) = 0$ for all $i \in \mathbb{N}$. □

Corollary 2. *Consider problem* **IECP**. *Suppose that the functions $c^k : \mathbb{R}^n \to \mathbb{R}, k \in \mathbf{p}$ and the functions $f^j, g^l : \mathbb{R}^n \to \mathbb{R}, j \in \mathbf{q}, l \in \mathbf{r}$ are at least once continuously differentiable. If $\hat{x}$ is a local minimizer for* **IECP** *and $MFCQ$ is satisfied at $\hat{x}$ there exist multipliers $\hat{\mu} \in \mathbb{R}_+^q$, $\hat{\nu} \in \Sigma_p := \{\nu | \nu \in \mathbb{R}_+^p, \sum_{j=0}^p \nu^j = 1\}$, and $\hat{\zeta} \in \mathbb{R}^r$ such that*

$$\sum_{k=1}^{p} \hat{\nu}^k \nabla c^k(\hat{x}) + \sum_{j=1}^{q} \hat{\mu}^j \nabla f^j(\hat{x}) + \sum_{l=1}^{r} \hat{\zeta}^l \nabla g^l(\hat{x}) = 0,$$

and

$$\sum_{k=1}^{p} \hat{\nu}^k \left[c^k(\hat{x}) - f^0(\hat{x}) \right] + \sum_{j=1}^{q} \hat{\mu}^j f^j(\hat{x}) = 0.$$

Proof. It is sufficient to show that $\hat{\mu}^0 > 0$ in (3.7). We proof by contraposition. Suppose $\hat{\mu}^0 = 0$, then there exist multipliers $\hat{\mu} \in \Sigma_q := \{\mu | \mu \in \mathbb{R}^q_+, \sum_{j=0}^q \mu^j = 1\}$, and $\hat{\zeta} \in \mathbb{R}^r$ such that

$$\sum_{j=1}^{q} \hat{\mu}^j \nabla f^j(\hat{x}) + \sum_{l=1}^{r} \hat{\zeta}^l \nabla g^l(\hat{x}) = 0.$$

Since $\hat{\mu} \in \Sigma_q$ and $MFCQ$ holds at $\hat{x}$ there exists $\tilde{h} \in \mathbb{R}^n$ such that for all $j \in \mathbf{q}$ with $\hat{\mu}^j > 0$, $\hat{\mu}^j \nabla f^j(\hat{x})^T \tilde{h} < 0$ and $\hat{\zeta}^l \nabla g^l(\hat{x})^T \tilde{h} = 0,\ l \in \mathbf{r}$. It follows

$$0 > \left(\sum_{j=1}^{q} \hat{\mu}^j \nabla f^j(\hat{x}) + \sum_{l=1}^{r} \hat{\zeta}^l \nabla g^l(\hat{x}) \right)^T \tilde{h} = 0^T \tilde{h} = 0,$$

which is a contradiction and therefore $\hat{\mu}^0 > 0$. □

Remark. *In the special case* $\mathbf{p} = \{1\}$ *Corollary 2 restates the Karush–Kuhn–Tucker (KKT) conditions [83].*

3.1.2. Optimality function for IECP

Proposition 2. *Suppose that* $g : \mathbb{R}^n \to \mathbb{R}^r$*, with* $r < n$*, is continuously differentiable and that the matrix* $g_x(x)$ *has maximum row rank for all* $x \in \mathbb{R}^n$*. Let the set valued function* $\mathcal{H}_E(\cdot)$ *be defined by*

$$\mathcal{H}_E(x) := \left\{ h \in \mathbb{R}^n | g_x(x)h = 0 \right\}.$$

$\mathcal{H}_E(\cdot)$ *is outer semicontinuous (Definition 32) and inner semicontinuous (Definition 33) and hence continuous.*

Proof. First we show that $\mathcal{H}_E(\cdot)$ is *outer semicontinuous* using Theorem 21(a).
Let $\hat{x} \in \mathbb{R}^n$ and let $\{x_i\}_{i=0}^{\infty}$ be a sequence such that $x_i \to \hat{x}$, as $i \to \infty$. Let $\hat{h}$ be a cluster point (Definition 35) of $\{\mathcal{H}_E(x_i)\}_{i=0}^{\infty}$. Since $\hat{h}$ is a cluster point of $\{\mathcal{H}_E(x_i)\}_{i=0}^{\infty}$ it follows that there exists a $\{h_i\}_{i=0}^{\infty}$, $h_i \in \mathcal{H}_E(x_i)$ with $h_i \to^K \hat{h}$, for a infinite subset $K \subset \mathbb{N}$. Since $g_x(x)h$ is continuous in x and h and $g_x(x_i)h_i = 0$ it follows that

$$\lim_{i \to^K \infty} g_x(x_i)h_i = g_x(\hat{x})\hat{h} = 0,$$

and hence $\overline{\text{Lim}}\mathcal{H}_E(x_i) \subset \mathcal{H}_E(\hat{x})$.
Next we show that $\mathcal{H}_E(\cdot)$ is *inner semicontinuous* using Theorem 21(c).
Suppose that $\hat{x} \in \mathbb{R}^n$ and $\hat{h} \in \mathcal{H}_E(\hat{x})$ are given and that $\{x_i\}_{i=0}^{\infty}$ is a sequence in $\mathbb{R}^n$ converging to $\hat{x}$. Clearly, we only need to show that there exist vectors $h_i \in \mathcal{H}_E(x_i)$ such that $h_i \to \hat{h}$, as $i \to \infty$. Since the matrix $g_x(\cdot)$ has maximum row rank, there

exists a continuous, $(n-r)\times n$, matrix-valued function $W(x)$ (e.g. whose rows form an orthogonal basis for the null space of $g_x(x)$) such that the matrix

$$A(x) = \left[\frac{W(x)}{g_x(x)}\right]$$

is continuous and nonsingular. For an $x \in \mathbb{R}^n$, let

$$\tilde{h}(x) = A^{-1}(x)A(\hat{x})\hat{h}.$$

$\tilde{h}(\cdot)$ is continuous and well defined so that obviously $\tilde{h}(x_i) \in \mathcal{H}_E(x_i)$ and

$$\lim_{i\to\infty} \tilde{h}(x_i) \to \hat{h},$$

thus $\underline{\text{Lim}}\mathcal{H}_E(x_i) \supset \mathcal{H}_E(\hat{x})$. □

Theorem 6 (Theorem 5.4.1 in [91]). *Suppose that $\phi : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is continuous, that $Y : \mathbb{R}^n \to 2^{\mathbb{R}^m}$ is outer semicontinuous and that the function $\overline{\psi} : \mathbb{R}^n \to \mathbb{R}$ defined by*

$$\overline{\psi}(x) := \max_{y\in Y(x)} \phi(x,y) \tag{3.10}$$

is well defined for all $x \in \mathbb{R}^n$. If, for every bounded set $X \subset \mathbb{R}^n$, there exists an $\alpha < \infty$ such that for all $x \in X$,

$$\| \arg\max_{y\in Y(x)} \phi(x,y)\| \le \alpha, \tag{3.11}$$

then $\overline{\psi}(\cdot)$ is upper semicontinuous.

Proof (The proof is taken from [91]). Let $\hat{x} \in \mathbb{R}^n$ be given. Let $\{x_i\}_{i=0}^{\infty}$ be an arbitrary sequence such that $x_i \to \hat{x}$, as $i \to \infty$, and let $y_i \in Y(x_i)$ be such that $\overline{\psi}(x_i) = \phi(x_i, y_i)$ for $i \in \mathbb{N}$. Since the sequence $\{x_i\}_{i=0}^{\infty}$ is bounded, it follows from our hypotheses that there exists an $\alpha < \infty$ such that $\|y_i\| \le \alpha$ for all $i \in \mathbb{N}$, and hence, since $\phi(\cdot,\cdot)$ is continuous, $\overline{\lim}\,\phi(x_i, y_i)$ exists. Suppose that $y_i, i \in K \subset \mathbb{N}$ are such that $\overline{\lim}\,\phi(x_i, y_i) = \lim_{i\to^K\infty} \phi(x_i, y_i)$ and $y_i \to^K y^*$, as $i \to \infty$. Then $y^* \in Y(\hat{x})$ because $Y(\cdot)$ is *outer semicontinuous* and hence

$$\overline{\psi}(\hat{x}) \ge \phi(\hat{x}, y^*) = \lim_{i\in K} \phi(x_i, y_i) = \overline{\lim}\,\psi(x_i).$$

□

Proposition 3 (Corollary 5.4.2 in [91]). *Suppose that $\phi : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is continuous, that $Y : \mathbb{R}^n \to 2^{\mathbb{R}^m}$ is continuous, and that the function $\overline{\psi} : \mathbb{R}^n \to \mathbb{R}$, defined by (3.10),*

is well defined for all $x \in \mathbb{R}^n$. If, for every bounded set $X \subset \mathbb{R}^n$, there exists an $\alpha < \infty$ such that (3.11) holds for all $x \in X$, then $\overline{\psi}(\cdot)$ as defined by (3.10) is continuous.

Proof (The proof is taken from [91]). Since by Theorem 6, $\overline{\psi}(\cdot)$ is *upper semicontinuous*, we only need to show that it is *lower semicontinuous* under the stronger assumption on $Y(\cdot)$. For the sake of contradiction, suppose that there is a point $\hat{x} \in \mathbb{R}^n$ and a sequence $x_i \to \hat{x}$, as $i \to \infty$ such that $\lim_{i\to\infty} \overline{\psi}(x_i)$ exists and

$$\lim \overline{\psi}(x_i) < \overline{\psi}(\hat{x}).$$

Suppose that $\overline{\psi}(\hat{x}) = \phi(\hat{x}, \hat{y})$ with $\hat{y} \in Y(\hat{x})$. Then, since $Y(\cdot)$ is continuous, there exist $y_i \in Y(x_i)$, $i \in \mathbb{N}$, such that $y_i \to \hat{y}$, as $i \to \infty$. Since $\phi(\cdot, \cdot)$ is continuous, $\lim_{i\to\infty} \phi(x_i, y_i) = \phi(\hat{x}, \hat{y})$. Hence there exists an i_0 such that $\phi(x_i, y_i) \geq \overline{\psi}(x_i)$, for all $i \geq i_0$, which contradicts the definition of $\overline{\psi}(x_i)$. □

Theorem 7 (Modification of Corollary 5.5.2 in [91])**.** *Consider the problem*

$$\min\{f^0(x) | f^j(x) \leq 0, j \in \mathbf{q}, Ax - b = 0, x \in C\} \tag{3.12}$$

where the functions $f^j : \mathbb{R}^n \to \mathbb{R}$, $j \in \bar{\mathbf{q}} := \{0, 1, \ldots, q\}$, are convex (and hence continuous), A is an $r \times n$ $(r < n)$ matrix of maximum row rank, $b \in \mathbb{R}^r$, and C is a convex subset of $\mathbb{R}^n$. Let $Z := \{z' \in \mathbb{R}^r | z' = Ax - b, x \in C\}$ be a subset of $\mathbb{R}^r$. Suppose that there exists an $x_0 \in C$, such that $Ax_0 - b = 0$ and $f^j(x_0) < 0$ for all $j \in \mathbf{q}$, the minimum in (3.12) is achieved and the zero vector $0_r \in \mathbb{R}^r$ is in the interior of Z. Let $f : \mathbb{R}^n \to \mathbb{R}^q$ be defined by

$$f(x) := (f^1(x), f^2(x), \ldots, f^q(x)), \tag{3.13}$$

let

$$\nu_p := \inf_{x\in C} \sup_{\substack{\mu\in\mathbb{R}^q_+ \\ \zeta\in\mathbb{R}^r}} f^0(x) + \langle \mu, f(x)\rangle + \langle \zeta, Ax - b\rangle \tag{3.14}$$

and let

$$\nu_d := \sup_{\substack{\mu\in\mathbb{R}^q_+ \\ \zeta\in\mathbb{R}^r}} \inf_{x\in C} f^0(x) + \langle \mu, f(x)\rangle + \langle \zeta, Ax - b\rangle$$

Then
(a)

$$\nu_p = \min\left\{f^0(x) | f^j(x) \leq 0, j \in \mathbf{q}, Ax - b = 0, x \in C\right\} \tag{3.15}$$

(b) $\nu_p = \nu_d$, and

(c) there exist $\hat{x} \in C, \hat{\mu} \in \mathbb{R}^q_+$, *and* $\hat{\zeta} \in \mathbb{R}^r$, *such that*

$$\begin{aligned} \nu_p &= \min_{x \in C} f^0(x) + \langle \hat{\mu}, f(x) \rangle + \langle \hat{\zeta}, Ax - b \rangle \\ &= \max_{\substack{\mu \in \mathbb{R}^q_+ \\ \zeta \in \mathbb{R}^r}} f^0(\hat{x}) + \langle \mu, f(\hat{x}) \rangle + \langle \zeta, A\hat{x} - b \rangle = \nu_d \end{aligned} \tag{3.16}$$

and

$$\langle \hat{\mu}, f(\hat{x}) \rangle = \langle \hat{\zeta}, A\hat{x} - b \rangle = 0. \tag{3.17}$$

Proof (Modification of the proof of Theorem 5.5.1 in [91]).
First, since, for any $x \in C$ such that $f^j(x) > 0$ for some $j \in \mathbf{q}$ or $Ax - b \neq 0$,

$$\sup_{\substack{\mu \in \mathbb{R}^q_+ \\ \zeta \in \mathbb{R}^r}} \langle \mu, f(x) \rangle + \langle \zeta, Ax - b \rangle = \infty,$$

it follows directly that the equality in (3.15) holds and that, if $\hat{x}$ is a solution of (3.12), then $\hat{x}$ is also a solution of (3.14), to which corresponds a $\hat{\mu}$ and $\hat{\zeta}$ satisfying (3.17).
For every $x \in \mathbb{R}^n$, let

$$G(x) := \{\bar{y} = (y^0, y, z) \in \mathbb{R} \times \mathbb{R}^q \times \mathbb{R}^r | y^j \geq f^j(x),\ j \in \bar{\mathbf{q}},\ z = Ax - b\}$$

and let $\mathbf{G} = G(C)$. We begin by showing (i) that $\mathbf{G}$ is convex and (ii) that

$$\nu_p = \inf_{\bar{y} \in \mathbf{G}} \sup_{\substack{\mu \in \mathbb{R}^q_+ \\ \zeta \in \mathbb{R}^r}} y^0 + \langle \mu, y \rangle + \langle \zeta, z \rangle. \tag{3.18}$$

(i) Suppose that $\bar{y}_1,\ \bar{y}_2 \in \mathbf{G}$ and $\lambda \in (0, 1)$. Then there exist $x_1, x_2 \in C$ such that $\bar{y}_1 \in G(x_1)$, $\bar{y}_2 \in G(x_2)$, and for every $j \in \bar{\mathbf{q}}$,

$$\lambda y_1^j + (1 - \lambda) y_2^j \geq \lambda f^j(x_1) + (1 - \lambda) f^j(x_2) \geq f^j(\lambda x_1 + (1 - \lambda) x_2)$$

and

$$\lambda z_1 + (1 - \lambda) z_2 = A(\lambda x_1 + (1 - \lambda) x_2) - b,$$

which shows that

$$\lambda \bar{y}_1 + (1 - \lambda) \bar{y}_2 \in G(\lambda x_1 + (1 - \lambda) x_2).$$

Since $\lambda x_1 + (1 - \lambda) x_2 \in C$, it follows that $\lambda \bar{y}_1 + (1 - \lambda) \bar{y}_2 \in \mathbf{G}$.

(ii) Let $F : \mathbb{R}^n \to \mathbb{R} \times \mathbb{R}^q \times \mathbb{R}^r$ be defined by

$$F(x) = \begin{pmatrix} f^0(x) \\ f(x) \\ Ax - b \end{pmatrix},$$

Then because $F(C) \subset \mathbf{G}$,

$$\begin{aligned} \nu_p = \inf_{x \in C} \sup_{\substack{\mu \in \mathbb{R}^q_+ \\ \zeta \in \mathbb{R}^r}} f^0(x) + \langle \mu, f(x) \rangle + \langle \zeta, Ax - b \rangle &= \inf_{\bar{y} \in F(C)} \sup_{\substack{\mu \in \mathbb{R}^q_+ \\ \zeta \in \mathbb{R}^r}} y^0 + \langle \mu, y \rangle + \langle \zeta, z \rangle \\ &\geq \inf_{\bar{y} \in \mathbf{G}} \sup_{\substack{\mu \in \mathbb{R}^q_+ \\ \zeta \in \mathbb{R}^r}} y^0 + \langle \mu, y \rangle + \langle \zeta, z \rangle. \end{aligned}$$

But, for every $\bar{y} \in \mathbf{G}$, there exists $\bar{y}' \in F(C)$ such that $y'^j \leq y^j$, for all $j \in \bar{\mathbf{q}}$ and $z' = z$. Hence (3.18) must hold.
Next for any $x' \in C$, $\mu \in \mathbb{R}^q_+$ and $\zeta \in \mathbb{R}^r$,

$$\inf_{x \in C} (f^0(x) + \langle \mu, f(x) \rangle + \langle \zeta, Ax - b \rangle) \leq f^0(x') + \langle \mu, f(x') \rangle + \langle \zeta, Ax' - b \rangle,$$

which leads to the conclusion that, for all $x' \in C$,

$$\nu_d = \sup_{\substack{\mu \in \mathbb{R}^q_+ \\ \zeta \in \mathbb{R}^r}} \inf_{x \in C} f^0(x) + \langle \mu, f(x) \rangle + \langle \zeta, Ax - b \rangle \leq \sup_{\substack{\mu \in \mathbb{R}^q_+ \\ \zeta \in \mathbb{R}^r}} f^0(x') + \langle \mu, f(x') \rangle + \langle \zeta, Ax' - b \rangle.$$

Hence

$$\nu_d \leq \inf_{x' \in C} \sup_{\substack{\mu \in \mathbb{R}^q_+ \\ \zeta \in \mathbb{R}^r}} f^0(x') + \langle \mu, f(x') \rangle + \langle \zeta, Ax' - b \rangle = \nu_p. \tag{3.19}$$

Next, we will show that the point $\bar{y}_* = (\nu^p, 0, 0) \in \mathbb{R} \times \mathbb{R}^q \times \mathbb{R}^r$ is a boundary point of the set $\mathbf{G}$, by showing that

$$\Big(\mathbf{G} - \{\bar{y}_*\}\Big) \cap \mathring{Q}_-^{r+q+1} = \emptyset, \tag{3.20}$$

where $\mathring{Q}_-^{r+q+1} := \{\bar{y} = (y^0, y, 0) \in \mathbb{R} \times \mathbb{R}^q \times \mathbb{R}^r | y^j < 0, j \in \bar{\mathbf{q}}\}$. Clearly the zero vector is an element of $\Big(\mathbf{G} - \{\bar{y}_*\}\Big) \cap Q_-^{r+q+1}$ with $Q_-^{r+q+1} := \{\bar{y} = (y^0, y, 0) \in \mathbb{R} \times \mathbb{R}^q \times \mathbb{R}^r | y^j \leq 0, j \in \bar{\mathbf{q}}\}$. Hence, if (3.20) does not hold, then there must exist a $\bar{y}_{**} \in \mathbf{G}$ such that

$$y^j_{**} - y^j_* < 0, \quad j \in \bar{\mathbf{q}},$$

and $z_{**} = 0$. Since $\bar{y}_{**} \in \mathbf{G}$, there exists an $x_{**} \in C$ such that $f^j(x_{**}) \leq y^j_{**}$, for all

$j \in \bar{\mathbf{q}}$ and $Ax_{**} - b = 0$. Since $y^j_{**} < 0$ for all $j \in \mathbf{q}, f_j(x_{**}) < 0$ for all $j \in \mathbf{q}$, and $y^0_{**} < y^0_* = \nu_p$. Since, by definition, $y^0_* = \nu_p$, we have a contradiction of the optimality of ν_p.
Since (3.20) holds and $\mathbf{G} - \{\bar{y}_*\}$ and $\mathring{Q}^{r+q+1}_-$ are convex, disjoint sets, their closures can be separated, i.e. there exists a nonzero vector $\bar{\pi} := (\pi^0, \pi, \pi_z) \in \mathbb{R} \times \mathbb{R}^q \times \mathbb{R}^r$ and an $\alpha \in \mathbb{R}$ such that

$$\langle \bar{y} - \bar{y}_*, \bar{\pi} \rangle \geq \alpha, \quad \forall \bar{y} \in \mathbf{G} \tag{3.21}$$

and

$$\langle \bar{y}, \bar{\pi} \rangle \leq \alpha, \quad \forall \bar{y} \in Q^{r+q+1}_-. \tag{3.22}$$

Because $0 \in \Big(\mathbf{G} - \{\bar{y}_*\}\Big) \cap Q^{r+q+1}_-$, it follows that $\alpha = 0$. Since the $\mathbb{R} \times \mathbb{R}^q \times \mathbb{R}^r$ unit vectors $-\bar{e}_j = (0, 0, \ldots, -1, \ldots, 0) \in Q^{r+q+1}_-$, for $j \in \bar{\mathbf{q}}$, it follows from (3.22) that

$$\langle -\bar{e}_j, \bar{\mu} \rangle = -\pi^j \leq 0, \quad \forall j \in \bar{\mathbf{q}},$$

which shows that $\pi^j \geq 0$ for all $j \in \bar{\mathbf{q}}$. We will now show that $\pi^0 > 0$. Suppose that $\pi^0 = 0$. Then (3.21) implies that

$$\sum_{j=1}^{q} \pi^j (y^j - 0) + \sum_{j=1}^{r} \pi^j_z (z^j - 0) = \sum_{j=1}^{q} \pi^j y^j + \sum_{j=1}^{r} \pi^j_z z^j \geq 0, \quad \forall \bar{y} \in \mathbf{G}. \tag{3.23}$$

By assumption, there exists an $x_0 \in C$ such that $f^j(x_0) < 0$, for all $j \in \mathbf{q}$ and $Ax_0 - b = 0$. Since $(f^0(x_0), f(x_0), 0) \in \mathbf{G}$, it follows from (3.23) that

$$\sum_{j=1}^{q} \pi^j f^j(x_0) \geq 0. \tag{3.24}$$

Since $\pi_j \geq 0$ for $j \in \bar{\mathbf{q}}$, we see that (3.24) can hold if and only if $\pi^j = 0$ for all $j \in \bar{\mathbf{q}}$. Thus for $\bar{\pi} \neq 0$, there has to be a $\pi_z \in \mathbb{R}^r$,$\pi_z \neq 0$ and $\langle z', \pi_z \rangle \geq 0$, for all $z' \in Z$. But this implies that the zero vector 0_r is a boundary point of the set Z which is a contradiction to the assumption, so we conclude that $\pi^0 > 0$.
We now define

$$\hat{\mu}^j := \pi^j / \pi^0, \quad j \in \mathbf{q},$$

and

$$\hat{\zeta}^j := \pi^j_z / \pi^0, \quad j \in \{1, \ldots, r\}.$$

Then, from (3.21), we obtain

$$\frac{1}{\pi^0} \langle \bar{\pi}, \bar{y} - \bar{y}_* \rangle = y^0 - y^0_* + \langle \hat{\mu}, y - y_* \rangle + \langle \hat{\zeta}, z - z_* \rangle \geq 0, \quad \forall \bar{y} \in \mathbf{G}. \tag{3.25}$$

Since $y_*^0 = \nu_p$, $y_* = 0$ and $z_* = 0$, (3.25) yields

$$y^0 + \langle \hat{\mu}, y \rangle + \langle \hat{\zeta}, z \rangle \geq \nu_p, \quad \forall \bar{y} \in \mathbf{G}.$$

which shows that the first equation in (3.16) holds. Let $\hat{x}$ be a solution of (3.12). Then $f^0(\hat{x}) = \nu_p$, $f(\hat{x}) \leq 0$ (componentwise) and $A\hat{x} - b = 0$. Hence

$$\max_{\mu \in \mathbb{R}_+^q} \langle \mu, f(\hat{x}) \rangle = 0,$$

and

$$\langle \zeta, A\hat{x} - b \rangle = 0, \quad \forall \zeta \in \mathbb{R}^r,$$

therefore the second equation in (3.16) follows by inspection. Next, since $F(C) \subset \mathbf{G}$, it follows that

$$\begin{aligned} \nu_d = \sup_{\substack{\mu \in \mathbb{R}_+^q \\ \zeta \in \mathbb{R}^r}} \inf_{\bar{y} \in F(x)} (y^0 + \langle \mu, y \rangle + \langle \zeta, z \rangle) &\geq \sup_{\substack{\mu \in \mathbb{R}_+^q \\ \zeta \in \mathbb{R}^r}} \inf_{\bar{y} \in \mathbf{G}} (y^0 + \langle \mu, y \rangle + \langle \zeta, z \rangle) \\ &\geq \inf_{\bar{y} \in \mathbf{G}} (y^0 + \langle \hat{\mu}, y \rangle + \langle \hat{\zeta}, z \rangle) \geq \nu_p. \end{aligned}$$

In view of (3.19), it follows that $\nu_p = \nu_d$, which completes the proof. □

Corollary 3 (Corollary 5.5.3 (b) in [91])**.** *Consider the minMax problem*

$$\min_{x \in \mathbb{R}^n} \{\psi(x) | Ax - b = 0\}, \tag{3.26}$$

where $\psi(x) := \max_{j \in \mathbf{q}} f^j(x)$, and suppose, that for all $j \in \mathbf{q}$, the functions $f^j : \mathbb{R}^n \to \mathbb{R}$ in (3.26) are convex, with at least one of them bounded from below, and that A is an $n \times r$ $(r < n)$ matrix of maximum rank. Then

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \{\psi(x) | Ax - b = 0\} &= \min_{x \in \mathbb{R}^n} \max_{\substack{\mu \in \Sigma_q \\ \zeta \in \mathbb{R}^r}} \langle \mu, f(x) \rangle + \langle \zeta, Ax - b \rangle \\ &= \max_{\substack{\mu \in \Sigma_q \\ \zeta \in \mathbb{R}^r}} \min_{x \in \mathbb{R}^n} \langle \mu, f(x) \rangle + \langle \zeta, Ax - b \rangle \end{aligned} \tag{3.27}$$

Furthermore, there exist an $\hat{x} \in \mathbb{R}^n$, a $\hat{\mu} \in \Sigma_q$, and a $\hat{\zeta} \in \mathbb{R}^r$, such that

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \langle \hat{\mu}, f(x) \rangle + \langle \hat{\zeta}, Ax - b \rangle &= \max_{\substack{\mu \in \Sigma_q \\ \zeta \in \mathbb{R}^r}} \langle \mu, f(\hat{x}) \rangle + \langle \zeta, A\hat{x} - b \rangle \\ &= \langle \hat{\mu}, f(\hat{x}) \rangle + \langle \hat{\zeta}, A\hat{x} - b \rangle. \end{aligned} \tag{3.28}$$

Proof (Modification of the proof of Theorem 5.5.3 (a) in [91]).
Consider the problem

$$\min_{(x^0,x)\in\mathbb{R}\times\mathbb{R}^n}\{x^0|f^j(x)-x^0\leq 0, j\in\mathbf{q},\ Ax-b=0\}. \tag{3.29}$$

For any $x\in\mathbb{R}^n$, let $\psi(x):=\max_{j\in\mathbf{q}} f^j(x)$. Then we see that $\hat{x}$ solves (3.26) if and only if $(\psi(\hat{x}),\hat{x})$ solves (3.29). Let $x_0\in\mathbb{R}^n$ with $Ax_0-b=0$ and $\epsilon>0$ be arbitrary, and let $x^0=\psi(x_0)+\epsilon$. Then we see that $f^j(x_0)-x^0\leq-\epsilon<0$ for all $j\in\mathbf{q}$, and, since all the functions in (3.29) are convex, it follows that all assumptions of Theorem 7 are satisfied. Hence it follows from Theorem 7 that, if $(\psi(\hat{x}),\hat{x})$ solves (3.29), then there exist multipliers $\hat{\mu}\in\mathbb{R}^q_+$, $\hat{\zeta}\in\mathbb{R}^r$ such that

$$\begin{aligned}\min_{x\in\mathbb{R}^n}\{\psi(x)|Ax-b=0\} &= \inf_{(x^0,x)\in\mathbb{R}\times\mathbb{R}^n}\sup_{\substack{\mu\in\mathbb{R}^q_+\\ \zeta\in\mathbb{R}^r}} x^0(1-\sum_{j=1}^q\mu^j)+\langle\mu,f(x)\rangle+\langle\zeta,Ax-b\rangle\\ &= \psi(\hat{x})(1-\sum_{j=1}^q\hat{\mu}^j)+\langle\hat{\mu},f(\hat{x})\rangle+\langle\hat{\zeta},Ax-b\rangle\\ &= \sup_{\substack{\mu\in\mathbb{R}^q_+\\ \zeta\in\mathbb{R}^r}}\inf_{(x^0,x)\in\mathbb{R}\times\mathbb{R}^n} x^0(1-\sum_{j=1}^q\mu^j)+\langle\mu,f(x)\rangle+\langle\zeta,Ax-b\rangle.\end{aligned} \tag{3.30}$$

Since $\inf_{x^0\in\mathbb{R}} x^0(1-\sum_{j=1}^q\mu^j)=-\infty$, whenever $(1-\sum_{j=1}^q\mu^j)\neq 0$, it follows that $\sum_{j=1}^q\hat{\mu}^j=1$ must hold in (3.30) for the values to be finite. Hence $\hat{\mu}\in\Sigma_q$, and (3.30) reduces to (3.27) and (3.28).

□

Theorem 8 (Theorem 2.2.24 in [91])**.** *Consider problem* **IECP**. *Suppose that the functions $c^k(\cdot),k\in\mathbf{p}$, $f^j(\cdot),j\in\mathbf{q}$, and $g(\cdot)$ are all continuously differentiable and that the matrix $g_x(x)$ is of maximum row rank for all $x\in\mathbb{R}^n$. Let $\gamma,\delta>0$, and let the optimality function $\theta:\mathbb{R}^n\to\mathbb{R}$ be defined by*

$$\begin{aligned}\theta(x):=\min_{h\in\mathcal{H}_E(x)}\Big\{\max\Big\{&\max_{k\in\mathbf{p}}\{c^k(x)-f^0(x)-\gamma\psi(x)_+ +\langle\nabla c^k(x),h\rangle\},\\ &\max_{j\in\mathbf{q}}\{f^j(x)-\psi(x)_+ +\langle\nabla f^j(x),h\rangle\}\Big\}+\frac{1}{2}\delta\|h\|^2\Big\},\end{aligned} \tag{3.31}$$

with

$$\psi(x)_+:=\max\{0,\psi(x)\}$$

for any $z\in\mathbb{R}^n$.

Then, (a) $\theta(x) \leq 0$ for all $x \in \mathbb{R}^n$; (b) $\theta(\cdot)$ is continuous; and (c) for any $x \in \mathbb{R}^n$ such that $\psi(x) \leq 0$ and $g(x) = 0$, $\theta(x) = 0$ if and only if there exist multipliers $\hat{\mu} \in \Sigma_q^0$, $\hat{\nu} \in \Sigma_p$, and $\hat{\zeta} \in \mathbb{R}^r$ such that (3.7) and (3.8) are satisfied.
Furthermore, an alternative expression for $\theta(x)$ is given by

$$\begin{aligned}\theta(x) = -\min_{\substack{\mu\in\Sigma_q^0\\ \nu\in\Sigma_p\\ \zeta\in\mathbb{R}^r}} \Bigg\{ & -\mu^0 \left[\sum_{k=1}^{p} \nu^k \left[c^k(x) - f^0(x) - \gamma\psi(x)_+\right]\right] \\ & - \sum_{j=1}^{q} \mu^j \left[f^j(x) - \psi(x)_+\right] \\ & + \frac{1}{2\delta} \left\| \mu^0 \left[\sum_{k=1}^{p} \nu^k \nabla c^k(x)\right] + \sum_{j=1}^{q} \mu^j \nabla f^j(x) + \sum_{l=1}^{r} \zeta^l \nabla g^l(x) \right\|^2 \Bigg\}. \end{aligned} \tag{3.32}$$

Proof. (a) Let $\omega(x,h)$ be given by

$$\begin{aligned}\omega(x,h) := \Bigg\{ \max \Bigg\{ & \max_{k\in\mathbf{p}} \{c^k(x) - f^0(x) - \gamma\psi(x)_+ + \langle \nabla c^k(x), h\rangle\}, \\ & \max_{j\in\mathbf{q}} \{f^j(x) - \psi(x)_+ + \langle \nabla f^j(x), h\rangle\} \Bigg\} + \frac{1}{2}\delta\|h\|^2 \Bigg\}. \end{aligned}$$

Obviously $\omega(x,0) \leq 0$, it follows for any $x \in \mathbb{R}^n$ that $\theta(x) \leq 0$.
(b) $\omega(\cdot,\cdot)$ is obviously continuous. From Proposition 2 it follows that $\mathcal{H}_E(\cdot)$ is continuous. Now, to fulfil the requirements of Proposition 3, to show that $\theta(\cdot)$ is continuous, it is sufficient to show that for every bounded set $X \subset \mathbb{R}^n$ there exists a constant $\alpha < \infty$ with

$$\| \arg \min_{h\in\mathcal{H}_E(x)} \omega(x,h)\| < \alpha,$$

for all $x \in X$.
First, because X is bounded and the functions $f^k(\cdot)$ and $c^k(\cdot)$ are all continuously differentiable, $\nabla f^k(\cdot)$ and $\nabla c^k(\cdot)$ are bounded as well. Looking at the problem

$$\min_{\tau\in\mathbb{R}} \omega(x, \tau h),$$

for a given $x \in X$ and $h \in \mathcal{H}_E(x)$ with $h \neq 0$ and without loss of generality $\|h\| = 1$, we see that $\hat{\tau} = \arg\min_{\tau\in\mathbb{R}} \omega(x,\tau h)$ is given by

$$\hat{\tau} = \arg\min_{\tau\in\mathbb{R}} \langle \nabla c^k(x), \tau h\rangle + \frac{1}{2}\delta\|\tau h\|^2$$

or

$$\hat{\tau} = \arg\min_{\tau\in\mathbb{R}} \langle \nabla f^j(x), \tau h\rangle + \frac{1}{2}\delta\|\tau h\|^2,$$

for the contributing $k \in \mathbf{p}$ or $j \in \mathbf{q}$. Therefore, $\hat{\tau}$ is either given by $\hat{\tau} = -\frac{\langle \nabla c^k, h\rangle}{\delta}$ or by $\hat{\tau} = -\frac{\langle \nabla f^j, h\rangle}{\delta}$. It follows that $\hat{\tau}$ is bounded and therefore there is a constant α with

$$\|\arg\min_{h\in\mathcal{H}_E(x)} \omega(x,h)\| < \alpha,$$

and hence $\theta(\cdot)$ is continuous.
(c) First $\omega(x,h)$ is obviously bounded from below with respect to h and convex with respect to h. We define the scalar functions $\tilde{f}^j_*(x,h)$, $j \in \mathbf{p}+\mathbf{q} := \{1,\ldots,p,p+1,\ldots,p+q\}$ by

$$\tilde{f}^j_*(x,h) = c^j(x) - f^0(x) - \gamma\psi(x)_+ + \langle \nabla c^j(x), h\rangle + \frac{1}{2}\delta\|h\|^2, \quad \forall j \in \mathbf{p},$$

and

$$\tilde{f}^{j+p}_*(x,h) = f^j(x) - \psi(x)_+ + \langle \nabla f^j(x), h\rangle + \frac{1}{2}\delta\|h\|^2, \quad \forall j \in \mathbf{q}.$$

We see that now

$$\theta(x) = \min_{h\in\mathbb{R}^n} \max_{j\in\mathbf{q}+\mathbf{p}} \left\{ \tilde{f}^j_*(x,h) | g_x(x)h = 0 \right\}. \tag{3.33}$$

By use of Corollary 3 and the fact that

$$\arg\min_{h\in\mathbb{R}} \langle \xi, h\rangle + \frac{1}{2}\delta\|h\|^2 = -\frac{\xi}{\delta}.$$

we see that (3.33) transforms to (3.32). By use of (3.32) we directly see that (c) holds. □

3.2. Semi-infinite inequality and finite equality constrained optimization problem

Definition 9 (Definition (1c), page 368 in [91])**.** *The problem*

$$\min\{\psi^0(x) | \psi^j(x) \le 0, j \in \mathbf{q} := \{1,\ldots,q\}, g(x) = 0\},$$

where the functions $\psi^j : \mathbb{R}^n \to \mathbb{R}$, $j \in \bar{\mathbf{q}} := \{0,1,\ldots,q\}$, *are of the form*

$$\psi^j(x) := \max_{y_j\in Y_j} \phi^j(x, y_j),$$

with the functions $\phi^j : \mathbb{R}^n \times \mathbb{R}^{m_j} \to \mathbb{R}$ *and the sets* $Y_j \subset \mathbb{R}^{m_j}$, $g : \mathbb{R}^n \to \mathbb{R}^r$, *with* $r < n$, *is called a semi-infinite inequality and finite equality constrained optimization problem (***SIECP***).*

In contrast to the preceding section $\psi(x)$ is now defined by

$$\psi(x) := \max_{j \in \mathbf{q}} \psi^j(x). \tag{3.34}$$

Definition 10 (Definition 3.2.1 in [91])**.** *Let* $X_{IE} := \{x \in \mathbb{R}^n | \psi^j(x) \leq 0,\ j \in \mathbf{q},\ g(x) = 0\}$. *We will say that* $\hat{x} \in X_{IE}$ *is a local minimizer for* **SIECP** *if there exist a* $\rho > 0$ *such that* $\psi^0(x) \geq \psi^0(\hat{x})$ *for all* $x \in X_{IE} \cap B(\hat{x}, \rho)$ *with*

$$g(x) := (g^1(x), g^2(x), \ldots, g^r(x)).$$

If $\psi^0(x) > \psi^0(\hat{x})$ *for all* $x \in X_{IE} \cap B(\hat{x}, \rho)$, $x \neq \hat{x}$, $\hat{x}$ *is called a strict local minimizer.*

Assumption 1 (Assumption 3.1.1 in [91])**.** *We will assume that,*
(i) for all $j \in \bar{\mathbf{q}}$, *the functons* $\phi^j(\cdot, \cdot)$ *are continuous and their gradients* $\nabla_x \phi^j(\cdot, \cdot)$ *exists and are continuous.*
(ii) the subsets $Y_j \subset \mathbb{R}^{m_j}$ *are compact, and*
(iii) the function $g(\cdot)$ *is continuously differentiable, and its Jacobian* $g_x(x)$ *has maximum row rank for all* $x \in \mathbb{R}^n$.

3.2.1. First order optimality conditions for SIECP

Theorem 9. *Suppose that* $\hat{x}$ *is a local minimizer for the problem* **SIECP**, *then* $\hat{x}$ *is a local minimizer for the problem*

$$\min\{\widehat{F}(x) | g(x) = 0\}, \tag{3.35}$$

where

$$\widehat{F} := \max\{\psi^0(x) - \psi^0(\hat{x}), \psi(x)\}. \tag{3.36}$$

Proof. First, since $\psi(\hat{x}) \leq 0$, by assumption, we see that $\widehat{F}(\hat{x}) = 0$. Next, let $\hat{\rho} > 0$ be the radius associated with $\hat{x}$. Then, for all $x \in B(\hat{x}, \hat{\rho}) \cap X_E$ with

$$X_E := \{x \in \mathbb{R}^n | g(x) = 0\}, \tag{3.37}$$

if $\psi(x) \geq 0$, then $\widehat{F}(x) \geq 0$, and if $\psi(x) \leq 0$, then $\psi^0(x) - \psi^0(\hat{x}) \geq 0$, which also implies that $\widehat{F}(x) \geq 0$. Hence $\hat{x}$ is a local minimizer for (3.35). □

Theorem 10. *Consider problem* **SIECP**. *Suppose that Assumption 1 is satisfied and that, for any $x \in \mathbb{R}^n$ such that both $\psi(x) = 0$ and $g(x) = 0$, there exists a sequence of vectors $\{x_i\}_{i=0}^{\infty}$ converging to x, such that $\psi(x_i) < 0$ and $g(x_i) = 0$ for all $i \in \mathbb{N}$. If $\hat{x}$ is a local minimizer for (3.35) such that $\psi(\hat{x}) \leq 0$, then $\hat{x}$ is a local minimizer for* **SIECP**.

Proof. Since Assumption 1 holds, with $Y_j(x) : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ and $Y_j(x) := Y_j$ is constant and thus continuous and Y_j is compact and thus bounded for all $j \in \bar{\mathbf{q}}$, it follows from Proposition 3 that $\psi^j(x)$, $j \in \bar{\mathbf{q}}$ are continuous and thus $\psi(x)$ is continuous as well.
First suppose that $\psi(\hat{x}) < 0$. Since $\psi(\cdot)$ is continuous, and $\widehat{F}(\hat{x}) = 0$, and $\hat{x}$ is a local minimizer of (3.35), there exists a $\hat{\rho} > 0$ such that, for all $x \in B(\hat{x}, \hat{\rho}) \cap X_E$, $\widehat{F}(x) \geq 0$ and $\psi(x) < 0$. It now follows by inspection that, for all $x \in B(\hat{x}, \hat{\rho}) \cap X_E$, $\psi^0(x) - \psi^0(\hat{x}) \geq 0$, which shows that $\hat{x}$ is a local minimizer for **SIECP**.
Next suppose that $\psi(\hat{x}) = 0$. Since $\hat{x}$ is a local minimizer of (3.35), it follows that there exists a $\rho > 0$ such that for all $x \in B(\hat{x}, \hat{\rho}) \cap X_E$, $\widehat{F}(x) \geq 0$. Now, if $x \in B(\hat{x}, \hat{\rho}) \cap X_E$ is such that $\psi(x) < 0$, then $\widehat{F}(x) \geq 0$ implies that $\psi^0(x) - \psi^0(\hat{x}) \geq 0$. If $x \in B(\hat{x}, \hat{\rho}) \cap X_E$ is such that $\psi(x) = 0$ then by assumption there exists a sequence of vectors $\{x_i\}_{i=0}^{\infty}$ converging to x, such that $\psi(x_i) < 0$ and $g(x_i) = 0$ for all $i \in \mathbb{N}$. Consequently there exist $i_0 \in \mathbb{N}$ so that for all $i > i_0$, $x_i \in B(\hat{x}, \hat{\rho}) \cap X_E$ with $\psi(x_i) < 0$, which implies that $\widehat{F}(x_i) = \psi^0(x_i) - \psi^0(\hat{x}) \geq 0$ for all $i > i_0$. It now follows from the continuity of $\psi^0(\cdot)$ that $\psi^0(x) - \psi^0(\hat{x}) \geq 0$. Hence we conclude that $\hat{x}$ is a local minimizer for **SIECP**. □

Theorem 11 (Theorem 3.2.16 in [91])**.** *Consider the problem* **SIECP** *defined in Definition 9. Suppose that Assumption 1 is satisfied. If $\hat{x} \in \mathbb{R}^n$ is a local minimizer for* **SIECP**, *then*
(a) first,

$$\mathrm{d}\widehat{F}(\hat{x}; h) \geq 0, \forall h \in \mathcal{H}_E(\hat{x}), \tag{3.38}$$

where

$$\mathcal{H}_E(x) := \{h \in \mathbb{R}^n | g_x(x)h = 0\}, \tag{3.39}$$

and $\mathrm{d}\widehat{F}(\hat{x}; h)$ is the directional derivative (see Theorem 25) of function $\widehat{F}(\cdot)$ at $\hat{x} \in \mathbb{R}^n$ in direction $h \in \mathbb{R}^n$ as given in Definition 24;
(b) and second, there exist a multiplier $\hat{\zeta} \in \mathbb{R}^r$ such that

$$-g_x(\hat{x})^T \hat{\zeta} \in \partial \widehat{F}(\hat{x}), \tag{3.40}$$

and hence there exists a

$$\hat{\mu} \in \Sigma_q^0 := \left\{ (\mu^0, \mu) |\ \mu^0 \in \mathbb{R}_+,\ \mu \in \mathbb{R}_+^q,\ \sum_{j=0}^{q} \mu^j = 1 \right\}$$

such that [1]

$$-g_x(\hat{x})^T\hat{\zeta} \in \sum_{j=0}^{q} \hat{\mu}^j \partial\psi^j(\hat{x}), \tag{3.41}$$

and

$$\sum_{j=1}^{q} \hat{\mu}^j \psi^j(\hat{x}) = 0, \tag{3.42}$$

where $\bar{\mathbf{q}} := \{0, 1, \ldots, q\}$.

Proof. (a) Since $\hat{x}$ is a local minimizer for **SIECP**, it must also be a local minimizer for the problem (3.35). Hence, for the sake of contradiction, suppose that there is a vector $h \in \mathcal{H}_E(\hat{x})$ such that $\mathrm{d}\widehat{F}(\hat{x}; h) < 0$. Since the matrix $g_x(\hat{x})$ is of maximum row rank, it follows from Corollary 6 that there exists a $t_h > 0$ and continuously differentiable function $s : [0, t_h] \to \mathbb{R}^n$ such that $s(0) = \hat{x}$, $\dot{s}(0) = h$, and $g(s(t)) = 0$ for all $t \in [0, t_h]$. Let $\sigma : [0, t_h] \to \mathbb{R}$ be defined by $\sigma(t) = \widehat{F}(s(t))$. Then by the Chain Rule Theorem 27, the directional derivative $\mathrm{d}\sigma(0; 1) = \mathrm{d}\widehat{F}(\hat{x}; h) < 0$, and hence there exists a $t' \in (0, t_h]$ such that $\sigma(t) < \sigma(0)$ for all $t \in (0, t')$. Consequently, for any $t \in (0, t'), g(s(t)) = 0$ and $\widehat{F}(s(t)) < \widehat{F}(\hat{x}) = 0$, which contradicts the fact that $\hat{x}$ is a local minimizer for the problem (3.2) and hence for the problem **SIECP**.
(b) Now, by Theorem 25 (c) for any $h \in \mathbb{R}^n$,

$$\mathrm{d}\widehat{F}(\hat{x}; h) = \max_{\xi \in \partial\widehat{F}(\hat{x})} \langle \xi, h \rangle$$

and

$$\partial\widehat{F}(\hat{x}) = \mathrm{conv} \bigcup_{j \in (\{0\} \cup \mathbf{q_A}(\hat{x}))} \left(\mathrm{conv} \bigcup_{y \in \widehat{Y}_j(\hat{x})} \{\nabla_x \phi^j(\hat{x}, y)\} \right), \tag{3.43}$$

with

$$\mathbf{q_A}(x) := \left\{ j \in \mathbf{q} | \psi^j(x) \geq 0 \right\},$$

and

$$\widehat{Y}_j(x) := \left\{ y \in Y_j | \phi^j(x, y) = \psi^j(x) \right\},$$

for $j \in \bar{\mathbf{q}}$.
It now follows from (3.38), (3.39) and Proposition 1 that

$$\partial\widehat{F}(\hat{x}) \cap \mathcal{H}_E^{\perp}(\hat{x}) \neq \emptyset.$$

[1] Note that (3.41) holds if and only if there exist vectors $\xi_j \in \partial\psi^j(\hat{x}),\ j \in \bar{\mathbf{q}}$, such that

$$\sum_{j=0}^{q} \hat{\mu}^j \xi_j = -g_x(\hat{x})^T\hat{\zeta}$$

.

Since the vectors $\nabla g^l(\hat{x}), l \in \mathbf{r}$, form a basis for $\mathcal{H}_E^{\perp}(\hat{x})$, it follows that there exists $\hat{\zeta} \in \mathbb{R}^r$ such that

$$-g_x(\hat{x})^T\hat{\zeta} \in \partial\widehat{F}(\hat{x}). \tag{3.44}$$

The expressions (3.41) and (3.42) follow by inspection. □

We now formulate an extension of Theorem 11, i.e. an alternative way of stating (3.40) with (3.43) that does not involve the active index set $\mathbf{q_A}(\hat{x})$ and the sets of maximizer $\widehat{Y}_j(\hat{x})$, $j \in \{0\} \cup \mathbf{q_A}(\hat{x})$. This reformulation of Theorem 11 is used for the construction of an optimality function in Theorem 13.

Theorem 12. *Suppose that Assumption 1 is satisfied and that $\hat{x} \in \mathbb{R}^n$ with $\psi(\hat{x}) \leq 0$ and $g(\hat{x}) = 0$. Then,*
(a) $\hat{x}$ satisfies (3.40) if and only if there exists a multiplier $\hat{\zeta} \in \mathbb{R}^r$ such that

$$\begin{pmatrix} 0 \\ -g_x(\hat{x})^T\hat{\zeta} \end{pmatrix} \in \overline{G}F(\hat{x}), \tag{3.45}$$

where $\overline{G}F(x) \subset \mathbb{R}^{n+1}$ has elements denoted by $\bar{\xi} = (\xi^0, \xi)^T$, with $\xi^0 \in \mathbb{R}$, $\xi \in \mathbb{R}^n$, and is defined by

$$\overline{G}F(x) := \operatorname{conv}\left(\bigcup_{j\in\bar{\mathbf{q}}} \overline{G}\psi^j(x)\right), \tag{3.46}$$

with

$$\overline{G}\psi^0(x) := \operatorname{conv}\left[\bigcup_{y_0\in Y_0}\left\{\begin{pmatrix}\psi^0(x) - \phi^0(x,y_0) + \gamma\psi(x)_+ \\ \nabla_x\phi^0(x,y_0)\end{pmatrix}\right\}\right], \tag{3.47}$$

$\gamma > 0$ an arbitrary parameter, $\psi_+(x) := \max\{0, \psi(x)\}$, and for $j \in \mathbf{q}$,

$$\overline{G}\psi^j(x) := \operatorname{conv}\left[\bigcup_{y_j\in Y_j}\left\{\begin{pmatrix}\psi(x)_+ - \phi^j(x,y_j) \\ \nabla_x\phi^j(x,y_j)\end{pmatrix}\right\}\right]; \tag{3.48}$$

(b) the set-valued map $\overline{G}F(\cdot)$, defined by (3.46), is continuous and bounded on bounded subsets of $\mathbb{R}^n$.

Proof. (a) " $\Rightarrow$ " Suppose that $\hat{x}$ satisfies (3.40). Since $\psi(\hat{x}) \leq 0$ and therefore

$$\psi^0(\hat{x}) - \phi^0(\hat{x}, y_0) + \gamma\psi(\hat{x})_+ = 0, \quad y_0 \in \widehat{Y}_0(\hat{x}),$$

and for every $j \in \mathbf{q_A}$,

$$\psi(\hat{x})_+ - \phi^j(\hat{x}, y_j) = 0, \quad y_j \in \widehat{Y}_j(\hat{x}),$$

it follows that the set

$$\widehat{G} := \left\{\bar{x} = (\xi^0, \xi) \in \mathbb{R}^{n+1} | \xi^0 = 0, \xi \in \partial F(\hat{x})\right\}$$

is a subset of $\overline{G}F(\hat{x})$. Since by (3.40) there exists a $\hat{\zeta} \in \mathbb{R}^r$ such that $\begin{pmatrix} 0 \\ -g_x(\hat{x})^T\hat{\zeta} \end{pmatrix} \in \widehat{G}$, it follows that (3.45) must hold.
" $\Leftarrow$ " Now suppose that (3.45) holds. Since

$$\psi^0(\hat{x}) - \phi^0(\hat{x}, y_0) + \gamma\psi(\hat{x})_+ \geq 0, \quad y_0 \in Y_0,$$

and for every $j \in \mathbf{q}$,

$$\psi(\hat{x})_+ - \phi^j(\hat{x}, y_j) \geq 0, \quad y_j \in Y_j,$$

because by assumption $\psi(\hat{x}) \leq 0$, it follows that the vector $\begin{pmatrix} 0 \\ -g_x(\hat{x})^T\hat{\zeta} \end{pmatrix}$ in $\mathbb{R}^{n+1}$ can be only a convex combination of vectors $\bar{\xi} = (\xi^0, \xi) \in \overline{G}F(\hat{x})$ such that $\xi^0 = 0$. Hence we conclude from (3.45) and (3.46) that

$$\begin{pmatrix} 0 \\ -g_x(\hat{x})^T\hat{\zeta} \end{pmatrix} \in \operatorname{conv} \bigcup_{j \in (\{0\} \bigcup \mathbf{q}_{\mathbf{A}}(\hat{x}))} \left(\operatorname{conv} \bigcup_{y \in \widehat{Y}_j(\hat{x})} \left\{ \begin{pmatrix} 0 \\ \nabla_x \phi^j(\hat{x}, y) \end{pmatrix} \right\} \right) \subset \overline{G}F(\hat{x}).$$

It now follows by inspection that (3.40) holds.
(b) The continuity and boundedness of $\overline{G}F(\cdot)$ follow directly from Theorem 23 and Corollary 5. □

For later purpose we now introduce the so called extended Mangasarian-Fromowitz constraint qualification ($EMFCQ$).

Definition 11. *Consider problem* **SIECP**. *We say that the extended Mangasarian-Fromovitz constraint qualification holds at $x \in \mathbb{R}^n$, if the vectors $\nabla g^l(x), l \in \mathbf{r}$, are linearly independent, and there exists $\tilde{h} \in \mathbb{R}^n$ such that (strictly feasible direction),*

$$\nabla g^l(x)^T\tilde{h} = 0, \quad l \in \mathbf{r}, \tag{3.49}$$

and

$$\mathrm{d}\psi^j(x; \tilde{h}) < 0, \quad j \in \mathbf{q}_{\mathbf{A}}(x). \tag{3.50}$$

Corollary 4. *If $\hat{x}$ is a local minimizer for problem (3.35) with $\psi(\hat{x}) \leq 0$ and $EMFCQ$ is satisfied at all points $x \in \mathbb{R}^n$ with $\psi(x) = 0$ and $g(x) = 0$, then $\hat{x}$ is a local minimizer*

for problem **SIECP**.

Proof. It is sufficient to show (Theorem 10), that if $x \in \mathbb{R}^n$ with $\psi(x) = 0$ and $g(x) = 0$ there exists a sequence of vectors $\{x_i\}_{i=0}^{\infty}$ converging to x, such that $\psi(x_i) < 0$ and $g(x_i) = 0$ for all $i \in \mathbb{N}$. Be $x \in \mathbb{R}^n$ with $\psi(x) = 0$ and $g(x) = 0$. By assumption x satisfies $EMFCQ$ and thus there is a direction $\tilde{h}$ such that $g_x(x)^T\tilde{h} = 0$ and $\mathrm{d}\psi(x;\tilde{h}) < 0$. It follows from Corollary 6 that there exists a $t_{\tilde{h}} > 0$ and continuously differentiable function $s : [0, t_{\tilde{h}}] \to \mathbb{R}^n$ such that $s(0) = x$, $\dot{s}(0) = \tilde{h}$, and $g(s(t)) = 0$ for all $t \in [0, t_{\tilde{h}}]$. Let $\sigma : [0, t_{\tilde{h}}] \to \mathbb{R}$ be defined by $\sigma(t) = \psi(s(t))$. Then by the Chain Rule Theorem 27, the directional derivative $\mathrm{d}\sigma(0;1) = \mathrm{d}\psi(x;\tilde{h}) < 0$, and hence there exists a $t' \in (0, t_{\tilde{h}}]$ such that $\sigma(t) < \sigma(0)$ for all $t \in (0, t')$. Consequently, for any $t \in (0, t'), g(s(t)) = 0$ and $\psi(s(t)) < \psi(\hat{x}) = 0$ and thus there exists a sequence of vectors $\{x_i\}_{i=0}^{\infty}$ converging to x, such that $\psi(x_i) < 0$ and $g(x_i) = 0$ for all $i \in \mathbb{N}$. □

3.2.2. Optimality function for SIECP

Theorem 13 (Theorem 3.2.18 in [91])**.** *Consider the problem* **SIECP** *(Definition 9) and suppose that Assumption 1 is satisfied. Let the optimality function $\theta : \mathbb{R}^n \to \mathbb{R}$ be defined by*

$$\theta(x) := \min_{h \in \mathcal{H}_E(x)} \max\{\tilde{\psi}^0(x, x+h) - \psi^0(x) - \gamma\psi(x)_+, \tilde{\psi}(x, x+h) - \psi(x)_+\}, \qquad (3.51)$$

where $\tilde{\psi}(x, x+h)$ is given by

$$\tilde{\psi}(x, x+h) := \max_{j \in \mathbf{q}} \max_{y_j \in Y_j} \{\phi^j(x, y_j) + \langle \nabla_x \phi^j(x, y_j), h\rangle + \frac{1}{2}\delta\|h\|^2\},$$

and $\tilde{\psi}^0(x, x+h)$ is given by

$$\tilde{\psi}^0(x, x+h) := \max_{y \in Y_0} \{\phi^0(x, y) + \langle \nabla_x \phi^0(x, y), h\rangle + \frac{1}{2}\delta\|h\|^2\},$$

$\psi(x)_+$ is given by

$$\psi(x)_+ := \max\{0, \psi(x)\},$$

and $\mathcal{H}_E(x)$ is defined as in (3.39) and $\gamma > 0$ and $\delta > 0$ act as parameters. Then,
(a) $\theta(x) \leq 0$ for all $x \in \mathbb{R}^n$,

(b) an alternative expression for $\theta(x)$ is given by

$$\begin{aligned}\theta(x) &= \min_{h\in\mathcal{H}_E(x)} \max_{\bar{\xi}\in\overline{G}F(x)} \{-\xi^0 + \langle \xi, h\rangle + \frac{1}{2}\delta\|h\|^2\} \\ &= -\min_{\substack{\bar{\xi}\in\overline{G}F(x) \\ \zeta\in\mathbb{R}^r}} \{\xi^0 + \frac{1}{2\delta}\|\xi + g_x(x)^T\zeta\|^2\},\end{aligned} \tag{3.52}$$

where $\bar{\xi} = (\xi^0, \xi) \in \mathbb{R}^{n+1}$, with $\xi^0 \in \mathbb{R}$, $\xi \in \mathbb{R}^n$, and $\overline{G}F(x)$ was defined in (3.46), (3.47) and (3.48),
(c) the function $\theta(\cdot)$ is continuous, and
(d) for any x such that $\psi(x) \leq 0$ and $g(x) = 0$, $\theta(x) = 0$ if and only if there exist multipliers $\hat{\mu} \in \Sigma_q^0$ and $\zeta \in \mathbb{R}^r$ such that (3.41) and (3.42) are satisfied. Therefore, if $\hat{x}$ is a local minimizer of **SIECP**, *then $\theta(\hat{x}) = 0$.*

Proof (Modification of the proof of Theorem 3.1.6 in [91]).
(a) Be $\omega(x, h)$ given by

$$\omega(x,h) := \max\{\tilde{\psi}^0(x, x+h) - \psi^0(x) - \gamma\psi(x)_+, \tilde{\psi}(x, x+h) - \psi(x)_+\}.$$

then

$$\omega(x,0) = \max\{\phi^0(x,y) - \psi^0(x) - \gamma\psi(x)_+, \left(\max_{j\in\mathbf{q}} \max_{y_j\in Y_j} \phi^j(x, y_j)\right) - \psi(x)_+\}.$$

Obviously $\omega(x, 0) \leq 0$, it follows for any $x \in \mathbb{R}^n$ that $\theta(x) \leq 0$.
(b) Next let,

$$\widetilde{G}F(x) := \bigcup_{j\in\bar{\mathbf{q}}} \widetilde{G}\psi^j(x),$$

with

$$\widetilde{G}\psi^0(x) := \bigcup_{y_0\in Y_0} \left\{ \begin{pmatrix} \psi^0(x) - \phi^0(x, y_0) + \gamma\psi(x)_+ \\ \nabla_x\phi^0(x, y_0) \end{pmatrix} \right\},$$

and for $j \in \mathbf{q}$,

$$\widetilde{G}\psi^j(x) := \bigcup_{y_j\in Y_j} \left\{ \begin{pmatrix} \psi(x)_+ - \phi^j(x, y_j) \\ \nabla_x\phi^j(x, y_j) \end{pmatrix} \right\}.$$

Then we can rewrite (3.51) as

$$\theta(x) = \min_{h\in\mathcal{H}_E(x)} \max_{\bar{\xi}\in\widetilde{G}F(x)} \left\{ -\xi^0 + \langle \xi, h\rangle + \frac{1}{2}\delta\|h\|^2 \right\}, \tag{3.53}$$

where as above, $\bar{\xi} := (\xi^0, \xi)$. Since the maximum over a set of scalars is equal to the maximum over the convex hull of these scalars and since $\overline{G}F(x) = \operatorname{conv} \widetilde{G}F(x)$, we conclude that (3.53) is equivalent to the first line in (3.52). Since by assumption the Jacobian $g_x(x)$ has maxmimum row rank for all $x \in \mathbb{R}^n$ and thus $\mathcal{H}_E(x)$ is a subspace of $\mathbb{R}^n$ we can apply Theorem 28 to the first line in (3.52), we conclude that the max and min in (3.52) can be interchanged and hence that

$$\theta(x) = \max_{\bar{\xi} \in \overline{G}F(x)} \min_{h \in \mathcal{H}_E(x)} \left\{ -\xi^0 + \langle \xi, h \rangle + \frac{1}{2}\delta \|h\|^2 \right\}.$$

Now consider the "inside" function

$$\nu(\bar{\xi}) := \min_{h \in \mathcal{H}_E(x)} \left\{ -\xi^0 + \langle \xi, h \rangle + \frac{1}{2}\delta \|h\|^2 \right\}, \tag{3.54}$$

where $\bar{\xi} = (\xi^0, \xi) \in \mathbb{R}^{n+1}$. We see that obviously (3.54) is equivalent to

$$\nu(\bar{\xi}) = \min_{h \in \mathbb{R}^n} \left\{ -\xi^0 + \langle \xi, h \rangle + \frac{1}{2}\delta \|h\|^2 \mid g_x(x)h = 0 \right\}. \tag{3.55}$$

Applying Corollary 3 we see that (3.55) is equivalent to

$$\nu(\bar{\xi}) = \max_{\zeta \in \mathbb{R}^r} \min_{h \in \mathbb{R}^n} \left\{ -\xi^0 + \langle \xi, h \rangle + \frac{1}{2}\delta \|h\|^2 + \langle \zeta, g_x(x)h \rangle \right\}.$$

Now consider the new "inside" function

$$\tilde{\nu}(\bar{\xi}, \zeta) := \min_{h \in \mathbb{R}^n} \left\{ -\xi^0 + \langle \xi, h \rangle + \frac{1}{2}\delta \|h\|^2 + \langle \zeta, g_x(x)h \rangle \right\}, \tag{3.56}$$

where $\bar{\xi} = (\xi^0, \xi) \in \mathbb{R}^{n+1}$ and $\zeta \in \mathbb{R}^r$. Solving the unconstrained minimization problem in (3.56) for h in terms of $\bar{\xi}$ and ζ, we obtain

$$\delta h = -\xi - g_x(x)^T \zeta,$$

and hence

$$\nu(\bar{\xi}) = \max_{\zeta \in \mathbb{R}} \left\{ -\xi_0 - \frac{1}{2\delta} \|\xi + g_x(x)^T \zeta\|^2 \right\}.$$

Substituting back into (3.54), we obtain (3.52).

(c) The function

$$\tilde{\omega}(\bar{\xi}, \zeta) := \xi^0 + \frac{1}{2\delta} \|\xi + g_x(x)^T \zeta\|^2,$$

where $\bar{\xi} = (\xi^0, \xi) \in \mathbb{R}^{n+1}$ and $\zeta \in \mathbb{R}^r$, is obviously continuous in $\bar{\xi}$ and ζ. From Theorem 12(b) we know that $\overline{G}F(\cdot)$ is continuous and bounded on bounded sets. To fulfil now

the requirements of Proposition 3 to show that $\theta(\cdot)$ is continuous it is sufficient to show that for every bounded set $X \subset \mathbb{R}^n$ there exist a constant $\alpha < \infty$ with

$$\| \arg\min_{\zeta \in \mathbb{R}^r} \tilde{\omega}(\bar{\xi}, \zeta) \| < \alpha,$$

for all $\bar{\xi} \in \overline{G}F(X)$.
First we know that g is continuously differentiable and that the Jacobian $g_x(x)$ has maximum rank for all $x \in \mathbb{R}^n$. Therefore, there exists a $\tilde{\alpha}$ with

$$\|\nabla g^j(x)\| \geq \tilde{\alpha} \quad , \forall x \in \bar{X},\ j \in \mathbf{r} := \{1, \ldots, r\},$$

where $\bar{X}$ denotes the closure of X. Now be $x \in X$ and $\bar{\xi} \in \overline{G}F(x)$, then there exists a vector $\kappa \in \mathbb{R}^r$ and a vector $\xi_\perp \in \mathbb{R}^n$ with

$$\xi = \xi_\perp + \sum_{j=1}^{r} \kappa^j \frac{\nabla g^j(x)}{\|\nabla g^j(x)\|},$$

and

$$\xi_\perp^T \nabla g^j(x) = 0, \quad \forall j \in \mathbf{r},$$

as well as $|\kappa^j| \leq \|\xi\|,\ j \in \mathbf{R}$. It follows that,

$$\begin{aligned} \hat{\zeta} &= \arg\min_{\zeta \in \mathbb{R}^r} \left\{ \xi^0 + \frac{1}{2\delta} \|\xi + g_x(x)^T \zeta\|^2 \right\} \\ &= \arg\min_{\zeta \in \mathbb{R}^r} \left\| \sum_{j=1}^{r} \kappa^j \frac{\nabla g^j(x)}{\|\nabla g^j(x)\|} + \zeta^j \nabla g^j(x) \right\|^2 . \end{aligned}$$

Therefore, $\hat{\zeta} = (\hat{\zeta}^1, \ldots, \hat{\zeta}^r)$ is given by

$$\hat{\zeta}^j = -\frac{\kappa^j}{\|\nabla g^j(x)\|}, \quad j \in \mathbf{r}.$$

Cause

$$|\hat{\zeta}^j| \leq \frac{\|\xi\|}{\tilde{\alpha}}, \qquad j \in \mathbf{r},$$

it follows that there exists a $\alpha > 0$ with

$$\| \arg\min_{\zeta \in \mathbb{R}^r} \tilde{\omega}(\bar{\xi}, \zeta) \| < \alpha,$$

for all $\bar{\xi} \in \overline{G}F(X)$ and hence $\theta(\cdot)$ is continuous.
(d) This follows directly by Theorem 12(a) with equation (3.52) second line. □

3.3. Method of Outer Approximations

In this section we present a discretization scheme to solve **SIECP** utilizing the concept of optimality functions, therefore consider **SIECP**.

Assumption 2 (Assumption 3.6.1 in [91])**.** *We will assume that*
(i) for all $j \in \bar{\mathbf{q}} := \{0, \dots, q\}$*, the functions* $\phi^j(\cdot,\cdot)$ *and their gradients* $\nabla_x \phi^j(\cdot,\cdot)$ *are Lipschitz continuous on bounded sets,*
(ii) the subsets $Y_j \subset \mathbb{R}^{m_j}$ *are compact, and*
(iii) the function $g(\cdot)$ *is continuously differentiable and its Jacobian* $g_x(x)$ *has maximum row rank for all* $x \in \mathbb{R}^n$*.*

Now for $j \in \bar{\mathbf{q}}$, let S_j be compact subsets of Y_j, and let $S = S_0 \times S_1 \times \cdots \times S_q$. Then we define the functions $\psi^j_{S_j} : \mathbb{R}^n \to \mathbb{R}$, $j \in \bar{\mathbf{q}}$, and $\psi_S : \mathbb{R}^n \to \mathbb{R}$ by

$$\psi^j_{S_j}(x) := \max_{y_j \in S_j} \phi^j(x, y_j), \tag{3.57}$$

$$\psi_S(x) := \max_{j \in \mathbf{q}} \psi^j_{S_j}(x), \tag{3.58}$$

and the corresponding optimization problem $\mathbf{P}_{\mathrm{S}}$ by

$$\min\{\psi^0_S(x) \mid \psi_S(x) \le 0, g(x) = 0\}. \tag{3.59}$$

Now we define a corresponding optimality function $\theta_S : \mathbb{R}^n \to \mathbb{R}$ by

$$\theta_S(x) := \min_{h \in \mathcal{H}_E(x)} \widetilde{F}_S(x, x+h), \tag{3.60}$$

where

$$\widetilde{F}_S(x, x+h) := \max\{\tilde{\psi}^0_S(x, x+h) - \psi^0_S(x) - \gamma\psi_S(x)_+, \tilde{\psi}_S(x, x+h) - \psi_S(x)_+\},$$

and

$$\tilde{\psi}_S(x, x+h) := \max_{j \in \mathbf{q}} \tilde{\psi}^j_{S_j}(x, x+h), \tag{3.61}$$

and

$$\tilde{\psi}^j_{S_j}(x, x+h) := \max_{y \in S_j}\{\phi^j(x,y) + \langle \nabla_x \phi^j(x,y), h\rangle + \frac{1}{2}\|h\|^2\}, \quad j \in \bar{\mathbf{q}}, \tag{3.62}$$

and

$$\psi_S(x)_+ := \max\{0, \psi_S(x)\},$$

where $\gamma > 0$.

Remark. *If $S_j, j \in \bar{\mathbf{q}}$ has finite cardinality then* $\mathbf{P}_{\mathrm{S}}$ *restates a corresponding* **IECP** *and under appropriate assumptions (3.60) restates the corresponding optimality function (3.31) with $\delta = 1$. If $S_j = Y_j, j \in \bar{\mathbf{q}}$ under appropriate assumptions (3.60) restates (3.51) with $\delta = 1$.*

Let N_0 be a strictly positive integer, and, for $N = N_0,\ N_0 + 1,\ N_0 + 2,\ N_0 + 3, \ldots$, let $Y_{j,N}$ be subsets, of finite cardinality, of the $Y_j,\ j \in \bar{\mathbf{q}}$, such that $Y_{j,N} \subset Y_{j,N+1}$ for all N and the closure of the set $\mathrm{Lim}_{N\to\infty} Y_{j,N}$ is equal to $Y_j,\ j \in \bar{\mathbf{q}}$.

Assumption 3 (Modification of Assumption 3.4.2 in [91])**.** *There exist a strictly positive-valued, strictly monotone decreasing function $\Delta : \mathbb{N} \to \mathbb{R}$ such that $\Delta(N) \to 0$, as $N \to \infty$, and constants $N_0 \in \mathbb{N}, K < \infty$, such that, for every $N \geq N_0$, $j \in \bar{\mathbf{q}}$, and $y \in Y_j$, there exits a $y' \in Y_{j,N}$ such that*

$$\|y - y'\| \leq K\Delta(N).$$

Lemma 1 (Lemma 3.4.3 in [91])**.** *Suppose that Assumptions 2 and 3 are satisfied and that, for all $N \in \mathbb{N}$, $\psi_{Y_N}(\cdot)$ is defined by*

$$\psi_{Y_N}(x) := \max_{j\in\mathbf{q}} \psi^j_{Y_{j,N}}(x),$$

with

$$\psi^j_{Y_{j,N}}(x) := \max_{y_j \in Y_{j,N}} \phi^j(x, y_j).$$

Let $S \subset \mathbb{R}^n$ a bounded subset, and let $L < \infty$ be a Lipschitz constant valid for the $\phi^j(\cdot,\cdot)$ and the $\nabla_x\phi^j(\cdot,\cdot)$, on $S \times Y_j$, $j \in \mathbf{q}$. Then there exists a constant $C < \infty$ such that for all $x \in S$, $N \in \mathbb{N}$, with $N \geq N_0$,

$$|\psi_{Y_N}(x) - \psi(x)| \leq C\Delta(N). \tag{3.63}$$

Proof (The proof is taken from [91]). First, because $Y_{j,N} \subset Y_j,\ j \in \mathbf{q}$, we always have that $\psi_{Y_N} \leq \psi(x)$. Next for any $x \in S$, there must exist a $j_x \in \mathbf{q}$ and a $y_x \in Y_{j_x}$, such that

$$\psi(x) = \phi^{j_x}(x, y_x),$$

since by assumption for $j \in \mathbf{q}$ the sets Y_j are compact and $\phi^j(x,\cdot)$ are continuous. By Assumption 3, there exists a $y'_x \in Y_{j_x,N}$ such that $\|y'_x - y_x\| \leq K\Delta(N)$. Hence

$$\psi_{Y_N}(x) \geq \phi^{j_x}(x, y'_x) \geq \phi^{j_x}(x, y_x) - LK\Delta(N) = \psi(x) - LK\Delta(N).$$

Hence (3.63) holds with $C = LK$. □

Lemma 2 (Lemma 3.4.24 in [91])**.** *Suppose that*
(a) Assumption 2 is satisfied,
(b) for $j \in \bar{\mathbf{q}}$, $\{Y_{j,N}\}_{N=N_0}^{\infty}$ is a set of compact subsets of Y_j, satisfying Assumption 3,
(c) $\{x_N\}_{N=N_0}^{\infty}$ is a sequence in $\mathbb{R}^n$, and
(d) the compact sets $\Omega_{j,N} \subset Y_j$, $j \in \bar{\mathbf{q}}$, $N \in \mathbb{N}$, $N \geq N_0$, are constructed recursively as follows: $\Omega_{j,N_0} = \{\hat{y}_{j,N_0}\}$, with $\hat{y}_{y,N_0} \in \widehat{Y}_{j,N_0}(x_0)$, and

$$\Omega_{j,N+1} = \Omega_{j,N} \cup \{\hat{y}_{j,N+1}\}, \quad j \in \bar{\mathbf{q}},$$

with

$$\hat{y}_{j,N+1} \in \widehat{Y}_{j,N+1}(x_{N+1}), \quad j \in \bar{\mathbf{q}},$$

where $\widehat{Y}_{j,N}(x)$ is given by

$$\widehat{Y}_{j,N}(x) := \left\{ y \in Y_{j,N} | \psi^j_{Y_{j,N}}(x) = \phi^j(x,y) \right\}. \tag{3.64}$$

Let $\Omega_N := \Omega_{0,N} \times \cdots \times \Omega_{q,N}$, and let $\psi_{\Omega_N}(x) := \max_{j \in \mathbf{q}} \psi^j_{\Omega_{j,N}}(x)$ with

$$\psi^j_{\Omega_{j,N}}(x) := \max_{y_j \in \Omega_{j,N}} \phi^j(x, y_j), \quad j \in \bar{\mathbf{q}}.$$

If $\hat{x}$ is an accumulation point of $\{x_N\}_{N=N_0}^{\infty}$, so that, for some infinite subset $K \subset \mathbb{N}$, $X_N \to^K \hat{x}$, as $N \to \infty$, then
(a)

$$\psi_{\Omega_N}(x_N) \to^K \psi(\hat{x}),$$

as $N \to \infty$, and
(b)

$$\psi^0_{\Omega_{0,N}}(x_N) \to^K \psi^0(\hat{x}),$$

as $N \to \infty$.

Proof (The proof is taken from [91]). Case (b) can be seen as special version of (a) with $\mathbf{q} = \{1\}$, therefore it is sufficient to show (a).
Now, for any $N \in K$, let $k(N) := \max\{N' \in K | N' < N\}$. Then, by construction, for any $N \in K$ and $j \in \mathbf{q}$, $\hat{y}_{j,k(N)} \in \Omega_{j,N}$ and hence, because $\Omega_{j,N} \subset Y_j$, for any $N \in K$,

$$\psi^j_{Y_j}(x_N) \geq \psi^j_{\Omega_{j,N}}(X_N) \geq \phi^j(X_N, \hat{y}_{j,k(N)}), \tag{3.65}$$

where $\psi^j_{Y_j}(\cdot)$ is given by (3.57) with $S_j = Y_j$. Now, because $x_N \to^K \hat{x}$, as $N \to \infty$,

and because the functions $\psi^j_{Y_j}(\cdot)$ are continuous (Proposition 3), $\psi^j_{Y_j}(x_N) \to \psi^j_{Y_j}(\hat{x})$, as $N \to \infty$. Next,

$$|\psi^j_{Y_{j,k(N)}}(x_{k(N)}) - \psi^j_{Y_j}(\hat{x})| \leq |\psi^j_{Y_{j,k(N)}}(x_{k(N)}) - \psi^j_{Y_j}(x_{k(N)})| + |\psi^j_{Y_j}(x_{k(N)}) - \psi^j_{Y_j}(\hat{x})|. \quad (3.66)$$

Now, in (3.66), $|\psi^j_{Y_j}(x_{k(N)}) - \psi^j_{Y_j}(\hat{x})| \to^K 0$, as $N \to \infty$, because $\psi^j_{Y_j}(\cdot)$ is continuous. Next, because of Lemma 1 and because $\psi^j_{Y_j}(\cdot)$ is continuous,

$$|\psi^j_{Y_{j,k(N)}}(x_{k(N)}) - \psi^j_{Y_j}(x_{k(N)})| \to^K 0, \quad \text{as} \quad N \to \infty.$$

Finally, because $\phi^j(\cdot, y)$ is uniformly continuous for y in a compact set and

$$\|x_N - x_{k(N)}\| \to^K 0, \text{as} \quad N \to \infty,$$

it follows that

$$|\phi^j(x_N, \hat{y}_{k(N)}) - \phi^j(x_{k(N)}, \hat{y}_{k(N)})| \to^K 0, \quad \text{as} \quad N \to \infty.$$

Hence because

$$\psi^j_{Y_{j,k(N)}}(x_{k(N)}) = \phi^j(x_{k(N)}, \hat{y}_{j,k(N)})$$

and because

$$\psi^j_{Y_{j,k(N)}}(x_{k(N)}) \to^K \psi_Y(\hat{x}),$$

it follows that

$$\phi^j(x_N, \hat{y}_{j,k(N)}) \to^K \psi^j_Y(\hat{x}), \quad \text{for all} \quad j \in \mathbf{q}.$$

It now follows from (3.65) that

$$\psi_{\Omega_N}(x_N) \to^K \psi_Y(\hat{x}), \quad \text{as} \quad N \to \infty,$$

with $\psi_Y(\hat{x}) = \psi(\hat{x})$. □

We now state the *Outer Approximations* discretization scheme to solve **SIECP**.

Algorithm 3 (Algorithm 3.6.4 in [91])**.**

Data: $N_0 \in \mathbb{N}$, $x_{N_0} \in \mathbb{R}^n$, *for each* $j \in \bar{\mathbf{q}}$, *a family of descrete subsets* $\{Y_{j,N}\}_{N=N_0}^{\infty}$, *satisfying Assumption 3 and* $\{\epsilon_N\}_{N=N_0}^{\infty}$, *with* $\epsilon_N \downarrow 0$.

Step 0. Set $N = N_0$, *and* $\Omega_{j,N-1} = \emptyset$, *for all* $j \in \bar{\mathbf{q}}$.

Step 1. For $j \in \bar{\mathbf{q}}$, compute points

$$\hat{y}_{j,N} \in \widehat{Y}_{j,N}(x_N), \tag{3.67}$$

and set

$$\Omega_{j,N} = \Omega_{j,N-1} \cup \{\hat{y}_{j,N}\}. \tag{3.68}$$

Step 2. Set $\Omega_N = \Omega_{0,N} \times \cdots \times \Omega_{q,N}$. Use an optimization algorithm on Problem $\mathbf{P}_{\Omega_N}$ (where $\mathbf{P}_{\Omega_N}$ corresponds to problem (3.59) with $S = \Omega_N$) to compute an x_{N+1} such that

$$\theta_{\Omega_N}(x_{N+1}) \geq -\epsilon_N \tag{3.69}$$

and

$$\psi_{\Omega_N}(x_{N+1}) \leq \epsilon_N, \quad \|g(x_{N+1})\| \leq \epsilon_N. \tag{3.70}$$

Step 3. Replace N by $N+1$, and goto Step 1.

Theorem 14 (Theorem 3.6.5 in [91])**.** *Suppose that Assumption 2 and Assumption 3 are satisfied. If $\hat{x}$ is an accumulation point of a sequence $\{x_i\}_{i=0}^{\infty}$ constructed by Algorithm 3 in solving problem* **SIECP***, then $\theta(\hat{x}) = 0$, $\psi(\hat{x}) \leq 0$ and $g(\hat{x}) = 0$ (where $\theta(\cdot)$ is given as in (3.51) with $\delta = 1$).*

Proof. Suppose that, for some infinite subset $K \subset \mathbb{N}$, $x_N \to^K \hat{x}$, as $N \to \infty$. Then by Lemma 2, $\psi_{\Omega_N}(x_N) \to^K \psi(\hat{x})$ and $\psi^0_{\Omega_{0,N}}(x_N) \to^K \psi^0(\hat{x})$, as $N \to \infty$.
Next, let $\tilde{\psi}_{\Omega_N} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ be defined by

$$\tilde{\psi}_{\Omega_N}(x, x+h) := \max_{j \in \mathbf{q}} \max_{y \in \Omega_{j,N}} \left\{ \phi^j(x,y) + \langle \nabla_x \phi^j(x,y), h \rangle + \frac{1}{2}\|h\|^2 \right\}, \tag{3.71}$$

and $\tilde{\psi}^0_{\Omega_{0,N}} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ be defined by

$$\tilde{\psi}^0_{\Omega_{0,N}}(x, x+h) := \max_{y \in \Omega_{0,N}} \left\{ \phi^0(x,y) + \langle \nabla_x \phi^0(x,y), h \rangle + \frac{1}{2}\|h\|^2 \right\}, \tag{3.72}$$

corresponding to (3.61) and respectively to (3.62) with $j = 0$, and $S = \Omega_N$, $S_j = \Omega_{j,N}$, $j \in \bar{\mathbf{q}}$. Then, because $\Omega_{j,N} \subset Y_j$, $j \in \bar{\mathbf{q}}$, we see that, for all $N \in \mathbb{N}$ and any $h \in \mathcal{H}_E(x_N)$,

$$\tilde{\psi}_{\Omega_N}(x_N, x+h) \leq \tilde{\psi}_Y(x_N, x+h), \tag{3.73}$$

where $\tilde{\psi}_Y$ corresponds to (3.61) with $S = Y := Y_1 \times \cdots \times Y_q$ and

$$\tilde{\psi}^0_{\Omega_{0,N}}(x_N, x+h) \leq \tilde{\psi}^0_{Y_0}(x_N, x+h), \tag{3.74}$$

where $\tilde{\psi}^0_{Y_0}$ corresponds to (3.62) with $j = 0$ and $S_0 = Y_0$. Hence for θ_{Ω_N} corresponding to (3.60) with $S = \Omega_N$ and θ_Y corresponding to (3.60) with $S = Y$, we see that

$$\begin{aligned}\theta_{\Omega_N}(x_N) = \min_{h \in \mathcal{H}_E(x_N)} \max \Big\{ & \tilde{\psi}^0_{\Omega_{0,N}}(x_N, x_N+h) - \psi^0_{\Omega_{0,N}}(x_N) - \gamma \psi_{\Omega_N}(x_N)_+, \\ & \tilde{\psi}_{\Omega_N}(x_N, x_N+h) - \psi_{\Omega_N}(x_N)_+ \Big\} \\ \leq \theta_Y(x_N) + \psi^0_{Y_0}(x_N) - \psi^0_{\Omega_{0,N}}(x_N) & + \max\{\gamma, 1\} \cdot [\psi_Y(x_N)_+ - \psi_{\Omega_N}(x_N)_+] \,.\end{aligned} \tag{3.75}$$

It now follows from the continuity of $\theta(\cdot)$, Lemma 2, and the test (3.69) that,

$$0 = \underline{\lim}\, \theta_{\Omega_N}(X_N) \leq \theta_Y(\hat{x}) = \theta(\hat{x}) \leq 0, \tag{3.76}$$

so that we must have $\theta(\hat{x}) = 0$, and due to the tests (3.70) we must have $\psi(\hat{x}) \leq 0$ and $g(\hat{x}) = 0$, which completes the proof. □

CHAPTER 4

Automatic Differentiation

In this chapter we give a brief introduction to Automatic Differentiation (AD) utilizing truncated Taylor series propagation in *forward* and *reverse* mode.
In [53], Griewank states :
"*Full higher derivative tensors in several variables can be evaluated directly by multivariate versions of the chain rule. This approach has been implemented in [82] and [21], and by several other authors. They have given particular thought to the problem of addressing the roughly* $p^d/d!$ *distinct elements of a symmetric tensor of order* $p \leq n$ *and degree* d *efficiently*[1]*. Since this problem cannot be solved entirely satisfactorily, we advocate propagating a family of roughly* $p^d/d!$ *univariate Taylor series instead.*"
Since the implemented numerical methods shall be generally applicable to allow further use, we adopt his advice and rely on truncated Taylor series propagation for AD, as well. We first introduce the basic concepts, i.e. the *forward* (Section 4.1) and the *reverse* mode (Section 4.2). These methods are implemented in the sophisticated AD package `CppAD` [19, 18], which is used for the numerical calculation of derivatives in this work.
Hereafter, we treat the special case of AD of solutions of parametrized nonlinear equations, i.e. implicitly defined functions in Section 4.3.

[1] Here, n denotes the dimension of the domain space and d denotes the order of the derivatives.

4.1. Forward mode of Automatic Differentiation

In this section we stay close to the presentation in [53]. Be

$$F : \mathbb{R}^n \to \mathbb{R}^m \tag{4.1}$$

a d-times continuously differentiable function with $0 \leq d \leq \infty$. We are interested in the k-th derivative, $k \leq d$ of $y = F(x)$ at a point $x \in \mathbb{R}^n$ and with $y \in \mathbb{R}^m$. We call y the *dependent* and x the *independent* variable. We will assume that the function $F(\cdot)$ can be evaluated utilizing a sequence of elemental functions ϕ^i, $i \in \{1, \ldots, l\}$, where ϕ^i is either $\phi^i : \mathbb{R}^2 \to \mathbb{R}$ or $\phi^i : \mathbb{R} \to \mathbb{R}$, by the following scheme:

Be $v^i \in \mathbb{R}$, $i \in \{1-n, \ldots, 0, 1, \ldots, l\}$ *intermediate* variables. The values of the *intermediate* variables v^{i-n}, $i \in \{1, \ldots, n\}$ are given by

$$v^{i-n} = x^i, \quad i \in \{1, \ldots, n\},$$

thus v^{i-n}, $i \in \{1, \ldots, n\}$ stores the values of the *independent* variable x. Now be $i \in \{1, \ldots, l\}$ then v^i is either given by

$$v^i = \phi^i(v^j)$$

or

$$v^i = \phi^i(v^j, v^k),$$

where $j, k < i$, such that for

$$y^{m+i} = v^{l+i}, \quad i \in \{1-m, \ldots, 0\},$$

one has $y = F(x)$ for all $x \in \mathbb{R}^n$. Since v^i, $i \in \{1, \ldots, l\}$ only depend on former *intermediate* variables v^j and v^k, $j, k < i$, all v^i can be successively evaluated using only elemental functions ϕ^i.

Assumption 4 (Elemental Differentiability, Assumption ED in [53])**.** *All elemental functions ϕ^i are d-times continuously differentiable on their open domains $\mathcal{D}_i$, i.e. $\phi^i \in \mathcal{C}^d(\mathcal{D}_i)$.*

Now consider a given univariate polynomial in t

$$x(t) = x_0 + x_1 t + x_2 t^2 + \cdots + x_d t^d \in \mathbb{R}^n. \tag{4.2}$$

We are interested in the resulting expansion

$$y(t) \equiv y_0 + y_1 t + y_2 t^2 + \cdots + y_d t^d = F(x(t)) + \mathcal{O}(t^{d+1}) \in \mathbb{R}^m, \tag{4.3}$$

since at $x^* \in \mathbb{R}^n$ for a "suitable" polynomial $x(t)$ with $x_0 = x^*$, y_k contains the "desired" derivative information of $F(\cdot)$ at x^*, which will be worked out further below.

Definition 12 (Taylor coefficient functions, Definition TC in [53])**.** *Under Assumption 4, let*

$$y_k = F_k(x_0, x_1, \ldots, x_k) \quad \textit{with} \quad F_k : \mathbb{R}^{n\times(k+1)} \to \mathbb{R}^m,$$

denote for $k \leq d$ the coefficient function defined by the relations (4.2) and (4.3). When more than $k+1$ vector arguments are supplied the extra ones are ignored and when fewer are supplied the missing ones default to zero.

Definition 13 (Approximation of intermediates, Definition AI in [53])**.** *For a given d-times continuously differentiable input path $x(t) : (-\epsilon, \epsilon) \to \mathbb{R}^n$ with $\epsilon > 0$, denote the resulting values of an intermediate variable v by $v(x(t))(-\epsilon, \epsilon) \to \mathbb{R}$ and define its Taylor polynomial by*

$$v(t) = v_0 + v_1 t + v_2 t^2 + \cdots + v_d t^d = v(x(t)) + o(t^d).$$

Since by assumption $F(\cdot)$ can be evaluated utilizing elemental functions ϕ^i, one can apply truncated polynomial arithmetics to the elemental evaluation scheme presented above to calculate the expansion of $y(t)$, if ϕ^i is given by elemental arithmetic operations. Now be v, u and w *intermediate* variables, the recurrences for arithmetic operations in respect of the approximation of *intermediates* are given in Table 4.1. A similar table can be found in [53].
All univariate elemental functions $\phi(u)$ of interest can be expressed as solutions of linear Ordinary Differential Equations (ODEs):

$$b(u)\phi'(u) - a(u)\phi(u) = c(u),$$

where the coefficient functions $a(u)$, $b(u)$ and $c(u)$ are given for a specific univariate elemental function $\phi(u)$ and the Taylor coefficients a_k, b_k and c_k of $a(u)$, $b(u)$ and $c(u)$, respectively can be calculated from the Taylor coefficients u_k of u.

Proposition 4 (Taylor polynomials of ODE solutions, Proposition 10.1 in [53])**.** *Pro-*

$v =$	Recurrence for $k = 1, \ldots, d$
$u + cw$	$v_k = u_k + cw_k$
$u \cdot w$	$v_k = \sum_{j=0}^{k} u_j w_{k-j}$
u/w	$v_k = \frac{1}{w_0}\left[u_k - \sum_{j=0}^{k-1} v_j w_{k-j}\right]$
u^2	$v_k = \sum_{j=0}^{k} u_j u_{k-j}$

Table 4.1.: Taylor Coefficient Propagation via Arithmetic Operations

vided $b_0 \equiv b(u_0) \neq 0$ *one has*

$$\tilde{v}_k = \frac{1}{b_0}\left[\sum_{j=1}^{k}(c_{k-j} + e_{k-j})\tilde{u}_j - \sum_{j=1}^{k-1} b_{k-j}\tilde{v}_j\right], \quad \text{for} \quad k \in \{1, \ldots, d\},$$

where

$$e_k \equiv \sum_{j=0}^{k} a_j v_{k-j}, \quad \text{for} \quad k \in \{0, \ldots, d-1\},$$

and

$$\tilde{v}_j \equiv j v_j, \quad j \in \{1, \ldots, d\},$$
$$\tilde{u}_j \equiv j u_j, \quad j \in \{1, \ldots, d\}.$$

A proof of this proposition is given in [53].

Table 4.2 gives a list of the coefficient functions $a(u)$, $b(u)$ and $c(u)$ for the standard univariate elemental functions and as well the resulting recurrences, which can be calculated using Proposition 4. A similar table can be found in [53].
This recurrence formulas can be applied to the elemental evaluation scheme in the same manner as the recurrence formulas for arithmetic operations.

We can summarize the *forward* mode approach of AD in the following algorithmic scheme.

$v =$	a	b	c	Recurrence for $k = 1, \ldots, d$
$\ln(u)$	0	u	1	$\tilde{v}_k = \frac{1}{u_0}\left[\tilde{u}_k - \sum_{j=1}^{k-1} u_{k-j}\tilde{v}_j\right]$
$\exp(u)$	1	1	0	$\tilde{v}_k = \left[\sum_{j=1}^{k} v_{k-j}\tilde{u}_j\right]$
u^r	r	u	0	$\tilde{v}_k = \frac{1}{u_0}\left[r\sum_{j=1}^{k} v_{k-j}\tilde{u}_j - \sum_{j=1}^{k-1} u_{k-j}\tilde{v}_j\right]$
$\sin(u)$	0	1	$cos(u)$	$\tilde{v}_k = \left[\sum_{j=1}^{k} \tilde{u}_j c_{k-j}\right]$
$\cos(u)$	0	-1	$sin(u)$	$\tilde{v}_k = \left[\sum_{j=1}^{k} -\tilde{u}_j c_{k-j}\right]$
$\sqrt{u}$	0	$\sqrt{u}$	$\frac{1}{2}$	$v_0 = \sqrt{u_0}$ $\tilde{v}_k = \frac{1}{v_0}\left[\frac{k}{2}u_k - \sum_{j=1}^{k-1} \tilde{v}_j v_{k-j}\right]$

Table 4.2.: Taylor Coefficient Propagation through Univariate Elementals

Algorithm 4.

Data: A d-times continuously differentiable function $F : \mathbb{R}^n \to \mathbb{R}^m$, which can be evaluated utilizing elemental functions ϕ_i, $i \in \{1, \ldots, l\}$ fulfilling Assumption 4, a univariate polynomial $x(t) = x_0 + x_1 t + x_2 t^2 + \cdots + x_d t^d \in \mathbb{R}^n$.

Step 0. Set $N = 1$, and the polynomials

$$v^{i-n}(t) = x^i(t), \quad i \in \{1, \ldots, n\}.$$

Step 1. Set the polynomials $v^N(t)$ according to the recurrence given by the elemental function ϕ^N.
Step 2. If $N < l$ set $N = N + 1$ and goto Step 1., else goto Step 3.
Step 3. Set the polynomials

$$y^{m+i}(t) = v^{l+i}(t), \quad i \in \{1 - m, \ldots, 0\},$$

and stop.

The following proposition gives a result, which enables the construction of any "desired" derivative tensor $\nabla^k F(x)$ of order k from the propagation of univariate Taylor polynomials of degree $k \leq d$ through the elemental evaluation scheme above.

Proposition 5 (Taylor to tensor conversion, Proposition 10.2 in [53])**.** *Let $F : \mathbb{R}^n \to \mathbb{R}^m$ be at least d-times continuously differentiable at some point $x \in \mathbb{R}^n$ and denote by $F_r(x, s)$ the r-th Taylor coefficient of the curve $F(x + ts)$ at $t = 0$ for some direction $s \in \mathbb{R}^n$. Then one has for any seed matrix $S = [s_j]_{j=1}^p \in \mathbb{R}^{n \times p}$ with $s_j \in \mathbb{R}^n$ and any multi-index $\mathbf{i} \in \mathbb{N}_+^p$ with $|\mathbf{i}| \leq d$ the identity*

$$\left.\frac{\partial^{|\mathbf{i}|} F(x + z^1 s_1 + z^2 s_2 + \cdots + z^p s_p)}{(\partial z^1)^{\mathbf{i}_1} (\partial z^2)^{\mathbf{i}_2} \ldots (\partial z^p)^{\mathbf{i}_p}}\right|_{z=0} = \sum_{|\mathbf{j}|=d} \gamma_{\mathbf{ij}} F_{|\mathbf{i}|}(x, S\mathbf{j}),$$

where the constant coefficients $\gamma_{\mathbf{ij}}$ are given by the finite sums

$$\gamma_{\mathbf{ij}} \equiv \sum_{0<\mathbf{k}\leq\mathbf{i}} (-1)^{|\mathbf{i}-\mathbf{k}|} \binom{\mathbf{i}}{\mathbf{k}} \binom{d\mathbf{k}/|\mathbf{k}|}{\mathbf{j}} \left(\frac{|\mathbf{k}|}{d}\right)^{|\mathbf{i}|},$$

$\mathbf{i}$, $\mathbf{j}$, $\mathbf{k} \in \mathbb{N}_+^p$ denote multi-indices, $z \in \mathbb{R}^p$ and for any $m \in \mathbb{R}^p$, $\mathbf{l} \in \mathbb{N}_+^p$

$$\binom{m}{\mathbf{l}} \equiv \binom{m^1}{\mathbf{l}^1} \binom{m^2}{\mathbf{l}^2} \cdots \binom{m^p}{\mathbf{l}^p},$$

and for $\alpha \in \mathbb{R}$, $l \in \mathbb{N}_+$,

$$\binom{\alpha}{l} \equiv \begin{cases} \dfrac{\alpha(\alpha-1)(\alpha-2)\ldots(\alpha-l+1)}{l!} & if \quad l > 0 \\ 1 & if \quad l = 0. \end{cases}$$

A proof of this proposition is given in [55].

4.2. Reverse mode of Automatic Differentiation

In this section we stay close to the presentation in [38]. For the matter of notational simplicity we treat the function of interest $F : \mathbb{R}^n \to \mathbb{R}$ to be smooth. It should be noted that now $m = 1$ in view of (4.1).

Theorem 15. *Let u be a univariate Taylor series in t*

$$u(t) = u_0 + u_1 t + u_2 t^2 + u_3 t^3 + \cdots \in \mathbb{R}$$

and let

$$v(t) \equiv v_0 + v_1 t + v_2 t^2 + v_3 t^3 + \cdots = f(u(t)) \in \mathbb{R}$$

for some smooth function $f : \mathbb{R} \to \mathbb{R}$. *Then for all* $k \geq 0$, $p > 0$ *we have*

$$\frac{\partial v_{k+p}}{\partial u_p} = \frac{\partial v_k}{\partial u_0}, \qquad \frac{\partial v_k}{\partial u_{k+p}} = 0.$$

A proof of this theorem is given in [38].

Now let $y \in \mathbb{R}$ be the *dependent* variable and $x \in \mathbb{R}^n$ the *independent* one with $y = F(x)$, as in Section 4.1.
We again assume that the function $F(\cdot)$ of interest can be evaluated utilizing elemental functions ϕ^i, $i \in \{1, \ldots, l\}$, where ϕ^i is either $\phi^i : \mathbb{R}^2 \to \mathbb{R}$ or $\phi^i : \mathbb{R} \to \mathbb{R}$.
Let u, v and w be *intermediate* variables with v is dependent on u, i.e. $v = f(u)$ where $f : \mathbb{R} \to \mathbb{R}$ is a smooth function.
We treat again the case that $x(t)$ is a univariate Taylor series in t with $u(t)$, $v(t)$, $w(t)$ are the corresponding approximations of *intermediates* and $y(t)$ is the resulting *dependent* one.
Be $s(t)$ any (truncated) Taylor series in t, i.e.

$$s(t) \equiv s_0 + s_1 t + s_2 t^2 + \dots \ ,$$

we define

$$[s(t)]_k := s_k.$$

In the following we are interested in the calculation of the Taylor coefficients of $F_x(x(t))$ for given Taylor series $x(t) = x_0 + x_1 t + x_2 t^2 + \dots$.

Since $\frac{\partial x(t)}{\partial x_0} = 1$, for $F_x(x(t))$ it holds that

$$F_x(x(t)) = F_x(x(t))\frac{\partial x(t)}{\partial x_0} = \frac{\partial y(t)}{\partial x_0} = \frac{\partial y_0}{\partial x_0} + \frac{\partial y_1}{\partial x_0} t + \frac{\partial y_2}{\partial x_0} t^2 + \dots \ . \tag{4.4}$$

Remark. *Having the Taylor coefficients of* $F_x(x(t))$ *for given Taylor series* $x(t)$ *one can apply Proposition 5 to* $F_x(\cdot)$ *and therefore one gains one derivative order.*

Under Assumption 4 and in view of Definition 12 and Definition 13, we define a Taylor coefficient function of *intermediates*.

Definition 14 (Taylor coefficient functions of intermediates)**.** *Under Assumption 4 and*

for given intermediate Taylor polynomial $v(t)$, let

$$y_k = F_k^v(v_0, v_1, \ldots, v_k) \quad \textit{with} \quad F_k : \mathbb{R}^{k+1} \to \mathbb{R},$$

denote for k the coefficient function defined by the relations (4.2) and (4.3) and Definition 13.

Now we can define the Taylor series $\bar{v}(t)$ by

$$\bar{v}(t) := \frac{\partial y(t)}{\partial v_0} \qquad \text{i.e.} \quad \bar{v}_k = \frac{\partial y_k}{\partial v_0} = \frac{\partial y_{k+p}}{\partial v_p} \quad \text{for} \quad k, p \geq 0.$$

Then for the Taylor series $\bar{u}(t) = \dfrac{\partial y(t)}{\partial u_0}$ we get

$$\bar{u}_p = \sum_{k \leq p} \frac{\partial y_p}{\partial v_k} \frac{\partial v_k}{\partial u_0} = \sum_{k \leq p} \frac{\partial y_{p-k}}{\partial v_0} \frac{\partial v_k}{\partial u_0} = \sum_{k \leq p} \bar{v}_{p-k} [f_u(u(t))]_k = [\bar{v}(t) * f_u(u(t))]_p, \qquad (4.5)$$

for $p \geq 0$ and where $\bar{v}(t) * f_u(u(t))$ denotes the polynomial product of $\bar{v}(t)$ and $f_u(u(t))$, namely

$$[\bar{v}(t) * f_u(u(t))]_k = \sum_{j=0}^{k} \bar{v}_j [f_u(u(t))]_{k-j}, \quad \text{for} \quad k \geq 0,$$

and $f_u(u(t))$ is considered as Taylor series.
Applying this relation to the elemental evaluation scheme we can associate to any univariate elemental function ϕ^i the reverse accumulation step

$$\bar{u}(t) += \bar{v}(t) * \phi_u^i(u(t)),$$

with $\phi_u^i(u(t))$ considered as Taylor series.
For the bivariate case with congruent thoughts one can associate to any bivariate elemental function ϕ^i the reverse accumulation steps

$$\bar{u}(t) += \bar{v}(t) * \phi_u^i(u(t), w(t))$$

and

$$\bar{w}(t) += \bar{v}(t) * \phi_w^i(u(t), w(t)).$$

The reverse accumulation steps for the elementary arithmetic operations are given in Table 4.3.

In the following we give a algorithmic scheme for the calculation of the Taylor coef-

$v =$	Reverse accumulation step
$u + cw$	$\bar{u}_k \mathrel{+}= \bar{v}_k$ $\bar{w}_k \mathrel{+}= c\bar{v}_k$
$u \cdot w$	$\bar{u}_k \mathrel{+}= \sum_{j=0}^{k} \bar{v}_j [w(t)]_{k-j}$ $\bar{w}_k \mathrel{+}= \sum_{j=0}^{k} \bar{v}_j [u(t)]_{k-j}$
u/w	$\bar{u}_k \mathrel{+}= \bar{v}_k \frac{1}{w_0}$ $\bar{w}_k \mathrel{+}= \sum_{j=0}^{k} \bar{v}_j \frac{\partial v_{k-j}}{\partial w_0}$ with $\frac{\partial v_k}{\partial w_0} = -\frac{1}{w_0^2}\left[u_k - \sum_{j=0}^{k-1} v_j w_{k-j}\right] - \frac{1}{w_0}\sum_{j=0}^{k-1} \frac{\partial v_j}{\partial w_0} w_{k-j}$
u^2	$\bar{u}_k \mathrel{+}= 2\sum_{j=0}^{k} \bar{v}_j [u(t)]_{k-j}$

Table 4.3.: Reverse Accumulation Steps for the Elementary Arithmetic Operations

ficients of $F_x(x(t))$.

Algorithm 5.

Data: A $(d+1)$-times continuously differentiable function $F : \mathbb{R}^n \to \mathbb{R}$, which can be evaluated utilizing elemental functions ϕ_i, $i \in \{1, \ldots, l\}$ fulfilling Assumption 4, a univariate polynomial $x(t) = x_0 + x_1 t + x_2 t^2 + \cdots + x_d t^d \in \mathbb{R}^n$ and a corresponding dependent polynomial $y(t) = y_0 + y_1 t + y_2 t^2 + \cdots + y_d t^d \in \mathbb{R}$ fulfilling relation (4.3). (Under assumption that the intermediate Taylor polynomials $v^i(t)$, $i \in \{1-n, \ldots, l\}$ are on hand.)

Step 0. Set $N = l - 1$, and the polynomials

$$\bar{v}^{i-n}(t) = 0, \quad i \in \{1, \ldots, n\},$$
$$\bar{v}^{i}(t) = 0, \quad i \in \{1, \ldots, l-1\},$$

and

$$\bar{v}^l(t) = 1, \quad \text{i.e.} \quad \bar{v}_0^l = 1,\ \bar{v}_i^l = 0,\ i \geq 1.$$

Step 1. Apply the reverse accumulation step to the $\bar{v}^i(t)$, where i is associated to the argument(s) of the elemental function ϕ^N.
Step 2. If $N > 1$ set $N = N - 1$ and goto Step 1., else goto Step 3.
Step 3. Set the polynomial $\bar{x}(t)$ to

$$\bar{x}^i(t) = \bar{v}^{i-n}(t), \quad i \in \{1, \ldots, n\},$$

and stop.

By now we assumed that $F : \mathbb{R}^n \to \mathbb{R}$, i.e. $m = 1$ in view of (4.1) for the reverse case. To apply the reverse mode to $F : \mathbb{R}^n \to \mathbb{R}^m$ with $m > 1$ we introduce the reverse seed vector $\omega \in \mathbb{R}^m$ and define $\widetilde{F} : \mathbb{R}^n \to \mathbb{R}$ by

$$\widetilde{F}(x) = \omega^1 F^1(x) + \cdots + \omega^m F^m(x), \quad x \in \mathbb{R}^n. \tag{4.6}$$

The reverse mode is now applied on $\widetilde{F}(x)$.

4.3. Automatic Differentiation of implicitly defined functions

In this section, suppose that $F : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^m$ is a d-times continuously differentiable function, which can be evaluated utilizing elemental functions ϕ_i, $i \in \{1, \ldots, l\}$ as above and therefore we can apply AD as described in Section 4.1 and 4.2 on it.
Now again, let $y \in \mathbb{R}^m$ be the *dependent* variable and $x \in \mathbb{R}^n$ the *independent* one. The relation between y and x may be implicitly given by the nonlinear equation,

$$F(y, x) = 0 \quad \text{with} \quad F : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^m. \tag{4.7}$$

Further, assume that for given $x^* \in \mathbb{R}^n$ there exist a $y^* \in \mathbb{R}^m$ such that (4.7) is fulfilled and in addition that $F_y(y^*, x^*)$ is nonsingular. According to the Implicit Function Theorem (Theorem 26), there exist $\rho_x, \rho_y > 0$ and a d-times continuously differentiable function $\Phi : B(x^*, \rho_x) \to B(y^*, \rho_y)$ such that $\Phi(x^*) = y^*$ and $F(\Phi(x), x) = 0$ for all $x \in B(x^*, \rho_x)$. In the following, we are interested in the calculation of (higher) derivatives of $\Phi(\cdot)$ at x^*.

Below, we present two approaches to calculate the desired derivatives.

- In the first one, AD is not applied on $\Phi(\cdot)$ at x^* directly but on Newton's method,

an *iterative process*, which can be used to numerically calculate for a given $x^* \in \mathbb{R}^m$ the corresponding $y^* \in \mathbb{R}^n$ such that (4.7) is fulfilled up to a given stopping accuracy, instead. By doing so, Gilbert notes in [51] that "*by good behavior, we mean that the derivatives will be calculated correctly, asymptotically*". Therefore we refer to it as *iterative* mode. The theoretical validation of this approach is described in Section 4.3.1.

- In the second one, the *direct* mode presented in Section 4.3.2, we directly calculate the Taylor coefficients of $\Phi(\cdot)$ at x^* according to Section 4.1 and Section 4.2, and by use of the Implicit Function Theorem.

4.3.1. Iterative mode

In this subsection, for a clearer presentation we restrict ourselves to first order sensitivities. For the case of higher derivatives we refer to [54]. We stay close to the presentation in [51].

In [51], Gilbert calls "*an iterative process a part of a computer program whose instructions are executed several times until a stopping criterion is reached*". Here, the *iterative process* $\phi : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^m$ of interest, has the following properties:

- On an open set $\mathcal{W}_y \times \mathcal{W}_x \subset \mathbb{R}^m \times \mathbb{R}^n$, $\phi(\cdot, \cdot)$ is continuously differentiable and there exists a $L > 0$ such that for $(y', x'), (y'', x'') \in \mathcal{W}_y \times \mathcal{W}_x$, the directional derivatives $\mathrm{d}\phi(y', x'; \cdot)$ and $\mathrm{d}\phi(y'', x''; \cdot)$ (Definition 24) satisfy

$$\max_{h \in \mathbb{R}^m \times \mathbb{R}^n} \frac{\|\mathrm{d}\phi(y', x'; h) - \mathrm{d}\phi(y'', x''; h)\|}{\|h\|} \leq L \|(y', x') - (y'', x'')\|. \tag{4.8}$$

- For given $x \in \mathcal{W}_x$, the next iterate y_{k+1} is given by

$$y_{k+1} = \phi(y_k, x), \quad k \geq 0, \tag{4.9}$$

  while x is constant.

- The initial iterate y_0 depends on x such that

$$y_{k+1} = \phi(y_k(x), x), \quad k \geq 0. \tag{4.10}$$

- It holds that

$$\lim_{k \to \infty} y_k(x) \to y_*(x), \tag{4.11}$$

thus

$$y_*(x) = \phi(y_*(x), x) \tag{4.12}$$

is a fixed point of $\phi(\cdot, x)$.

- For all $k \geq 0$ it holds that $y_k(x) \in \mathcal{W}_y$ and $y^*(x) \in \mathcal{W}_y$.

Proposition 6 (Proposition 1 in [51]). *Suppose that $\phi(\cdot,\cdot)$ is an iterative process as above. Suppose also that the initial iterate y_0 is a differentiable function of x on $\mathcal{W}_x$. If the spectral radius ρ of $\mathrm{d}\phi(y'', x''; \cdot)$, where the direction h is restricted to $h \in \mathcal{W}_y \times \{0\}$, denoted by $\mathrm{d}\phi_y(y'', x''; \cdot)$, satisfies*

$$\rho(\mathrm{d}\phi_y(y'', x''; \cdot)) < \tau < 1,$$

where ρ of $\mathrm{d}\phi_y(y'', x''; \cdot)$ is given by

$$\rho(\mathrm{d}\phi_y(y'', x''; \cdot)) := \lim_{k\to\infty} \left(\max_{h \in \mathcal{W}_y \times \{0\}} \frac{\|\mathrm{d}\phi_y^k(y'', x''; h)\|}{\|h\|} \right)^{\frac{1}{k}},$$

and $\mathrm{d}\phi_y^k(y'', x''; h)$ denotes the k-th function iteration of $\mathrm{d}\phi_y(y'', x''; \cdot)$ on h,

then

(i) the convergence of the sequence $\{y_k\}_{k\geq 0}$ is asymptotically linear; that is, there exist an index k_0 such that $\|y_{k+1} - y_*\| \leq \tau \|y_k - y_*\|$, for all indices $k \geq k_0$ and

(ii) the sequence of derivatives $\left\{ \frac{\mathrm{d}y_k(x)}{\mathrm{d}x} \right\}_{k\geq 0}$ converges to $\frac{\mathrm{d}y_*(x)}{\mathrm{d}x}$.

A proof of this proposition is given in [51].

In the following, we will apply the preceding Proposition 6 to Newton's method for the calculation of a corresponding $y^* \in \mathbb{R}^m$ for given $F : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^m$ as above and given $x^* \in \mathbb{R}^n$ such that (4.7) is fulfilled, i.e.

$$F(y^*, x^*) = 0. \tag{4.13}$$

Applying Newton's method to (4.13), we get for successive iterates $y_k \in \mathbb{R}^m$

$$y_{k+1} = y_k - F_y(y_k, x^*)^{-1} F(y_k, x^*) \quad \text{for} \quad k \geq 0.$$

Therefore, the *iterative process* $\tilde{\phi} : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^m$ of interest is given by

$$\tilde{\phi}(y, x) := y - F_y(y, x)^{-1} F(y, x).$$

The Jacobian of $\tilde{\phi}_y(\cdot, x^*)$ at $y^* \in \mathbb{R}^m$ for given $x^* \in \mathbb{R}^n$ in respect of y is given by

$$\begin{aligned}\tilde{\phi}_y(y^*, x^*) &= 1 + F_y(y^*, x^*)^{-1} F_{yy}(y^*, x^*) F_y(y^*, x^*)^{-1} F(y^*, x^*) - F_y(y^*, x^*)^{-1} F_y(y^*, x^*) \\ &= 0.\end{aligned}$$

It follows that the directional derivative $\mathrm{d}\tilde{\phi}_y(y^*, x^*; \cdot))$ at (y^*, x^*) for any direction $(h_y, h_x) \in \mathbb{R}^m \times \{0\}$ is given by

$$\mathrm{d}\tilde{\phi}_y(y^*, x^*; h) = \tilde{\phi}_y(y^*, x^*) h_y = 0.$$

For the spectral radius ρ of $\mathrm{d}\tilde{\phi}_y(y^*, x^*; \cdot)$, it immediately follows that

$$\rho(\mathrm{d}\tilde{\phi}_y(y^*, x^*; \cdot)) = 0.$$

Consequently, the sequence $\left\{ \frac{\mathrm{d}y_k(x)}{\mathrm{d}x} \right\}_{k \geq 0}$ will converge to $\frac{\mathrm{d}y_*(x)}{\mathrm{d}x}$ assuming that the *iterative process* $\tilde{\phi} : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^m$ fulfills the conditions (4.8)–(4.12) and the initial iterate y_0 is a differentiable function of x in some environment around x^*.

In summery and view of [54] under appropriate conditions on the *iterative process*, we can apply AD as presented in Section 4.1 and 4.2 directly to the *iterative process* for the solution of (4.7) for given $x^* \in \mathbb{R}^n$ and can expect that the desired derivatives are asymptotically correct.

4.3.2. Direct mode

In the *direct* mode approach, the task is to calculate for a given $F : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^m$ as above at $x^* \in \mathbb{R}^n$ for a given univariate polynomial

$$x(t) = x_0 + x_1 t + x_2 t^2 + \cdots + x_d t^d \in \mathbb{R}^n, \tag{4.14}$$

in t, with

$$x(0) = x^*, \quad \text{i.e. } x_0 = x^*,$$

such that there exist a corresponding $y^* \in \mathbb{R}^m$ with

$$F(y^*, x^*) = 0$$

and $F_y(y^*, x^*)$ is nonsingular, the resulting expansion

$$y(t) \equiv y_0 + y_1 t + y_2 t^2 + \cdots + y_d t^d = \Phi(x(t)) + \mathcal{O}(t^{d+1}) \in \mathbb{R}^m. \tag{4.15}$$

Obviously y_0 is given by $y_0 = y^*$.
Since the requirements on the Implicit Function Theorem (Theroem 26) are fulfilled, there exist $\rho_x, \rho_y > 0$ such that $\Phi : B(x^*, \rho_x) \to B(y^*, \rho_y)$ exists in some environment around x^* and $\Phi(\cdot)$ is d-times continuously differentiable.
Therefore, the k-th total derivative with $k \leq d$ of $F(y(t), x(t))$ in respect of t, for all t such that $x(t) \in B(x^*, \rho_y)$, is well defined and it holds that

$$\left.\frac{\mathrm{d}^k F(\Phi(x(t)), x(t))}{\mathrm{d}t^k}\right|_{t=0} = 0. \tag{4.16}$$

We first present the calculation of y_1. Hereafter, we give a scheme to successively calculate higher order Taylor coefficients y_i with $2 \leq i \leq d$.

We start with (4.16) and $k = 1$. Therefore, we have that

$$\left.\frac{\mathrm{d}F(\Phi(x(t)), x(t))}{\mathrm{d}t}\right|_{t=0} = F_y(y^*, x^*) \left.\frac{\mathrm{d}\Phi(x(t))}{\mathrm{d}t}\right|_{t=0} + F_x(y^*, x^*) \left.\frac{\mathrm{d}x(t)}{\mathrm{d}t}\right|_{t=0} = 0. \tag{4.17}$$

According to (4.15), we have

$$\left.\frac{\mathrm{d}\Phi(x(t))}{\mathrm{d}t}\right|_{t=0} = \left.\frac{\mathrm{d}y(t)}{\mathrm{d}t}\right|_{t=0} = y_1. \tag{4.18}$$

From (4.17) and (4.18), we get

$$y_1 = -F_y(y^*, x^*)^{-1} F_x(y^*, x^*) \left.\frac{\mathrm{d}x(t)}{\mathrm{d}t}\right|_{t=0}.$$

Now, consider two "input" polynomials $\tilde{y}^{(1)}(t)$ and $\tilde{x}^{(1)}(t)$ given by

$$\tilde{y}^{(1)}(t) = y_0 \in \mathbb{R}^m \tag{4.19}$$

and

$$\tilde{x}^{(1)}(t) = x_0 + x_1 t \in \mathbb{R}^n. \tag{4.20}$$

For the resulting expansion of $F(\tilde{y}^{(1)}(t), \tilde{x}^{(1)}(t))$ given by

$$\widetilde{F}^{(1)}(t) \equiv \widetilde{F}_0^{(1)} + \widetilde{F}_1^{(1)} t = F(\tilde{y}^{(1)}(t), \tilde{x}^{(1)}(t)) + \mathcal{O}(t^2) \in \mathbb{R}^m,$$

one directly sees that

$$F_x(y^*, x^*) \left.\frac{\mathrm{d}x(t)}{\mathrm{d}t}\right|_{t=0} = \widetilde{F}_1^{(1)},$$

since $\frac{\mathrm{d}x(t)}{\mathrm{d}t}\Big|_{t=0} = x_1$. Therefore, y_1 can be easily calculated by

$$y_1 = -F_y(y^*, x^*)^{-1} \widetilde{F}_1^{(1)}.$$

Now assume that we have calculated y_0, y_1,...,y_{k-1} and we are interested in calculating y_k for $1 < k \leq d$.
For this case the "input" polynomials $\tilde{y}^{(k)}(t)$ and $\tilde{x}^{(k)}(t)$ shall now be given by

$$\tilde{y}^{(k)}(t) = y_0 + y_1 t + \cdots + y_{k-1} t^{k-1} \in \mathbb{R}^m \tag{4.21}$$

and

$$\tilde{x}^{(k)}(t) = x_0 + x_1 t + \cdots + x_k t^k \in \mathbb{R}^n. \tag{4.22}$$

For the resulting expansion of $F(\tilde{y}^{(k)}(t), \tilde{x}^{(k)}(t))$ given by

$$\widetilde{F}^{(k)}(t) \equiv \widetilde{F}_0^{(k)} + \widetilde{F}_1^{(k)} t + \cdots + \widetilde{F}_k^{(k)} t^k = F(\tilde{y}^{(k)}(t), \tilde{x}^{(k)}(t)) + \mathcal{O}(t^{k+1}) \in \mathbb{R}^m,$$

it directly follows with (4.16) and Taylor's Theorem that

$$k! \cdot \widetilde{F}_k^{(k)} + F_y(y^*, x^*) \left.\frac{\mathrm{d}^k \Phi(x(t))}{\mathrm{d}t^k}\right|_{t=0} = \left.\frac{\mathrm{d}^k F(\Phi(x(t)), x(t))}{\mathrm{d}t^k}\right|_{t=0} = 0,$$

where $k!$ denotes the k-th factorial.
Since by (4.15) $\left.\frac{\mathrm{d}^k \Phi(x(t))}{\mathrm{d}t^k}\right|_{t=0} = k! \cdot y_k$, it follows that y_k can be calculated by

$$y_k = -F_y(y^*, x^*)^{-1} \widetilde{F}_k^{(k)}.$$

This leads to following iterative algorithm for the successive calculation of y_k, $1 \leq k \leq d$.

Algorithm 6.

Data: A d-times continuously differentiable function $F : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^m$, which can be evaluated utilizing elemental functions ϕ_i fulfilling Assumption 4, $i \in \{1, \ldots, l\}$ as above, $x^* \in \mathbb{R}^n$, $y^* \in \mathbb{R}^m$ such that $F(y^*, x^*) = 0$ and $F_y(y^*, x^*)$ is nonsingular, the inverse Jacobian matrix $F_y(y^*, x^*)^{-1}$, a univariate polynomial $x(t) = x_0 + x_1 t + x_2 t^2 + \cdots + x_d t^d \in \mathbb{R}^n$ with $x(0) = x_0 = x^*$, $k \in \{1, \ldots, d\}$.

Step 0. Set $N = 1$, and $y_0 = y^*$.

Step 1. Set "input" polynomials to

$$\tilde{y}^{(N)}(t) = \sum_{i=0}^{N-1} y_i t^i \in \mathbb{R}^m$$

and

$$\tilde{x}^{(N)}(t) = \sum_{i=0}^{N} x_i t^i \in \mathbb{R}^n$$

Step 2. Calculate corresponding expansion $\widetilde{F}^{(N)}(t)$ *of* $F(\tilde{y}^{(N)}(t), \tilde{x}^{(N)}(t))$ *given by*

$$\widetilde{F}^{(N)}(t) \equiv \widetilde{F}_0^{(N)} + \cdots + \widetilde{F}_N^{(N)} t^N = F(\tilde{y}^{(N)}(t), \tilde{x}^{(N)}(t)) + \mathcal{O}(t^{N+1}) \in \mathbb{R}^m.$$

Step 3. Calculate y_N *by*

$$y_N = -F_y(y^*, x^*)^{-1} \widetilde{F}_N^{(N)}.$$

Step 4. If $N < k$ *set* $N = N + 1$*, and goto Step 1, else stop.*

Beside the *forward* mode of AD as presented above and summarized in Algorithm 6, one may be interested in directly applying the *reverse* mode of AD to the implicit function given by relation (4.7), which is presented in the following.

Reverse mode

For the matter of notational simplicity, we treat again the case that $F : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^m$ is smooth.
For the remaining conditions beeing as above, at the point of interest, i.e. $x^* \in \mathbb{R}$, the requirements on the Implicit Function Theorem (Theorem 26) are fulfilled.
Therefore, there exists in some environment $B(x^*, \rho_x)$, $\rho_x > 0$, around x^* a $\Phi : \mathbb{R}^n \to \mathbb{R}^m$ with $F(y, x) = 0$, where $y = \Phi(x) \in \mathbb{R}^m$ for all $x \in B(x^*, \rho_x) \subset \mathbb{R}^n$.
We also assume that we have an additional function $G : \mathbb{R}^m \to \mathbb{R}$, $z = G(y)$ with $z \in \mathbb{R}$ and $y \in \mathbb{R}^m$ such that the resulting function of interest is given by $G \circ \Phi : B(x^*, \rho_x) \to \mathbb{R}$ in some environment around the point of interest x^* and we have already calculated the input seed $\bar{y}(t) = \frac{\partial z(t)}{\partial y_0}$ where $z(t)$ is considered as Taylor series, i.e. the implicit defined function is part of some evaluation scheme.
One sees from relations (4.4) and (4.5) that the task in applying the *reverse* mode to the implicit function given by relation (4.7), for the general case that the implicit defined function is part of some evaluation scheme as clarified above, is to calculate $\bar{x}(t) = \frac{\partial z(t)}{\partial x_0}$ for given Taylor series $x(t), y(t)$ such that $x_0 = x^*$, $y_0 = y^*$ and relation (4.7) is fulfilled

for $t \in \mathbb{R}$ such that $x(t) \in B(x^*, \rho_x)$, i.e.

$$\bar{x}_p = \left[\frac{\partial z(t)}{\partial x_0}\right]_p = [\bar{y} * \Phi_x(x(t))]_p = \left[\bar{y} * \frac{\partial y(t)}{\partial x_0}\right]_p, \quad \text{for} \quad p \geq 0. \tag{4.23}$$

It should be noted that since $y(t) \in \mathbb{R}^m$ for $t \in \mathbb{R}$ is not a scalar, the convention in this thesis is that $\bar{y}(t)$ is regarded as row vector with each entry is a Taylor series, i.e.

$$\bar{y}(t) := \left(\frac{\partial z(t)}{\partial y_0^k}\right)^T_{k=1,\ldots,m} = \begin{pmatrix} \frac{\partial z(t)}{\partial y_0^1} & \frac{\partial z(t)}{\partial y_0^2} & \cdots & \frac{\partial z(t)}{\partial y_0^m} \end{pmatrix}.$$

Accordingly, $\Phi_x(x(t)) = \frac{\partial y(t)}{\partial x_0}$ is regarded as matrix with each entry is a Taylor series, i.e.

$$\frac{\partial y(t)}{\partial x_0} := \left(\frac{\partial y^k(t)}{\partial x_0^l}\right)_{k=1,\ldots,m,\ l=1,\ldots,n} = \begin{pmatrix} \frac{\partial y^1(t)}{\partial x_0^1} & \frac{\partial y^1(t)}{\partial x_0^2} & \cdots & \frac{\partial y^1(t)}{\partial x_0^n} \\ \frac{\partial y^2(t)}{\partial x_0^1} & \frac{\partial y^2(t)}{\partial x_0^2} & \cdots & \frac{\partial y^2(t)}{\partial x_0^n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y^m(t)}{\partial x_0^1} & \frac{\partial y^m(t)}{\partial x_0^2} & \cdots & \frac{\partial y^m(t)}{\partial x_0^n} \end{pmatrix}.$$

Here, the polynomial product of $\bar{y}(t) * \frac{\partial y(t)}{\partial x_0}$ is defined by

$$\left[\bar{y}(t) * \frac{\partial y(t)}{\partial x_0}\right]_p := \sum_{j=0}^{p} \bar{y}_j \left[\frac{\partial y(t)}{\partial x_0}\right]_{p-j}, \quad \text{for} \quad p \geq 0, \tag{4.24}$$

where $\bar{y}_p$ denotes the row vector

$$\bar{y}_p := \left(\bar{y}_p^k\right)^T_{k=1,\ldots,m} = \begin{pmatrix} \bar{y}_p^1 & \bar{y}_p^2 & \cdots \bar{y}_p^m \end{pmatrix}$$

and $\left[\frac{\partial y(t)}{\partial x_0}\right]_p$ is accordingly defined for $p \geq 0$.

Since this expression is not evaluable directly, we first calculate for

$$h(t) := F(x(t), y(t)) \in \mathbb{R}^m$$

and given input seed Taylor series $\tilde{h}(t) \in \mathbb{R}^m$ the Taylor coefficients of $\tilde{x}(t)$ and $\tilde{y}(t)$ defined by

$$\tilde{x}_p := \left[\tilde{h}(t) * \frac{\partial h(t)}{\partial x_0}\right]_p \quad \text{for} \quad p \geq 0 \tag{4.25}$$

and

$$\tilde{y}_p := \left[\tilde{h}(t) * \frac{\partial h(t)}{\partial y_0}\right]_p \quad \text{for} \quad p \geq 0, \tag{4.26}$$

where

$$\frac{\partial h(t)}{\partial x_0} = F_x(y(t), x(t)) \quad \text{and} \quad \frac{\partial h(t)}{\partial y_0} = F_y(y(t), x(t)).$$

Due to Taylor's Theorem, we first observe that the i-th component of $h(t)$, namely $h^i(t)$, $i \in \{1, \ldots, m\}$, is equal to

$$\begin{aligned} h^i(t) = F^i &+ (F^i_x x_1 + F^i_y y_1)t \\ &+ \frac{1}{2}(F^i_x x_2 + F^i_y y_2 + x_1^T F^i_{xx} x_1 + y_1^T F^i_{yy} y_1 + 2x_1^T F^i_{xy} y_1)t^2 + \dots \ , \end{aligned} \tag{4.27}$$

where we have omitted the arguments of

$$F^i \equiv F^i(y^*, x^*), \quad F^i_x \equiv F^i_x(y^*, x^*), \quad F^i_y \equiv F^i_y(y^*, x^*), \quad F^i_{xx} \equiv F^i_{xx}(y^*, x^*),$$
$$F^i_{yy} \equiv F^i_{yy}(y^*, x^*) \quad \text{and} \quad F^i_{xy} \equiv F^i_{xy}(y^*, x^*),$$

for the case of notational simplicity as in the remainder of the section.
F^i_{xx}, F^i_{yy} and F^i_{xy} denote the second derivative matrix of the i-th component of F in respect of x and y, respectively, i.e.

$$F^i_{xy} := \left(\frac{\partial^2 F^i}{\partial x_k \partial y_l}\right)_{k=1,\ldots,m,\ l=1,\ldots,n} = \begin{pmatrix} \frac{\partial^2 F^i}{\partial x_1 \partial y_1} & \frac{\partial^2 F^i}{\partial x_1 \partial y_2} & \cdots & \frac{\partial^2 F^i}{\partial x_1 \partial y_m} \\ \frac{\partial^2 F^i}{\partial x_2 \partial y_1} & \frac{\partial^2 F^i}{\partial x_2 \partial y_2} & \cdots & \frac{\partial^2 F^i}{\partial x_2 \partial y_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 F^i}{\partial x_n \partial y_1} & \frac{\partial^2 F^i}{\partial x_n \partial y_2} & \cdots & \frac{\partial^2 F^i}{\partial x_n \partial y_m} \end{pmatrix}$$

and F^i_{xx}, F^i_{yy} are accordingly defined.
From (4.27) and because of $\frac{\partial x(t)}{\partial x_0} = \frac{\partial y(t)}{\partial y_0} = 1$ the i-th component of $\frac{\partial h(t)}{\partial x_0}$ and $\frac{\partial h(t)}{\partial y_0}$ are given by

$$\begin{aligned} \frac{\partial h^i(t)}{\partial x_0} &= F^i_x + (x_1^T F^i_{xx} + y_1^T F^i_{yx})t + \dots \ , \\ \frac{\partial h^i(t)}{\partial y_0} &= F^i_y + (x_1^T F^i_{xy} + y_1^T F^i_{yy})t + \dots \ , \end{aligned} \tag{4.28}$$

for $i \in \{1, \ldots, m\}$.

Next, we calculate the Taylor coefficients of $\frac{\partial y(t)}{\partial x_0}$.
We only calculate the coefficients up to order one, i.e $\frac{\partial y_0}{\partial x_0}$ and $\frac{\partial y_1}{\partial x_0}$. This is because higher terms are not used in this work and also the calculation of them gets more and more complex.
We start with the calculation of $\frac{\partial y_0}{\partial x_0}$.

From (4.16) it follows that

$$F(y(t), x(t))|_{t=0} = 0.$$

Again reasoning the Implicit Function Theorem (Theorem 26), we have that

$$\left.\frac{\partial F(y(t), x(t))}{\partial x_0}\right|_{t=0} = 0.$$

Therefore

$$F_x + F_y \frac{\partial y_0}{\partial x_0} = 0.$$

Since from (4.4) and (4.28) it holds that $F_x = \frac{\partial h_0}{\partial x_0}$, it follows that

$$\frac{\partial y_0}{\partial x_0} = -F_y^{-1} \frac{\partial h_0}{\partial x_0}. \tag{4.29}$$

Next, we calculate $\frac{\partial y_1}{\partial x_0}$.
From relation (4.16) we have that

$$\left.\frac{\mathrm{d}F(y(t), x(t))}{\mathrm{d}t}\right|_{t=0} = F_x(y(t), x(t))|_{t=0} \left.\frac{\partial x(t)}{\partial t}\right|_{t=0} + F_y(y(t), x(t))|_{t=0} \left.\frac{\partial y(t)}{\partial t}\right|_{t=0} = 0.$$

Again reasoning the Implicit Function Theorem (Theorem 26) it follows that

$$\left[\frac{\partial}{\partial x_0}\left(\frac{\mathrm{d}}{\mathrm{d}t} F^i(y(t), x(t))\right)\right]\Bigg|_{t=0} = (x_1^T F_{xx}^i + y_1^T F_{yx}^i) + (x_1^T F_{xy}^i + y_1^T F_{yy}^i)\frac{\partial y_0}{\partial x_0} + F_y^i \frac{\partial y_1}{\partial y_0} = 0. \tag{4.30}$$

Then from (4.4), (4.28) and (4.30) we get

$$\frac{\partial y_1}{\partial x_0} = -F_y^{-1}\left(\frac{\partial h_1}{\partial x_0} + \frac{\partial h_1}{\partial y_0}\frac{\partial y_0}{\partial x_0}\right). \tag{4.31}$$

Now, with this preliminary work, we focus on the calculation of (4.23), where we restrict the concern to $p \in \{0, 1\}$.

With $p = 0$ from (4.24) and (4.29) one sees that (4.23) gets

$$\bar{x}_0 = \bar{y}_0 \frac{\partial y_0}{\partial x_0} = -\bar{y}_0 F_y^{-1} \frac{\partial h_0}{\partial x_0}. \tag{4.32}$$

For $p = 1$ from (4.24), (4.29) and (4.31) one sees that (4.23) gets

$$\bar{x}_1 = \bar{y}_0 \frac{\partial y_1}{\partial x_0} + \bar{y}_1 \frac{\partial y_0}{\partial x_0} = -\bar{y}_0 F_y^{-1}\left(\frac{\partial h_1}{\partial x_0} + \frac{\partial h_1}{\partial y_0}\frac{\partial y_0}{\partial x_0}\right) - \bar{y}_1 F_y^{-1}\frac{\partial h_0}{\partial x_0}. \tag{4.33}$$

We introduce the auxiliary Taylor series $\tilde{\mu}(t)$ which we define by

$$\tilde{\mu}_p^T := -\left(\bar{y}_p F_y^{-1}\right)^T = -F_y^{-T}\bar{y}_p^T. \tag{4.34}$$

Obviously (4.32) and (4.33) is equal to

$$\bar{x}_0 = \tilde{\mu}_0 \frac{\partial h_0}{\partial x_0} \tag{4.35}$$

$$\bar{x}_1 = \tilde{\mu}_0 \frac{\partial h_1}{\partial y_0}\frac{\partial y_0}{\partial x_0} + \tilde{\mu}_0 \frac{\partial h_1}{\partial x_0} + \tilde{\mu}_1 \frac{\partial h_0}{\partial x_0}. \tag{4.36}$$

This leads to following algorithm for the calculation of $\bar{x}_0$ and $\bar{x}_1$:

Algorithm 7.

Data: $p \in \{1,2\}$, *a d-times continuously differentiable function* $F : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^m$, *which can be evaluated utilizing elemental functions* ϕ_i *fulfilling Assumption 4,* $i \in \{1,\ldots,l\}$ *as above and* $d \geq p$, $x^* \in \mathbb{R}^n$, $y^* \in \mathbb{R}^m$ *such that* $F(y^*, x^*) = 0$ *and* $F_y(y^*, x^*)$ *is nonsingular, the inverse Jacobian matrix* $F_y(y^*, x^*)^{-1}$, *a univariate polynomial* $x(t) = x_0 + \cdots + x_{p-1}t^{p-1} \in \mathbb{R}^n$ *with* $x(0) = x_0 = x^*$, *the corresponding polynomial* $y(t) = y_0 + \cdots + y_{p-1}t^{p-1} \in \mathbb{R}^m$ *with* $y(0) = y_0 = y^*$ *(e.g., calculated by Algorithm 6), the Taylor coefficients of the input seed* $\bar{y}(t)$ *up to order* $p-1$.

Step 0. Calculate $\tilde{\mu}_k$, $k \in \{0,..,p-1\}$ *according to (4.34).*
Step 1. Calculate $\tilde{x}(t)$ *and* $\tilde{y}(t)$ *for given input seed Taylor series* $\tilde{h}(x) = \tilde{\mu}(t)$ *up to order* $p-1$ *according to (4.25) and (4.26).*

Step 2. Set $\bar{x}_0 = \tilde{x}_0$ *(compare (4.35) and (4.25)). If* $p = 1$ *stop, else goto Step 3.*
Step 3. Calculate $\tilde{x}_0'$ *for given input seed Taylor coefficient* $\tilde{h}_0^T = -F_y^{-T}\left(\tilde{y}_1 + \bar{y}_1\right)^T$ *according to (4.25) since*

$$\begin{aligned}
\tilde{y}_1 + \bar{y}_1 &= \tilde{y}_1 - (-\bar{y}_1 F_y^{-1} F_y) \\
&= \tilde{y}_1 - \tilde{\mu}_1 F_y \\
&= \tilde{y}_1 - \tilde{\mu}_1 \frac{\partial h_0}{\partial y_0} \\
&= \tilde{\mu}_0 \frac{\partial h_1}{\partial y_0} + \tilde{\mu}_1 \frac{\partial h_0}{\partial y_0} - \tilde{\mu}_1 \frac{\partial h_0}{\partial y_0} \\
&= \tilde{\mu}_0 \frac{\partial h_1}{\partial y_0},
\end{aligned}$$

and therefore

$$\tilde{x}_0' = -(\tilde{y}_1 + \bar{y}_1)F_y^{-1}\frac{\partial h_0}{\partial x_0} = \tilde{\mu}_0 \frac{\partial h_1}{\partial y_0}\frac{\partial y_0}{\partial x_0}$$

(compare (4.26), (4.28), (4.29), (4.34), (4.36) and Step 1.).

Step 4. Calculate $\bar{x}_1 = \tilde{x}_1 + \tilde{x}'_0$ *(compare (4.25) and (4.36)) and stop.*

In summery, the *direct* mode approach offers an alternative tracktable method for the calculation of the Taylor coefficients of $y(t)$ for given Taylor series $x(t)$ subject to (4.7) and as well for the calculation of the *reverse* mode Taylor coefficients $\bar{x}(t)$ at least up to order one as presented above. In contrast to the *iterative* mode approach for given x^* and exact y^* such that (4.7) is fulfilled, the Taylor Coefficients of $y(t)$ and $\bar{x}(t)$ are truncation error free. Opposite to the *iterative* mode, for the *reverse* mode the full Jacobian F_y of $F(\cdot, x^*)$ at y^* has to be calculated. On the other hand, applying AD to an *iterative process* directly, the performance depends on the number of iterations of it, whereas the direct mode approach is independent of them.

CHAPTER 5

Calculating numerical solutions of Ordinary Differential Equations and Sensitivity Generation for Ordinary Differential Equations

5.1. Calculating numerical solutions of Ordinary Differential Equations

In this section we treat the numerical solution of an *initial value problem* (IVP) given by the *ordinary differential equation* (ODE)

$$\dot{y}(t) = F(y(t), t, p_0), \quad t \in [t^{\text{init}}, t^{\text{end}}], \tag{5.1}$$

with initial condition

$$y(t^{\text{init}}) = y_{\text{I}}, \tag{5.2}$$

where $F : \mathbb{R}^m \times \mathbb{R} \times \mathbb{R}^p \to \mathbb{R}^m$ is a continuously differentiable function on $\mathbb{R}^m \times [t^{\text{init}}, t^{\text{end}}] \times \{p_0\}$. The function $y : [t^{\text{init}}, t^{\text{end}}] \to \mathbb{R}^m$ denotes the solution of the IVP in (5.1) and (5.2) on the time horizon $[t^{\text{init}}, t^{\text{end}}] \subset \mathbb{R}$ for given initial state $y(t^{\text{init}}) = y_{\text{I}} \in \mathbb{R}^m$ and given parameter $p_0 \in \mathbb{R}^p$, which is considered fixed and therefore this argument in $y(t)$ is omitted. The first derivative with respect to the time for the solution $y(t)$ is denoted by $\dot{y}(t)$, i.e.

$$\dot{y}(t') := \left.\frac{\mathrm{d}y(t)}{\mathrm{d}t}\right|_{t=t'}.$$

For conditions which ensure the existence and uniqueness of a solution $y(t)$ of an IVP as given in (5.1) and (5.2) we refer to, e.g. [78].

Here, we present the theoretical aspects for the numerical implementation and as well the numerical implementation itself of a *Backward Differentiation Formula* (BDF) method, more precisely a variable step variable order BDF method based on Nordsiek array interpolation to solve IVPs numerically, which has been implemented as a part of the algorithmic framework developed for this thesis.
The core integrator is similar to the `EPISODE` BDF method by Byrne and Hindmarsh [32], but with the step size selection strategy of Calvo and Rández [36] and the capability to generate higher-order sensitivities with respect to the initial state vector y_{I} and possibly a parameter vector p_0 based on the sophisticated framework of *Internal Numerical Differentiation* [1, 2] as presented in Section 5.2.1 and 5.2.2.
For a rigorous presentation of the convergence theory of BDF methods we refer to the textbooks [9, 42, 58, 78].
All BDF methods are based on implicit *Backward Differentiation Formulas*, which were first introduced by Curtiss and Hirschfelder [41] and are specially suited for *stiff* ODEs. In [59], Hairer and Wanner define: "*stiff equations are equations where certain implicit methods, in particular BDF, perform better, usually tremendously better, than explicit ones.*"
Since the algorithmic framework is intended to be used for (bio)chemical kinetic systems, it is expected that the systems of interest may be *stiff* and therefore a *stiff* method is implemented rather than an explicit one like, e.g. explicit Runge-Kutta methods (see for example the textbooks [42, 58, 78]). For a detailed discussion of *stiffness* we refer to the textbooks [9, 42, 58, 78].

BDF methods belong to the family of *linear multistep methods*. The general form of a $k-$th order *linear multistep method* at step n can be expressed as

$$\sum_{i=0}^{k} \alpha_{n_i} y_{n-i} = h \sum_{i=0}^{k} \beta_i F_{n-i}, \tag{5.3}$$

where

$$F_{n-i} := F(y_{n-i}, t_{n-i}, p_0)$$

and y_{n-i} denotes the approximate solution at $t = t_{n-i}$ with

$$t_{n-k} < t_{n-k+1} < \cdots < t_n$$

and α_i, β_i are the method's coefficients for $i \in \{0, \ldots, k\}$. The step size is denoted by h. Ascher and Petzold remark [9]: "*The method is called linear because, unlike general Runge-Kutta, the expression in (5.3) is linear in F.*"

5.1.1. Classical linear multistep form of the BDF method

The conceptual idea of the $k-$th order BDF integration method is, to calculate at step n for given step size h_n and already computed approximations y_{n-i}, $i \in \{1, \ldots, k\}$, at time points $t_{n-i} \in \mathbb{R}$, the successive approximation y_n of the exact solution $y(t_n)$ of the IVP in (5.1,5.2) at time point t_n, by the following strategy.
The strategy is based on the construction of an interpolation polynomial $\pi_{n,k}(t)$ of degree k or less in $t \in \mathbb{R}$, such that

$$\pi_{n,k}(t_{n-i}) = y_{n-i}, \quad i \in \{0, 1, \ldots, k\} \tag{5.4}$$

and

$$\dot{\pi}_{n,k}(t_n) = F(y_n, t_n, p_0) \tag{5.5}$$

are fulfilled. These conditions implicitly define y_n.
As Byrne and Hindmarsh state in [32]: "*This set of conditions can be rephrased in the classical linear multistep form*

$$h_n \dot{y}_n = -\sum_{i=0}^{k} \alpha_{n_i} y_{n-i}, \quad \dot{y}_n := F(y_n, t_n, p_0), \tag{5.6}$$

such that the solution of (5.6) for y_n is necessary and sufficient for the existence of $\pi_{n,k}(t)$ with (5.4) and (5.5)."
Here h_n denotes the step size of step n. The sizes of each step do not have to be the same and satisfy

$$0 < \min\{h_j\} \quad \text{and} \quad \max\{h_j\} < H$$

such that for

$$t_j := t^{\text{init}} + \sum_{i=1}^{j} h_i,$$

$t^{\text{init}} =: t_0 < t_1 < \ldots t_{\text{final}} = t^{\text{end}}$ defines a strict partition of $[t^{\text{init}}, t^{\text{end}}]$, where H denotes the maximum step size.

5.1.2. Predictor-corrector scheme in Nordsieck representation

In 1962, Nordsieck invented an integration method [84], which calculates at step n the next approximation y_n to the true solution $y(t)$ of the IVP in (5.1) and (5.2), by use of a stored approximation to a scaled Taylor series of order k around t_{n-1} of the true solution $y(t)$. Here, this approximation is called Nordsieck array of order k and is given

by

$$z_{n-1} := \left(y_{n-1} \quad h_n y_{n-1}^{(1)} \quad h_n^2 y_{n-1}^{(2)}/2 \quad \dots \quad h_n^k y_{n-1}^{(k)}/k! \right),$$

where in general

$$y_{n-1}^{(i)} \neq \left. \frac{\mathrm{d}^i y(t)}{\mathrm{d}t^i} \right|_{t=t_{n-1}},$$

but

$$y_{n-1}^{(i)} := \left. \frac{\mathrm{d}^i \pi_{n-1,k}(t)}{\mathrm{d}t^i} \right|_{t=t_{n-1}},$$

for $i \in \{0, \dots, k\}$. Precisely, z_{n-1} stores the scaled Taylor coefficients of the Taylor series of $\pi_{n-1,k}(t)$ around t_{n-1}.

Osborne [85] and Skeel [108] showed "*that every Nordsiek method is equivalent to a multistep formula and that the order of this method is at least k*" [58].

Now, we present a practical *predictor-corrector scheme* utilizing Nordsieck arrays to calculate the next approximation y_n according to the BDF solution strategy presented in Section 5.1.1 as developed in [32].

In view of the solution strategy in Section 5.1.1, a practical k-th order BDF integration method calculates at step n the desired interpolation polynomial $\pi_{n,k}(t)$, defined by (5.4) and (5.5), by help of the former interpolation polynomial $\pi_{n-1,k}(t)$, calculated at step $n-1$, which is well defined by the conditions

$$\pi_{n-1,k}(t_{n-i}) = y_{n-i}, \quad i \in \{1, \dots, k\} \tag{5.7}$$

and

$$\dot{\pi}_{n-1,k}(t_{n-1}) = F(y_{n-1}, t_{n-1}, p_0).$$

Here, the complete polynomial $\pi_{n-1,k}(t)$ is coded in the Nordsieck array z_{n-1}.

The first step in constructing the desired polynomial $\pi_{n,k}(t)$[1] is to calculate the predictor array $z_{n(0)}$, which is defined by

$$z_{n(0)} := z_{n-1} A[k] \tag{5.8}$$

and $A[k]$ is the $(k+1) \times (k+1)$ Pascal triangle matrix, namely

$$A^{ij}[k] := \begin{cases} 0, & i < j \\ \frac{i!}{j!(i-j)!} & i \geq j. \end{cases}$$

[1] Again, the complete polynomial $\pi_{n,k}(t)$ can be coded in the Nordsieck array z_n. Thus, the task is to construct z_n, instead.

Therefore, $A[6]$ is for example given by

$$A[6] = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 2 & 1 & 0 & 0 & 0 & 0 \\ 1 & 3 & 3 & 1 & 0 & 0 & 0 \\ 1 & 4 & 6 & 4 & 1 & 0 & 0 \\ 1 & 5 & 10 & 10 & 5 & 1 & 0 \\ 1 & 6 & 15 & 20 & 15 & 6 & 1 \end{pmatrix}.$$

Obviously $z_{n(0)}$ stores the scaled Taylor coefficients of the Taylor series of $\pi_{n-1,k}(t)$ around t_n. Hence, the Nordsieck array $z_{n(0)}$ is called the *predictor array.*

On the other hand, the Nordsieck array of interest z_n is given by

$$z_n := \left(y_n \quad h_n y_n^{(1)} \quad h_n^2 y_n^{(2)}/2 \quad \dots \quad h_n^k y_n^{(k)}/k! \right), \tag{5.9}$$

with

$$y_n^{(i)} := \left. \frac{\mathrm{d}^i \pi_{n,k}(t)}{\mathrm{d}t^i} \right|_{t=t_n}.$$

In [32], Byrne and Hindmarsh derived a fundamental relation between $z_{n(0)}$ and z_n, which we present now.
First, we define the polynomial $\Delta_n(t)$ of degree k or less by

$$\Delta_n(t) := \pi_{n,k}(t) - \pi_{n-1,k}(t).$$

From (5.4) and (5.7) one sees, that

$$\Delta_n(t_{n-i}) = 0, \quad \text{for all} \quad i \in \{1, \dots, k\} \tag{5.10}$$

and

$$\Delta_n(t_n) = \pi_{n,k}(t_n) - \pi_{n-1,k}(t_n).$$

If one denotes the entries of the Nordsieck array $z_{n(0)}$ by

$$z_{n(0)} = \left(y_{n(0)} \quad h_n y_{n(0)}^{(1)} \quad h_n^2 y_{n(0)}^{(2)}/2 \quad \dots \quad h_n^k y_{n(0)}^{(k)}/k! \right)$$

and by the fact that the entries of $z_{n(0)}$ are the scaled Taylor coefficients of the Taylor series of $\pi_{n-1,k}(t)$ around t_n, one can conclude that

$$\Delta_n(t_n) = y_n - y_{n(0)} =: e_n, \tag{5.11}$$

where e_n is called the *corrector vector*.
From (5.10) and (5.11) and using the polynomial interpolation in Lagrange form, it immediately follows that $\Delta_n(t)$ is uniquely defined by

$$\Delta_n(t_n) = \prod_{i=1}^{k} \frac{t - t_{n-i}}{t_n - t_{n-i}} e_n. \tag{5.12}$$

We introduce the auxiliary quantities

$$x_n := \frac{t - t_n}{h_n} \quad \text{and} \quad \xi_{n,i} := \frac{t_n - t_{n-i}}{h_n} \quad \text{with} \quad i \in \{1, \ldots, k\} \tag{5.13}$$

and the scalar auxiliary polynomial

$$\Lambda_n(x) = \prod_{i=1}^{k} \left(1 + \frac{x}{\xi_{n,i}}\right). \tag{5.14}$$

Using these quantities, (5.12) transforms to

$$\Delta_n(t) = \Delta_n(t_n + h_n x) = \Lambda_n(x) e_n.$$

We denote the i-th coefficient of the scalar polynomial $\Lambda_n(x)$ by l_n^i, i.e.

$$\Lambda_n(x) = \sum_{i=0}^{k} l_n^i x^i. \tag{5.15}$$

From the fact that the entries in z_n and $z_{n(0)}$ are the scaled Taylor coefficients of the Taylor series of $\pi_{n,k}(t)$ and $\pi_{n-1,k}(t)$, respectively, around t_n and with

$$\begin{aligned} \frac{h_n^i}{i!} \left.\frac{\mathrm{d}^i \pi_{n,k}(t)}{\mathrm{d}t^i}\right|_{t=t_n} - \frac{h_n^i}{i!} \left.\frac{\mathrm{d}^i \pi_{n-1,k}(t)}{\mathrm{d}t^i}\right|_{t=t_n} &= \frac{h_n^i}{i!} \left.\frac{\mathrm{d}^i \Delta_n(t)}{\mathrm{d}t^i}\right|_{t=t_n} \\ &= \frac{1}{i!} \left.\frac{\mathrm{d}^i \Lambda_n(x_n(t))}{\mathrm{d}t^i}\right|_{t=t_n} = l_n^i e_n, \end{aligned} \tag{5.16}$$

for $i \in \{0, \ldots, k\}$, and x_n is regarded as function of t, it follows that

$$z_n = z_{n(0)} + e_n l_n, \tag{5.17}$$

where l_n denotes the $1 \times (k+1)$ row vector

$$l = \begin{pmatrix} l_n^0 & l_n^1 & \dots & l_n^k \end{pmatrix}.$$

Thus, by knowledge of the *corrector vector* e_n and the coefficient row vector l_n, the *predictor array* $z_{n(0)}$ can be easily corrected to yield the desired Nordsieck vector z_n. The coefficient row vector l_n is uniquely defined by the relations (5.14) and (5.15). We first note that obviously $l_n^0 = 1$. We calculate the remaining coefficients by the recursive scheme presented in Algorithm 8, which can be easily verified by induction.

Algorithm 8.

Data: *Order of BDF method k, Auxiliar quantities $\xi_{n,q}$ for $q \in \{1, \dots, k\}$.*

Step 0. *Set $l_n^0 = 1$. Set $i = 0$.*
Step 1. *Set $i = i + 1$. If $i > k$ stop.*
Step 2. *Set $j = i - 1$. Set*

$$l_n^i = \prod_{q=1}^{i} \xi_{n,q}^{-1}.$$

Step 3. *If $j < 1$ goto Step 1. Else set*

$$l_n^j = l_n^j + l_n^{j-1} \xi_{n,i}^{-1},$$

set $j = j - 1$ and goto Step 3.

Remark. *It can directly be deduced from Algorithm 8 that l_n^1 is given by*

$$l_n^1 = \sum_{i=1}^{k} \frac{1}{\xi_{n,i}} \tag{5.18}$$

and l_n^k is given by

$$l_n^k = \prod_{i=1}^{k} \frac{1}{\xi_{n,i}}.$$

The calculation of the *corrector vector* e_n is presented in Section 5.1.5.

5.1.3. Estimation of the local error

In [32], Byrne and Hindmarsh state at the beginning of the section "Error Estimation":

"*The algorithm described so far is of little or no use without an accompanying algorithm for the selection of order k and step size h throughout the integration. This selection algorithm is based on the local discretization error*".

Therefore, in this section, we will derive formulas for the estimation of the local discretization error, based on the corrector vector e_n. In the following, we will denote the local discretization error *local truncation error*.

Definition 15. *The local truncation error $LTE_n(k)$ of the $k-$th order BDF method at step n is defined by*

$$LTE_n(k) := y(t_n) - \tilde{y}_n, \tag{5.19}$$

where $y(t_n)$ is the exact solution of IVP in (5.1) and (5.2) at time point t_n, and $\tilde{y}_n$ is the numerical one, i.e. obtained by (5.6) but with y_{n-i}, $i \in \{1, \ldots, k\}$ given by the exact solution at time points t_{n-i}, $i \in \{1, \ldots, k\}$, i.e.

$$y_{n-i} = y(t_{n-i}), \quad i \in \{1, \ldots, k\},$$

for this case.

Definition 16. *The local error $LE_n(k)$ of the $k-$th order BDF method at step n is defined by*

$$LE_n(k) := y(t_n) + \frac{1}{\alpha_{n_0}} \left[h_n \dot{y}(t_n) + \sum_{i=1}^{k} \alpha_{n_i} y(t_{n-i}) \right], \tag{5.20}$$

where $y(t_{n-i})$, $i \in \{0, \ldots, k\}$ is the exact solution of IVP in (5.1) and (5.2) at time points t_{n-i} and $\dot{y}(t_n) = F(y(t_n), t_n, p)$.

The following lemma gives a relation between the *local truncation error* $LTE_n(k)$ and the *local error* $LE_n(k)$ of the $k-$th order BDF method at step n. The lemma gives essentially the same result as Lemma 2.2 in [58] but formulated not only for uniform step sizes.

Lemma 3. *Consider the IVP in (5.1) and (5.1) with $F(\cdot, t_n, p)$ continuously differentiable and let $y(t)$ be its exact solution. Then it holds for some $s \in [0, 1]$ that for the $k-$th order BDF method at step n*

$$LTE_n(k) = \left(\mathbb{1} + \frac{h_n}{\alpha_0} F_y(s(y(t) - \tilde{y_n}), t_n, p_0) \right)^{-1} LE_n(k).$$

Proof. According to (5.6) $\tilde{y}_n$ is implicitly given by

$$\tilde{y}_n = -\frac{1}{\alpha_{n_0}} \left[h_n F(\tilde{y}_n, t_n, p_0) + \sum_{i=1}^{k} \alpha_{n_i} y_{n-i} \right]. \tag{5.21}$$

Using (5.19) and (5.20) we see that (5.21) is equivalent to

$$LE_n(k) = LTE_n(k) + \frac{h_n}{\alpha_{n_0}} \left[F(y(t_n), t_n, p) - F(\tilde{y}_n, t_n, p) \right].$$

From the Mean-Value Theorem it follows that

$$\begin{aligned} LE_n(k) &= LTE_n(k) + \frac{h_n}{\alpha_{n_0}} F_y(\tilde{y}_n + s(y(t_n) - \tilde{y}_n), t_n, p) LTE_n(k) \\ &= \left(\mathbb{1} + \frac{h_n}{\alpha_{n_0}} F_y(\tilde{y}_n + s(y(t_n) - \tilde{y}_n), t_n, p) \right) LTE_n(k), \end{aligned}$$

for some $s \in [0, 1]$. □

Therefore, the *local truncation error* $LTE_n(k)$ and the *local error* $LE_n(k)$ for the $k-$th order BDF method at step n are essentially the same, i.e. for small enough step size h_n $LTE_n(k) \approx LE_n(k)$.

Lemma 4. *For the local error $LE_n(k)$ of the $k-$th order BDF method at step n it holds that*

$$LE_n(k) = \frac{\prod_{i=1}^{k} \xi_{n,i}}{(k+1)! \alpha_{n_0}} h_n^{k+1} \frac{\mathrm{d}^{(k+1)} y(t_n)}{\mathrm{d}t^{(k+1)}} + \mathcal{O}(H^{k+2}),$$

where $\xi_{n,i}$, $i \in \{0, \ldots, k\}$ is defined as in (5.13).

Proof. Let $\tilde{\pi}_{n,k}(t)$ be an interpolation polynomial of degree k defined by the conditions

$$\tilde{\pi}_{n,k}(t_{n-i}) = y(t_{n-i}) \quad \text{for all} \quad i \in \{0, 1, \ldots, k\}.$$

From (5.6) we see that

$$-h_n \dot{\tilde{\pi}}_{n,k}(t_n) = \sum_{i=0}^{q} \alpha_{n_i} y(t_{n-i}).$$

Therefore for the $LE_n(k)$ it holds that

$$LE_n(k) = \frac{h_n}{\alpha_{n_0}} \left[\dot{y}(t_n) - \dot{\tilde{\pi}}_{n,k}(t_n) \right].$$

Let $\tilde{\pi}_{n,k+1}(t)$ be another interpolation polynomial of degree $k+1$ defined by the condi-

tions

$$\tilde{\pi}_{n,k+1}(t_{n-i}) = y(t_{n-i}), \quad \text{for all} \quad i \in \{0, 1, \ldots, k\}. \tag{5.22}$$

and

$$\dot{\tilde{\pi}}_{n,k+1}(t_n) = \dot{y}(t_n). \tag{5.23}$$

With $\Delta(t) := \tilde{\pi}_{n,k+1}(t) - \tilde{\pi}_{n,k}(t)$ it obviously follows that

$$LE_n(k) = \frac{h_n}{\alpha_{n_0}} \dot{\Delta}(t_n). \tag{5.24}$$

Since by construction $\tilde{\pi}_{n,k+1}(t_{n-i}) = \tilde{\pi}_{n,k}(t_{n-i})$ for all $i \in \{0, 1, \ldots, k\}$, it holds, that

$$\Delta(t_{n-i}) = 0 \quad \text{for all} \quad i \in \{0, 1, \ldots, k\}.$$

Thus, it follows from the polynomial interpolation in Lagrange form that

$$\Delta(t) = c \prod_{i=0}^{k} (t - t_{n-i}), \tag{5.25}$$

for some constant vector c.
The constant vector c is the leading coefficient of $\Delta(t)$ and thus of $\tilde{\pi}_{n,k+1}$, as well. Therefore it follows that

$$c = \frac{1}{(k+1)!} \frac{\mathrm{d}^{(k+1)} \tilde{\pi}_{n,k+1}(t_n)}{\mathrm{d}t^{(k+1)}}. \tag{5.26}$$

An upper bound for the error term $\|y(t) - \tilde{\pi}_{n,k+1}(t)\|$ of the Hermite polynomial interpolation for (5.22) and (5.23) is given by

$$\|y(t) - \tilde{\pi}_{n,k+1}(t)\| \leq c'(t - t_n)^2 \prod_{i=1}^{k} (t - t_{n-i}), \tag{5.27}$$

where c' is some constant. Therefore, it immediately follows that

$$c = \frac{1}{(k+1)!} \frac{\mathrm{d}^{(k+1)} \tilde{\pi}_{n,k+1}(t_n)}{\mathrm{d}t^{(k+1)}} = \frac{1}{(k+1)!} \frac{\mathrm{d}^{(k+1)} y(t_n)}{\mathrm{d}t^{(k+1)}} + \mathcal{O}(H). \tag{5.28}$$

From (5.25), we also know that

$$\dot{\Delta}(t_n) = c \prod_{i=1}^{k} (t_n - t_{n-k}) = c h_n^k \prod_{i=1}^{k} \xi_{n,i}, \tag{5.29}$$

and therefore, from (5.24), (5.28) and (5.29), we see, that

$$LE_n(k) = \frac{\prod_{i=1}^{k} \xi_{n,i}}{(k+1)!\alpha_{n_0}} h_n^{k+1} \frac{\mathrm{d}^{(k+1)} y(t_n)}{\mathrm{d}t^{(k+1)}} + \mathcal{O}(H^{k+2}). \tag{5.30}$$

□

Now, from Lemma 3 and Lemma 4, it obviously follows, that, under above conditions,

$$\begin{aligned} y(t_n) - \tilde{y}_n = LTE_n(k) &= LE_n(k) + \mathcal{O}(H^{k+2}) \\ &= \frac{\prod_{i=1}^{k} \xi_{n,i}}{(k+1)!\alpha_{n_0}} h_n^{k+1} \frac{\mathrm{d}^{(k+1)} y(t_n)}{\mathrm{d}t^{(k+1)}} + \mathcal{O}(H^{k+2}). \end{aligned} \tag{5.31}$$

In the following, we will develop an estimation of the asymptotic part of the *local error* $LE_n(k)$, based on the corrector vector e_n.

For this task, we define a second polynomial $\hat{\pi}_{n-1,k}(t)$ of degree k or less by

$$\hat{\pi}_{n-1,k}(t_{n-i}) = y(t_{n-i}) \quad \text{for all} \quad i \in \{1, \ldots, k\} \quad \text{and} \quad \dot{\hat{\pi}}_{n-1,k}(t_{n-1}) = y(t_{n-1}).$$

Therefore, it yields (by definition of $\hat{\pi}_{n-1,k}(t_{n-i})$), that

$$y_{n(0)} = \hat{\pi}_{n-1,k}(t_n).$$

Again, we define a second polynomial $\hat{\pi}_{n-1,k+1}(t)$ of degree $k+1$ or less, given by

$$\hat{\pi}_{n-1,k+1}(t_{n-i}) = y(t_{n-i}) \quad \text{for all} \quad i \in \{0, 1, \ldots, k\} \quad \text{and} \quad \dot{\hat{\pi}}_{n-1,k+1}(t_{n-1}) = y(t_{n-1}).$$

Now we define the polynomial $\widetilde{\Delta}(t)$ by

$$\widetilde{\Delta}(t) := \hat{\pi}_{n-1,k+1}(t) - \hat{\pi}_{n-1,k}(t).$$

Then, by definition of $\hat{\pi}_{n-1,k+1}(t)$ and $\hat{\pi}_{n-1,k+1}(t)$, it immediately follows that

$$y(t_n) - y_{n(0)} = \hat{\pi}_{n-1,k+1}(t_n) - \hat{\pi}_{n-1,k}(t_n) = \widetilde{\Delta}(t_n)$$

and

$$\widetilde{\Delta}(t_{n-i}) = 0 \quad \text{for all} \quad i \in \{1, \ldots, k\} \quad \text{and} \quad \dot{\widetilde{\Delta}}(t_{n-1}) = 0.$$

Thus, it follows from the Hermite polynomial interpolation that

$$\widetilde{\Delta}(t) = \tilde{c}(t - t_{n-1})^2 \prod_{i=2}^{k}(t - t_{n-k}),$$

for a constant vector $\tilde{c}$. With similar thoughts as above (compare (5.26), (5.27) and (5.28)), it follows that

$$\tilde{c} = \frac{1}{(k+1)!}\frac{\mathrm{d}^{(k+1)}\hat{\pi}_{n-1,k+1}(t_n)}{\mathrm{d}t^{(k+1)}} = \frac{1}{(k+1)!}\frac{\mathrm{d}^{(k+1)}y(t_n)}{\mathrm{d}t^{(k+1)}} + \mathcal{O}(H).$$

Thus, we can conclude that

$$y(t_n) - y_{n(0)} = \widetilde{\Delta}(t_n) = \tilde{c}h_n^{k+1}\prod_{i=1}^{k}\xi_{n,i} = \frac{h_n^{k+1}}{(k+1)!}\frac{\mathrm{d}^{(k+1)}y(t_n)}{\mathrm{d}t^{(k+1)}}\prod_{i=1}^{k}\xi_{n,i} + \mathcal{O}(H^{k+2}). \quad (5.32)$$

Now, we subtract (5.31) from (5.32) and we get

$$e_n := \tilde{y}_n - y_{n(0)} = \left(1 - \frac{1}{\alpha_{n_0}}\right)\frac{h_n^{k+1}}{(k+1)!}\frac{\mathrm{d}^{(k+1)}y(t_n)}{\mathrm{d}t^{(k+1)}}\prod_{i=1}^{k}\xi_{n,i} + \mathcal{O}(H^{k+2}). \quad (5.33)$$

From (5.30) and (5.33) we get the desired *local error* estimator

$$LE_n(k) = \frac{1}{\alpha_{n_0}}\frac{1}{(1 - 1/\alpha_{n_0})}e_n + \mathcal{O}(H^{k+2}), \quad (5.34)$$

which is correct within $\mathcal{O}(H^{k+2})$, if the past values are exactly known and the *corrector iteration* is solved exactly. Generally, the past values are not known exactly.
Byrne and Hindmarsh state in [32]: "*However, the errors in the past values can be taken into account [49] if we assume that the global errors in $y(t_n) - y_n$ at order k satisfy the expansion*

$$y_n - y(t_n) = d_k(t_n)h^k + d_{k+1(t_n)}h^{k+1} + \mathcal{O}(H^{k+2}), \quad (5.35)$$

with functions d_k and d_{k+1} which satisfy differential equations of the form

$$\dot{d_k}(t) = F_y d_k(t) + \phi_k(t).$$

This assumption is valid for constant h [111] and, at least under some circumstances, for nonconstant h [50]. Here $\phi_k(t)$ is the "principal error function", for which $\alpha_{n_0}^{-1}h^{k+1}\phi_k(t_n)$ is the principal term of the local error $LE_n(k)$."

Therefore, to take the errors in the past values into account, instead of (5.34) Byrne and

Hindmarsh [32] use

$$\begin{aligned} LE_n(k) &= \frac{1}{\alpha_{n_0}} \frac{1}{(1-1/\tilde{\alpha}_{n_0})} e_n + \mathcal{O}(H^{k+2}) \\ &= \frac{1}{\alpha_{n_0}} \left[1 + \prod_{i=2}^{k} \left(\frac{t_n - t_{n-i}}{t_{n-1} - t_{n-i}} \right) \right]^{-1} e_n + \mathcal{O}(H^{k+2}) \end{aligned} \tag{5.36}$$

with

$$\tilde{\alpha}_{n_0} := -\prod_{i=2}^{k} \left(\frac{t_{n-1} - t_{n-i}}{t_n - t_{n-i}} \right), \tag{5.37}$$

as estimate of the *local error* $LE_n(k)$, which is correct within $\mathcal{O}(H^{k+2})$, if the global errors in $y(t_n) - y_n$ satisfy (5.35). We use the same estimator (5.36) of the *local error* $LE_n(k)$, too.

For later purpose, we additionally introduce estimates of the *local error* $LE_n(k'-1)$ at order $k'-1$ and of the *local error* $LE_n(k'+1)$ at order $k'+1$, for the case that the current order of the method is k'.

First, we give an estimate of the *local error* $LE_n(k'-1)$. Since the current order of the method is k' the last entry of the Nordiesk array z_n gives an estimate of $\frac{h_n^{k'}}{k'!} \frac{\mathrm{d}^{k'} y(t_n)}{\mathrm{d}t^{k'}}$, i.e.

$$\frac{h_n^{k'}}{k'!} \frac{\mathrm{d}^{k'} y(t_n)}{\mathrm{d}t^{k'}} \approx \frac{h_n^{k'} y_n^{(k')}}{k'!}.$$

By construction of z_n, this estimate is correct within $\mathcal{O}(H^{k'+1})$, if the nodes in (5.4) are correct within $\mathcal{O}(H^{k'+1})$, as stated in [32]. Therefore, from (5.30) an estimate of $LE_n(k'-1)$ is given by

$$LE_n(k'-1) = \frac{\prod_{i=1}^{k'-1} \xi_{n,i}}{\alpha_{n_0}^{(k'-1)}} \frac{h_n^{k'} y_n^{(k')}}{k'!} + \mathcal{O}(H^{k'+1}), \tag{5.38}$$

where $\alpha_{n_0}^{(k'-1)}$ denotes the coefficient a_{n_0} of the *classical linear multistep form* in (5.6) of order $k'-1$ (i.e. $k \hat{=} k'-1$ in (5.6)).

Next, we give an estimate of the *local error* $LE_n(k'+1)$. We start with

$$e_n = \hat{c}_n h_n^{k'+1} \frac{\mathrm{d}^{(k'+1)} y(t_n)}{\mathrm{d}t^{(k'+1)}} + \mathcal{O}(H^{k'+2}), \tag{5.39}$$

where

$$\hat{c}_n := \left(1 - \frac{1}{\tilde{\alpha}_{n_0}}\right) \frac{\prod_{i=1}^{k'} \xi_{n,i}}{(k'+1)!},$$

which is essentially the same as (5.33), but again corrected by taking the past values into account, as above.

We again follow the ideas in [32] and consider a combination of e_n and e_{n-1}, such that the combination is asymptotically $\mathcal{O}(H^{k'+2})$. This combination has to be proportional to $e_n - Q_n e_{n-1}$, with

$$Q_n := \frac{\hat{c}_n}{\hat{c}_{n-1}} \left(\frac{h_n}{h_{n-1}}\right)^{k'+1},$$

such that the combination is asymptotically $\mathcal{O}(H^{k'+2})$. Now, if one neglects the $\mathcal{O}(H^{k'+2})$ terms in (5.39) and $y(t) \in \mathcal{C}^{k'+2}$, one gets the relation

$$\begin{aligned} h_n^{k'+2} \frac{\mathrm{d}^{(k'+2)} y(t_n)}{\mathrm{d}t^{(k'+2)}} &= h_n^{k'+1} \frac{\mathrm{d}^{(k'+1)} y(t_n)}{\mathrm{d}t^{(k'+1)}} - h_n^{k'+1} \frac{\mathrm{d}^{(k'+1)} y(t_{n-1})}{\mathrm{d}t^{(k'+1)}} + \mathcal{O}(H^{k'+3}) \\ &= \hat{c}_n^{-1} \left(e_n - Q_n e_{n-1}\right) + \mathcal{O}(H^{k'+3}). \end{aligned} \tag{5.40}$$

Now, from (5.30) and (5.40) we get, as estimate of the *local error* $LE_n(k'+1)$, the following expression

$$LE_n(k'+1) = \frac{\xi_{n,k'+1}}{(k'+2)\alpha_{n_0}^{(k'+1)}} \frac{1}{(1 - 1/\tilde{\alpha}_{n_0})} \left(e_n - Q_n, e_{n-1}\right) + \mathcal{O}(H^{k'+3}), \tag{5.41}$$

where $\alpha_{n_0}^{(k'+1)}$ denotes the coefficient a_{n_0} of the *classical linear multistep form* in (5.6) of order $k'+1$ (i.e. $k \hat{=} k'+1$ in (5.6)).

Remark. *Since we have neglected the terms $\mathcal{O}(H^{k'+2})$ in (5.39), the estimate in (5.41) might be inaccurate, as also stated in [32]. But this estimate is only used for the estimation of a new step size h_{n+1} for the next integration step, if the order at the $(n+1)$-th step changes from $k' \to k'+1$.*

The order and step size selection strategy is presented in Section 5.1.6. In the following, we give some notes on scaling and, as well, on the influence of the numerical accuracy on the numerical behavior of the implemented BDF method.

5.1.4. Scaling and numerical accuracy principles of the implemented BDF method

For the implemented BDF method, one integration step at order k is considered successful, if the norm of the estimated error $LE_n(k)$ is below some user given threshold ϵ,

i.e.

$$\|LE_n(k)\|_{W_n} \leq \epsilon. \tag{5.42}$$

Since the components of the solution might have different orders of magnitude, it is common [32, 86, 29, 94, 16, 62], to use a weighted norm $\|\cdot\|_{W_n}$ in (5.42).
Here, we use the weighted norm

$$\|LE_n(k)\|_{W_n} := \frac{1}{m}\sqrt{\sum_{i=1}^{m}\frac{LE_n^i(k)}{W_n^i}},$$

where $LE_n^i(k)$ denotes the i-th component of the estimated error $LE_n(k)$ and W_n^i, for $i \in \{1,\ldots,m\}$, are scaling factors.
The user can choose between two different scaling factors. The first one, also used, e.g. in [29, 62], is given by

$$\widetilde{W}_n^i := |y_n^i| + \frac{atol^i}{\epsilon}, \tag{5.43}$$

and the second one is given by

$$\overline{W}_n^i := \max\{|y_n^i|, \overline{W}_{n-1}^i, atol^i\}, \tag{5.44}$$

where $atol \in \mathbb{R}^m_{>0}$, with

$$\mathbb{R}^m_{>0} := \left\{(x_1,\ldots,x_m)^T | \mathbb{R} \ni x_i > 0 \forall i \in \{1,\ldots,m\}\right\},$$

is given by the user and $i \in \{1,\ldots,m\}$. The second one is called "new Deuflhard-scaling" [16]. We use $\|\cdot\|_{\widetilde{W}_n}$ for the calculation of the results in Section 8.

In [101], Shampine investigated the influence of limiting precision in differential equation solvers. He points out, that if

$$\epsilon < u_\mathrm{r}\|y_n\|_{W_n},$$

then the user given threshold for the acceptance of the *local error* $LE_n(k)$ is less than a unit roundoff of $\|y_n\|_{W_n}$, where u_r is the unit roundoff of the machine. Therefore, the user has clearly asked for too much accuracy. Here, the strategy to tackle this situation, is to relax the user given threshold by

$$\epsilon_\mathrm{relax} := \max\{\epsilon, 2u_\mathrm{r}\|y_n\|\} \tag{5.45}$$

and to demand

$$\|LE_n(k)\|_{W_n} \leq \epsilon_\mathrm{relax} \tag{5.46}$$

in place of (5.42), for the acceptance of the step.

5.1.5. Calculation of the corrector vector e_n

For the calculation of the *correct vector* e_n we start with the first column of relation (5.17), just as Byrne and Hindmarsh in [32], i.e.

$$h_n y_n^{(1)} = h_n y_{n(0)}^{(1)} + e_n l_n^1 = h_n y_{n(0)}^{(1)} + (y_n - y_{n(0)}) l_n^1. \tag{5.47}$$

Since by definition of z_n, $y_n^{(1)} = \dot{y}_n$ and therefore, (5.47) is equivalent to

$$h_n \dot{y}_n = h_n y_{n(0)}^{(1)} + (y_n - y_{n(0)}) l_n^1, \tag{5.48}$$

with

$$\dot{y}_n = F(y_n, t_n, p_0),$$

according to (5.6).
Since, by construction, (5.48) is equivalent to the *classical linear multistep form* in (5.6), we can identify l_n^1 with

$$l_n^1 = -\alpha_{n_0}.$$

Therefore, we can deduce with (5.18) that

$$\alpha_{n_0}^{(k-1)} = -\sum_{i=1}^{k-1} \frac{1}{\xi_{n,i}}$$

and

$$\alpha_{n_0}^{(k+1)} = -\sum_{i=1}^{k+1} \frac{1}{\xi_{n,i}}.$$

This is helpful, for the calculation of (5.38) and (5.41).
For the calculation of the vector y_n, we define the nonlinear mapping $G^n : \mathbb{R}^m \to \mathbb{R}^m$ by

$$G^n(u) = (u - y_{n(0)}) - \frac{h_n}{l_n^1} \left(F(u, t_n, p_0) - y_{n(0)}^{(1)} \right). \tag{5.49}$$

Obviously, the vector of interest y_n is a root of $G^n(u)$, i.e. $G^n(y_n) = 0$.
For the calculation of this root of $G^n(u)$, we use a simplified Newton method, i.e. the iteration scheme, we use, for the $j+1$-th approximation $u_{n,j+1}$, of this root of $G^n(u)$, is given by

$$u_{n,j+1} = u_{n,j} + \Delta u_{n,j}, \tag{5.50}$$

where $\Delta u_{n,j}$ is given by

$$\Delta u_{n,j} = -\widetilde{P}_n^{-1}(u_{n,j})G^n(u_{n,j}) \tag{5.51}$$

and $\widetilde{P}_n^{-1}(u)$ denotes an approximation to the inverse of the Jacobian $G_u^n(u)$. Here, the approximation $\widetilde{P}_n^{-1}(u)$ is kept constant during one pass of the simplified Newton method, i.e. $\widetilde{P}_n^{-1}(u_j) = \widetilde{P}_n^{-1}$ and possibly on successive ones.

We use $\pi_{n-1,k}(t_n)$, i.e. $y_{n(0)}$, as start value $u_{n,0}$ for this simplified Newton method. Instead of calculating $\Delta u_{n,j}$ by (5.51), we solve the linear equation

$$\widetilde{P}_n \Delta u_{n,j} = -G^n(u_{n,j}), \tag{5.52}$$

using LU decomposition of the matrix $\widetilde{P}_n$. For this task, we use the software package `ATLAS` [130, 131], which provides a tuned interface to `LAPACK` [6, 5] and `BLAS` [72, 48]. Especially, `ATLAS` provides the possibility to store the LU decomposition of the matrix $\widetilde{P}_n$ and to reuse it for the solution of successive linear equations, if no more than the right hand side of a successive linear equation is changing.

The modified Newton method obeys the following local contraction theorem as given in [27] by Bock.

Theorem 16 (Local contraction, Theorem 3.1.44 in [27] (modified))**.** *Let $G : D \to \mathbb{R}^m$ be a continuously differentiable function, where D is an open subset of $\mathbb{R}^m$. Let $\widetilde{P}^{-1}$ denote an approximation to the inverse of the Jacobian $G_u(u_0)$ at the start value $u_0 \in D$. In addition, for all $\tau \in [0,1]$ and for all $u, u + \Delta u \in D$, with $\Delta u := -\widetilde{P}^{-1}G(u)$, there exist bounds $\omega < \infty$ and $\kappa < 1$, such that*

$$\|\widetilde{P}^{-1}\left(G_u(u + \tau\Delta u) - G_u(u)\right)\Delta u\| \leq \omega\tau\|\Delta u\|^2 \tag{5.53}$$

and

$$\|\widetilde{P}^{-1}R(u)\| \leq \kappa\|\Delta u\|, \tag{5.54}$$

with

$$R(u) := G(u) + G_u(u)\Delta u,$$

where $R(u)$ is called the residuum.

Then, if the start value u_0 satisfies

$$\delta_0 := \frac{\omega\|\Delta u_0\|}{2} + \kappa < 1 \tag{5.55}$$

and

$$D_0 := B(u_0, \frac{\|\Delta u_0\|}{1-\delta_0}) \subset D, \tag{5.56}$$

it follows that the iterates $u_{j+1} = u_j + \Delta u_j$ *are well defined and stay in* D_0 *and the series* $\{u_j\}_{j=0}^{\infty}$ *converges to a fixed point, denoted by* $u^* \in D_0$*, with* $\Delta u^* = -\widetilde{P}^{-1}G(u^*) = 0$*. Furthermore, it holds, that an a priori estimate is given by*

$$\|u_j - u^*\| \leq \delta_0^j \frac{\|\Delta u_0\|}{1-\delta_0} \tag{5.57}$$

and the convergence is linear with

$$\|\Delta u_j\| \leq \left(\frac{\omega\|\Delta u_{j-1}\|}{2} + \kappa\right)\|\Delta u_{j-1}\| =: \delta_j \|\Delta u_{j-1}\|.$$

Proof. By assumption $u_0, u_1 \in D_0$. Now assume that $u_{j+1} \in D_0$, then it holds that

$$\begin{aligned}
\|\Delta u_{j+1}\| &= \|\widetilde{P}^{-1}\left(G(u_{j+1}) - G(u_j) - G_u(u_j)\Delta u_j\right) + \widetilde{P}^{-1}R(u_j)\| \\
&= \|\widetilde{P}^{-1}\int_0^1 \left(G_u(u_j + \tau\Delta u_j) - G_u(u_j)\right)\Delta u_j \mathrm{d}\tau + \widetilde{P}^{-1}R(u_j)\| \\
&\leq \int_0^1 \tau\omega\|\Delta u_j\|^2 \mathrm{d}\tau + \kappa\|\Delta u_j\| = \left(\frac{\omega\|\Delta u_j\|}{2} + \kappa\right)\|\Delta u_j\| =: \delta_j\|\Delta u_j\|.
\end{aligned} \tag{5.58}$$

It should be noted that $\delta_{j+1} \leq \delta_j$, which can be seen from (5.55) and (5.58).
It follows that

$$\|u_{j+2} - u_0\| \leq \sum_{i=0}^{j+1}\|\Delta u_j\| \leq \|\Delta u_0\|\sum_{i=0}^{j+1}\delta_0^i < \frac{\|\Delta u_0\|}{1-\delta_0}$$

and therefore, $u_{j+2} \in D_0$.
It also holds that

$$\|u_{j+p} - u_j\| \leq \sum_{i=0}^{p-1}\|\Delta u_{j+i}\| < \delta_0^j\frac{\|\Delta u_0\|}{1-\delta_0}$$

and thus $\{u_j\}_{j=0}^{\infty}$ is a Cauchy sequence and converges in D_0, where u^* denotes the limit point.
It remains to show that u^* is a fixed point.

First, by assumption (5.54) and

$$\lim_{j\to\infty}\|\Delta u_j\| = 0, \tag{5.59}$$

it follows that

$$\lim_{j\to\infty} \|\widetilde{P}^{-1}R(u_j)\| = 0.$$

Second, the *residuum* $R(u)$ is a composition of continuous functions and thus continuous itself. Therefore, by definition of $R(u)$ and (5.59), it follows that $R(u^*) = G(u^*)$. Finally, $\widetilde{P}^{-1}R(u_j)$ is continuous, too. Thus,

$$\lim_{j\to\infty} \|\widetilde{P}^{-1}R(u_j)\| = \|\widetilde{P}^{-1}G(u^*)\| = 0$$

and therefore, u^* is fixed point, which completes the proof. □

In [14, 80], it is noted that:

- The Lipschitz constant ω measures the nonlinearity of $G(u)$, since for given $\widetilde{P}^{-1}$, (5.53) can be replaced by the Lipschitz condition

  $$\|G_u(u'') - G_u(u')\| \leq \gamma\|u'' - u'\|,$$

  where $u', u'' \in D$. Then γ gives an overestimate of ω, with $\omega \leq \|\widetilde{P}^{-1}\|\gamma$.

- The condition (5.54) can be replaced by

  $$\|\mathbb{1} - \widetilde{P}^{-1}G_u(u)\|_{\text{operator}} \leq \kappa < 1, \quad \text{for all} \quad u \in D,$$

  where $\|\cdot\|_{\text{operator}}$ is the operator norm, induced by the norm $\|\cdot\|$. Thus, κ gives an measure for the quality of the approximate inverse $\widetilde{P}^{-1}$.

As well known, the problem to determine a root of a nonlinear function $G(u)$, as above, is *affine invariant*. This means, for the nonlinear equation of interest $G^n(u)$, that, for given non singular $m \times m$ matrices D^n_{L} and D^n_{R}, the problem of finding a root of $G^n(u)$ is equivalent to problem

$$\widehat{G}^n(\hat{u}) := D^n_{\mathrm{L}}G^n(D^n_{\mathrm{R}}\hat{u}) = 0, \quad \text{with} \quad u = D^n_{\mathrm{R}}\hat{u}, \tag{5.60}$$

since

$$G^n(u) = 0 \Leftrightarrow D^n_{\mathrm{L}}G^n(u) = 0 \Leftrightarrow \widehat{G}^n(\hat{u}) = 0, \quad \text{with} \quad u = D^n_{\mathrm{R}}\hat{u}.$$

Additionally, the simplified Newton method in (5.50) and (5.50) is affine invariant, as well. Therefore, we are free to solve $\widehat{G}^n(\hat{u}) = 0$, for given non singular matrices D^n_{L} and D^n_{R}, instead of (5.50) and then recover u from $u = D^n_{\mathrm{R}}\hat{u}$. For the simplified Newton

method this means, that now, instead of (5.52), we have to solve the linear equation

$$\widehat{P}_n \Delta\hat{u}_{n,j} = -\widehat{G}^n(\hat{u}_{n,j}), \tag{5.61}$$

where

$$\widehat{P}_n := D_{\mathrm{L}}^n \widetilde{P}_n D_{\mathrm{R}}^n,$$

and

$$\widehat{G}^n(\hat{u}) := D_{\mathrm{L}}^n G^n(D_{\mathrm{R}}^n \hat{u}).$$

As stated in [27], the assumptions in Theorem 16 and the statement itself are invariant against scaling with a non singular matrix D_{L}^n. Particularly, the constants ω and κ are invariant against a scaling with D_{L}^n, too. The effect of a scaling with a non singular matrix D_{R}^n is clarified further below.

In [80], Minh gives a brief summary on the error analysis in the context of solving linear equations and in the context of Newton-like methods. Based on this error analysis, he investigates a suitable choosing of the matrices D_{L}^n and D_{R}^n for application in a BDF method.

Concretely, he considers the linear equation $Pu = b$ with square matrix P, right-hand side b and solution vector u.

Then, for a perturbed linear equation, where the matrix P is perturbed by ΔP and the right-hand side b is perturbed by Δb, the solution u gets perturbed by Δu, such that

$$(P + \Delta P)(u + \Delta u) = b + \Delta b,$$

holds.

For the case $\left(\kappa(P)\frac{\|\Delta P\|_{\text{operator}}}{\|P\|_{\text{operator}}}\right) < 1$, he gives an upper bound on the relative error $\|\Delta u\|/\|u\|$ by

$$\frac{\|\Delta u\|}{\|u\|} \leq \frac{\kappa(P)}{1 - \left(\kappa(P)\frac{\|\Delta P\|_{\text{operator}}}{\|P\|_{\text{operator}}}\right)} \left(\frac{\|\Delta b\|}{\|b\|} \frac{\|\Delta P\|_{\text{operator}}}{\|P\|_{\text{operator}}}\right), \tag{5.62}$$

where $\kappa(P)$ is the condition number of P, given by

$$\kappa(P) := \frac{\|P\|_{\text{operator}}}{\|P^{-1}\|_{\text{operator}}}.$$

Although the bound in (5.62) is very pessimistic, one clearly sees that a suitable scaling minimizes the condition number $\kappa(P)$. Minh notes that an optimal scaling implicitly depends on the solution u of the linear equation itself, as also noted in [61]. This is obvioulsy computational inpracticable for a BDF method.

Therefore, on heuristic base, we use a similar strategy as presented by Minh in [80]. This is, we use diagonal matrices D_{L}^n and D_{R}^n to scale rows and columns of the matrix $\widetilde{P}_n$, if $\widehat{P}_n$ has to be refactored by LU decomposition at integration step n, to safe computational time.

First, it should be noted, that, due to condition $u = D_{\mathrm{R}}^n \hat{u}$, the relative error

$$\frac{\|\Delta\hat{u}\|}{\|\hat{u}\|} = \frac{\| \left(D_{\mathrm{R}}^n\right)^{-1} \Delta u\|}{\| \left(D_{\mathrm{R}}^n\right)^{-1} u\|}$$

is measured in a different norm. For the implemented method the norm $\|\cdot\|$ corresponds to

$$\|u\| = \frac{1}{m}\sqrt{\sum_{i=1}^{m} u^2}.$$

It is desirable that the norm $\|D_{\mathrm{R}^{-1}} u\|$ of u is "compatible" with the norm presented in Section 5.1.4. On the other hand, no additional error should be introduced by the row and column scaling of $\widetilde{P}_n$.

Therefore, like Minh [80], we use integer powers of machine base as elements for the scaling matrices D_{L}^n and D_{R}^n.

Minh states [80]: "*To avoid scaling roundoff error, integer powers of machine base are chosen for elements of D_{L}^n and D_{R}^n. In fact, if a scaling number has such a form, then the mantissa of its floating-point representation is exactly* 1*, i.e., there arises no roundoff error, when converting the original scaling number into its floating-point form. Moreover, the multiplication is faster because, to multiply a scaling number with a matrix entry, one has only to add two integers, namely the exponents of the scaling number and of the entry matrix.*"

Functions, recommended by the IEEE-754 standard for floating-point arithmetic [39], are `scalb` for multiplying 2^n and `logb` for computing the logarithm of base 2, which we use and which are used by Minh [80], as well.

Similar to [80], we have chosen D_{R}^n to be

$$D_{\mathrm{R}}^n := \begin{pmatrix} 2^{\alpha_{\mathrm{col}}^{n,1}} & 0 & \dots & 0 \\ 0 & 2^{\alpha_{\mathrm{col}}^{n,2}} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & 2^{\alpha_{\mathrm{col}}^{n,m}} \end{pmatrix}, \quad \text{where} \quad \alpha_{\mathrm{col}}^{n,i} := \left[\log_2(\widehat{W}_n^i)\right], \tag{5.63}$$

for $i \in \{1, \dots, m\}$ and $\widehat{W}_n^i$ is given as in (5.43) or (5.44) but y_n^i is replaced by $y_{n(0)}$, since y_n^i is not available yet and (5.63) is still "compatible" with the norm presented in Section 5.1.4. Here, $[x]$ denotes the closest integer to $x \in \mathbb{R}$.

Like [80], we have chosen D_{L}^n to be

$$D_{\mathrm{L}}^n := \begin{pmatrix} 2^{\alpha_{\mathrm{row}}^{n,1}} & 0 & \dots & 0 \\ 0 & 2^{\alpha_{\mathrm{row}}^{n,2}} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & 2^{\alpha_{\mathrm{row}}^{n,m}} \end{pmatrix}, \quad \text{where} \quad \alpha_{\mathrm{row}}^{n,i} := -\left[\log_2\left(\sum_{j=1}^{m} |(\widetilde{P}_n D_{\mathrm{R}}^n)^{i,j}|\right)\right]^+, \tag{5.64}$$

for $i \in \{1, \dots, m\}$ and where $[x]^+$ denotes the integer part of $x \in \mathbb{R}$, if $x \leq 0$, and otherwise the integer part of $x + 1$. Therefore, $[\,\cdot\,]^+$ defines a "round up" operator.
It should be noted, that, after scaling $\widetilde{P}_n$ with D_{L}^n and D_{R}^n, $\widehat{P}_n$ is row equilibrated. Since in [121] it is shown, that, for given square matrix P and for the condition number $\kappa^{(S)}(P)$ defined as

$$\kappa^{(S)}(P) = \frac{\|P\|_\infty}{\mathrm{glb}_{pq}(P)},$$

with

$$\mathrm{glb}_{pq}(P) := \min_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_q},$$

the condition number of $D_{\mathrm{L}}^n P$, namely $\kappa^{(S)}(D_{\mathrm{L}}^n P)$, is minimal, if all rows of the matrix $D_{\mathrm{L}}^n P$ have the same 1-norm, this seems an effective strategy. Minh [80] notes that he could reduce, by this approach, condition numbers in range $10^{14} - 10^{18}$ to in range $10^6 - 10^8$.

As stopping criterion for the simplified Newton method we either demand that the iterates have converged or, that $\|\Delta\hat{u}_j\|$, with $j < N_{\mathrm{Newton}}$, does not interfere with the *local error* test in (5.42).
We regard the simplified Newton method to be converged, if, within N_{Newton} iterates, a reduction of $\|\hat{u}_{n,N_{\mathrm{Newton}}} - \hat{u}_n^*\|$ by a given factor γ_{Newton} is achieved, i.e.

$$\|\hat{u}_{n,N_{\mathrm{Newton}}} - \hat{u}_n^*\| \leq \gamma_{\mathrm{Newton}} \|\hat{u}_{n,0} - \hat{u}_n^*\|$$

and for each iteration it holds that

$$\delta_{n,j} \leq (\gamma_{\mathrm{Newton}})^{\frac{1}{N_{\mathrm{Newton}}}} := \hat{\delta}_{\mathrm{C}} \quad \text{for} \quad j \in \{1, \dots, N_{\mathrm{Newton}} - 1\},$$

where $\delta_{n,j}$ is approximated by

$$\delta_{n,j} = \frac{\Delta\hat{u}_{n,j}}{\Delta\hat{u}_{n,j-1}}$$

and $\hat{u}_n^*$ denotes an exact solution of the nonlinear equation in (5.60).
On the other hand, we regard, that a $\|\Delta\hat{u}_{n,j}\|$, with $j < N_{\mathrm{Newton}}$, does not interfere with

the *local error* test, if

$$\|\Delta\hat{u}_{n,j}\| \leq \epsilon_{\text{Newton}} \left| \frac{1}{\alpha_{n_0}} \frac{1}{(1-1/\tilde{\alpha}_{n_0})} \right|^{-1} \epsilon, \tag{5.65}$$

where ϵ_{Newton} is a user given threshold with $\epsilon_{\text{Newton}} < 1$. Here, we use $\epsilon_{\text{Newton}} = 0.01$ for the calculation of the numerical results in Section 8.
If a $\|\Delta\hat{u}_{n,j}\|$, with $j < N_{\text{Newton}}$, does not interfere with the *local error* test, we stop the simplified Newton method (compare (5.34), (5.42) and consider the scaling matrix D^n_{R} in (5.63)).
Obviously, (5.65) can only be achived if

$$\delta_{n,j} \leq \left(\frac{\epsilon_{\text{Newton}} \left| \frac{1}{\alpha_{n_0}} \frac{1}{(1-1/\tilde{\alpha}_{n_0})} \right|^{-1} \epsilon}{\|\Delta\hat{u}_{n,0}\|} \right)^{\frac{1}{N_{\text{Newton}}-1}} := \hat{\delta}_{n,\text{I}} \quad \text{for} \quad j \in \{1, \ldots, N_{\text{Newton}} - 1\}.$$

Therefore, if

$$\delta_j > \max\{\hat{\delta}_{\text{C}}, \hat{\delta}_{n,\text{I}}\},$$

or no convergence of the simplified Newton method can be achieved within N_{Newton}, we consider the simplified Newton method to be failed.
Further, if

$$\|\Delta\hat{u}_{n,j}\| < u_{\text{r}}\|\hat{u}_{n,j+1}\|,$$

this is, $\|\Delta\hat{u}_{n,j}\|$ is less than a unit roundoff of $\|\hat{u}_{n,j+1}\|$, where, again, u_{r} is the unit roundoff of the machine, the user has clearly asked for too much accuracy [101].
Therefore, we also stop the simplified Newton method in the j-th iteration , if

$$\|\Delta\hat{u}_{n,j}\| \leq 2u_{\text{r}}\|\hat{u}_{n,j+1}\|. \tag{5.66}$$

There are mainly two reasons that the simplified Newton method fails.
First, the approximation $\widehat{P}_n$ to the Jacobian $\hat{G}_{\hat{u}}(\hat{u}_{n,0})$ might be too insufficient in view of assumption (5.54) and second, the start value $\hat{u}_{n,0}$ might be to bad.
As noted above, $\widehat{P}_n$ is kept for successive intgeration steps as long as convergence by the simplified Newton method can be achieved. Therefore, if the simplified Newton method fails, we first update $\widehat{P}_n$.
The exact Jacobian of $G^n(y_{n(0)})$ is given by

$$P_n(y_{n(0)}) = \mathbb{1} - \frac{h_n}{l^1_n} F_y(y_{n_{n(0)}}, t_n, p_0). \tag{5.67}$$

Since the computation of the Jacobian $F_y(y_{n(0)}, t_n, p_0)$ is more likely to be computational expensive than the LU decomposition of $\widehat{P}_n$, we store the Jacobian $F_y^n = F_y(y_{n(0)}, t_n, p_0)$ for successive integration steps, where F_y^n denotes the stored Jacobian. Then, if an update of $\widehat{P}_n$ has to be performed, we first calculate (5.67) with an former stored Jacobian $F_y^n = F_y^{n-1}$ instead of $F_y(y_{n(0)}, t_n, p_0)$ but current coefficients h_n and l_n^1. Now, we rescale $\widetilde{P}_n$ with updated scaling matrices D_{L}^n and D_{R}^n and finally restart the simplified Newton method.
If, again, the simplified Newton method does not converge, we recalculate $\widetilde{P}_n$ with an updated Jacobian $F_y(y_{n(0)}, t_n, p_0)$, too.
If still convergence can not be achieved, the step size of the BDF method, at the current integration step, is reduced by a factor of 1/2. The procedure of step size change is, discussed in Section 5.1.6.
In summary, we have the following algorithm for the calculation of the corrector vector e_n.

Algorithm 9.

Data: *Stored LU decomposition of $\widehat{P}_{n-1}$, stored Jacobian F_y^{n-1}, stored scaling matrices D_{L}^{n-1} and D_{R}^{n-1}, step size h_n, coefficient l_n^1, parameter vector p_0, current integration time t_n, user given error threshold ϵ, user given threshold $\epsilon_{\mathrm{Newton}}$, user given factor γ_{Newton}, coefficient $\alpha_{n_0} = -l_n^1$, coefficient $\tilde{\alpha}_{n_0}$, start value $y_{n(0)} = \pi_{n-1,k}(t_n)$, number of allowed Newton iterations N_{Newton}.*

Step 0. *If no stored data is available (if the integration starts) goto Step 2. Else goto Step 1.*

Step 1. *Set $\widehat{P}_n := \widehat{P}_{n-1}$, set $F_y^n = F_y^{n-1}$, set $D_{\mathrm{R}}^n = D_{\mathrm{R}}^{n-1}$, set $D_{\mathrm{R}}^n = D_{\mathrm{R}}^{n-1}$ and got Step 4.*
Step 2. *Set $F_y^n := F_y(y_{n_{n(0)}}, t_n, p_0)$.*
Step 3. *Set*

$$\widetilde{P}_n = \mathbb{1} - \frac{h_n}{l_n^1} F_y^n.$$

Calculate and store scaling matrices D_{L}^n and D_{R}^n according to (5.63) and (5.64). Set

$$\widehat{P}_n = D_{\mathrm{L}}^n \widetilde{P}_n D_{\mathrm{R}}^n.$$

Calculate and store LU decomposition of $\widehat{P}_n$.

Step 4. *Set* $j = 0$*, set*

$$\hat{u}_{n,0} = (D_{\mathrm{R}}^{n})^{-1}\, y_{n(0)}.$$

Step 5. *Calculate* $\Delta\hat{u}_{n,j}$ *according to (5.61) using the stored LU decomposition of* $\widehat{P}_n$*, set*

$$\hat{u}_{n,j+1} = \hat{u}_{n,j} + \Delta\hat{u}_{n,j}$$

and set $j = j + 1$*.*

Step 6. *If* $\|\Delta\hat{u}_{n,j-1}\| < u_{\mathrm{r}}\|\hat{u}_{n,j}\|$ *or*

$$\|\Delta\hat{u}_{n,j-1}\| \leq \epsilon_{\mathrm{Newton}} \left| \frac{1}{\alpha_{n_0}} \frac{1}{(1 - 1/\tilde{\alpha}_{n_0})} \right|^{-1} \epsilon$$

consider the simplified Newton method to be successful with $e_n = D_{\mathrm{R}}^{n}\hat{u}_{n,j}$ *and stop.*

Step 7. *If* $j > 1$ *and if*

$$\delta_{n,j-1} > \max\{\hat{\delta}_C, \hat{\delta}_{n,I}\},$$

then,

- *if Step 3 has not been performed, goto Step 3,*
- *else, if Step 2 has not been performed, goto Step 2,*
- *if Step 2 has been performed, set* $e_n = D_{\mathrm{R}}^{n}\hat{u}_{n,j}$*, consider the Newton method to be failed and stop.*

Step 8. *If* $j < N_{\mathrm{Newton}}$ *goto Step 5, else consider the simplified Newton method to be successful with* $e_n = D_{\mathrm{R}}^{n}\hat{u}_{n,j}$ *and stop.*

In the next section we present strategies for the selection of step size and order of the implemented BDF method.

5.1.6. Strategies for the selection of step size and order of the BDF method

If the error test in (5.45) or the convergence of the simplified Newton method in Section 5.1.5 fails, the step size h_n, of the current integration step n, has to be reduced.
While the step size h_n gets reduced by a factor of 0.5, if the simplified Newton method in Section 5.1.5 fails, a new step size h'_n is estimated for the case that the error test in (5.45) is not fulfilled, for the current integration step n at order k. The new step size

has to be estimated with the the goal that the new step size h'_n successfully passes the local error test in (5.45).
On the other hand, if a step is accepted a new step size h_{n+1} gets estimated for the successive integration step $n+1$. Again, this is based on the goal that the new step size h_{n+1} successfully passes the local error test in (5.45).
The estimators for new step sizes h'_n or h_{n+1}, respectively, are based on the asymptotic behavior of the *local error* formula (5.30). In many codes like `EPISODE` [32], `DASSL` [86] or `CVODE` [62] the estimate is based on the assumption that the former steps have been taken at same step size h, which leads to the step size factor estimate

$$\theta^n_{\text{classical}}(k) := \left(\frac{\epsilon_{\text{relax}}}{\|LE_n(k)\|_{W_n}} \right)^{\frac{1}{k+1}}, \tag{5.68}$$

for the current integration order k of the BDF method at the n-th integration step. This estimate is used both for the reduction of the step size in case of a failure of the error test in (5.45) and for the estimation of a new step size h_{n+1} for the successive intgeration step $n+1$.
In [102], Shampine and Bogacki investigated the behavior of the *local error* $LE_n(k)$ on uniform and variable grids.
They showed that the behavior of the *local error* $LE_n(k)$ on uniform grids differs from the behavior on variable ones and therefore (5.68), based on the assumption that the former steps have been taken at same step size h, might not be appropriate for all cases.
In [36], Calvo and Rández suggested new heuristic factors, which are based on the behavior of the *local error* $LE_n(k)$ on variable grids, namely $\theta^n_{\text{rejected}}(k)$, given by

$$\theta^n_{\text{rejected}}(k) := \begin{cases} \sqrt{(\mu_n)^{\frac{1}{2}+\frac{1}{k+1}}} & \text{if } 0.05 \le \mu_n \le 1 \\ \nu_n \, (\mu_n)^{\frac{1}{k+1}} + (1-\nu_n) \, (\mu_n)^{\frac{1}{2}} & \text{if } \mu_n < 0.05, \end{cases} \tag{5.69}$$

for the calculation of a new step size

$$h'_n = \theta_{\text{rejected}}(k) h_n. \tag{5.70}$$

The factor μ_n is defined by

$$\mu_n := \frac{c_{\text{safety}} \epsilon_{\text{relax}}}{\|LE_n(k)\|_{W_n}},$$

where c_{safety} is some safety factor, which is set to $c_{\text{safety}} = 0.9$ for the calculation of the numerical results in Section 8 and

$$\nu_n := \min \left\{ \left(\frac{\prod_{i=2}^{k} \xi_{n,i}}{l_n^1 \prod_{i=2}^{k} (\xi_{n,i} - 1)} \right)^{\frac{1}{2}} (\mu_n)^{\frac{k-1}{2k+2}}, 1 \right\}.$$

Second, $\theta_{\text{newStep}}(k)$ is given by

$$\theta_{\text{newStep}}(k) := \left(\frac{\sum_{i=1}^{k} 1/i}{\sum_{i=1}^{k} 1/\xi_{n,i}} \frac{\prod_{i=2}^{k} \xi_{n,i}}{k!} \right)^{\frac{1}{(k+1)}} (\mu_n)^{\frac{1}{(k+1)}},$$

for the estimation of a new step size

$$h_{n+1} = \theta_{\text{newStep}}(k) h_n, \tag{5.71}$$

after a step acceptance at integration step n.
The factor $\theta_{\text{newStep}}(k)$ is not used directly for the estimation of a new step size h_{n+1} in (5.71), but the factor

$$\hat{\theta}_{\text{newStep}}(k) := \min\{\theta_{\text{newStep}}(k), \theta^{\max}_{\text{newStep}}(k)\}, \tag{5.72}$$

where $\theta^{\max}_{\text{newStep}}(k)$ gives an upper bound on the factor $\theta_{\text{newStep}}(k)$ for the k-th order BDF method.
This bounds are neccessary such that the k-th order BDF method retains *zero stable*, an important prerequisite for the convergence of the BDF method (see e.g. Theorem 5.8 in [58]). Several authors [109, 56, 34, 33, 35] have constructed such bounds in the context of BDF methods and in the context of *linear multistep methods*, as well.
In [33], Calvo et al. determined upper bounds such that any combination of BDF forumlas with $k \leq k^{\max}_{\text{order}}$ retains *zero stable*. They have excluded the case $k \leq k^{\max}_{\text{order}} = 6$, since as they state:

"*(. . .) it has been shown by Philippe [87] that even on a uniform grid zero stability cannot hold if any combination of BDF formulas with* $m \leq 6$ *is allowed, but for arbitrary combinations of BDF formulas with* $m \leq 5$ *the stability is maintained.*"

Their results are given in Table 5.1. It should be noted that there might exist looser

	$k^{\max}_{\text{order}} = 3$	$k^{\max}_{\text{order}} = 4$	$k^{\max}_{\text{order}} = 5$
$\theta^{\max}_{\text{newStep}}(2) =$	2.391	2.137	2.117
$\theta^{\max}_{\text{newStep}}(3) =$	1.476	1.321	1.255
$\theta^{\max}_{\text{newStep}}(4) =$	–	1.101	1.088
$\theta^{\max}_{\text{newStep}}(5) =$	–	–	0.964

Table 5.1.: Upper bounds on $\theta^{\max}_{\text{newStep}}(k)$, for $k \in \{1, \ldots, 5\}$ given by Calvo et al. in [33]

bounds, such that any combination of BDF forumlas with $k \leq k^{\max}_{\text{order}}$ retains *zero stable*, too, and therefore these bounds are too pessimistic.
In [35], Calvo et al. introduce the concept of A_0-stability. There, for the scalar test equation $y = \lambda y$, (5.6) is equivalently rewritten in the matrix form

$$U_n = \Omega_n U_{n-1},$$

where $U_n = (y_n, y_{n-1}, \ldots, y_{n-k+1})^T \in \mathbb{R}$ and Ω_n is the so called *propagation matrix* given by

$$\Omega_n^{ij} = \begin{cases} -\frac{\alpha_{n_j}}{\alpha_{n_0} - z_n} & \text{if} \quad i = 1, \\ \delta_{i,j+1} & \text{if} \quad i \leq 2, \end{cases}$$

where $z_n = h_n \lambda$ and $\delta_{i,j}$ is the Kronecker delta. Then, since the coefficients α_{n_i} are rational functions of the step size ratios h_i/h_{i-1} for $i \in \{n-k+2, \ldots, n\}$, Calvo et al. determine bounds on the step size ratios h_i/h_{i-1}, such that the spectral radius of the propagation matrix is ≤ 1, here with $z_n = 0$, associated to *zero stability*.
They state [35]: "*Although the stability derived from the spectral radius condition does not guarantee boundedness of products of propagation matrices, it turns out to be a realistic criterion to compare the stability of different methods*".
The resulting bounds $\theta^{\max}_{\text{newStep}}(k)$ on the factors $\theta_{\text{newStep}}(k)$ for $k \in \{1, \ldots, 5\}$, such that the propagation matrix has a spectral radius smaller than one, are given in Tabel 5.2.

$\theta^{\max}_{\text{newStep}}(2) = 2.414$	$\theta^{\max}_{\text{newStep}}(3) = 1.618$
$\theta^{\max}_{\text{newStep}}(4) = 1.280$	$\theta^{\max}_{\text{newStep}}(5) = 1.127$

Table 5.2.: Upper bounds on $\theta^{\max}_{\text{newStep}}(k)$, for $k \in \{1, \ldots, 5\}$ given by Calvo et al. in [35]

Again the order $k = 6$ is excluded in [35].
For the implemented BDF method we use the heuristic factors (5.69) and (5.72) together with the upper bounds given in Tabel 5.2 instead of (5.68). Since we additionally allow for the 6−th order BDF method, we set the additional upper bound $\theta^{\max}_{\text{newStep}}(6) = 1.064$ in the implemented BDF method, which works well in practice for the examples in Chapter 8.

According to [101], if $h'_n < u_{\text{r}} t_{n-1}$, where again u_{r} is the unit roundoff of the machine, then

$$t_{n-1} \equiv t_n$$

in terms of numerical accuracy. Particularly, this means that the integration gets stalled. Therefore, if

$$h'_n < 4u_{\text{r}} t_{n-1}, \tag{5.73}$$

where h'_n is calculated with (5.70) we set

$$h'_n = 4u_{\text{r}} t_{n-1} \tag{5.74}$$

and accept the step, if (5.46) might not be fulfilled. Of course, this mechanism is also implemented for the step size reduction due to a failure of the simplified Newton method. With the same thoughts as above, the estimate of the next step size h_{n+1} is modfied by

$$h_{n+1} = \max\{\hat{\theta}_{\text{newStep}}(k) h_n, 4u_{\text{r}} t_n\}.$$

Next, we discuss the adaption of the Nordiesk array $z_n(h_n)$ if the step size h_n changes by a factor of θ, i.e

$$h_n \to h_n \theta. \tag{5.75}$$

Obviously the desired Nordsiek array $z_n(h_n\theta)$ is given by

$$z_n(h_n\theta) := \left(y_n \quad \theta \left[h_n y_n^{(1)} \right] \quad \theta^2 \left[h_n^2 y_n^{(2)}/2 \right] \quad \dots \quad \theta^k \left[h_n^k y_n^{(k)}/k! \right] \right),$$

compare (5.9). Therefore, $z_n(h_n\theta)$ can be calculated by

$$z_n(h_n\theta) = z_n(h_n)\vec{\theta},$$

where $\vec{\theta}$ denotes the vector

$$\vec{\theta} = \left(1 \quad \theta^1 \quad \dots \quad \theta^k \right)^T.$$

If, for $k+1$ successive integration steps at order k of the BDF method, the algorithm passes without any step failures due to a failure in (5.46) or due to a failor of the simplified Newton method, we allow an order change from $k \to k-1$ or $k \to k+1$ of the BDF method.
For that, using the local error formulas in (5.38) and (5.41) we first estimate the step sizes

$$h_{n+1}(k) := \max\{\hat{\theta}_{\text{newStep}}(k)h_n, 4u_{\text{r}}t_n\},$$

$$h_{n+1}(k-1) := \max\{\hat{\theta}_{\text{newStep}}(k-1)h_n, 4u_{\text{r}}t_n\}$$

and

$$h_{n+1}(k+1) := \max\{\hat{\theta}_{\text{newStep}}(k+1)h_n, 4u_{\text{r}}t_n\}$$

and choose the consecutive order k' for the next integration step $n+1$ of the BDF method, such that the next step size h_{n+1} is maximal, i.e.

$$h_{n+1} := \max\{h_{n+1}(k), h_{n+1}(k-1), h_{n+1}(k+1)\},$$

with corresponding order k'. If $h_{n+1} = 4u_{\text{r}}t_n$, the order is retained.
If the order of of the BDF method changes from $k \to k'$ at finished integration step n, the current Nordiesk array has to be modfied, i.e. $z_n(k) \to z_n(k')$.
For the case $k' = k-1$, the Nordiesk array $z_n(k)$ represents the polynomial $\pi_{n,k}(t)$ and $z_n(k')$ represents the polynomial $\pi_{n,k-1}(t)$, respectively. Again, as presented in [32], we define a difference polynomial, here $\Delta_n^{\downarrow}(t) := \pi_{n,k}(t) - \pi_{n,k-1}(t)$ and therefore $\Delta_n^{\downarrow}(t)$ satisfies

$$\Delta_n^{\downarrow}(t_{n-i}) = 0 \quad \text{for all} \quad i \in \{0, 1, \ldots, k-2\} \quad \text{and} \quad \dot{\Delta}_n^{\downarrow}(t_n) = 0.$$

Additionally, the leading coefficient of $\Delta_n^{\downarrow}(t)$ has to be the same as the leading coefficient of $\pi_{n,k}(t)$, i.e. $y_n^{(k)}/k!$ and therefore

$$\Delta_n^{\downarrow}(t) = (t-t_n)^2 \left[\prod_{i=1}^{k-1}(t-t_{n-i})\right] \frac{y_n^{(k)}}{k!}. \tag{5.76}$$

Again, we perform a change of variables by (5.13) in (5.76) and thus

$$\Delta_n^{\downarrow}(t) = \Delta_n^{\downarrow}(t_n + h_n x_n) = d_n(x_n) h_n^k \frac{y_n^{(k)}}{k!}, \quad \text{with} \quad d_n(x) := x^2 \prod_{i=1}^{k-2}(x + \xi_{n,i}).$$

Again, consider the coefficients d_n^i of $d_n(x)$, i.e.

$$d_n(x) = \sum_{i=0}^{k} d_n^i x^i.$$

With similar thoughts as in (5.16), one can easily see, that $d_n^i h_n^k y_n^{(k)}/k!$ has to be subtracted from column i of the Nordiesk array $z_n(k)$ for $i \in \{2, \dots k-1\}$ to yield $z_n(k')$, whereby column k is deleted in $z_n(k')$.
The coefficients d_n^i for $i \in \{2, \dots k-1\}$ are calculated by Algorithm 10, which can be easily verified by induction.

Algorithm 10.

Data: *Order of BDF method k, Auxiliar quantities $\xi_{n,q}$ for $q \in \{1, \dots, k\}$.*

Step 0. *Set $d_n^0 = 0$. Set $d_n^1 = 0$ Set $i = 1$.*
Step 1. *Set $i = i + 1$. If $i > k$ stop.*
Step 2. *Set $j = i - 1$. Set*

$$d_n^i = 1.$$

Step 3. *If $j < 2$ goto Step 1. Else set*

$$d_n^j = d_n^{j-1} + d_n^j \xi_{n,i-2},$$

set $j = j - 1$ and goto Step 3.

On the other hand, if $k' = k + 1$, $z_n(k')$ corresponds to a polynomial $\pi_{n,k+1}(t)$ of degree $k + 1$ or less, such that

$$\pi_{t_{n-i}}(t) = y_{n-i} \quad \text{for all} \quad i \in \{0, 1, \dots, k\} \quad \text{and} \quad \dot{\pi}_{n,k+1}(t_n) = \dot{y}_n. \tag{5.77}$$

Obviously, in view of the construction strategy in (5.4) and (5.5), $\pi_{n,k}(t)$ already satisfies the interpolation conditions in (5.77) and hence $\pi_{n,k}(t) = \pi_{n,k+1}(t)$. Therefore, as Byrne and Hindmarsh state [32]: "*Thus columns 0 to k of $z_n(k)$ need no adjustment, and column $k + 1$ of $z_n(k')$ is* 0".
If both the order and the step size have to be changed, then first the order ist changed and the step size is changed afterwards.

5.1.7. Initilization of Nordsieck arrays and estimation of start step size h_1

The implemented BDF method is initialized as one step method, i.e. the order of the initial step $n = 1$ is $k = 1$.
The start Nordiesk array z_0 is initialized as

$$z_0 := \begin{pmatrix} y_{\mathrm{I}} & h_{\mathrm{init}} F(y_{\mathrm{I}}, t^{\mathrm{init}}, p_0) \end{pmatrix}, \tag{5.78}$$

where h_{init} is estimated with the goal, that y_1 satisfies the *local error* test in (5.46).
For that, with similar thoughts as presented in [29], based on the asymptotic behavior of the *local error* given by (5.30) we demand, that

$$\left\| \frac{h_{\mathrm{init}}^2}{2} \ddot{y}(t^{\mathrm{init}}) \right\|_{W_n} = \epsilon R_{\mathrm{SE}}, \tag{5.79}$$

with

$$\ddot{y}(t') := \left. \frac{\mathrm{d}^2 y(t)}{\mathrm{d}t^2} \right|_{t=t'},$$

and R_{SE} is some user given constant $0 < R_{\mathrm{SE}} \leq 1$. We have chosen $R_{\mathrm{SE}} = 1$, which worked well for the examples in Section 8. The constant ϵ is the user given threshold introduced in Section 5.1.4.
If y_{I} is the zero vector, then the estimate of the initial step size h_1 fails. In this case we set $h_1 = h_{\max}$, where $h_{\max}$ is some user given maximal step size. For the examples in Section 8 we always set $h_{\max}$ to be $h_{\max} = 1000$, which worked well for these cases.

5.1.8. Calculation of the solution vector at t^{end}

If, for the first time, the n-th integration step, is accepted with $t_n \geq t^{\mathrm{end}}(1 - 2u_{\mathrm{r}})$, where, again, u_{r} is the unit roundoff of the machine at order k of the BDF method, the integration procedure is stopped and the desired solution vector $y_{t^{\mathrm{end}}}$ at time point t^{end} is calculated by

$$y_{t^{\mathrm{end}}} = \sum_{i=0}^{k} \left(\frac{y_{t^{\mathrm{end}}} - t_n}{h_n} \right)^i \left[h_n^i y_n^{(i)} / i! \right], \tag{5.80}$$

using the entries of the current Nordsieck array z_n.

5.1.9. Algorithmic scheme of the implemented BDF method

In this section, we present the overall algorithmic scheme of the implemented BDF method in Algorithm 11.

Algorithm 11.

Data: *Right hand side function* $F : \mathbb{R}^m \times \mathbb{R} \times \mathbb{R}^p \to \mathbb{R}^m$*, initial time* t^{init}*, end time* t^{end}*, start vector* $y_{\text{I}} \in \mathbb{R}^m$*, parameter vector* $p_0 \in \mathbb{R}^p$*, maximum number of integration steps* $N_{\max}$*, maximum number of allowed simplified Newton steps, user given error threshold* ϵ*, user given Newton accuracy* ϵ_{Newton}*, user given Newton factor* γ_{Newton}*, user given weight vector atol* $\in \mathbb{R}^m_+$*, maximum step size* $h_{\max}$*.*

Step 0. *Set* $n = 1$*, set* $n_{\text{SO}} = 1$*. Set initial Nordsiek array* z_0 *according to (5.78). Set* h_1 *according to (5.79). Set* $t_0 = t^{\text{init}}$ *and set* $t_1 = t^{\text{init}} + h_1$*. Set order* $k = 1$*.*

Step 1. *Calculate predictor Nordsiek array* $z_{n(0)}$ *according to (5.8).*
Step 2. *Calculate the coefficients* $\xi_{n,i}$ *for* $i \in \{1, \ldots, k\}$*. Calculate the coefficient vector* l_n *according to Algorithm 8. Set* $\alpha_{n_0} = -l_n^1$*. Calculate* $\tilde{\alpha}_{n_0}$ *according to (5.37).*

Step 3. *Calculate* e_n *with Algorithm 9. If Algorithm 9 reports a failure and* $h_n > 4u_{\text{r}}t_{n-1}$ *goto Step 4, else goto Step 5.*

Step 4. *Set* $n_{\text{SO}} = 0$*. If* $0.5h_n > 4u_{\text{r}}t_{n-1}$*, set* $h_n = 0.5h_n$*, set* $t_n = t_{n-1} + h_n$*, rescale* z_{n-1} *according to (5.75) with* $\theta = 0.5$*, else set* $\theta = 4u_{\text{r}}t_{n-1}/h_n$*, set* $h_n = 4u_{\text{r}}t_{n-1}$*, set* $t_n = t_{n-1} + h_n$*, rescale* z_{n-1} *according to (5.75) with* θ*. Goto Step 1 and force in Step 3 a recalculation of* $\widehat{P}_n$ *and scaling matrices* D^n_{L} *and* D^n_{R} *using an updated Jacobian* F^n_y*.*

Step 5. *Calculate vector* W_n *according to (5.43) or (5.44) (defined by user), calculate local error vector* $LE_n(k)$ *according to (5.36) calculate* ϵ_{relax} *according to (5.45).*

Step 6. *If*

$$\|LE_n(k)\|_{W_n} \leq \epsilon_{\text{relax}}$$

or $h_n \leq 4u_{\text{r}}t_{n-1}$*, then goto Step 8, else goto Step 7.*

Step 7. *Set* $n_{\text{SO}} = 0$*. Calculate*

$$\theta = \max\{\theta^n_{\text{rejected}}(k), 4u_{\text{r}}t_{n-1}/h_n\},$$

where $\theta^n_{\text{rejected}}(k)$ *is calculated according to (5.69). Set* $h_n = \theta h_n$*, set* $t_n = t_{n-1} + h_n$*, rescale* z_{n-1} *according to (5.75) with* θ *and goto Step 1.*

Step 8. *Set* $n_{\mathrm{SO}} = n_{\mathrm{SO}} + 1$. *Calculate the Nordsiek array* z_n *according to (5.17). If*

$$t_n \geq t^{\mathrm{end}}(1 - 2u_{\mathrm{r}})$$

goto Step 9, else goto Step 10.

Step 9. *Calculate* $y_{t^{\mathrm{end}}}$ *according to (5.80) and stop.*
Step 10. *If* $n_{\mathrm{SO}} < k + 1$, *set* $n = n + 1$, *set*

$$h_n = \min\{\max\{\hat{\theta}_{\mathrm{newStep}}(k)h_{n-1}, 4u_{\mathrm{r}}t_{n-1}\}, h_{\max}\},$$

set $\theta = h_n/h_{n-1}$, $t_n = t_{n-1} + h_n$, *rescale* z_{n-1} *according to (5.75) with* θ *and goto Step 1, else goto Step 11.*

Step 11. *Calculate*

$$h_{n+1}(k) := \max\{\hat{\theta}_{\mathrm{newStep}}(k)h_n, 4u_{\mathrm{r}}t_n\},$$

$$h_{n+1}(k-1) := \max\{\hat{\theta}_{\mathrm{newStep}}(k-1)h_n, 4u_{\mathrm{r}}t_n\}$$

and

$$h_{n+1}(k+1) := \max\{\hat{\theta}_{\mathrm{newStep}}(k+1)h_n, 4u_{\mathrm{r}}t_n\}.$$

Calculate

$$k' = \arg \max_{k'' \in \{k-1,k,k+1\}} h_n(k'').$$

If $h_{n+1}(k') = h_{n+1}(k)$, *set* $h_{n+1} = h_{n+1}(k)$, *set* $n = n + 1$, *set* $\theta = h_n/h_{n-1}$, $t_n = t_{n-1} + h_n$, *rescale* z_{n-1} *according to (5.75) with* θ *and goto Step 1, else goto Step 12.*

Step 12. *Set* $n_{\mathrm{SO}} = 0$. *If* $h_{n+1}(k') = h_{n+1}(k-1)$ *calculate coefficient vector* d_n *using Algorithm 10, substract* $d_n^i h_n^k y_n^{(k)}/k!$ *from column* i *of the Nordiesk array* z_n *for* $i \in \{2, \ldots k-1\}$, *delete column* k *of* z_n, *set* $k = k-1$, *set* $h_{n+1} = h_{n+1}(k)$, *set* $n = n+1$, *set* $\theta = h_n/h_{n-1}$, $t_n = t_{n-1} + h_n$, *rescale* z_{n-1} *according to (5.75) with* θ *and goto Step 1, else goto Step 13.*

Step 13. *Augment* z_n *by column* $k + 1$ *and set column* $k + 1$ *to the zero vector, set* $k = k + 1$, *set* $h_{n+1} = h_{n+1}(k)$, *set* $n = n + 1$, *set* $\theta = h_n/h_{n-1}$, $t_n = t_{n-1} + h_n$, *rescale* z_{n-1} *according to (5.75) with* θ *and goto Step 1.*

The algorithm is implemented in `C++`.
For the evaluation of the Jacobian and as well for the sensitivity generation, presented in Section 5.2, we use the AD package `CppAD` (see Chapter 4).
Further, each summation contained in Algorithm 8 and Algorithm 10 is performed using compensated summation [61].

5.2. Sensitivity Generation for Ordinary Differential Equations

Beside calculating an approximate solution $y_{t^{\text{end}}}$ of IVP in (5.1) and (5.2) at time point t^{end}, we are interested in calculating approximations to (higher) derivatives of $y(t^{\text{end}})$ with respect to the initial condition $y_{\text{I}} \in \mathbb{R}^m$ and possibly with respect to a parameter vector $p_0 \in \mathbb{R}^p$, assuming $F(y(t), t, p_0)$, as given in (5.1), has as many partial derivatives as needed.
In the following, these derivatives of $y(t^{\text{end}})$ are called *sensitivities* , since these are the solution of the corresponding forward sensitivity differential equation, e.g. given by

$$\dot{s}_{y^i}(t) = F_y(y(t), t, p_0) s_{y^i}(t), \tag{5.81}$$

with initial condition

$$s_{y^i}(t^{\text{init}}) = \vec{e}_i,$$

for the first order *sensitivities* with respect to the i-th component of the initial condition vector $y_{\text{I}} \in \mathbb{R}^m$, where $\vec{e}_i$ represents the canonical unit vector with a 1 in the i-th coordinate and 0's elsewhere, for $i \in \{1, \ldots, m\}$.
The corresponding first order *sensitivities* with respect to the j-th component of the parameter vector $p_0 \in \mathbb{R}^p$ are given by

$$\dot{s}_{p^j}(t) = F_y(y(t), t, p_0) s_{p^j}(t) + \frac{\partial F(y(t), t, p_0)}{\partial p^j} \quad \text{with} \quad s_{p^j}(t^{\text{init}}) = 0, \tag{5.82}$$

for $j \in \{1, \ldots, p\}$.
For more theoretical details we refer to the textbooks, e.g. [58, 78].
In principal, for calculating the desired *sensitivities*, one can augment the initial IVP as given in (5.1) and (5.2) by the desired set of sensitivity differential equations and then solve the resulting system as a whole, using the integration method presented in Section 5.1.
This approach leads to a high dimensional nonlinear mapping $G^n(u)$ in (5.49) for the calculation of the corrector vector e_n at step n of the integration process. Thus, depending on the set of desired *sensitivities*, the computation of a root of $G^n(u)$ can get computational very expensive.

Therefore, for the calculation of the desired *sensitivities*, we follow the principal of *Internal Numerical Differentiation*, developed by Bock [25, 26] and applied to a BDF integration method, based on a modified divided differences interpolation scheme [23, 24], by Albersmeyer and Bock [2, 1].
In [1], Albersmeyer states: "*The idea of Internal Numerical Differentiation (IND) [25, 26] is to freeze the adaptive components of the integrator and to differentiate not the whole adaptive integrator code, but the adaptively generated discretization scheme (fixing the used step sizes, orders, iteration matrices and number of Newton-like iterations). This scheme can be interpreted as a sequence of differentiable mappings, each leading from the solution at one timepoint of the discretization grid via intermediate values to the next. Hence it can be differentiated, for example, using finite differences, the complex step method or the techniques of automatic differentiation. This leads to numerical schemes for the computation of the sensitivities that are strongly intertwined with the computation of the nominal solution*".

Since the numerical schemes for the computation of the *sensitivities* are strongly intertwined with the computation of the nominal solution, information, generated by the calculation of the nominal solution, can be efficiently reused for the calculation of the *sensitivities*, as stated in [1] and therefore this approach offers a computational efficient alternative.
Here, IND is applied to the intgration method presented in Section 5.1, which is based on Nordsieck array polynomial interpolation, using the AD techniques as presented in Chapter 4, which are based on Taylor Series propagation.
Again, as in in Chapter 4, this results in two basic operation modes for the *sensitivity* generation. The first one, presented in Section 5.2.1, is based on the *forward* mode of AD (Section 4.1), whereas the second one, presented in Section 5.2.2, is based on the *reverse* mode of AD (Section 4.2).
Preliminary to the discussion of both operation modes, we first state the preconditions on applying the principal of IND using the AD techniques of Chapter 4, which arise from the implemented BDF method.
Since all adaptive components of the integration method are "frozen" in applying the IND principle for the calculation of the desired *sensitivities*, we only have to consider, after a successful calculation of a desired approximation $y_{t^{\text{end}}}$ to $y(t^{\text{end}})$ at t^{end}, the underlying elementary operation sequence. In other words, the performed arithmetic operations whithin the integration process. Whereas, the discretization scheme is considered fixed as well as the order of the BDF method at each step n and the integration steps n' after which an order change is performed.
Because during the integration process nonlinear equations have to be solved, this non-

linear equations have to be tackled by AD, too. As elaborate in Section 4.3, there are in principal two ways in doing so.
For the first one (Section 4.3.1), the simplified Newton method (Algorithm 9) gets tackled by AD, directly. Again, thereby all adaptive elements of Algorithm 9 gets frozen, i.e. the number of performed simplified Newton steps, the scaling matrices D_{L}^n, D_{R}^n and the scaled approximation to the Jacobian $\widehat{P}_n$ at each integration step n, to stick to the IND principle.
For the second one (Section 4.3.2) the nonlinear equation is treated as an elementary AD operation itself.
In general, during the calculation of the desired approximate solution, $y_{t^{\mathrm{end}}}$, the underlying elementary operation sequence for integration has to be stored. Here, the operation sequence gets stored via a list in which a sequence of identifiers gets saved. One identifier corresponds to a distinct elementary integration operation.
In the following, we give an overview of these elementary integration operations together with the corresponding identifiers:

- Initialization of the Nordsieck array z_0 by

$$z_0 := \begin{pmatrix} y_{\mathrm{I}} & h_{\mathrm{init}} F(y_{\mathrm{I}}, t^{\mathrm{init}}, p_0) \end{pmatrix},$$

  where h_{init} is considered to be a constant. Identifier: “iaO”.

- Calculation of the predictor array $z_{n(0)}$ at integration step n by

$$z_{n(0)} = z_{n-1} A[k],$$

  for constant matrix $A[k]$ and where the current order k is considered to be a constant, as well. Identifier: “paO”.

- Calculation of the corrector vector e_n at integration step n by following elementary operations:
  - Calculation of $\hat{u}_{n,0}$ by

$$\hat{u}_{n,0} = (D_{\mathrm{R}}^n)^{-1} y_{n(0)}.$$

  - Calculation of $\Delta\hat{u}_{n,j}$ by solving the linear equation

$$\widehat{P}_n \Delta\hat{u}_{n,j} = -D_{\mathrm{L}}^n \left[(D_{\mathrm{R}}^n \hat{u}_{n,j} - y_{n(0)}) - \frac{h_n}{l_n^1} \left(F(D_{\mathrm{R}}^n \hat{u}_{n,j}, t_n, p_0) - y_{n(0)}^{(1)} \right) \right].$$

  - Calculation of $\hat{u}_{n,j+1} = \hat{u}_{n,j} + \Delta\hat{u}_{n,j}$.

– Calculation of e_n by

$$e_n = D_{\mathrm{R}}^n \hat{u}_{n,j}.$$

Again, the number of performed simplified Newton iterations N_{pN}, the scaling matrices D_{L}^n, D_{L}^n, the scaled approximation to the Jacobian $\widehat{P}_n$, the integration time t_n at integration step n, the step size h_n and the coefficient l_n^1 are considered to be constants. Identifier: "cvO".

- Correction of the predictor array $z_{n(0)} \to z_n$ by

$$z_n = z_{n(0)} + e_n l_n,$$

where the row vector l_n is considered to be constant. Identifier: "caO".

- Rescaling of the Nordiesk array z_n to adjust to the new step size h_{n+1} at integration step $n+1$, i.e. $z_n(h_n) \to z_n(h_{n+1})$ by

$$z_n(h_{n+1}) = z_n(h_n)\vec{\theta}_n,$$

where $\vec{\theta}_n$ denotes the constant vector

$$\vec{\theta}_n = \begin{pmatrix} 1 & \frac{h_{n+1}}{h_n} & \dots & \left(\frac{h_{n+1}}{h_n}\right)^k \end{pmatrix}^T,$$

and the current order k is considered to be a constant. Identifier: "raO".

- Change of order $k \to k+1$ by augmenting the current Nordsiek array z_n by an additional column, which is initialized by the zero vector. Identifier: "ouO".

- Change of order $k \to k-1$ by

$$z_n(k-1) = z_n(k) - z_n^{\downarrow},$$

where the Nordsieck array $z_n^{\downarrow}$ is given by

$$z_n^{\downarrow} = \begin{pmatrix} 0 & 0 & d_n^2 h_n^k \frac{y_n^{(k)}}{k!} & \dots & d_n^{k-1} h_n^k \frac{y_n^{(k)}}{k!} & 0 \end{pmatrix}$$

and the coefficients d_n^i, for $i \in \{2, \dots, k-1\}$, are considered constant.
After that, the last column of the Nordiesk array $z_n(k-1)$ is deleted. Identifier: "odO".

- Calculation of the desired approximate solution $y_{t^{\text{end}}}$ to $y(t^{\text{end}})$ at t^{end} by

$$y_{t^{\text{end}}} = \sum_{i=0}^{k} \left(\frac{y_{t^{\text{end}}} - t_n}{h_n} \right)^i \left[h_n^i y_n^{(i)} / i! \right],$$

  where the current order k is considered to be a constant. Identifier: "caO".

Now, if the calculation of *sensitivities* is required, the implemented integration method stores the concrete sequence of elementary integration operations, which are performed for the calculation of an approximated solution $y_{t^{\text{end}}}$, so that the AD techniques of Chapter 4 can be applied.
We assume that a specific operation sequence always starts with the elementary integration operation "iaO" and stops with the operation "caO".
First, we present the application of the *forward* mode of AD to the sequence of elementary integration operations in Section 5.2.1 and hereafter we present the application of the *reverse* mode of AD in Section 5.2.1.

5.2.1. Forward mode of sensitivity generation

In view of the principles of IND and for a given sequence of elementary integration operations, we can understand $y_{t^{\text{end}}}$ to be a function of the initial condition vector $y_{\text{I}} \in \mathbb{R}^m$ and of the parameter vector $p_0 \in \mathbb{R}^p$, which can be evaluated utilizing elemental functions as in Chapter 4. In accordance to the *forward* mode of AD as presented in Section 4.1, we are interested to calculate the Taylor coefficients of $y_{t^{\text{end}}}(t)$, namely $[y_{t^{\text{end}}}(t)]_i$, $i \in \{0, 1, \ldots, d\}$ up to the desired order d of sensitivities for given input polynomials

$$y_{\text{I}}(t) = y_{\text{I}} + y_{I,1} t + \cdots + y_{I,d} t^d$$

and

$$p_0(t) = p_0 + p_{0,1} t + \cdots + p_{0,d} t^d.$$

Here, the meaning of t is given as in Chapter 4 and does not correspond to the integration time as in (5.1) and (5.2).
With this conceptional preparations, we can apply the recurrences given in Table 4.1 and Table 4.2 to the elemental evaluation scheme, which is embedded in the stored integration operation sequence.
Obviously, instead of applying the recurrences to the elemental functions within the operation sequence, we can calculate to each elementary integration operation a corresponding recurrence formula and in turn can apply this one to the elementary integration operation itself.

For example the recurrence for the calculation of the Taylor polynomial $z_{n(0)}(t)$ of the predictor array $z_{n(0)}$ (Identifier: "paO") is given by

$$\left[z_{n(0)}(t)\right]_{k'} = \left[z_{n-1}(t)\right]_{k'} A[k],$$

where $k' \in \{1, \ldots, d\}$ corresponds to the k'-th order Taylor coefficient and k corresponds to the order of the BDF method.
In summary, we get the following algorithm for the calculation of the desired Taylor coefficients of $y_{t^{\text{end}}}(t)$.

Algorithm 12.

Data: $F : \mathbb{R}^m \times \mathbb{R} \times \mathbb{R}^p \to \mathbb{R}^m$ *which is d-times continuous differentiable on* $\mathbb{R}^m \times [t^{\text{init}}, t^{\text{end}}] \times \mathbb{R}^p$*, list of elementary integration operations, Taylor input polynomial* $y_{\text{I}}(t)$ *of order d, Taylor input polynomial* $p_0(t)$ *of order d.*

Step 0. Set $N = 2$*. Calculate the Taylor coefficients of* $z_0(t)$ *up to order d, according to the recurrence corresponding to the elementary operation "iaO", i.e. the initialization of the Nordsieck array* z_0*.*
Step 1. If the $N+1$*-th entry of the list of elementary operations is "caO", i.e. the final calculation of the desired approximate solution* $y_{t^{\text{end}}}$*, then calculate the Taylor coefficients of* $y_{t^{\text{end}}}(t)$ *up to order d, according to the corresponding recurrence of the elementary operation "caO" and stop.*
Else set $N = N + 1$ *and goto Step 2.*
Step 2. Calculate the intermediate Taylor coefficients up to order d, given by the recurrence corresponding to the elementary integration operation of the N*-th entry of the list of elementary operations. Goto Step 1.*

In [1], Albersmeyer state: "*The first oder iterative forward IND scheme is equal to the "staggered corrector method" that was proposed later by Feehery et al. [44], provided that the staggered corrector method uses the same number of Newton iterations for the solution of the corrector equation in the variational DAE*[2] *as used for the nominal solution. Otherwise the IND principle would be violated.*"
Because the only conceptional difference between the integration scheme of Albersmeyer and the one, which is used here, lies in the usage of a Nordsiek array based polynomial interpolation scheme instead of a modified divided differences polynomial interpolation

[2] Differential algebraic equation, see e.g. [9, 42, 59, 78]

scheme, here, his statement is true, as well, if the simplified Newton method is tackled by AD directly, i.e. the iterative mode, presented in Section 4.3.1, is used.
Albersmeyer also describes a procedure to simultaneously calculate the nominal solution and the *sensitivities* in *forward* mode, such that the step length of the discretization scheme of the BDF method is controlled by the nominal trajectory and by the *sensitivities*, as well.
Here, it should only be mentioned that this possibility is also incorporated into the developed BDF method for first order *sensitivities*. For more details we refer to [1].

5.2.2. Reverse mode of sensitivity generation

With the same preliminary thoughts as in Section 5.2.1, it is straight forward to apply the reverse accumlation rules as given in Section 4.2 to the stored sequence of elementary integration operations.
The approximate solution vector $y_{t^{\text{end}}}$ is in general no scalar. Therefore, to apply the accumulation rules we first introduce the reverse seed vector $\omega \in \mathbb{R}^m$ with

$$\tilde{y}_{t^{\text{end}}} = \omega^1 y^1_{t^{\text{end}}} + \cdots + \omega^m y^m_{t^{\text{end}}},$$

equivalently to (4.6). Formally, one can apply now the reverse accumlation rules to the evaluation scheme of $\tilde{y}_{t^{\text{end}}}$, regarding the principles of IND.
In the same manner as in Section 5.2.1, it is possible to derive accumulation rules for each elementary integration operation. Again, as example we give the accumulation rule for the elementary integration operation with identifier "paO", which is given by

$$[\bar{z}_{n-1}(t)]_{k'} += \left[\bar{z}_{n(0)}(t)\right]_{k'} A[k]^T,$$

where $k' \in \{1, \ldots, d\}$ corresponds to the k'-th order Taylor coefficient, k corresponds to the order of the BDF method and the accumulation has to be understood component wise.
In summary, we get the following algorithm for the calculation of the desired reverse Taylor coefficients of $\bar{y}_I(t)$ and $\bar{p}_0(t)$.

Algorithm 13.

Data: Reverse seed vector $\omega \in \mathbb{R}^m$, $F : \mathbb{R}^m \times \mathbb{R} \times \mathbb{R}^p \to \mathbb{R}^m$ which is $(d+1)$-times continuous differentiable on $\mathbb{R}^m \times [t^{\text{init}}, t^{\text{end}}] \times \mathbb{R}^p$, list of elementary integration operations, Taylor input polynomial $y_{\text{I}}(t)$ of order d, Taylor input polynomial $p_0(t)$ of order d, corresponding Taylor polynomial $y_{t^{\text{end}}}(t)$ of order d (e.g. calculated with Algorithm 12). (Under assumption that the underlying intermediate Taylor polynomials within the

elementary integration operations are on hand.)

Step 0. Set N to the number of entries in the list of elementary integration operations. Set all reverse intermediate Taylor polynomials to zero. Set

$$\left[\bar{y}^i_{t^{\mathrm{end}}}(t)\right]_0 = \omega^i, \quad \left[\bar{y}^i_{t^{\mathrm{end}}}(t)\right]_j = 0, \quad i \in \{1,\dots,m\}, j \in \{1,\dots,d\}.$$

Step 1. If $N > 0$, apply the reverse accumulation step associated to the N-th entry of the list of elementary operations up to order d of Taylor coefficients and set $N = N - 1$. Else, stop with final reverse Taylor coefficients of $\bar{y}_I(t)$ and $\bar{p}_0(t)$.

It should be noted, that if Algorithm 13 is applied with the principal, that the elementary operation "cvO" (the calculation of the corrector vector e_n by solving a nonlinear equation) is handled as an elementary AD operation itself, d is restricted to be $d \leq 1$ as in Section 4.3.2. Therefore, in this mode it is only possible to calculate sensitivities up to order 2, which is sufficient for the purposes in this thesis.

5.2.3. Calculation of sensitivities with respect to the end time t^{end}

Here, we present the calculation of (higher) derivatives of $y(t^{\mathrm{end}})$ with respect to the end time t^{end}.

- First, we present a reformulation of IVP in (5.1) and (5.2) with the goal that the end time t^{end} enters the reformulated IVP as parameter. This approach is also used in `MUSCOD` [74].

- Second, we give a direct way to calculate (higher or mixed) derivatives of $y(t^{\mathrm{end}})$ with respect to the end time t^{end}.

Obviously, IVP in (5.1) and (5.2) fulfills following integral equation,

$$y(t^{\mathrm{end}}) = y(t^{\mathrm{init}}) + \int_{t^{\mathrm{init}}}^{t^{\mathrm{end}}} F(y(t), t, p_0)\mathrm{d}t. \tag{5.83}$$

By substitution of the time variable t with

$$\tau = \frac{t - t^{\mathrm{init}}}{t^{\mathrm{end}} - t^{\mathrm{init}}},$$

(5.83) is equivalent to

$$\bar{y}(1) = \bar{y}(0) + \int_0^1 F(\bar{y}(\tau), \bar{\tau}(\tau), p_0)(t^{\text{end}} - t^{\text{init}})\mathrm{d}\tau,$$

where

$$\bar{\tau}(\tau) = \tau(t^{\text{end}} - t^{\text{init}}) + t^{\text{init}}$$

and

$$\bar{y}(\tau) = y(\bar{\tau}(\tau)).$$

Thus, instead of solving IVP in (5.1) and (5.2), it is equivalent to solve IVP

$$\frac{\mathrm{d}\bar{y}(\tau)}{\mathrm{d}\tau} = F(\bar{y}(\tau), \bar{\tau}(\tau), p_0)(t^{\text{end}} - t^{\text{init}}), \quad \tau \in [0,1] \tag{5.84}$$

with initial condition

$$\bar{y}(0) = y_{\text{I}}. \tag{5.85}$$

The advantage of the reformulated IVP in (5.84) and (5.85) over IVP in (5.1) and (5.2) is, that now t^{end} enters the right hand side in (5.84) as parameter and therefore the calculation of (higher) derivatives of $y(t^{\text{end}})$ with respect to t^{end} can be treated with the methods presented in Section 5.2.1 and Section 5.2.2 using IVP in (5.84) and (5.85).

For the direct way, we restrict ourself to derivatives at a max of second order, since no higher derivatives are needed for the purposes of this thesis.
Obviously, the first derivative of $y(t^{\text{end}})$ with respect to t^{end} is given by the right hand side $F(y(t^{\text{end}}), t^{\text{end}}, p_0)$ at time point t^{end}, directly.
Therefore, the second derivative of $y(t^{\text{end}})$ with respect to t^{end} is obviously given by

$$\begin{aligned}\frac{\mathrm{d}^2 y}{\mathrm{d}t^2}(t^{\text{end}}) &= \frac{\mathrm{d}F}{\mathrm{d}t}(y(t^{\text{end}}), t^{\text{end}}, p_0) \\ &= \frac{\partial F}{\partial t}(y(t^{\text{end}}), t^{\text{end}}, p_0) + F_y(y(t^{\text{end}}), t^{\text{end}}, p_0)F(y(t^{\text{end}}), t^{\text{end}}, p_0),\end{aligned}$$

which can be easily calculated by the AD techniques presented in Chapter 4.
Mixed derivatives of second order, namely

$$\frac{\mathrm{d}^2 y}{\mathrm{d}y^i \mathrm{d}t}(t^{\text{end}}) = F_y(y(t^{\text{end}}), t^{\text{end}}, p_0)s_{y^i}(t^{\text{end}}), \quad i \in \{1, \ldots, m\}$$

and

$$\frac{\mathrm{d}^2 y}{\mathrm{d}p^j \mathrm{d}t}(t^{\text{end}}) = F_y(y(t^{\text{end}}), t^{\text{end}}, p_0)s_{p^j}(t^{\text{end}}) + \frac{\partial F}{\partial p^j}(y(t^{\text{end}}), t^{\text{end}}, p_0), \quad j \in \{1, \ldots, p\},$$

are directly given by the right hand side of the corresponding sensitivity differential equation in (5.81) and (5.82), respectively.

CHAPTER 6

Nonlinear Programming

This chapter treats the solution of following nonlinear programming (NLP) problem given by

$$\min_{x \in \mathbb{R}^n} f^0(x) \tag{6.1}$$

subject to

$$g_L \leq g(x) \leq g_U \tag{6.2}$$

and

$$x_L \leq x \leq x_U, \tag{6.3}$$

where $f^0 : \mathbb{R}^n \to \mathbb{R}$ and $\tilde{g} : \mathbb{R}^n \to \mathbb{R}^r$ with $r < n$ are sufficiently smooth functions, $x_L, x_U \in \mathbb{R}^n$ and $g_L, g_U \in \mathbb{R}^r$ with $g_L \leq g_U$ component wise. (The notation in this chapter is not related to previous sections!)

In this thesis such problems are solved with the sophisticated NLP package `IPOPT` [124, 127].
`IPOPT` is a so called "Primal-Dual Interior Point filter line search algorithm for large-scale nonlinear programming" [127], whose theoretical background is sketched in the next section.

6.1. Interior Point method

Using *slack variables* [83] problem (6.1) s.t. (6.2) and (6.3) can be transformed to a NLP problem of following form

$$\min_{x' \in \mathbb{R}^{n'}} f'^0(x') \tag{6.4}$$

subject to

$$g'(x') = 0 \tag{6.5}$$

and

$$x'^i \geq 0, \quad i \in \mathcal{I}, \tag{6.6}$$

where $f'^0 : \mathbb{R}^{n'} \to \mathbb{R}$ and $g' : \mathbb{R}^{n'} \to \mathbb{R}^{r'}$ with $r' < n'$ are sufficiently smooth functions and $\mathcal{I}$ denotes the set of indices for which the components x'^i of x' are bounded below. In the following, we refer to NLPs of that form.
Interior Point (IP) methods are also called *barrier methods*, since a sequence of related barrier sub problems, here given by

$$\min_{x' \in \mathbb{R}^{n'}} f'^0_\mu(x') := f'^0(x') - \mu \sum_{i \in \mathcal{I}} \ln(x'^i) \tag{6.7}$$

subject to

$$g'(x') = 0, \tag{6.8}$$

with barrier parameter $\mu \to 0$ and $\mu > 0$ are solved.
As Wächter states in [124], *barrier methods* base on earlier work by Fiacco and McCormick [45].
The barrier term $b(\mu) := \mu \sum_{i \in \mathcal{I}} \ln(x'^i)$ forces each solution $\hat{x}'_\mu$ of sub problem (6.7) s.t. (6.8) (for corresponding barrier parameter μ) to lie in the strictly feasible region, i.e. $\hat{x}'^i_\mu > 0$ for all $i \in \mathcal{I}$ and thus the name IP method.
For $\mu \to 0$ the contribution of $b(\mu)$ tends to zero, i.e. $b(\mu) \to 0$. Therefore, under appropriate assumptions, $\lim_{\mu \to 0} \hat{x}'_\mu$ converges to a point $\hat{x}'$, at which the fist order optimality conditions (Corollary 2) in respect to the original problem (problem (6.4) s.t. (6.5) and (6.6)) are fulfilled. For a very good review article on IP methods we refer to [47].
From Corollary 2 (with $\mathbf{p} = \{1\}$ and $\mathbf{q} = \emptyset$), it follows that the so called KKT conditions for subproblem (6.7) s.t. (6.8) are given by

$$\begin{aligned} \nabla f'^0_\mu(x') + \zeta^T g'_{x'}(x') &= 0, \\ g'(x') &= 0, \end{aligned} \tag{6.9}$$

where $\zeta \in \mathbb{R}^{r'}$ are the so called Lagrangian multipliers associated to the equality constraints in (6.8). It should be noted that (6.9) is solely a nonlinear equation, which can be tackled by Newton's method.
However, as Wächter states in [124]: "*... the system (6.9) is not defined at a solution $\hat{x}'$ of NLP (6.4) s.t. (6.5) and (6.6) with an active bound $\hat{x}'^i = 0$, and the radius of convergence of Newton's method applied to (6.9) converges to zero as $\mu \to 0$ [123].*"
To remedy this obstacle, instead of solving (6.9) (via Newton's method), which leads to so called *primal-methods*, (6.9) is modified by introducing an auxiliary vector $z \in \mathbb{R}^{n'}$, defined by

$$z^i := \frac{\mu}{x'^i}, \quad \text{for all} \quad i \in \mathcal{I}, \tag{6.10}$$

and $z^i := 0$, if $i \notin \mathcal{I}$.
Here, the KKT conditions (6.9) transform to

$$\begin{aligned} \nabla f'^0(x') + \zeta^T g'_{x'}(x') - z &= 0 \\ g'(x') &= 0 \\ x'^i z'^i - \mu &= 0, \quad \text{for all} \quad i \in \mathcal{I}. \end{aligned} \tag{6.11}$$

Equations (6.11) are called the *primal-dual equations* and hence a method, relying on the solution of (6.11) (via Newton's method), is called a *primal-dual method*.
One further has to demand

$$x'^i > 0, \quad z'^i > 0, \quad \text{for all} \quad i \in \mathcal{I}, \tag{6.12}$$

such that a solution of (6.11) corresponds to a critical point of problem (6.7) s.t. (6.8).
Available IP algorithms, which rely on the solution of the *primal-dual equations*, are e.g. `LOQO` [122], `KNITRO` [31] and `IPOPT`. Here, we use `IPOPT`, since `IPOPT` is an open source package (and therefore free of charge).
For the algorithm implemented in `IPOPT` global convergence is ensured by a filter approach first introduce by Fletcher and Leyffer in [46]. A detailed description of the *primal-dual* IP algorithm `IPOPT` can be found in [124, 127]. The global convergence behavior is investigated in [126, 124] and the local one in [125, 124]. `IPOPT` uses an external linear solver. One can choose between several linear solvers including `MA27`, `MA57` [64], `MUMPS` [3, 4] or `PARDISO` [97, 98, 99, 96].
Following functions have to be supplied to `IPOPT`:

- the cost function $f^0 : \mathbb{R}^n \to \mathbb{R}$,
- the constraints $g : \mathbb{R}^n \to \mathbb{R}^r$,

- the gradient of the cost function, i.e. $\nabla f^0(x)$,
- the gradient of the constraints, i.e. $\nabla g(x)$,
- and the hessian of the Lagrangian function $f^0(x)x + \lambda^T g(x)$ given by

$$\sigma_f \nabla^2 f(x) + \sum_{i=1}^{r} \lambda_i \nabla^2 g_i(x), \tag{6.13}$$

  where σ_f is an additional factor, which is introduced by `IPOPT`.

`IPOPT` can be interfaced via the programming language `C++`.

CHAPTER 7

Numerical Calculation of Robust Optimal Experimental Designs

In this chapter we treat the numerical calculation of the optimal design for model discrimination worked out in Chapter 2. According to (2.3) and (2.10) we first state the optimization problem of interest,

$$\max_{\xi\in\Xi}\min_{\theta_1\in\Theta_1}\mathcal{I}(2:1,\mathcal{O}_1;\xi,\theta_1)$$

subject to

$$\begin{gathered} y_{\mathrm{I,min}} \leq y_{\mathrm{I}} \leq y_{\mathrm{I,max}}, \\ 0 \leq c_i \leq c_{i,\max}, \quad i \in \{1,...,n-1\}, \\ 0 \leq t^1 \leq t^2 \leq ... \leq t^n, \\ t^n = T^{\mathrm{end}}, \end{gathered}$$

where $\mathcal{I}(2:1,\mathcal{O}_1;\xi,\theta_1)$ is given by

$$\mathcal{I}(2:1,\mathcal{O}_1;\xi,\theta_1) = \frac{1}{2}\sum_{i=1}^{n}\mathcal{H}(t^i)\widetilde{\mathcal{H}}(c_i)\cdot\overline{\mathcal{I}}_i(\xi,\theta_1), \tag{7.1}$$

with

$$\overline{\mathcal{I}}_i(\xi,\theta_1) := \sum_{k=1}^{m}\left(\frac{\left(v_2^k(y_{2,i},t^i,\theta_2)\right)^2+\left(y_{2,i}^k-y_{1,i}^k\right)^2}{\left(v_1^k(y_{1,i},t^i,\theta_1)\right)^2}-2\log\left(\frac{v_2^k(y_{2,i},t^i,\theta_2)}{v_1^k(y_{1,i},t^i,\theta_1)}\right)\right)-m,$$

for $i \in \{1, ..., n\}$ and $\xi := (y_{\mathrm{I}}, t, c)$ as in Chapter 2. This is obviously equivalent to

$$\max_{\xi' \in \Xi'} \min_{\theta_1 \in \Theta_1} \mathcal{I}(2:1, \mathcal{O}_1; \xi', \theta_1) \tag{7.2}$$

subject to

$$\begin{aligned} y_{\mathrm{I,min}} \leq y_{\mathrm{I}} \leq & y_{\mathrm{I,max}}, \\ 0 \leq c_i \leq c_{i,\max}, \quad & i \in \{1, ..., n-1\}, \\ 0 \leq \Delta t^i \leq t^i_{\max}, \quad & i \in \{1, ..., n\}, \\ \sum_{i=1}^{n} \Delta t^i = & T^{\mathrm{end}}, \end{aligned} \tag{7.3}$$

with $\xi' := (y_{\mathrm{I}}, \Delta t, c)$ and where $\mathcal{I}(2:1, \mathcal{O}_1; \xi', \theta_1)$ is given by

$$\mathcal{I}(2:1, \mathcal{O}_1; \xi', \theta_1) := \frac{1}{2} \sum_{i=1}^{n} \mathcal{H}(\Delta t^i) \widetilde{\mathcal{H}}(c_i) \cdot \overline{\mathcal{I}}_i(\xi, \theta_1), \tag{7.4}$$

with

$$t^i = \sum_{j=1}^{i} \Delta t^j \quad \text{for} \quad i \in \{1, ..., n\}, \tag{7.5}$$

and $t^i_{\max}$ for $i \in \{1, ..., n\}$ are "carefully" chosen (e.g. $t^i_{\max} \geq \Delta T$ for $i \in \{1, ..., n\}$). In the following for simplicity we substitute $\xi' \to \xi$ and $\Xi' \to \Xi$.

The objective function $\mathcal{I}(2:1, \mathcal{O}_1; \xi, \theta_1)$ depending on the Heaviside-functions $\mathcal{H}(\Delta t^i)$, $i \in \{1, ..., n\}$, for a given ΔT (see Chapter 2) and $\widetilde{\mathcal{H}}(c_i)$, $i \in \{1, ..., n-1\}$, is discontinuous with respect to Δt^i, $i \in \{1, ..., n\}$ and c_i, $i \in \{1, ..., n-1\}$, respectively. Therefore it is not possible to apply the *Outer Approximations* scheme of Chapter 3 to (7.2) subject to (7.3), directly. However, instead of solving problem (7.2) subject to (7.3), we solve a related problem with smoothed objective function $\widetilde{\mathcal{I}}_{(\alpha,C)}(2:1, \mathcal{O}_1; \xi, \theta_1)$, with respect to Δt^i, $i \in \{1, ..., n\}$ and c_i, $i \in \{1, ..., n-1\}$ and depending on smoothing parameters $\alpha > 0$, $C > 0$.

In Section 7.1 the theoretical aspects of this smoothing approach are discussed. In Section 7.2 we present the application of the *Outer Approximations* scheme of Chapter 3 to a smoothed approximation $\mathbf{P}_{(\alpha,C)}$ of problem (7.2) subject to (7.3) as well as the numerical implementation of it. The resulting subproblem $\mathbf{P}_{\Omega_N}$ of Algorithm 3 is treated in Section 7.3. Finally, in Section 7.4 we discuss a homotopy approach to numerically stabilize the *Outer Approximations* scheme.

7.1. Smoothing of the objective function $\mathcal{I}(2:1,\mathcal{O}_1;\xi,\theta_1)$

In the following, for reasons of notational simplicity we treat c_i, $i = \{1,...,n-1\}$ as scalar values.
The first step to deal with optimization problem (7.2) subject to (7.3) is to replace the discontinuous Heaviside-functions $\mathcal{H}(\Delta t^i)$, $i \in \{1,...,n\}$, and $\widetilde{\mathcal{H}}(c_i)$, $i \in \{1,...,n-1\}$ in (7.2) by approximating functions

$$\mathcal{H}'(\alpha;\Delta t^i) : \mathbb{R}_{>0} \times \mathbb{R} \to \mathbb{R}, \quad i \in \{1,...,n\},$$

and respectively

$$\widetilde{\mathcal{H}}'(C;c_i) : \mathbb{R}_{>0} \times \mathbb{R} \to \mathbb{R}, \quad i \in \{1,...,n-1\},$$

which are twice continuously differentiable with respect to Δt^i, $i \in \{1,...,n\}$ and c_i, $i \in \{1,...,n-1\}$.
In contrast to $\mathcal{H}(\Delta t^i)$, the continuous approximations $\mathcal{H}'(\alpha;\Delta t^i)$ depend on the smoothing parameter $\alpha > 0$ for all $i \in \{1,...,n\}$. We also assume that $\mathcal{H}'(\alpha;\Delta t^i)$ are continuous with respect to α for all $i \in \{1,...,n\}$. Analogously, $\widetilde{\mathcal{H}}'(C;c_i)$ depend on the smoothing parameter $C > 0$ for $i \in \{1,...,n-1\}$. Again, we assume that $\widetilde{\mathcal{H}}'(C;c_i)$ are continuous with respect to C for all $i \in \{1,...,n-1\}$. Further, $\mathcal{H}'(\alpha;\Delta t^i)$ have to fulfill the following condition (as in [132]):

$$\mathcal{H}'(\alpha;\Delta t_i) := \begin{cases} 0 & \text{for} \quad \Delta t^i \leq \Delta T - \alpha, \\ 0 \leq \mathcal{H}'(\alpha;\Delta t_i) \leq 1 & \text{for} \quad \Delta T - \alpha \leq \Delta t^i \leq \Delta T + \alpha, \\ 1 & \text{for} \quad \Delta t^i \geq \Delta T + \alpha, \end{cases} \tag{7.6}$$

for $i \in \{1,...,n\}$. A graphical scheme of the continuous approximations $\mathcal{H}'(\alpha;\Delta t^i)$, $i \in \{1,...,n\}$ is shown in Figure 7.1.
Moreover, for $\widetilde{\mathcal{H}}'(C;c_i)$ we require that:

$$\widetilde{\mathcal{H}}'(C;c_i) := \begin{cases} 1 & \text{for} \quad c_i \leq (C - \alpha_C) \\ 0 \leq \widetilde{\mathcal{H}}'(c_i) \leq 1 & \text{for} \quad (C - \alpha_C) \leq c_i \leq (C + \alpha_C) \\ 0 & \text{for} \quad c_i \geq (C + \alpha_C) \end{cases}, \tag{7.7}$$

for $i \in \{1,...,n-1\}$ and $\alpha_C := \rho C$ with constant $0 < \rho < 1$. A graphical scheme of the continuous approximations $\widetilde{\mathcal{H}}'(C;c_i)$, $i \in \{1,...,n-1\}$ is shown in Figure 7.2.

Remark. *Condition (7.7) assures that* $\widetilde{\mathcal{H}}'(C;0) = 1$ *for all* $i \in \{1,...,n-1\}$.

Under the assumption that $\overline{\mathcal{I}}_i(\xi,\theta_1)$ is continuous on $\Xi \times \Theta_1$ for all $i \in \{1,...,n\}$, the

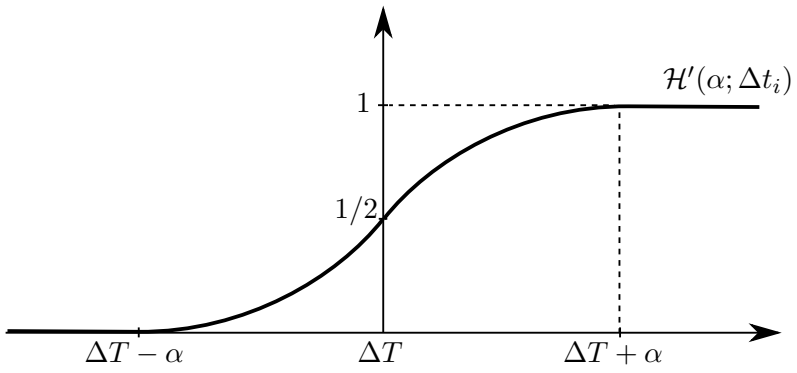


Figure 7.1.: A graphical scheme of the continuous approximations $\mathcal{H}'(\alpha;\Delta t^i)$ for $i \in \{1,...,n\}$.

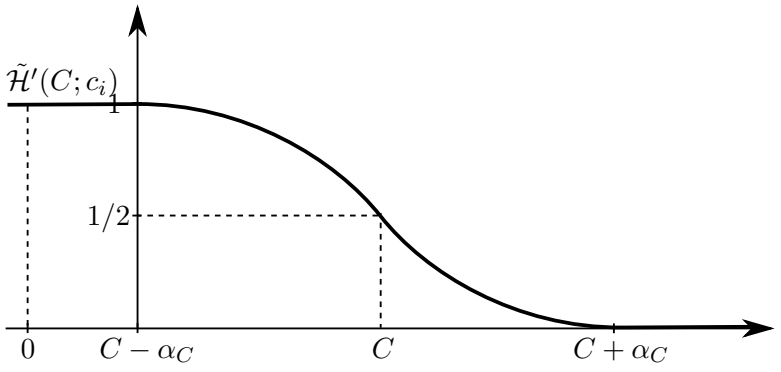


Figure 7.2.: A graphical scheme of the continuous approximations $\widetilde{\Theta}(C;c_i)$ for $i \in \{1,...,n-1\}$.

resulting approximation $\widetilde{\mathcal{I}}_{\alpha,C}(2:1,\mathcal{O}_1;\xi,\theta_1)$ of $\mathcal{I}(2:1,\mathcal{O}_1;\xi,\theta_1)$, given by

$$\widetilde{\mathcal{I}}_{(\alpha,C)}(2:1,\mathcal{O}_1;\xi,\theta_1) := \frac{1}{2}\sum_{i=1}^{n} \mathcal{H}'(\Delta t_i)\widetilde{\mathcal{H}}'(c_i)\cdot \overline{\mathcal{I}}_i(\xi,\theta_1),$$

is continuous on $\Xi \times \Theta_1$ and with respect to $\alpha > 0$ and $C > 0$.
For the sake of notational simplicity we define

$$\mathcal{I}(\xi) := \min_{\theta_1\in\Theta_1} \mathcal{I}(2:1,\mathcal{O}_1;\xi,\theta_1), \tag{7.8}$$

and

$$\widetilde{\mathcal{I}}_{(\alpha,C)}(\xi) := \min_{\theta_1\in\Theta_1} \widetilde{\mathcal{I}}_{(\alpha,C)}(2:1,\mathcal{O}_1;\xi,\theta_1). \tag{7.9}$$

Remark. *Under appropriate assumptions (in view of Proposition 3), for any $\alpha > 0$ and $C > 0$, $\widetilde{\mathcal{I}}_{(\alpha,C)}(\xi)$ is continuous on Ξ.*

We will draw the conclusion that this smoothing approach is a valid one by the fact that under some mild assumptions for a sequence of smoothing parameters $\{(\alpha_k, C_k)\}_{k=1}^{\infty}$ with $(\alpha_k, C_k) \to (0,0)$, as $k \to \infty$, a convergent subsequence of the sequence $\{\xi_k\}$, given by $\xi_k := \arg\max_{\xi\in\Xi} \widetilde{\mathcal{I}}_{(\alpha_k,C_k)}(\xi)$ subject to (7.3) (assuming existence), converges to the maximum value of $\mathcal{I}(\xi)$ subject to (7.3), i.e. for any $K \subset \mathbb{N}$ with $\xi_k \to^K \bar{\xi}$ and $\widetilde{\mathcal{I}}_{(\alpha_k,C_k)}(\xi_k) \to^K \overline{\mathcal{I}}$ it follows that $\overline{\mathcal{I}} = \max_{\xi\in\Xi} \mathcal{I}(\xi)$ subject to (7.3).
This statement will be proven in the remainder of this section. The proof is guided by the results in [132], where a similar (but finite and unconstrained) setting is investigated.

We first introduce some definitions and assumptions.

Definition 17 (Set of *strictly Δt-feasible* indices)**.** *For every $\xi \in \Xi$ we denote by*

$$I^{\Delta t}(\xi) := \{i | i \in \{1,...,n\},\ \Delta t^i > \Delta T\}$$

the set of strictly Δt-feasible indices.

Definition 18 (Set of *strictly c-feasible* indices)**.** *For every $\xi \in \Xi$ we denote by*

$$I^{c}(\xi) := \{i | i \in \{1,...,n-1\},\ c_i < 0\} \cup \{n\}$$

the set of strictly c-feasible indices.

Further, we say that the *Δt-discontinuity* is *active* at $\xi \in \Xi$, if $\Delta t_i = \Delta T$ for at least one $i \in \{1,...,n\}$.

Definition 19 (Set of *Δt-discontinuity* indices)**.** *For every $\xi \in \Xi$ we denote by*

$$E^{\Delta t}(\xi) := \{i | i \in \{1,...,n\},\ \Delta t^i = \Delta T\}$$

the set of Δt-discontinuity indices at $\xi \in \Xi$.

Equivalently, we say that the *c-discontinuity* is *active* at $\xi \in \Xi$, if $c_i = 0$ for at least one $i \in \{1,...,n-1\}$.

Definition 20 (Set of *c-discontinuity* indices)**.** *For every $\xi \in \Xi$ we denote by*

$$E^{c}(\xi) := \{i | i \in \{1,...,n-1\},\ c_i = 0\} \cup \{n\}$$

the set of c-discontinuity indices at $\xi \in \Xi$.

Definition 21 (Set of contributing terms of the sum)**.** *For every $\xi \in \Xi$ we denote by*

$$\Upsilon(\xi) := \left(I^{\Delta t}(\xi) \cup E^{\Delta t}(\xi)\right) \cap \left(I^{c}(\xi) \cup E^{c}(\xi)\right)$$

the set of indices contributing to the terms of the sum in (7.4).

Obviously, $\mathcal{I}(2:1,\mathcal{O}_1;\xi,\theta_1)$ is given by

$$\mathcal{I}(2:1,\mathcal{O}_1;\xi,\theta_1) = \frac{1}{2}\sum_{i\in\Upsilon(\xi)} \overline{\mathcal{I}}_i(\xi,\theta_1), \tag{7.10}$$

for all $\xi \in \Xi$ and $\theta_1 \in \Theta_1$.

Assumption 5 (Strictly-feasibly reachable assumption)**.** *For every $\xi' = (y_1', \Delta t', c') \in \Xi$ satisfying (7.3) there exists a sequence $\{\xi_k\}_{k=1}^{\infty} \subset \Xi$ such that $\xi_k \to \xi'$, as $k \to \infty$ with $\xi_k := (y_1', \Delta t_k, c')$, $E^{\Delta t}(\xi_k) = \emptyset$, such that*

$$I^{\Delta t}(\xi_k) = I^{\Delta t}(\xi') \cup E^{\Delta t}(\xi'),$$

and ξ_k satisfies (7.3).

Remark. *Assumption 5 is satisfied if $T^{\text{end}} > 0$, $n > 1$,*

$$T^{\text{end}} \neq k\Delta T \quad \text{for all} \quad k \in \mathbb{N},$$

and $t^i_{\max}$ for $i \in \{1,...,n\}$ in (7.3) are "carefully" chosen.

Assumption 6 (Nonnegativity assumption)**.** *Since the KL-Distance is always nonnegative we further assume that*

$$\overline{\mathcal{I}}_i(\xi,\theta_1) \geq 0,$$

for all $i \in \{1,...,n\}$, $\xi \in \Xi$ and $\theta_1 \in \Theta_1$.

Assumption 7 (Continuity assumption)**.** *We assume that $\overline{\mathcal{I}}_i(\xi,\theta_1)$ is continuous on $\Xi \times \Theta_1$ for all $i \in \{1,...,n\}$ and Θ_1 is compact .*

Definition 22. *Let Σ be a set. A collection of subsets*

$$\Lambda_1, \Lambda_2, ..., \Lambda_r \subset \Sigma$$

satisfying

$$\bigcup_{i=1}^{r} \Lambda_i = \Sigma,$$

and

$$\Lambda_i \cap \Lambda_j = \emptyset \quad \forall i \neq j,$$

is called a partition of Σ.

Now consider a *partition* of Ξ where Λ_i, $i \in \{1,...,r_1\}$ are defined by

$$\Lambda_i := \{\xi \in \Xi |\ t^l < \Delta T\ \forall l \in L_i^{\Delta t};\ t^l \geq \Delta T\ \forall l \in U_i^{\Delta t}\},$$

so that for each $i \in \{1,...,r_1\}$, $L_i^{\Delta t}$ and $U_i^{\Delta t}$ form a different *partition* of the set $\{1,...,n\}$. Since all different *partitions* of the set $\{1,...,n\}$ shall be covered, r_1 equals the cardinality of the set of all subsets of $\{1,...,n\}$.
By (7.10), Assumption 7 and Proposition 3, $\mathcal{I}(\xi)$ as given in (7.8) is continuous on Λ_i with respect to Δt for all $i \in \{1,...,r_1\}$.
Assumption 5 assures that for $\xi \in \Lambda_i$ with $i \in \{1,...,r_1\}$ and ξ is satisfying (7.3) there exists a sequence $\{\xi_k\}_{k=1}^{\infty} \subset \mathring{\Lambda}_i$, such that for all $k \in \mathbb{N}_{>0}$ ξ_k satisfies (7.3) and $\xi_k \to \xi$, as $k \to \infty$. ($\mathring{\Lambda}_i$ denotes the interior of Λ_i for $i \in \{1,...,r_1\}$.)
Since $\mathcal{H}'(\alpha;\Delta t^i) = \frac{1}{2}$ for all $\alpha > 0$ and $i \in E^{\Delta t}(\xi)$ (by construction of $\mathcal{H}'(\alpha;\Delta t^i)$), both the existence of an "interior sequence" and the continuity of $\mathcal{I}(\xi)$ in Λ_i for all $i = \{1,...,n\}$ are important requirements for the above mentioned statement to hold.

In the following, we first give an illustrative counter example for the case that such an "interior sequence" does not exists. Hereafter, we proceed with the proof of the above mentioned statement on the validity of the smoothing approach.

7.1.1. A counter example to the strictly-feasibly reachable assumption

Consider the optimization problem

$$\min_{(x,y)\in\mathbb{R}^2} F(x,y),$$

subject to

$$x + y = 0,$$

where $F(x,y) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is given by

$$F(x,y) := \begin{cases} 100(x^2+y^2), & \text{if} \quad x \geq 0 \quad \text{and} \quad y \geq 0, \\ 100((x-\frac{1}{2})^2+(y+\frac{1}{2})^2)+10, & \text{if} \quad x \geq 0 \quad \text{and} \quad y < 0, \\ 100((x+\frac{1}{2})^2+(y-\frac{1}{2})^2)+10, & \text{if} \quad x < 0 \quad \text{and} \quad y \geq 0, \\ 60, & \text{if} \quad x < 0 \quad \text{and} \quad y < 0. \end{cases}$$

Clearly, the function $F(x,y)$ subject to $x + y = 0$ has a global minimum at $(\hat{x},\hat{y}) := (0,0)$. With this example we treat an analogy of Δt-*discontinuities*, where an "interior sequence" as discussed above for the partition $x, y \geq 0$ does not exists. A plot of the function value of $F(\cdot,\cdot)$ is shown in Figure 7.3.

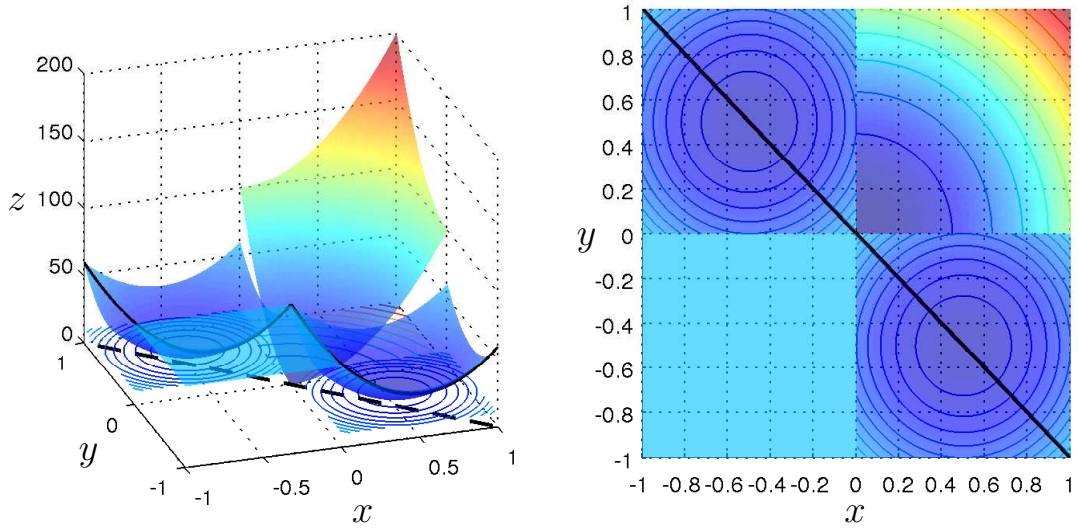


Figure 7.3.: In the left figure a surface plot of the function $F(\cdot,\cdot)$ as in Section 7.1.1 is shown. The black line and also the black dashed line indicate the constraint $x + y = 0$. On the right one a contourplot of the same function is shown. Again the black line indicates the constraint $x + y = 0$.

A smoothed approximation $\widetilde{F}_\alpha(x,y)$ of $F(x,y)$ (according to the smoothing approach presented above) is given by

$$\begin{aligned}\widetilde{F}_\alpha(x,y) :=& \mathcal{H}'(\alpha;x)\mathcal{H}'(\alpha;y)\left(100\left(x^2+y^2\right)\right) + \\ & \mathcal{H}'(\alpha;x)\left(1-\mathcal{H}'(\alpha;y)\right)\left(100\left(\left(x-\frac{1}{2}\right)^2+\left(y+\frac{1}{2}\right)^2\right)+10\right) + \\ & \left(1-\mathcal{H}'(\alpha;x)\right)\mathcal{H}'(\alpha;y)\left(100\left(\left(x+\frac{1}{2}\right)^2+\left(y-\frac{1}{2}\right)^2\right)+10\right) + \\ & \left(1-\mathcal{H}'(\alpha;x)\right)\left(1-\mathcal{H}'(\alpha;y)\right)60.\end{aligned}$$

Here, we define $\mathcal{H}'(\alpha;t)$ to be

$$\mathcal{H}'(\alpha;t) := \begin{cases} 0, & \text{for} \quad t \le -\alpha, \\ -\frac{1}{4}\cdot\left(\frac{t}{\alpha}\right)^3 + \frac{3}{4}\cdot\frac{t}{\alpha} + \frac{1}{2}, & \text{for} \quad -\alpha \le t \le \alpha, \\ 1, & \text{for} \quad \alpha \le t, \end{cases} \tag{7.11}$$

which satisfies the required condition (7.6). The approximation (7.11) of the Heaviside-function is proposed in [118] as stated in [132].
A plot of the function $\widetilde{F}_\alpha(x,y)$ subject to $x + y = 0$ is shown in Figure 7.4.

For each $\alpha > 0$ the value at $(\hat{x},\hat{y})$ is $\widetilde{F}_\alpha(\hat{x},\hat{y}) = 45$. A sequence of local minima generated by a local search method for decreasing smoothing parameter α might converge to the local minimum at $(\hat{x},\hat{y})$ (in some cases), i.e. $(x_\alpha, y_\alpha) \to (\hat{x},\hat{y})$ as $\alpha \to 0$ but with $\lim_{\alpha\to 0} \widetilde{F}_\alpha(x_\alpha, y_\alpha) = 45$ instead of $F(\hat{x},\hat{y}) = 0$. The approximation scheme fails to converge in the above mentioned sense. Therefore, the example confirms the necessity of an "interior sequence" as discussed above.

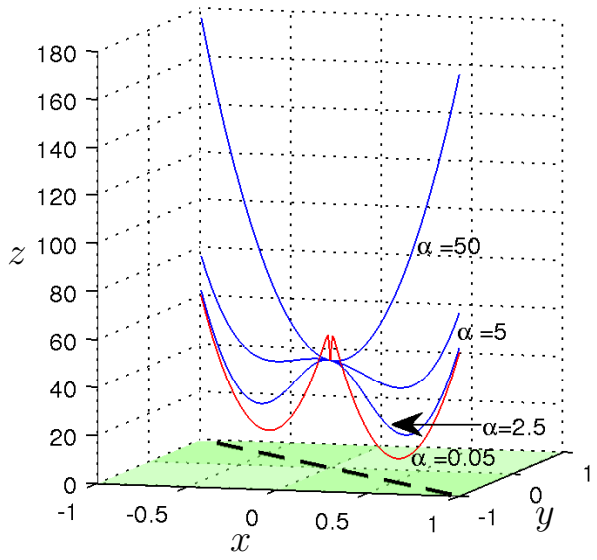


Figure 7.4.: A plot of the smoothed function $\widetilde{F}_\alpha(\cdot,\cdot)$ as in Section 7.1.1 for several values of α and $x + y = 0$ is shown. The black dashed line indicates the constraint $x + y = 0$.

7.1.2. Theoretical validation of the smoothing approach

Lemma 5. *Let $\bar{\xi} \in \Xi$. Then, there exist positive numbers δ, $\bar{\alpha}$ and $\overline{C}$ such that*

$$\begin{aligned}\widetilde{\mathcal{I}}_{(\alpha,C)}(2:1,\mathcal{O}_1;\xi,\theta_1) &= \frac{1}{2}\sum_{i=1}^{n} \mathcal{H}'(\alpha;\Delta t^i)\widetilde{\mathcal{H}}'(C;c_i)\cdot\overline{\mathcal{I}}_i(\xi,\theta_1)\\ &= \frac{1}{2}\sum_{i\in\Upsilon(\bar{\xi})} \mathcal{H}'(\alpha;\Delta t^i)\widetilde{\mathcal{H}}'(C;c_i)\cdot\overline{\mathcal{I}}_i(\xi,\theta_1),\end{aligned}$$

for all $\xi \in B(\bar{\xi},\delta)$, $0 < \alpha < \bar{\alpha}$ and $0 < C < \overline{C}$.

Proof. Since for $\Upsilon(\bar{\xi}) = \{1,...,n\}$ the statement is obvious, we only consider $\Upsilon(\bar{\xi}) \neq$

$\{1, ..., n\}$. Let $i \notin \Upsilon(\bar{\xi})$. Then, either

$$\Delta\bar{t}^i < \Delta T \quad \Leftrightarrow \quad 0 < \Delta T - \Delta\bar{t}^i \tag{7.12}$$

or

$$\bar{c}_i > 0. \tag{7.13}$$

Let $\gamma_\xi(i)$ be defined for the case that $\xi \in \Xi$ with $\Upsilon(\xi) \neq \{1, ..., n\}$ and $i \notin \Upsilon(\xi)$ by

$$\gamma_\xi(i) := \min\left(\{\max\{0, \Delta T - \Delta t^i\}, \max\{0, c_i\}\} \setminus \{0\}\right).$$

Obviously, $\gamma_\xi(i)$ is well defined and $\gamma_\xi(i) > 0$. Further, for that case let $\gamma(\xi)$ be defined by

$$\gamma(\xi) := \min_{i \notin \Upsilon(\xi)} \{\gamma_\xi(i)\},$$

and again $\gamma(\xi) > 0$. Be $\gamma := \frac{\gamma(\bar{\xi})}{2}$. By continuity of $\Delta T - \Delta t$ and c there exists a $\delta_1 > 0$ such that for every $\xi \in B(\bar{\xi}, \delta_1)$ and $i \notin \Upsilon(\bar{\xi})$ either

$$\Delta T - \Delta t^i > \gamma \quad \Leftrightarrow \quad \Delta t^i < \Delta T - \gamma,$$

or

$$c_i > \gamma.$$

Therefore, by construction of $\mathcal{H}'(\alpha; \Delta t^i)$ and $\widetilde{\mathcal{H}}'(C; c_i)$, for $0 < \alpha < \gamma$, $0 < C < \frac{\gamma}{2}$ and $\xi \in B(\bar{\xi}, \delta_1)$ we have that either

$$\mathcal{H}'(\alpha; \Delta t^i) = 0,$$

or

$$\widetilde{\mathcal{H}}'(C; c_i) = 0,$$

for all $i \notin \Upsilon(\bar{\xi})$. Thus, for $\delta = \delta_1$, $\bar{\alpha} = \gamma$ and $\overline{C} = \frac{\gamma}{2}$ it follows that

$$\widetilde{\mathcal{I}}_{(\alpha,C)}(2:1, \mathcal{O}_1; \xi, \theta_1) = \frac{1}{2} \sum_{i \in \Upsilon(\bar{\xi})} \mathcal{H}'(\alpha; \Delta t^i) \widetilde{\mathcal{H}}'(C; c_i) \cdot \overline{\mathcal{I}}_i(\xi, \theta_1),$$

for all $\xi \in B(\bar{\xi}, \delta)$, $0 < \alpha < \bar{\alpha}$ and $0 < C < \overline{C}$. □

Lemma 6. *Let $\bar{\xi} \in \Xi$ with $E^{\Delta t}(\bar{\xi}) = \emptyset$. Then, there exist positive numbers $\bar{\alpha}$ and $\overline{C}$ such that*

$$\widetilde{\mathcal{I}}_{(\alpha,C)}(\bar{\xi}) = \mathcal{I}(\bar{\xi}),$$

for all $0 < \alpha < \bar{\alpha}$ *and* $0 < C < \overline{C}$.

Proof. Since $E^{\Delta t}(\bar{\xi}) = \emptyset$, for all $i \in \{1, ..., n\}$ either

$$\Delta\bar{t}^i > \Delta T,$$

or

$$\Delta\bar{t}^i < \Delta T.$$

Therefore, for $0 < \alpha < \alpha_1$ with

$$\alpha_1 := \min_{i\in\{1,...,n\}} \{|\Delta T - \Delta\bar{t}^i|\},$$

it follows by construction of $\mathcal{H}'(\alpha; \Delta\bar{t}^i)$ that either

$$\mathcal{H}(\Delta\bar{t}^i) = \mathcal{H}'(\alpha; \Delta\bar{t}^i) = 1,$$

or

$$\mathcal{H}(\Delta\bar{t}^i) = \mathcal{H}'(\alpha; \Delta\bar{t}^i) = 0.$$

For $i \in E^c(\bar{\xi})$, it holds that

$$\widetilde{\mathcal{H}}(\bar{c}_i) = \widetilde{\mathcal{H}}'(C; \bar{c}_i) = 1,$$

for all $C > 0$. For $i \notin E^c(\bar{\xi})$, one sees with similar thoughts as above that there exists a $C_1 > 0$ such that either

$$\widetilde{\mathcal{H}}(\bar{c}_i) = \widetilde{\mathcal{H}}'(C; \bar{c}_i) = 0$$

or

$$\widetilde{\mathcal{H}}(\Delta\bar{c}_i) = \widetilde{\mathcal{H}}'(C; \bar{c}_i) = 1,$$

for all $C < C_1$. It follows with $\bar{\alpha} = \alpha_1$ and $\overline{C} = C_1$ that

$$\widetilde{\mathcal{I}}_{(\alpha,C)}(2:1,\mathcal{O}_1;\bar{\xi},\theta_1) = \frac{1}{2}\sum_{i\in\Upsilon(\bar{\xi})} \overline{\mathcal{I}}_i(\bar{\xi},\theta_1) = \mathcal{I}(2:1,\mathcal{O}_1;\bar{\xi},\theta_1),$$

for all $0 < \alpha < \bar{\alpha}$, $0 < C < \overline{C}$ and $\theta_1 \in \Theta_1$. Therefore

$$\widetilde{\mathcal{I}}_{(\alpha,C)}(\bar{\xi}) = \mathcal{I}(\bar{\xi}),$$

for all $0 < \alpha < \bar{\alpha}$ and $0 < C < \overline{C}$. □

Lemma 7. *Let* $\{(\alpha_k, C_k)\}_{k=1}^{\infty} \subset \mathbb{R}_{>0} \times \mathbb{R}_{>0}$ *and* $\{\xi_k\}_{k=1}^{\infty} \subset \Xi$ *with* $(\alpha_k, C_k) \to (0,0)$ *and*

$\xi_k \to \bar{\xi}$ *as* $k \to \infty$. *Then, we have that*

$$\overline{\lim}\, \widetilde{\mathcal{I}}_{(\alpha_k, C_k)}(\xi_k) \leq \mathcal{I}(\bar{\xi}). \tag{7.14}$$

Proof. By Lemma 5 for sufficient large k, it holds that

$$\begin{aligned}
\widetilde{\mathcal{I}}_{(\alpha,C)}(2:1, \mathcal{O}_1; \xi_k, \theta_1) =& \frac{1}{2} \sum_{i=1}^{n} \mathcal{H}'(\alpha_k; \Delta t^i) \widetilde{\mathcal{H}}'(C_k; c_i) \cdot \overline{\mathcal{I}}_i(\xi_k, \theta_1) \\
=& \frac{1}{2} \sum_{i \in \Upsilon(\bar{\xi})} \mathcal{H}'(\alpha_k; \Delta t^i) \widetilde{\mathcal{H}}'(C_k; c_i) \cdot \overline{\mathcal{I}}_i(\xi_k, \theta_1) \\
\leq& \frac{1}{2} \sum_{i \in \Upsilon(\bar{\xi})} \overline{\mathcal{I}}_i(\xi_k, \theta_1),
\end{aligned} \tag{7.15}$$

for all $\theta_1 \in \Theta_1$. The last inequality in (7.15) follows from Assumption 6. Therefore, for sufficiently large k

$$\begin{aligned}
\widetilde{\mathcal{I}}_{(\alpha_k, C_k)}(\xi_k) &= \min_{\theta_1 \in \Theta_1} \widetilde{\mathcal{I}}_{(\alpha,C)}(2:1, \mathcal{O}_1; \xi_k, \theta_1) \\
&\leq \min_{\theta_1 \in \Theta_1} \left(\frac{1}{2} \sum_{i \in \Upsilon(\bar{\xi})} \overline{\mathcal{I}}_i(\xi_k, \theta_1) \right) =: \widetilde{\mathcal{I}}_{\bar{\xi}}(\xi_k).
\end{aligned}$$

Because of the continuity of $\widetilde{\mathcal{I}}_{\bar{\xi}}(\cdot)$ (Proposition 3 and Assumption 7) and the fact that $\widetilde{\mathcal{I}}_{\bar{\xi}}(\bar{\xi}) = \mathcal{I}(\bar{\xi})$ it follows that (7.14) holds.

□

Theorem 17. *Let* $\{(\alpha_k, C_k)\}_{k=1}^{\infty}$ *with* $(\alpha_k, C_k) \to (0, 0)$ *as* $k \to \infty$. *Suppose that for every* k *there exists a point* $\xi_k \in \Xi$ *such that*

$$\widetilde{\mathcal{I}}_{(\alpha_k, C_k)}(\xi_k) = \max_{\xi \in \Xi} \widetilde{\mathcal{I}}_{(\alpha_k, C_k)}(\xi) \tag{7.16}$$

subject to (7.3) holds and that there exists a convergent subsequence, i.e. there is a subset $K \subset \mathbb{N}$ *such that*

$$\xi_k \to^K \bar{\xi} \quad \textit{and} \quad \widetilde{\mathcal{I}}_{(\alpha_k, C_k)}(\xi_k) \to^K \overline{\mathcal{I}}. \tag{7.17}$$

Further, assume that

$$\hat{\xi} = \arg\max_{\xi \in \Xi} \mathcal{I}(\xi), \tag{7.18}$$

subject to (7.3) exists. Then,

$$\overline{\mathcal{I}} = \mathcal{I}(\hat{\xi}), \tag{7.19}$$

Proof. By Lemma 7

$$\bar{\mathcal{I}} \leq \mathcal{I}(\bar{\xi}) \leq \widehat{\mathcal{I}},$$

where

$$\widehat{\mathcal{I}} := \mathcal{I}(\hat{\xi}).$$

We proof by contradiction. Therefore assume that

$$\widehat{\mathcal{I}} - \bar{\mathcal{I}} = \mu > 0 \tag{7.20}$$

holds.

In case $E^{\Delta t}(\hat{\xi}) = \emptyset$, it follows by Lemma 6 that

$$\widetilde{\mathcal{I}}_{(\alpha_k,C_k)}(\hat{\xi}) = \mathcal{I}(\hat{\xi}) = \widehat{\mathcal{I}},$$

for sufficient large k.
Otherwise, by assumption (7.17) for sufficient large k it holds that

$$\widetilde{\mathcal{I}}_{(\alpha_k,C_k)}(\xi_k) < \overline{\mathcal{I}} + (\mu/2).$$

Since by (7.20) obviously $\overline{\mathcal{I}} + (\mu/2) < \widehat{\mathcal{I}}$, it follows for sufficient large k that

$$\widetilde{\mathcal{I}}_{(\alpha_k,C_k)}(\xi_k) < \widetilde{\mathcal{I}}_{(\alpha_k,C_k)}(\hat{\xi}).$$

This is a contradiction to (7.16). Therefore, it holds that $\mu = 0$.

In case $E^{\Delta t}(\hat{\xi}) \neq \emptyset$, it follows from Assumption 5 that there is a sequence $\{\xi_s\}_{s=1}^{\infty}$ satisfying (7.3) with $y_{\mathrm{I},s} = \hat{y}_{\mathrm{I}}$ and $c_s = \hat{c}$, which converges to $\hat{\xi}$ such that $I^{\Delta t}(\xi_s) = I^{\Delta t}(\hat{\xi}) \cup E^{\Delta t}(\hat{\xi})$ and $E^{\Delta t}(\xi_s) = \emptyset$, for all $s \in \mathbb{N}_{>0}$.
It follows that

$$\mathcal{I}(\xi_s) = \min_{\theta_1 \in \Theta_1} \frac{1}{2} \sum_{i \in \Upsilon(\xi_s)} \overline{\mathcal{I}}_i(\xi_s, \theta_1) = \min_{\theta_1 \in \Theta_1} \frac{1}{2} \sum_{i \in \Upsilon(\hat{\xi})} \overline{\mathcal{I}}_i(\xi_s, \theta_1) = \widetilde{\mathcal{I}}_{\hat{\xi}}(\xi_s),$$

where $\widetilde{\mathcal{I}}_{\hat{\xi}}(\cdot)$ is defined as in Lemma 7. By continuity of $\widetilde{\mathcal{I}}_{\hat{\xi}}(\cdot)$ and the fact that $\widetilde{\mathcal{I}}_{\hat{\xi}}(\hat{\xi}) = \mathcal{I}(\hat{\xi})$, it follows that

$$\lim_{s \to \infty} \widetilde{\mathcal{I}}_{\hat{\xi}}(\xi_s) = \mathcal{I}(\hat{\xi}).$$

Consequently, for sufficient large s

$$\widehat{\mathcal{I}} - \mu/4 < \mathcal{I}(\xi_s). \tag{7.21}$$

Let $\bar{s}$ be the value for which (7.21) holds. It holds that $E^{\Delta t}(\xi_{\bar{s}}) = \emptyset$ and therefore by Lemma 6 one has that

$$\widehat{\mathcal{I}} - \mu/4 < \mathcal{I}(\xi_{\bar{s}}) = \widetilde{\mathcal{I}}_{(\alpha_k, C_k)}(\xi_{\bar{s}}),$$

for sufficient large k. Otherwise, by assumption (7.17) it follows for sufficient large k that

$$\widetilde{\mathcal{I}}_{(\alpha_k, C_k)}(\xi_k) < \overline{\mathcal{I}} + (\mu/4).$$

Since by (7.20), it obviously holds that

$$\overline{\mathcal{I}} + (\mu/2) < \widehat{\mathcal{I}}.$$

It follows for sufficient large k that

$$\widetilde{\mathcal{I}}_{(\alpha_k, C_k)}(\xi_k) < \widetilde{\mathcal{I}}_{(\alpha_k, C_k)}(\xi_{\bar{s}}).$$

This is a contradiction to (7.16). Therefore, it holds that $\mu = 0$.

Overall, assumption (7.20) results in a contradiction and therefore (7.19) subject to (7.3) has to hold. □

Remark. *The Heaviside-functions $\mathcal{H}(\Delta t^i)$ in (7.4) can also be approximated by smoothing functions $\widetilde{\mathcal{H}}'(C; \Delta t^i)$ satisfying condition (7.7) instead of (7.6) for all $i \in \{1, ..., n\}$. In this case Assumption 5 is not necessary and Theorem 17 still is valid. Otherwise, if $\widetilde{\mathcal{H}}(c_i)$ in (7.4) is approximated by smoothing functions satisfying condition (7.6) for $i \in \{1, ..., n-1\}$, Theorem 17 does not hold anymore. The fact of the matter is that Assumption 5 due to the constraints in (7.3) can not be fulfilled anymore.*

In the remainder of this chapter, we substitute

$$\widetilde{\mathcal{I}}_{(\alpha, C)}(2:1, \mathcal{O}_1; \xi, \theta_1) \to \widetilde{\mathcal{I}}(2:1, \mathcal{O}_1; \xi, \theta_1),$$

for reasons of notational simplicity.

The smoothed optimization problem

$$\max_{\xi \in \Xi} \min_{\theta_1 \in \Theta_1} \widetilde{\mathcal{I}}(2:1, \mathcal{O}_1; \xi, \theta_1) \tag{7.22}$$

subject to

$$\begin{aligned} y_{\mathrm{I,min}} \leq y_{\mathrm{I}} \leq & y_{\mathrm{I,max}}, \\ 0 \leq c_i \leq c_{i,\max}, \quad & i \in \{1, ..., n-1\}, \\ 0 \leq \Delta t^i \leq t^i_{\max}, \quad & i \in \{1, ..., n\}, \\ \sum_{i=1}^{n} \Delta t^i = & T^{\mathrm{end}}, \end{aligned} \tag{7.23}$$

with $\xi := (y_{\mathrm{I}}, \Delta t, c)$ and where $\widetilde{\mathcal{I}}(2:1, \mathcal{O}_1; \xi, \theta_1)$ is given by

$$\widetilde{\mathcal{I}}(2:1, \mathcal{O}_1; \xi, \theta_1) = \frac{1}{2} \sum_{i=1}^{n} \mathcal{H}'(\alpha; \Delta t_i) \widetilde{\mathcal{H}}'(C; c_i) \cdot \overline{\mathcal{I}}_i(\xi, \theta_1),$$

with smoothing parameters $\alpha > 0$ and $C > 0$, is equivalent to optimization problem:

$$\max_{(\tau, \xi) \in \mathbb{R} \times \Xi} \tau$$

subject to

$$\begin{aligned} y_{\mathrm{I,min}} \leq y_{\mathrm{I}} \leq & y_{\mathrm{I,max}}, \\ 0 \leq c_i \leq c_{i,\max}, \quad & i \in \{1, ..., n-1\}, \\ 0 \leq \Delta t^i \leq t^i_{\max}, \quad & i \in \{1, ..., n\}, \\ \sum_{i=1}^{n} \Delta t^i = & T^{\mathrm{end}}, \\ \min_{\theta_1 \in \Theta_1} \widetilde{\mathcal{I}}(2:1, \mathcal{O}_1; \xi, \theta_1) - \tau \geq & 0, \end{aligned}$$

which we denote by $\mathbf{P}_{(\alpha,C)}$ in the following.

Remark. *As mentioned above, under appropriate assumptions $\widetilde{\mathcal{I}}_{(\alpha,C)}(\xi)$ (Definition 7.9) is continuous on Ξ. Therefore, under mild assumptions reasoning the Theorem of Weierstraß [57] problem (7.22) subject to (7.23) has a solution. Since this problem is equivalent to $\mathbf{P}_{(\alpha,C)}$, under same assumptions $\mathbf{P}_{(\alpha,C)}$ has a solution. With similar thoughts and under mild assumptions, one can see that problem (7.2) subject to (7.3) has a solution, as well.*

7.2. Applying the Outer Approximations scheme to $\mathbf{P}_{(\alpha,C)}$

In the following, we apply the *Outer Approximations* scheme of Section 3.3 to optimzation problem $\mathbf{P}_{(\alpha,C)}$. We assume that Assumption 2 and Assumption 3 with respect to

optimization problem $\mathbf{P}_{(\alpha,C)}$ are satisfied. Further, we assume that $\mathbf{P}_{(\alpha,C)}$ has a solution. In view of the notational framework introduced in Chapter 3, i.e. according to Definition 9 and with $x := (\tau, \xi) \in \mathbb{R} \times \Xi$, optimization problem $\mathbf{P}_{(\alpha,C)}$ can be formulated as[1]:

$$\psi^0(x) = \phi^0(x, y_0) = -\tau, \tag{7.24}$$

where the inequality constraints $\psi^j(x)$ for $j \in \mathbf{q} = \{1, \ldots, 2m + 2m(n-1) + 2n + 1\}$ are defined by

$$\begin{aligned} \psi^j(x) = \phi^j(x, y_j) &= y^j_{\mathrm{I,min}} - y^j_{\mathrm{I}}, \quad j \in \{1, \ldots, m\} =: \mathbf{q}_1, \\ \psi^j(x) = \phi^j(x, y_j) &= y^{j-m}_{\mathrm{I}} - y^{j-m}_{\mathrm{I,max}}, \quad j \in \{m+1, \ldots, 2m\} =: \mathbf{q}_2, \end{aligned} \tag{7.25}$$

for the constraints on the initial species concentration;
second by

$$\psi^j(x) = \phi^j(x, y_j) = -c^{j_2}_{j_1} \tag{7.26}$$

for $j \in \{2m + 1, ..., 2m + m(n-1)\} =: \mathbf{q}_3$, where

$$2m + (j_1 - 1)m + j_2 = j$$

and

$$\psi^j(x) = \phi^j(x, y_j) = c^{j_2}_{j_1} - c^{j_2}_{j_1,\max} \tag{7.27}$$

for $j \in \{2m + m(n-1) + 1, \ldots, 2m + 2m(n-1)\} =: \mathbf{q}_4$, where

$$2m + m(n-1) + (j_1 - 1)m + j_2 = j,$$

with $j_1 \in \{1, \ldots, n-1\}$ and $j_2 \in \{1, \ldots, m\}$ for the constraints on the perturbation vectors c_i for $i \in \{1, \ldots, n-1\}$;
third by

$$\psi^j(x) = \phi^j(x, y_j) = -\Delta t^{j-2m-2m(n-1)} \tag{7.28}$$

for

$$j \in \{2m + 2m(n-1) + 1, \ldots, 2m + 2m(n-1) + n\} =: \mathbf{q}_5,$$

and

$$\psi^j(x) = \phi^j(x, y_j) = \Delta t^{j-2m-2m(n-1)-n} - t^{j-2m-2m(n-1)-n}_{\max} \tag{7.29}$$

for

$$j \in \{2m + 2m(n-1) + n + 1, \ldots, 2m + 2m(n-1) + 2n\} =: \mathbf{q}_6,$$

[1]Here, the meaning of y_j, $j \in \{0, \ldots, 2m + 2m(n-1) + 2n + 1\}$, as in Chapter 3 does interfere with the notation of Chapter 2, but from the context it is obvious what is meant.

for the constraints on Δt^i for $i = \{1, \ldots, n\}$;
and fourth by

$$\begin{aligned}\psi^j(x) &= \max_{y_j \in Y_j} \phi^j(x, y_j) \\ &= \max_{\theta_1 \in \Theta_1} \left(\tau - \widetilde{\mathcal{I}}(2:1, \mathcal{O}_1; \xi, \theta_1)\right), \quad \text{where} \quad j = 2m + 2m(n-1) + 2n + 1,\end{aligned} \tag{7.30}$$

with n denoting the number of possible measurement time points and m denotes the number of different species types as in Chapter 2. The inequality constraints in (7.24)–(7.29) do not depend on robustification vectors y_j for $j \in \{1, \ldots, 2m + 2m(n-1) + 2n\}$. Therefore, we arbitrarily set $y_j \in [0, 1]$ for $j \in \{1, \ldots, 2m + 2m(n-1) + 2n\}$ in obedience to Assumption 2. Additionally, the equality constraint with $r = 1$ is given by

$$g(x) = \sum_{i=1}^{n} \Delta t^i - T^{\text{end}}. \tag{7.31}$$

Remark. *In terms of Corollary 4, it is important that problem $\mathbf{P}_{(\alpha,C)}$ satisfies EMFCQ (Definition 11) at all points with $\psi(x) = 0$ ($\psi(x)$ is defined as in (3.34)) and $g(x) = 0$, since then EMFCQ ensures that a point $(\tau, \xi) \in \mathbb{R} \times \Xi$ with $\psi(x) \leq 0$ and $g(x) = 0$, satisfying the necessary optimality condition in Theorem 11 and Theorem 12, is indeed a critical point of problem $\mathbf{P}_{(\alpha,C)}$.*

Theorem 18. *Consider problem (7.24)–(7.31). Assume Assumption 2 and Assumption 3 are satisfied. Be $x' = (\tau', \xi') \in \mathbb{R} \times \Xi$ with $\psi(x') = 0$ and $g(x') = 0$, where $\psi(x)$ is defined as in (3.34). Assume that $c^j_{i,\max} > 0$ for $j \in \{1, \ldots, m\}$, $i \in \{1, \ldots, n-1\}$, $t^{i'}_{\max} > 0$ for $i' \in \{1, \ldots, n\}$ and $y_{\mathrm{I},\min} \neq y_{\mathrm{I},\max}$ component-wise.*
Further, consider the design vector $\Delta t'$ in $\xi' = (y'_{\mathrm{I}}, \Delta t', c')$. Assume that there is at least one $i \in \{1, \ldots, n\}$ such that $0 < \Delta t^i$ (i.e. $\exists j \in \mathbf{q}_5 | \psi^j(x) < 0$) and there is at least one $i \in \{1, \ldots, n\}$ such that $\Delta t^i < t^i_{\max}$ (i.e. $\exists j \in \mathbf{q}_6 | \psi^j(x) < 0$), then EMFCQ is satisfied at x'.

Proof. Clearly, for proving the above statement we have to find a $\tilde{h} \in \mathbb{R} \times \Xi$ such that (3.49) and (3.50) are fulfilled.
Here, we use the convention that $\tilde{h} = (\tilde{h}_\tau, \tilde{h}_{y_{\mathrm{I}}}, \tilde{h}_{\Delta t}, \tilde{h}_c)$ with $\tilde{h}_\tau \in \mathbb{R}$, $\tilde{h}_{y_{\mathrm{I}}} \in \mathbb{R}^m$, $\tilde{h}_{\Delta t} \in \mathbb{R}^n$ and $\tilde{h}_c \in \mathbb{R}^{m(n-1)}$, where the design variable c^j_i corresponds to the $((i-1)m + j)$-th entry in $\tilde{h}_c$.
Additionally, entry $\tilde{h}^j_{\Delta t}$ of $\tilde{h}_{\Delta t}$ corresponds to inequality constraints $\psi^i(x)$ with $i = (2m + 2m(n-1) + j)$ (i.e. $i \in \mathbf{q}_5$) and $\psi^{i'}(x)$ with $i' = (2m + 2m(n-1) + n + j)$ (i.e. $i' \in \mathbf{q}_6$).

For $x'' \in \mathbb{R} \times \Xi$ we define a partition (Definition 22) of $\mathbf{q}_5$ by $\mathbf{q}_5^{\mathrm{A}}(x'') \cup \mathbf{q}_5^{\mathrm{I}}(x'') = \mathbf{q}_5$ such that $\psi^j(x'') = 0$ for all $j \in \mathbf{q}_5^{\mathrm{A}}(x'')$ and $\psi^j(x'') < 0$ for all $j \in \mathbf{q}_5^{\mathrm{I}}(x'')$. Equivalently, we define a partition of $\mathbf{q}_6$ by $\mathbf{q}_6^{\mathrm{A}}(x'') \cup \mathbf{q}_6^{\mathrm{I}}(x'') = \mathbf{q}_6$.
Now, for x' it holds that either $\mathbf{q}_5^{\mathrm{A}}(x') = \emptyset$ or $\mathbf{q}_5^{\mathrm{A}}(x') \neq \emptyset$.

- If it holds that $\mathbf{q}_5^{\mathrm{A}} \neq \emptyset$, set $\tilde{h}^i_{\Delta t} = 1$ with $i = j - 2m - 2m(n-1)$ for all $j \in \mathbf{q}_5^{\mathrm{A}}$. Clearly, for $j \in \mathbf{q}_5^{\mathrm{A}}$, $\mathrm{d}\psi^j(x; \tilde{h}) < 0$. By assumption, it holds that $\mathbf{q}_5^{\mathrm{I}} \neq \emptyset$. Now, for $j \in \mathbf{q}_5^{\mathrm{I}}$ set $\tilde{h}^i_{\Delta t} = -\nu$ with $i = j - 2m - 2m(n-1)$ and $\nu > 0$ such that $\nabla g(x)^T \tilde{h} = 0$ ($g(x)$ only depends on Δt). By assumption $t_{\max} > 0$, it follows that $\mathrm{d}\psi^j(x; \tilde{h}) < 0$ for $j \in \mathbf{q}_6$ with $\psi^j(x) = 0$.
- If it holds that $\mathbf{q}_5^{\mathrm{A}} = \emptyset$ and $\mathbf{q}_6^{\mathrm{A}} = \emptyset$, set $\tilde{h}_{\Delta t}$ such that $g(x) = 0$.
- If it holds that $\mathbf{q}_5^{\mathrm{A}} = \emptyset$ and $\mathbf{q}_6^{\mathrm{A}} \neq \emptyset$, set $\tilde{h}_{\Delta t}$ such that $\mathrm{d}\psi^j(x; \tilde{h}) < 0$ for all $j \in \mathbf{q}_6^{\mathrm{A}}$ and $g(x) = 0$.

Set $\tilde{h}_{y_{\mathrm{I}}}$ and $\tilde{h}_c$ such that $\mathrm{d}\psi^j(x; \tilde{h}) < 0$ for all $j \in \bigcup_{k=1}^{4} \mathbf{q}_k$. By Assumption 2,

$$\omega = \max_{\theta_1 \in \Theta_1} \left(-\nabla_\xi \widetilde{\mathcal{I}}(2:1, \mathcal{O}_1; \xi, \theta_1)^T \tilde{h}_\xi \right)$$

exists.
Set $\tilde{h}_\tau = -2|\omega| - 1$. By Theorem 25, it follows for $j = 2m + 2m(n-1) + 2n + 1$ that

$$\mathrm{d}\psi^j(x; \tilde{h}) = -2|\omega| - 1 + \max_{\theta_1 \in \Theta_1} \left(-\nabla_\xi \widetilde{\mathcal{I}}(2:1, \mathcal{O}_1; \xi, \theta_1)^T \tilde{h}_\xi \right) < -|\omega| - 1 < 0.$$

Overall, it follows that there exists a $\tilde{h} \in \mathbb{R} \times \Xi$ such that $\nabla g(x')^T \tilde{h} = 0$ and $\mathrm{d}\psi^j(x'; \tilde{h}) < 0$ for all $j \in \mathbf{q}_{\mathbf{A}}(x')$. Therefore, $EMFCQ$ is satisfied. □

Remark. *For problem (7.24)–(7.31) at all $x \in \mathbb{R} \times \Xi$ with $g(x) = 0$ and*

$$0 < T^{\mathrm{end}} < \sum_{i=1}^{n} t^i_{\max},$$

it never happens that $\psi^j(x) = 0$ for all $j \in \mathbf{q}_5$ or for all $j \in \mathbf{q}_6$, respectively.

At each iteration N of the *Outer Approximations* scheme (Algorithm 3) two inner steps are performed. First, the calculation of new worst case "robustification vectors" $\hat{y}_{j,N} \in \hat{Y}_{j,N}(x_N)$ for $j \in \bar{\mathbf{q}}$, which depend on the current approximate solution x_N at iteration N. The set $\widehat{Y}_{j,N}(x_N)$ is defined in (3.64). These "robustification vectors" are used to augment the finite set of robustification vectors

$$\Omega_{j,N} = \Omega_{j,N-1} \cup \{\hat{y}_{j,N}\}, \quad j \in \bar{\mathbf{q}},$$

for which an **IECP** approximation to the original **SIECP** problem is solved in the second step to generate a new approximate solution x_{N+1}. These two inner steps are repeated until the finite set is extended such that a sufficient approximation of the entire robustification space $Y = Y_0 \times Y_1 ... \times Y_q$ is achieved.
In our implementation we redefine the restrictions $\hat{y}_{j,N} \in \widehat{Y}_{j,N}(x_N)$, $j \in \bar{\mathbf{q}}$ for the calculation of augmenting "robustification vectors" at iteration N in Algorithm 3 (Step 1.) by demanding

$$\hat{y}_{j,N} \in \widehat{Y}_j(x_N) := \{y_{j,N} | y_{j,N} = \arg\max_{y_j \in Y_j} \phi^j(x_N, y_j)\}, \quad j \in \bar{\mathbf{q}}. \tag{7.32}$$

If one can find a global solution of $\max_{y_j \in Y_j} \phi^j(x_N, y_j)$, $j \in \bar{\mathbf{q}}$, this of course does not influence the convergence properties of the *Outer Approximations* scheme (compare Theorem 14). The reason, that the compact sets Y_j are approximated by subsets of finite cardinality, namely by $Y_{j,N}$ for $j \in \bar{\mathbf{q}}$ at each iteration N of the *Outer Approximations* scheme, is that the calculation of elements of the sets $\widehat{Y}_{j,N}(x_N)$, $j \in \bar{\mathbf{q}}$, is a finite operation and therefore theoretically realizable. But this procedure can lead to high computational costs, specially if Y_j, $j \in \bar{\mathbf{q}}$ are high dimensional. Since in general it seems impossible to find a global solution of (7.32), our approach has to be rated as heuristic.
For problem (7.24)-(7.31) only inequality constraint (7.30) comprises a non trivial robustification space Y_j with $j = 2m + 2m(n-1) + 2n + 1$, the parameter space Θ_1 of the composite alternative hypothesis.
For the calculation of (7.32) in Step 1. of Algorithm 3, at iterate N and in view of problem (7.24)-(7.31) it follows that

$$\hat{\theta}_{1,N} = \hat{y}_{j,N} \in \widehat{Y}_j(x_N) = \{\hat{\theta}_1 | \hat{\theta}_1 = \arg\min_{\theta_1 \in \Theta_1} \widetilde{\mathcal{I}}(2:1, \mathcal{O}_1; \xi_N, \theta_1)\}, \tag{7.33}$$

with $j = 2m+2m(n-1)+2n+1$. Here, we use a simple random search approach coupled to a local optimization method, i.e. we have randomly generated P different start values in Θ_1 from which we have started a local optimization method. The best value out of the P trials is chosen to augment the set $\Omega_{j,N-1}$ with $j = 2m + 2m(n-1) + 2n + 1$.
Of course there are more sophisticated approaches to approximately search for a global minimum. For a review see e.g. [8], but at this point an efficient calculation of Step 1. of Algorithm 3 is not our primary goal.
For the local parameter optimization (Step 1. in Algorithm 3) we use the same optimization method as for Step 2. in Algorithm 3.
The procedure to calculate a robust optimal design is to solve iteratively both inner steps of the *Outer Approximations* scheme until ϵ_N reaches the level of desired accuracy

ϵ. This means that

$$\epsilon_N \leq \epsilon, \quad \forall N > N',$$

for the optimality function of the current **IECP** approximation it holds that

$$\theta_{\Omega_N}(x_N) \geq -\epsilon,$$

for the constraint functions it holds that

$$\psi_{\Omega_N}(x_N) \leq \epsilon \quad \text{and} \quad \|g(x_N)\| \leq \epsilon,$$

and the iterate x_N of Algorithm 3 has "numerically" converged for $N > N'$.
In our implementation we use a fixed $\epsilon_N = \epsilon$ for a desired final accuracy ϵ at every iterate of Step 2. in Algorithm 3. In that way, Step 1. of Algorithm 3 gives a worst case estimate of the KL divergence $\min_{\theta_1 \in \Theta_1} \widetilde{\mathcal{I}}(2:1, \mathcal{O}_1; \xi_N, \theta_1)$ for the current design ξ_N at iteration N, up to the desired accuracy ϵ. This single step might already be sufficient for practical application with real experiments.
As stopping criterion we use:

Algorithmic Stop Criterion. *Stop after Step 1. of Algorithm 3, if*

$$\delta \geq \min_{\theta_1 \in \Omega_{j,N-1}} \widetilde{\mathcal{I}}(2:1, \mathcal{O}_1; \xi_N, \theta_1) - \min_{\theta_1 \in \Theta_1} \widetilde{\mathcal{I}}(2:1, \mathcal{O}_1; \xi_N, \theta_1),$$

with $j = 2m + 2m(n-1) + 2n + 1$ *and where* δ *is a small positive constant. Then consider*

$$\hat{\theta}_1 := \arg \min_{\theta_1 \in \Theta_1} \widetilde{\mathcal{I}}(2:1, \mathcal{O}_1; \xi_N, \theta_1)$$

and ξ_N *as (approximate) solutions of the Maxmin problem at iteration* N *of the Outer Approximations scheme,*
else goto Step 2. and calculate a new design x_{N+1}*.*

This stop criterion is also used in [103, 92]. We call the distance Δ_{RG} given by,

$$\Delta_{\mathrm{RG}} := \min_{\theta_1 \in \Omega_{j,N-1}} \widetilde{\mathcal{I}}(2:1, \mathcal{O}_1; \xi_N, \theta_1) - \min_{\theta_1 \in \Theta_1} \widetilde{\mathcal{I}}(2:1, \mathcal{O}_1; \xi_N, \theta_1), \tag{7.34}$$

robustification gap, with $j = 2m + 2m(n-1) + 2n + 1$ at iteration N.

7.3. Numerical solution of subproblem P_{Ω_N}

Subproblem $\mathbf{P}_{\Omega_N}$ for problem (7.24)-(7.31) differs from (7.24)-(7.31) only in the inequality constraint (7.30), namely $\psi^j(x)$ with $j = 2m + 2m(n-1) + 2n + 1$.

Instead of

$$\psi^j(x) = \max_{\theta_1 \in \Theta_1} \left(\tau - \widetilde{\mathcal{I}}(2:1, \mathcal{O}_1; \xi, \theta_1) \right) \quad \text{for} \quad j = 2m + 2m(n-1) + 2n + 1,$$

as in the **SIECP** case of problem (7.24)-(7.31), for the **IECP** approximation $\mathbf{P}_{\Omega_N}$, $\psi^j(x)$ with $j = 2m + 2m(n-1) + 2n + 1$ is replaced by

$$\psi^j(x) = \max_{\theta_1 \in \Theta_{1,N}} \left(\tau - \widetilde{\mathcal{I}}(2:1, \mathcal{O}_1; \xi, \theta_1) \right), \quad j = 2m + 2m(n-1) + 2n + 1,$$

where $\Theta_{1,N} := \Omega_{j,N} = \{\hat{\theta}_{1,1}, ..., \hat{\theta}_{1,\tilde{N}}\}$ with $j = 2m + 2m(n-1) + 2n + 1$ at iteration N with $\tilde{N} = N - N_0$ and N_0 as given in Algorithm 3.
Therefore, the **IECP** approximation $\mathbf{P}_{\Omega_N}$ for problem (7.24)-(7.31) can be equivalently formulated as

$$\max_{(\tau, \xi) \in \mathbb{R} \times \Xi} \tau$$

subject to

$$\begin{aligned}
y_{\mathrm{I,min}} \leq y_{\mathrm{I}} \leq y_{\mathrm{I,max}}, \\
0 \leq \Delta t^i \leq t^i_{\max}, \quad i \in \{1, ..., n\}, \\
0 \leq c_i \leq c_{i,\max}, \quad i \in \{1, ..., n-1\}, \\
\sum_{i=1}^{n} \Delta t^i = T^{\mathrm{end}}, \\
\widetilde{\mathcal{I}}(2:1, \mathcal{O}_1; \xi, \hat{\theta}_{1,l}) - \tau \geq 0 \quad l \in \{1, ..., \tilde{N}\},
\end{aligned}$$

which is in this form solvable by a nonlinear programming algorithm for equality and inequality constrained optimization problems.

We have implemented the resulting optimization problem in a multiple shooting setup (see for example [113, 28, 27]). The idea of the multiple shooting method is to subdivide the whole integration interval $[0, T^{\mathrm{end}}]$ into several subintervals by introducing auxiliary multiple shooting node variables $s_{j,i,l}$ for $i \in \{1, \ldots, n\},\ j \in \{1, 2\}$ and $l \in \{1, \ldots, \tilde{N}\}$, on each of which an independent initial value problem is solved. In our implementation, each end point of a subinterval corresponds to one measurement time point. Matching conditions, which enter the optimization problem as additional equality constraints, assure continuity of the state trajectory from one subinterval to the next.
To incorporate the perturbations c, matching conditions

$$s_{j,i,l} - y_j(t^{i-1}, t^i, s_{j,i-1,l}, \hat{\theta}_{j,l}) = 0, \quad i \in \{1, \ldots, n\},\ j \in \{1, 2\},\ l \in \{1, \ldots, \tilde{N}\},$$

where $s_{j,0,l} = y_{\mathrm{I}}$ and $\hat{\theta}_{2,l} = \theta_2$ for $j \in \{1,2\}$ and $l \in \{1,\dots,\tilde{N}\}$ are modified to

$$\begin{aligned} &s_{j,i,l} - y_j(t^{i-1}, t^i, s_{j,i-1,l}, \hat{\theta}_{j,l}) = c_i, \quad i \in \{1,\dots,n-1\},\ j \in \{1,2\},\ l \in \{1,\dots,\tilde{N}\}, \\ &s_{j,n,l} - y_j(t^{n-1}, t^n, s_{j,n-1,l}, \hat{\theta}_{j,l}) = 0, \quad j \in \{1,2\},\ l \in \{1,\dots,\tilde{N}\}. \end{aligned} \tag{7.35}$$

A graphical scheme of the multiple shooting setup is shown in Figure 7.5.

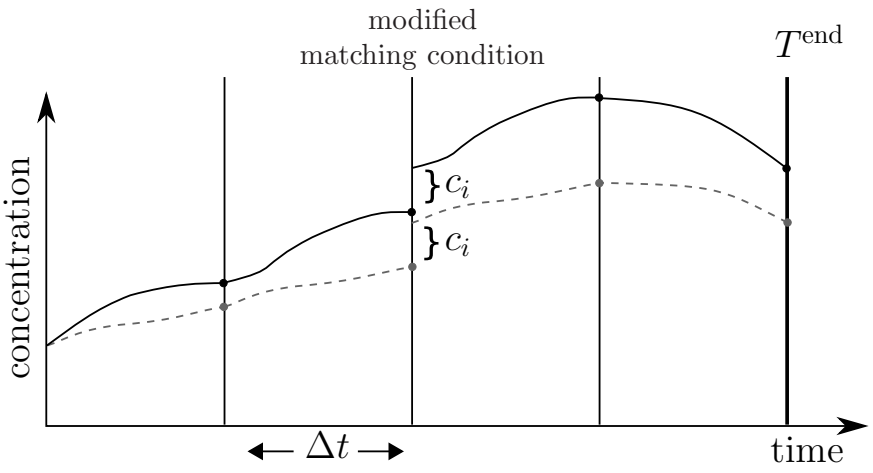


Figure 7.5.: Scheme of the multiple shooting setup for computing the experimental design. One dot denotes the concentration at one measurement time point. The black solid line denotes model 1 and the gray dashed one model 2.

Instead of evaluating $\widetilde{\mathcal{I}}(2:1,\mathcal{O}_1;\xi,\hat{\theta}_{1,l})$ by use of the values $y_j(t^{i-1},t^i,s_{j,i-1,l},\hat{\theta}_{j,l})$ with $i \in \{1,\dots,n\}$, $j \in \{1,2\}$ and $l \in \{1,\dots,\tilde{N}\}$, which are given by the solution of the initial value problem (2.4), $\widetilde{\mathcal{I}}(2:1,\mathcal{O}_1;\xi,\hat{\theta}_{1,l})$ is evaluated by use of the auxiliary multiple shooting node variables $s_{i,j,l}$ for $i \in \{1,\dots,n\}$, $j \in \{1,2\}$ and $l \in \{1,\dots,\tilde{N}\}$, by replacing the values $y_j(t^{i-1},t^i,s_{j,i-1,l},\hat{\theta}_{j,l})$ in $\widetilde{\mathcal{I}}(2:1,\mathcal{O}_1;\xi,\hat{\theta}_{1,l})$ with $s_{i,j,l}$, respectively. The dependency of $\widetilde{\mathcal{I}}(2:1,\mathcal{O}_1;\xi,\hat{\theta}_{1,l})$ on $s_{i,j,l}$ for $i \in \{1,\dots,n\}$, $j \in \{1,2\}$ and $l \in \{1,\dots,\tilde{N}\}$ is indicated by $\widetilde{\mathcal{I}}(2:1,\mathcal{O}_1;s_{\cdot,1,l},s_{\cdot,2,l},\hat{\theta}_{1,l})$ for $j \in \{1,2\}$ and $l \in \{1,\dots,\tilde{N}\}$.

The overall optimization problem can be stated as

$$\max_{(\tau,\xi)\in\mathbb{R}\times\Xi} \tau \tag{7.36}$$

subject to

$$\begin{aligned}
s_{j,i,l} - y_j(t^{i-1}, t^i, s_{j,i-1,l}, \hat{\theta}_{j,l}) &= c_i, \quad i \in \{1, ..., n-1\},\ j \in \{1,2\},\ l \in \{1, \dots, \tilde{N}\},\\
s_{j,n,l} - y_j(t^{n-1}, t^n, s_{j,n-1,l}, \hat{\theta}_{j,l}) &= 0, \quad j \in \{1,2\},\ l \in \{1, \dots, \tilde{N}\},\\
s_{j,0,l} &= y_{\mathrm{I}}, \quad j \in \{1,2\},\ l \in \{1, \dots, \tilde{N}\},\\
\frac{\mathrm{d}y_j}{\mathrm{d}t} &= f_j^{\mathrm{rhs}}(y, \hat{\theta}_{j,l}), \quad j \in \{1,2\},\ l \in \{1, \dots, \tilde{N}\},\\
y_{\mathrm{I,min}} \leq & y_{\mathrm{I}} \leq y_{\mathrm{I,max}},\\
0 \leq \Delta t^i &\leq t^i_{\max}, \quad i \in \{1, \dots, n\},\\
0 \leq c_i &\leq c_{i,\max}, \quad i \in \{1, \dots, n-1\},\\
s_{j,i,l,\min} \leq s_{j,i,l} &\leq s_{j,i,l,\max}, \quad i \in \{1, \dots, n\},\ j \in \{1,2\},\ l \in \{1, \dots, \tilde{N}\},\\
\sum_{i=1}^{n} \Delta t^i &= T^{\mathrm{end}},\\
\widetilde{\mathcal{I}}(2:1, \mathcal{O}_1; s_{\cdot,1,l}, s_{\cdot,2,l}, \hat{\theta}_{1,l}) - \tau &\geq 0 \quad l \in \{1, \dots, \tilde{N}\},
\end{aligned} \tag{7.37}$$

where $y_j(t^{i-1}, t^i, s_{j,i-1,l}, \hat{\theta}_{j,l})$ is the solution of the differential equation

$$\frac{\mathrm{d}y_j}{\mathrm{d}t} = f_j^{\mathrm{rhs}}(y_j, t, \hat{\theta}_{j,l}),$$

with initial state $s_{j,i-1,l}$ and integration interval $[t^{i-1}, t^i]$ and $\hat{\theta}_{2,l} = \theta_2$ for all $i \in \{1, \dots, n\}$, $j \in \{1,2\}$, $l \in \{1, \dots, \tilde{N}\}$.

Remark. *Under same assumptions as in Theorem 18 problem (7.36) subject to (7.37) satisfies MFCQ (Definition 8) for all $(\tau, \xi) \in \mathbb{R} \times \Xi$ with ξ satisfying (7.37), as can be seen with the same thoughts as in proof of Theorem 18.*

We have implemented this problem within the *Interior Point* optimization package `IPOPT` using the `C++` interface of `IPOPT`. A brief introduction to the theory of `IPOPT` is given in Chapter 6. All derivatives up to second order, which are used for the calculations of the Hessian (6.13), needed for a robust performance of `IPOPT` are calculated by automatic differentiation using `CppAD` [19, 18]. `CppAD` implements Automatic Differentiation by use of Taylor series propagation as presented in Chapter 4. For the solution of the ODEs as well as for the calculation of sensitivities we use the BDF integration method developed in Chapter 5, which is also implemented in `C++`.

Remark. *Since the solutions $y_j(t^{i-1}, t^i, s_{j,i-1,l}, \hat{\theta}_{j,l})$ of the ODEs $\frac{\mathrm{d}y_j}{\mathrm{d}t} = f_j^{rhs}(y_j, t, \hat{\theta}_j^l)$ for $i \in \{1, \dots, n\}$, $j \in \{1,2\}$ and $l \in \{1, \dots, \tilde{N}\}$ only enter the matching conditions (7.35) in problem (7.36) s.t. (7.37), the sensitivities of second order, which are needed*

to evaluate the Hessian in (6.13), can be efficiently computed using the reverse mode presented in Section 5.2.2, whereby the Lagrange multipliers associated to the matching conditions (7.35) are used for weighting the reverse seed vector.

7.4. Stabilizing homotopy method for subsequent $\mathbf{P}_{\Omega_{N+1}}$

Solving the subsequent optimization problems $\mathbf{P}_{\Omega_{N+1}}$ with an *Interior Point* code like `IPOPT` initialized with primal and dual variables of the previous problem or with primal variables only, one often observes that the new solution may differ significantly from the previous. This is due to the fact that the solution of the previous problem $\mathbf{P}_{\Omega_N}$ is infeasible for $\mathbf{P}_{\Omega_{N+1}}$ and thus the algorithm tries to find a feasible state before it proceeds to find a new optimum. This behavior is not desired in the context of an *Outer Approximations* algorithm, because convergence of the algorithm may be slowed down significantly. This circumstance originates from a jumping between vicinities of distinct local maxima of problem $\mathbf{P}_{(\alpha,C)}$. The discretization $\Omega_{j,N}$ for $j = (2m + 2m(n-1) + 2n + 1)$ of the robustification space Θ_1 may not be equally adequate for different local maxima. To overcome this problem we have implemented a heuristic homotopy method to gradually introduce the additional constraint

$$g_{\tilde{N}+1}(\tau, \xi, s) := \widetilde{\mathcal{I}}(2:1, \mathcal{O}_1; s_{\cdot,1,\tilde{N}+1}, s_{\cdot,2,\tilde{N}+1}, \hat{\theta}_{1,\tilde{N}+1}) - \tau \geq 0$$

of problem $\mathbf{P}_{\Omega_{N+1}}$. We replace $g_{\tilde{N}+1}(\tau, \xi, s)$ by

$$\tilde{g}_{\tilde{N}+1}(\tau, \xi, s) := \widetilde{\mathcal{I}}(2:1, \mathcal{O}_1; s_{\cdot,1,\tilde{N}+1}, s_{\cdot,2,\tilde{N}+1}, \hat{\theta}_{1,\tilde{N}+1}) - \tau + (1 - \kappa_{\mathrm{H}})\rho_{\mathrm{H}} \geq 0,$$

with homotopy parameter $\kappa_{\mathrm{H}} \in [0, 1]$ and ρ_{H} is a constant which has to be set such that $\tilde{g}_{\tilde{N}+1}(\tau, \xi, s)$ is inactive for $\kappa_{\mathrm{H}} = 0$ at the initial design ξ_N.
We choose ρ_{H} to be

$$\rho_H := K \cdot \left(\min_{\theta_1 \in \Omega_{j,N}} \widetilde{\mathcal{I}}(2:1, \mathcal{O}_1; \xi_N, \theta_1) - \min_{\theta_1 \in \Theta_1} \widetilde{\mathcal{I}}(2:1, \mathcal{O}_1; \xi_N, \theta_1) \right),$$

where $j = (2m + 2m(n-1) + 2n + 1)$. K is a save guard factor, we set empirically to $K = 1.4$, which worked well in practice for our examples. For $\kappa_{\mathrm{H}} = 0$ the augmented optimization problem should be easily solvable within a few iterations by performing a warm start from the solution of the previous problem. By increasing the homotopy parameter to $\kappa_{\mathrm{H}} = 1$, the additional constraint is gradually introduced, which leads to a sequence of easily solvable subproblems whose solutions stay in the vicinity of the solution of the previous problem $\mathbf{P}_{\Omega_N}$. A similar homotopy strategy can be found e.g. in [93] (in the context of finite optimization).

CHAPTER 8

Numerical results

We have applied the algorithm developed in Chapter 7 to two example problems for which we present results in the following sections, namely on models describing glycolytic oscillations in Section 8.1 and on models describing signal sensing in *dictyostelium discoideum* in Section 8.2. In the following we treat model 1 as null hypothesis and model 2 as alternative hypothesis.
Here, we replace the Heaviside-functions $\mathcal{H}(t^i)$ and $\widetilde{\mathcal{H}}(c_i)$ in (7.1) by parametrized hyperbolic tangent functions of the form

$$\mathcal{H}'(t^i) = \frac{\tanh(\frac{6(\Delta t_i - b_1)}{a_1}) + 1}{2} \quad \text{and} \quad \widetilde{\mathcal{H}}'(c_i) = \frac{\tanh(-\frac{6(c_i - b_2)}{a_2}) + 1}{2}.$$

The parameters a_j for $j \in \{1, 2\}$ characterize the width of the transition region between 0 and 1. The parameters b_j for $j \in \{1, 2\}$ determine the center of the transition region (see Figure 8.1). By setting the parameters in an adequate way, arbitrarily close approximations of the Heaviside-functions can be generated.
Additionally, in cooperation with Marcel Rehberg we have applied the algorithmic framework to design a Circadian Rhythm to set its period in a robust optimal way, which is presented in Section 8.3. For all examples N_0 is set to $N_0 = 1$ in Algorithm 3.

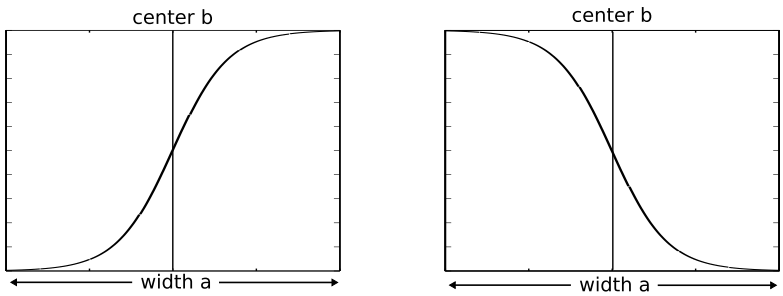


Figure 8.1.: Switching functions: the left switching function is used to guarantee that only one measurement is done at a time point, the right one is used to guarantee that if a perturbation is done at a time point no measurement is done at the same time point.

8.1. Discriminating design for two models describing glycolytic oscillations

In the first test case for model discrimination we implemented the following models for glycolytic oscillations as described in [52].

Model 1 is an allosteric enzyme model with positive feedback under cooperativity and linear product sink. The differential equations for model 1 are given by

$$\begin{aligned}\frac{\mathrm{d}\alpha_1}{\mathrm{d}t} &= \nu - \sigma\phi(\alpha_1, \gamma_1),\\ \frac{\mathrm{d}\gamma_1}{\mathrm{d}t} &= q_1\sigma\phi(\alpha_1, \gamma_1) - k_s\gamma_1,\\ \phi(\alpha_1, \gamma_1) &= \frac{\alpha_1(1+\alpha_1)(1+\gamma_1)^2}{L_1 + (1+\alpha_1)^2(1+\gamma_1)^2}.\end{aligned}$$

Model 2 is an allosteric model with positive feedback in the absence of cooperativity and the product sink is represented by Michaelis-Menten kinetics. The differential equations for Model 2 are given by

$$\begin{aligned}\frac{\mathrm{d}\alpha_2}{\mathrm{d}t} &= \nu - \phi(\alpha_2, \gamma_2),\\ \frac{\mathrm{d}\gamma_2}{\mathrm{d}t} &= q_2\phi(\alpha_2, \gamma_2) - \frac{r_s\gamma_2}{\mu + \gamma_2},\\ \phi(\alpha_2, \gamma_2) &= \frac{\alpha_2(1+\gamma_2)}{L_2 + (1+\alpha_2)(1+\gamma_2)}.\end{aligned}$$

The species concentration of the substrate are denoted by α_j and the ones of the product are denoted by γ_j for $j \in \{1, 2\}$, respectively.
For both models the inflow parameter ν is the same and fixed to the value $\nu = 0.22$. It represents the inflow of substrate to the experimental system, a continuously stirred tank reactor (CSTR).
The parameters σ, q_1, k_s and L_1 of model 1 are regarded as known. Their values are given in Table 8.1. The parameters q_2, r_s, μ and L_2 of model 2 are regarded as unknown and subject to robustification. For the permitted parameter range see Table 8.1.

Model 1				Model 2			
σ	q_1	k_s	L_1	q_2	r_s	μ	L_2
0.92	2.01	0.11	17206.10	$[10^{-7}, 100]$	$[10^{-7}, 100]$	$[10^{-7}, 100]$	$[100, 300]$

Table 8.1.: Parameter values for the glycolytic oscillation models.

For simplicity we consider the homoscedastic case with equal variances, i.e. $v_1 = v_2 = \sigma^2$. In this case $\widetilde{\mathcal{I}}(2:1, \mathcal{O}_1; \xi, \theta_1)$ reduces to,

$$\widetilde{\mathcal{I}}(2:1, \mathcal{O}_1; \xi, \theta_1) = \sum_{i=1}^{n} \mathcal{H}'(t^i) \widetilde{\mathcal{H}}'(c_i) \left((\alpha_{1,i} - \alpha_{2,i})^2 + (\gamma_{1,i} - \gamma_{2,i})^2 \right). \tag{8.1}$$

For this test case the homotopy strategy as presented in Section 7.4 is only applied if the robustification gap $\Delta_{\mathrm{RG}} < 0.1$, then the successive problem $\mathbf{P}_{\Omega_{N+1}}$ is calculated by use of the homotopy strategy with 30 homotopy steps, i.e. $\kappa_h = h/30$, $h \in \{1, ..., 30\}$. Otherwise, problem $\mathbf{P}_{\Omega_{N+1}}$ is solved without homotopy strategy. For each subsequent problem $\mathbf{P}_{\Omega_{N+1}}$ the solution of problem $\mathbf{P}_{\Omega_N}$ is used as initial guess.

We first present a robust design without the possibility to perturb the system by adding species at later time points.
The design is calculated within a fixed time window i.e. $T^{\mathrm{end}} = 400$. 100 equally spaced possible measurement points are defined in the initial state of the optimization procedure, the distance vector Δt between the time points is subject to design and each entry is restricted to $\Delta t^i \in [10^{-7}, 10^{19}]$, $i \in \{1, \ldots, 100\}$. The perturbation vectors c_i are set to $c_i = 0$ for $i \in \{1, \ldots, 99\}$ and are fixed to model the fact that no species perturbation is allowed.
The initial species concentrations which are also subject to experimental design are restricted to $\alpha_{\mathrm{I}} \in [10^{-7}, 25]$ and $\gamma_{\mathrm{I}} \in [10^{-7}, 25]$. The initial values were set to $\alpha_{\mathrm{I}} = 15$ and $\gamma_{\mathrm{I}} = 2$. The parameters of the switching functions $\mathcal{H}'(t^i)$ are chosen as $a_1 = 20.0$ and $b_1 = 10.0$. The parameters of the switching functions $\widetilde{\mathcal{H}}'(c_i)$ are chosen as $a_2 = 0.05$ and

$b_2 = 0.025$. The algorithmic settings are summarized in Table 8.2.

Optimization settings			Integrator settings	
P	δ	IPOPT-tol: Step 1./Step 2.	relTol/absTol	relTolSens/absTolSens
5	10^{-6}	$10^{-10}/10^{-8}$	$10^{-12}/10^{-12}$	$10^{-12}/10^{-12}$

Table 8.2.: On the left hand side the optimization settings are listed comprising the IPOPT stopping tolerances for Step 1. and Step 2. of Algorithm 3 and on the right hand side the integration tolerances for the nominal trajectory and the first order sensitivities are listed. We use the IPOPT option "honor_original_bounds=no" for Step 1. and Step 2. of Algorithm 3.

A plot of the functions α_1, α_2 and γ_1, γ_2 in the initial state and for the solution of problem $\mathbf{P}_{\Omega_1}$ are shown in Figure 8.2.

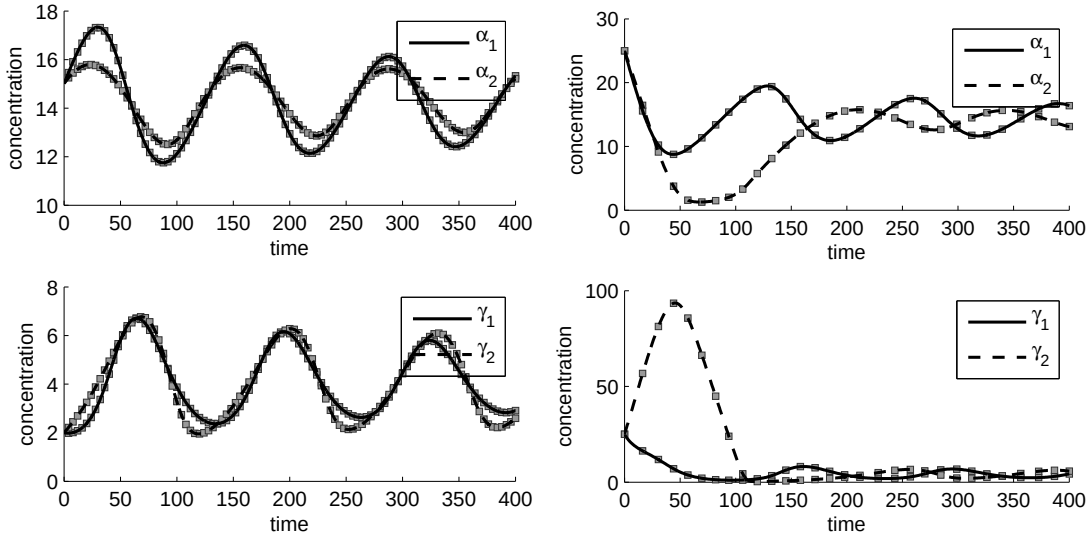


Figure 8.2.: The model functions α_1, α_2 and γ_1, γ_2 are shown before the optimization procedure (left) and after the optimization procedure of problem $\mathbf{P}_{\Omega_1}$ (right) for the glycolytic design setup without the possibility to perturb the system. One square represents one measurement time point.

A plot for the same functions with the same solution design as for problem $\mathbf{P}_{\Omega_1}$ after the next robustification step is shown in Figure 8.3. The final design is also shown in Figure 8.3.

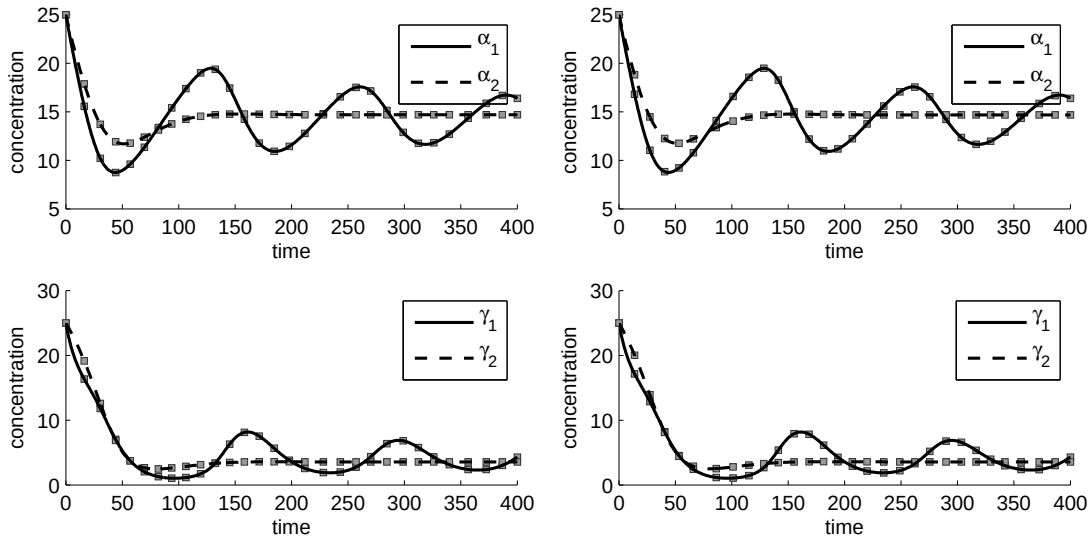


Figure 8.3.: The model functions α_1, α_2 and γ_1, γ_2 are shown for the same solution design as for problem $\mathbf{P}_{\Omega_1}$ after the next robustification step (left) and for the final design (right) for the glycolytic design setup without the possibility to perturb the system. One square represents one measurement time point.

A plot of the robustification gap Δ_{RG} and as well for the objective value of problem $\mathbf{P}_{\Omega_N}$ for each iteration N of Algorithm 3 are shown in Figure 8.4.
A selection of design variables as solutions of problem $\mathbf{P}_{\Omega_N}$ is shown in Figure 8.5(left).

In a second scenario we additionally allow for species perturbations. In this new scenario at the 21-th 41-th, 61-th and 81-th measurement time points, the system can get perturbed by additional species quantities. The free vectors c_i, $i \in \{21, 41, 61, 81\}$ are constrained by $c_i \in [10^{-7}, 10]$. The initial values are set to $c_i = 1$. The remaining conditions are as before. However, we change the time vector bound constraints for

$$i \in \{1, 6, 11, 21, 26, 31, 41, 46, 51, 61, 66, 71, 81\}$$

to $\Delta t^i \in [8, 10^{19}]$ and the initial state to $\Delta t^i = 15$. The bounds for the remaining entries are as before, and the remaining measurement time points are equally spaced.
A plot of the functions α_1, α_2 and γ_1, γ_2 in the initial state and for the solution of problem $\mathbf{P}_{\Omega_1}$ are shown in Figure 8.6. A plot for the same functions with the same solution design as for problem $\mathbf{P}_{\Omega_1}$ after the next robustification step is shown in Figure 8.7. The final design is also shown in Figure 8.7.

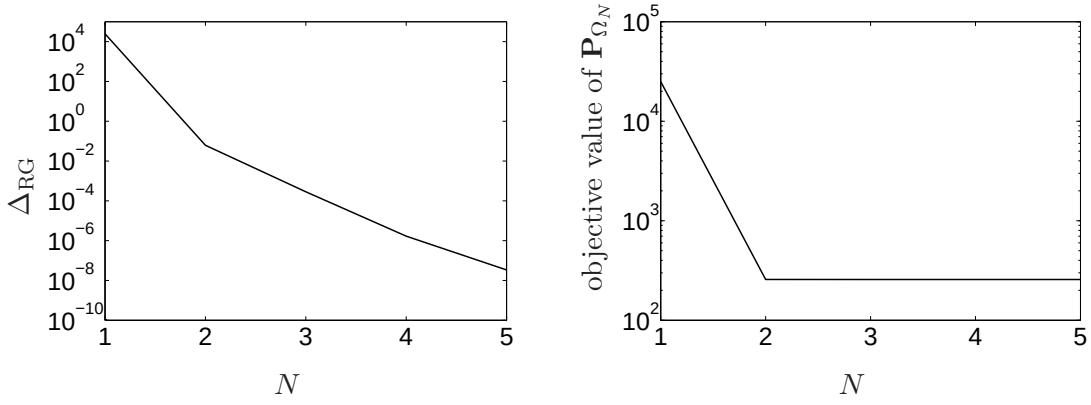


Figure 8.4.: In the left figure the robustification gap Δ_{RG} is plotted versus the number of iterations N of Algorithm 3 and in the right figure the objective value of problem $\mathbf{P}_{\Omega_N}$ is shown for the glycolytic design setup without the possibility to perturb the system.

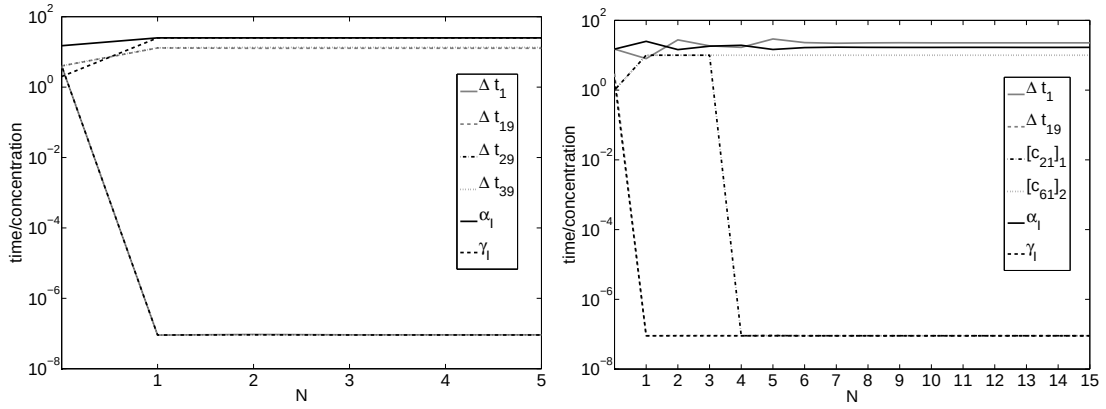


Figure 8.5.: A selection of design variables as solutions of problem $\mathbf{P}_{\Omega_N}$ for the glycolytic design setup without the possibility to perturb the system (left) and with the possibility to perturb the system (right) are shown.

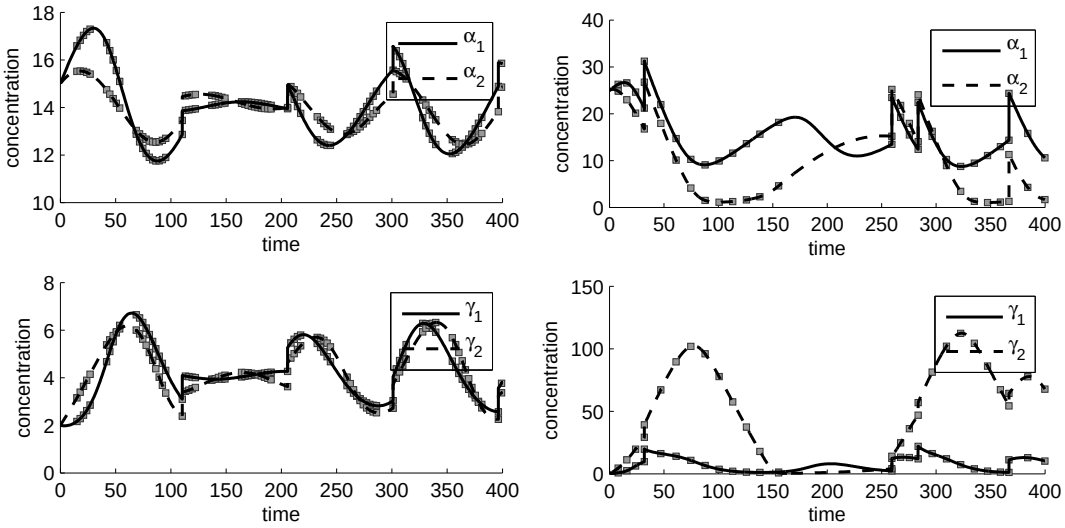


Figure 8.6.: The model functions α_1, α_2 and γ_1, γ_2 are shown before the optimization procedure (left) and after the optimization procedure of problem $\mathbf{P}_{\Omega_1}$ (right) for the glycolytic design setup with the possibility to perturb the system. One square represents one measurement time point.

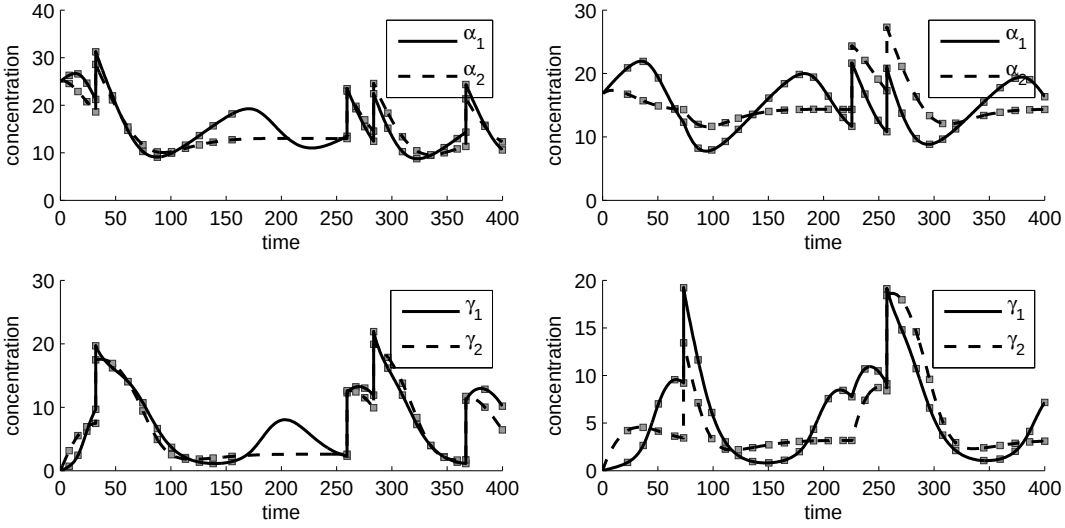
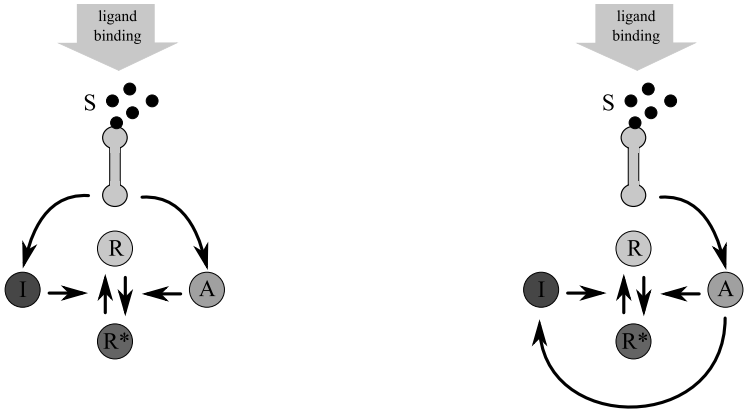


Figure 8.7.: The model functions α_1, α_2 and γ_1, γ_2 are shown for the same solution design as for problem $\mathbf{P}_{\Omega_1}$ after the next robustification step (left) and for the final design (right) for the glycolytic design setup with the possibility to perturb the system. One square represents one measurement time point.

A plot of the robustification gap Δ_{RG} and the objective value of problem $\mathbf{P}_{\Omega_N}$ for each iteration N of Algorithm 3 are shown in Figure 8.8. A selection of design variables as solutions of problem $\mathbf{P}_{\Omega_N}$ is shown in Figure 8.5(right).

8.2. Discriminating design for two models describing signal sensing in dictyostelium discoideum

The second test case is the discrimination of two models describing the chemotactic response in the amoeba *dictyostelium discoideum* as presented in [79] using the framework presented in Chapter 7. The two models describe the adaption mechanism observed when amoebae encounter the chemoattractant cAMP [76], see Figure 8.9.



(a) model 1 (b) model 2

Figure 8.9.: Two models of the signal system of the Dictyostelium amoeba.

For both models, a chemotaxis response regulator R gets activated (R^*) by an activator enzyme A, when a cAMP ligand S appears. But the deactivating mechanism determined by the interaction with an inhibitor molecule I differs for both models. Both models comprise mass action kinetics in form of ODE.

In model 1 the activator enzyme as well as the inhibitor enzyme are regulated by the external signal, which is proportional to the cAMP concentration S. The overall model

in this case is given by

$$\begin{aligned}\frac{\mathrm{d}A_1}{\mathrm{d}t} &= -k_{-a}A_1 + k_a S_1\\ \frac{\mathrm{d}I_1}{\mathrm{d}t} &= -k_{-i}I_1 + k_{i_1} S_1\\ \frac{\mathrm{d}R_1^*}{\mathrm{d}t} &= -(k_r A_1 + k_{-r} I_1)R_1^* + k_r R_T A_1,\end{aligned} \tag{8.2}$$

where k_{-a}, k_a, k_{-i}, k_{i_1}, k_r and k_{-r} are the mass action rate constants and $R_T := R^* + R$ is the total amount of the response regulator.

In model 2 the inhibitory molecule I is activated through the indirect action of activator A instead of direct activation by sensing ligand binding. The overall model in this case is given by,

$$\begin{aligned}\frac{\mathrm{d}A_2}{\mathrm{d}t} &= -k_{-a}A_2 + k_a S_2\\ \frac{\mathrm{d}I_2}{\mathrm{d}t} &= -k_{-i}I_2 + k_{i_2} A_2\\ \frac{\mathrm{d}R_2^*}{\mathrm{d}t} &= -(k_r A_2 + k_{-r} I_2)R_2^* + k_r R_T A_2,\end{aligned} \tag{8.3}$$

where k_{-a}, k_a, k_{-i}, k_{i_2}, k_r and k_{-r} are the mass action rate constants and $R_T := R^* + R$ is the total amount of the response regulator.
For modeling details we refer to [79]. We have extended these systems of ordinary differential equations by an additional state corresponding to the cAMP ligand S with $\mathrm{d}S/\mathrm{d}t = 0$. By allowing species concentration perturbations c only to the state S we can mimic a piecewise constant control of the system by the cAMP ligand S.

The experimental design parameters are the initial species concentrations of the four states namely, A_I, I_I, R_I^*, S_I, the measurement time points t and the species concentration perturbation c with respect to S. We discard the condition that either a measurement or a perturbation can be performed since in that setting by use of the perturbations c we mimic a piecewise constant input control S and therefore that restriction seems unnatural. Again for simplicity we consider the homoscedastic case with equal variances i.e. $v_1 = v_2 = \sigma^2$, where $\widetilde{\mathcal{I}}(2:1, \mathcal{O}_1; \xi, \theta_1)$ reduces now to

$$\widetilde{\mathcal{I}}(2:1, \mathcal{O}_1; \xi, \theta_1) = \sum_{i=1}^{n} \mathcal{H}'(t^i)\left((A_{1,i} - A_{2,i})^2 + (I_{1,i} - I_{2,i})^2 + (R_{1,i}^* - R_{2,i}^*)^2\right). \tag{8.4}$$

The parameters k_{-a}, k_a, k_{-i}, k_{i_1}, k_r, k_{-r} and R_T are regarded as known and fixed, their

values are given in Table 8.3.

k_{-a}	k_a	k_{-i}	k_{i_1}	k_r	k_{-r}	R_T
2.0	3.0	0.1	1.0	1.0	1.0	23/30

Table 8.3.: Parameter values for the fix values within model 1 and model 2.

Parameter k_{i_2} is regarded as unknown and subject to robustification. The range of the parameter k_{i_2} is set to $k_{i_2} \in [0, 2]$.
The optimal design is calculated within a fixed time window with $T^{\text{end}} = 100$. 100 equally spaced possible measurement points are defined in the initial state of the optimization procedure. The distance vector Δt between time points is subject to design and each entry is restricted to $\Delta t^i \in [10^{-7}, 10^{19}]$ for $i \in \{1, \ldots, 100\}$.
The free perturbation vectors c_i for $i \in \{11, 21, 31, 41, 51, 61, 71, 81, 91\}$ are not restricted. The initial values are set to $c_i = 0$ for $i \in \{11, 21\}$, $c_{31} = 0.3$, $c_i = -0.48$, $i \in \{41, 61, 81\}$ and $c_i = 0.48$ for $i \in \{51, 71, 91\}$.
The initial species concentrations which are also subject to the experimental design are restricted to $S_{\text{I}} \in [0.01, 0.5]$, $A_{\text{I}} \in [10^{-7}, 1]$, $I_{\text{I}} \in [10^{-7}, 1]$ and $R^*_{\text{I}} \in [10^{-7}, 1]$. The initial values are set to $S_{\text{I}} = 0.2$, $A_{\text{I}} = 1.0$, $I_{\text{I}} = 10^{-4}$ and $R^*_{\text{I}} = 10^{-4}$. The multiple shooting intermediate variables for the species S are restricted to $s_i \in [0.01, 0.5]$ to restrict the piecewise constant control to this interval. The parameters of the switching functions $\mathcal{H}'(t^i)$ are chosen as $a_1 = 5.0$ and $b_1 = 2.5$. The algorithmic settings are summarized in Table 8.4.

Optimization settings			Integrator settings	
P	δ	`IPOPT`-tol: Step 1./Step 2.	relTol/absTol	relTolSens/absTolSens
5	10^{-8}	$10^{-10}/10^{-11}$	$10^{-14}/10^{-14}$	$10^{-14}/10^{-14}$

Table 8.4.: On the left hand side the optimization settings are listed comprising the `IPOPT` stopping tolerances for Step 1. and Step 2. of Algorithm 3 and on the right hand side the integration tolerances for the nominal trajectory and the first order sensitivities are listed. We use the `IPOPT` option "`honor_original_bounds=no`" for Step 1. and Step 2. of Algorithm 3.

With these design conditions we start the optimization procedure twice. First by use of the homotopy strategy for successive problems $\mathbf{P}_{\Omega_{N+1}}$ with 10 homotopy steps.
Since the "discriminating power" of the experimental setup is very low in this case, i.e. the deviation between the two models is small, we plot the distance functions $(S_1 - S_2)$, $(A_1 - A_2)$, $(I_1 - I_2)$ and $(R^*_1 - R^*_2)$ for the initial state and for the solution of problem

$\mathbf{P}_{\Omega_1}$ in Figure 8.10. A plot for the same functions with the same solution design as for problem $\mathbf{P}_{\Omega_1}$ after the next robustification step is shown in Figure 8.11. The final design is also shown in Figure 8.11.
A plot of the robustification gap Δ_{RG} for each iteration N of Algorithm 3 is shown in Figure 8.12 (left). A plot of the objective value of problem $\mathbf{P}_{\Omega_N}$ for each iteration N of Algorithm 3 is shown in Figure 8.13 (left). A selection of design variables as solutions of problem $\mathbf{P}_{\Omega_N}$ is shown in Figure 8.14 (left).

Secondly, we calculate the design without the homotopy strategy. We experience huge jumps in the final objective value of problem $\mathbf{P}_{\Omega_N}$ for subsequent iterations N of Algorithm 3. This is due to the fact that the final design of the former problem $\mathbf{P}_{\Omega_N}$ is an infeasible starting point for the successive problem $\mathbf{P}_{\Omega_{N+1}}$ in the *Interior Point* solution strategy. First the optimizer tries to force the iterates back into the feasible region and afterwards the new central path leads to a different locally optimal design.
For this case a plot of the robustification gap Δ_{RG} for each iteration N of Algorithm 3 is shown in Figure 8.12 (right). A plot of the objective value of problem $\mathbf{P}_{\Omega_N}$ for each iteration N of Algorithm 3 is shown in Figure 8.13 (right). A selection of design variables as solutions of problem $\mathbf{P}_{\Omega_N}$ is shown in Figure 8.14 (right).
As one can clearly see, the homotopy strategy helps to considerably stabilize Algorithm 3.

8.3. Optimal design of Circadian Rhythm

For this example no experimental design is calculated, but a cellular oscillator is designed such that its period is set in an robust optimal way. For that purpose, the algorithm developed in this thesis is used. The calculated results originate from a cooperative work with Marcel Rehberg. Similar results with slight different setting are published in [73]. The mathematical model of the circadian oscillator, which is used here, is taken from [75] and describes the circadian system in the fruit fly Drosophila. More precisely, it models the transcriptional network of the proteins TIM and PER.

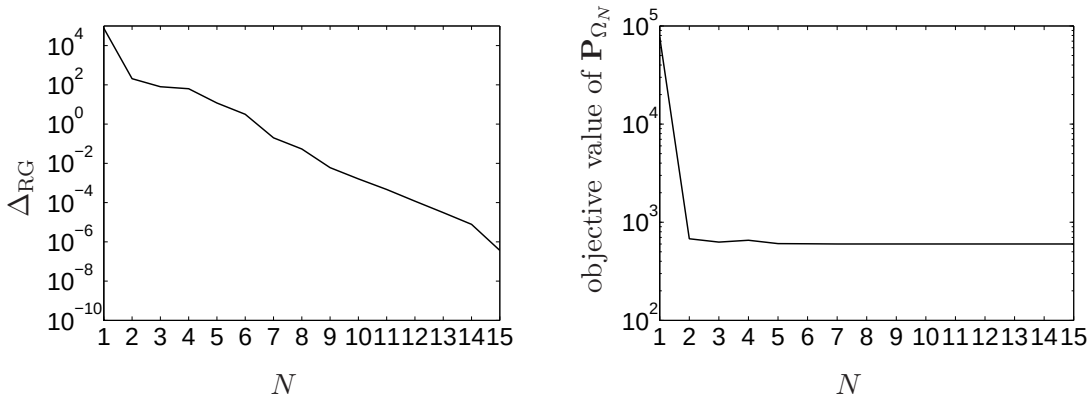


Figure 8.8.: In the left figure the robustification gap Δ_{RG} is plotted versus the number of iterations N of Algorithm 3 and in the right figure the objective value of problem $\mathbf{P}_{\Omega_N}$ is shown for the glycolytic design setup with the possibility to perturb the system.

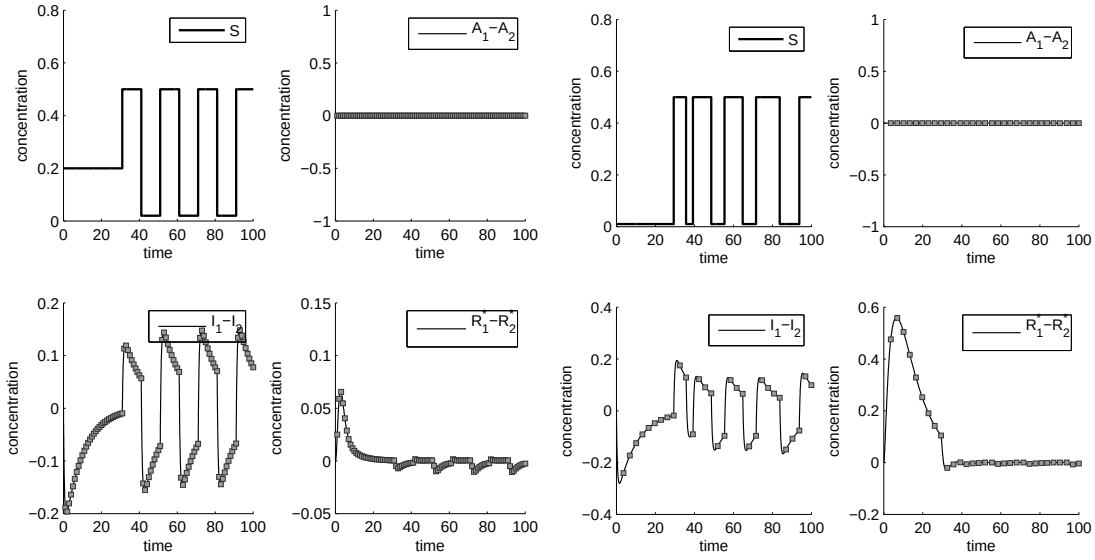


Figure 8.10.: The model variable distance functions $(S_1 - S_2)$, $(A_1 - A_2)$, $(I_1 - I_2)$ and $(R_1^* - R_2^*)$ are shown before the optimization procedure (left) and after the optimization procedure of problem $\mathbf{P}_{\Omega_1}$ (right) for two models describing signal sensing in *dictyostelium discoideum*. One square represents one measurement time point.

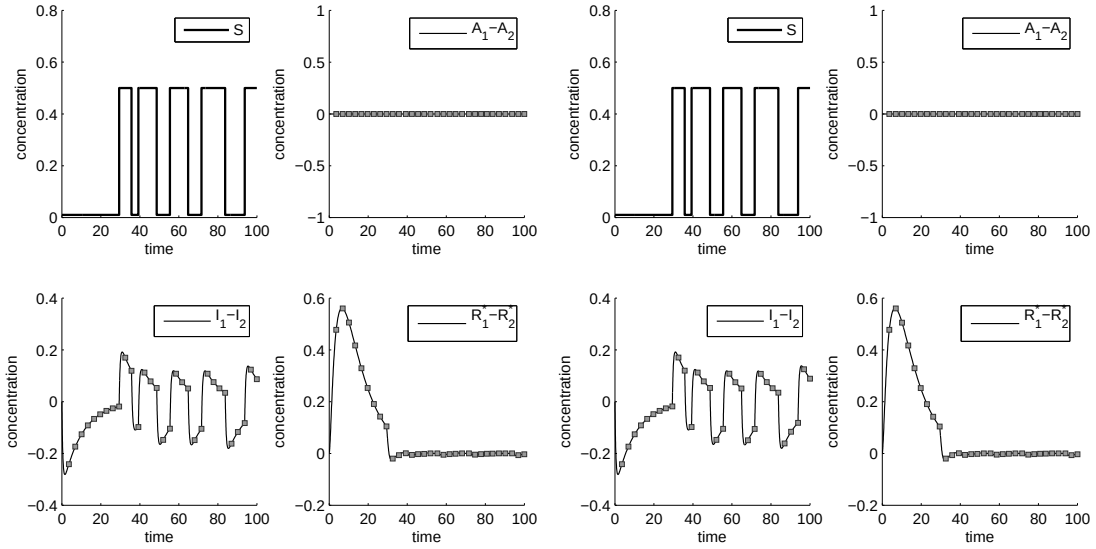
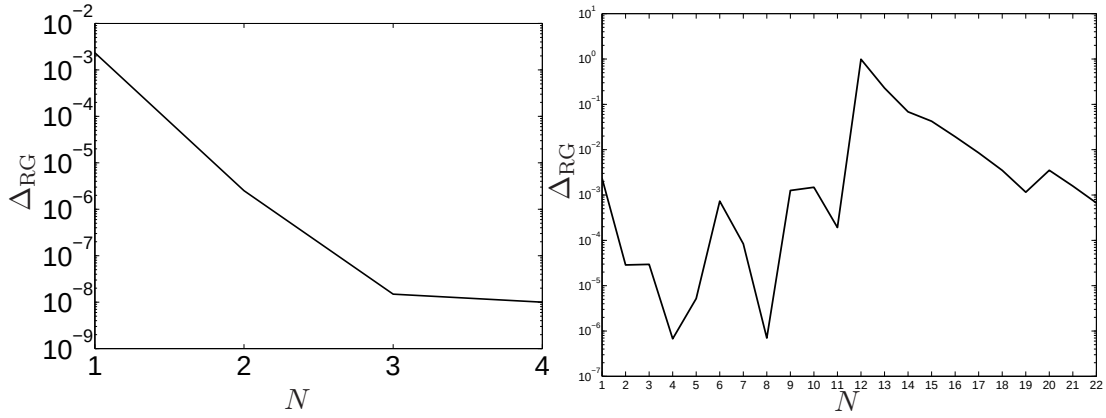
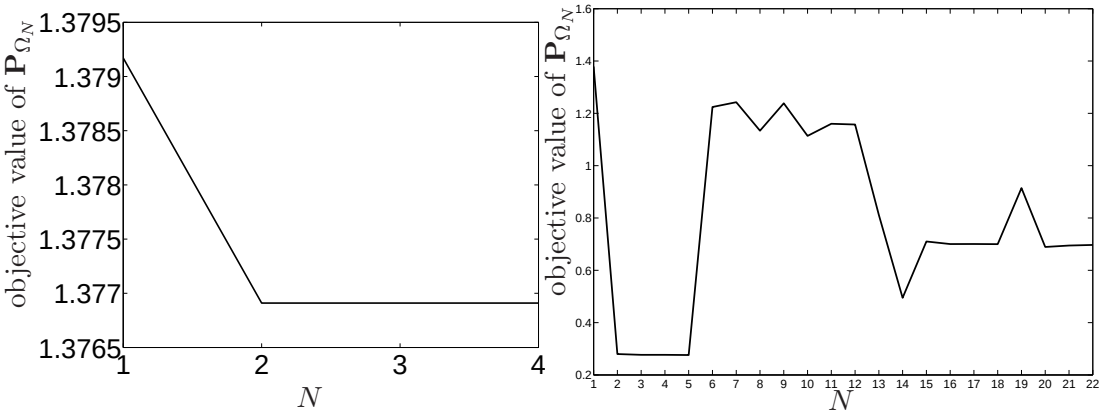


Figure 8.11.: The model variable distance functions $(S_1 - S_2)$, $(A_1 - A_2)$, $(I_1 - I_2)$ and $(R_1^* - R_2^*)$ are shown for the same solution design as for problem $\mathbf{P}_{\Omega_1}$ after the next robustification step and for the final design (right) for two models describing signal sensing in *dictyostelium discoideum*. One square represents one measurement time point.



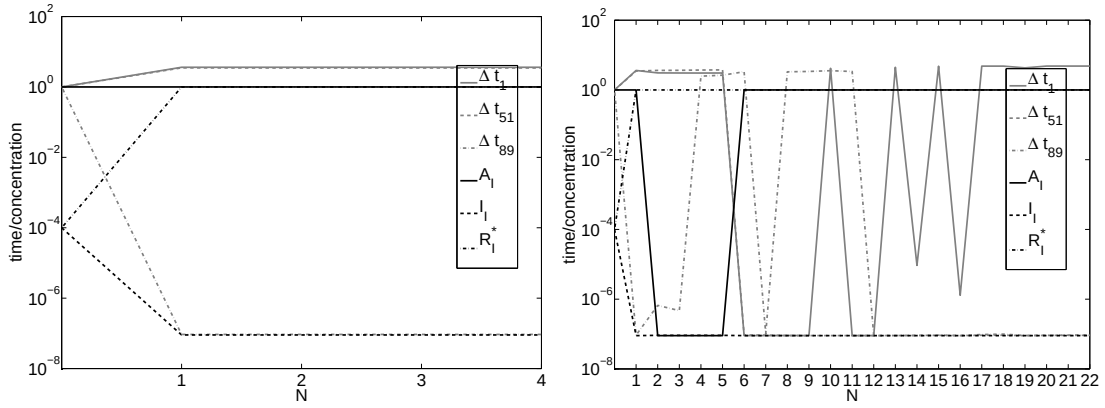
(a) with homotopy strategy (b) without homotopy strategy

Figure 8.12.: The robustification gap Δ_{RG} is plotted versus the number of iterations N of Algorithm 3 for the setup with two models describing signal sensing in *dictyostelium discoideum*, with homotopy strategy (left) and without homotopy strategy (right).



(a) with homotopy strategy (b) without homotopy strategy

Figure 8.13.: The objective value of problem $\mathbf{P}_{\Omega_N}$ is plotted versus the number of iterations N of Algorithm 3 for the setup with two models describing signal sensing in *dictyostelium discoideum*, with homotopy strategy (left) and without homotopy strategy (right).



(a) with homotopy strategy (b) without homotopy strategy

Figure 8.14.: A selection of design variables calculated as solution of problem $\mathbf{P}_{\Omega_N}$ is plotted versus the number of iterations N of Algorithm 3 for the setup with two models describing signal sensing in *dictyostelium discoideum*, on the left with homotopy strategy and on the right without homotopy strategy.

The model system is given by

$$
\begin{aligned}
\frac{\mathrm{d}y^1}{\mathrm{d}t} =& v_{sP}\frac{(K_{IP})^n}{(K_{IP})^n + (y^{10})^n} - v_{mP}\frac{y^1}{K_{mP}+y_1} - k_d y^1 \\
\frac{\mathrm{d}y^2}{\mathrm{d}t} =& k_{sP}y_1 - v_{1P}\frac{y^2}{K_{1P}+y^2} + v_{2P}\frac{y^3}{K_{2P}+y^3} - k_d y^2 \\
\frac{\mathrm{d}y^3}{\mathrm{d}t} =& v_{1P}\frac{y^2}{K_{1P}+y^2} - v_{2P}\frac{y_3}{K_{2P}+y^3} - v_{3P}\frac{y^3}{K_{3P}+y^3} + \\
& v_{4P}\frac{y^4}{K_{4P}+y^4} - k_d y^3 \\
\frac{\mathrm{d}y^4}{\mathrm{d}t} =& v_{3P}\frac{y^3}{K_{3P}+y^3} - v_{4P}\frac{y^4}{K_{4P}+y^4} - k_3 y^4 y^8 + k_4 y^9 - \\
& v_{dP}\frac{y^4}{K_{dP}+y^4} - k_d y^4 \\
\frac{\mathrm{d}y^5}{\mathrm{d}t} =& V_{sT}\frac{(K_{IT})^n}{(K_{IT})^n + (y^{10})^n} - v_{mT}\frac{y^2}{K_{mT}+y^5} - k_d y^5 \\
\frac{\mathrm{d}y^6}{\mathrm{d}t} =& k_{sT}y^5 - v_{1T}\frac{y^6}{K_{1T}+y^6} + v_{2T}\frac{y^7}{K_{2T}+y^7} - k_d y^6 \\
\frac{\mathrm{d}y^7}{\mathrm{d}t} =& v_{1T}\frac{y^6}{K_{1T}+y^6} - v_{2T}\frac{y^7}{K_{2T}+y^7} - v_{3T}\frac{y^7}{K_{3T}+y^7} + \\
& v_{4T}\frac{y_8}{K_{4T}+y_8} - k_d y_7 \\
\frac{\mathrm{d}y^8}{\mathrm{d}t} =& v_{3T}\frac{y_7}{K_{3T}+y_7} - v_{4T}\frac{y_8}{K_{4T}+y_8} - k_3 y_4 y_8 + k_4 y_9 - \\
& v_{dT}\frac{y_8}{K_{dT}+y_8} + -k_d y_8 \\
\frac{\mathrm{d}y^9}{\mathrm{d}t} =& k_3 y_4 y_8 - k_4 y_9 - k_1 y_9 + k_2 y_{10} - k_{dC} y_9 \\
\frac{\mathrm{d}y^{10}}{\mathrm{d}t} =& k_1 y_9 - k_2 y_{10} - k_{dN} y_{10},
\end{aligned}
\tag{8.5}
$$

where y^1 is linked to "per mRNA" and y^5 is linked to "tim mRNA". The model comprises three phosphorylation states, namely 0, 1 and 2 of "PER protein", which are associated to y^2, y^3 and y^4, respectively. It also comprises in the same manner three phosphorylation states of "TIM protein", which are associated to y^6, y^7 and y^8, respectively. The "PER-TIM-complex" in the cytoplasm is associated to y^9, whereas the "PER-TIM-complex" in the nucleus is associated to y^{10}.

i	1	2	3	4	5
$y^i_{\mathrm{I,min}}$	1.0391	0.2983	0.2624	0.1697	1.0391
y^i_{I}	1.5587	0.4474	0.3936	0.2545	1.5587
$y^i_{\mathrm{I,max}}$	2.0783	0.5965	0.5248	0.3393	2.0783

i	6	7	8	9	10
$y^i_{\mathrm{I,min}}$	0.2985	0.2638	0.1819	0.0953	0.3730
y^i_{I}	0.4477	0.3957	0.2728	0.1429	0.5595
$y^i_{\mathrm{I,max}}$	0.5969	0.5276	0.3637	0.1905	0.7460

Table 8.5.: The initial values of y_{I} and the bounds $y_{\mathrm{I,min}}$, $y_{\mathrm{I,max}}$ are shown for the mathematical model of the circadian oscillator in Drosophila.

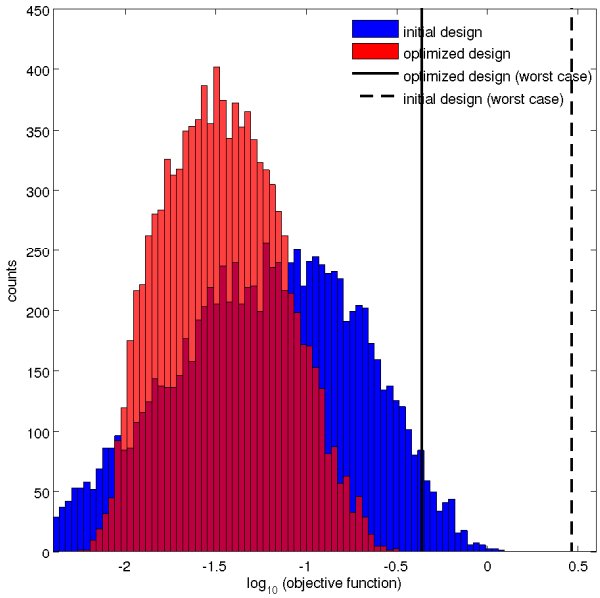


Figure 8.15.: The figure shows histograms of the objective function values on a logarithmic scale for a sampled set of robustification vectors θ_{r}. First, a histogram is shown for the sample scenario with initial design in blue and second for the optimized design in red for the model of a circadian oscillator. In any case, the vertical black line gives a worst case estimate of the objective function value on a logarithmic scale for the specific design.

According to [110] and [73], the parameters can be grouped into two sets. The first one is associated to the local parameters of the circadian system, i.e. parameters which only affect the circadian clock. These parameters are listed in Table 8.6. The second one comprises parameter which affect other cellular processes, as well. All these global parameters are listed in Table 8.7. The local parameters $\theta_\mathrm{d} \subset \Theta_\mathrm{d} \in \mathbb{R}^{p_\mathrm{d}}$ with

$$\begin{aligned}\theta_\mathrm{d} = \big(&n, K_{IP}, K_{IT}, v_{1P}, v_{1T}, v_{2P}, v_{2T}, v_{3P}, v_{3T}, v_{4P},\\ &v_{4T}, K_{1P}, K_{1T}, K_{2P}, K_{2T}, K_{3P}, K_{3T}, K_{4P}, K_{4T}\big)^T\end{aligned}$$

are used as design variables, whereby the optimized design should be robust against the global ones $\theta_\mathrm{r} \subset \Theta_\mathrm{r} \in \mathbb{R}^{p_\mathrm{r}}$ with

$$\begin{aligned}\theta_\mathrm{r} = \big(&v_{sP}, v_{sT}, v_{mP}, v_{mT}, v_{dP}, v_{dT}, K_{mP}, K_{mT}, K_{dP}, K_{dT},\\ &k_{sP}, k_{sT}, k_1, k_2, k_3, k_4, k_d, k_{dC}, k_{dN}\big)^T.\end{aligned}$$

The initial species concentration vector y_I is considered as design vector, too.
For an appropriate objective function $\mathcal{J}(y(t))$, the considered design optimization problem is given by

$$\min_{y_\mathrm{I}, \theta_\mathrm{d}} \max_{\theta_\mathrm{r}} \mathcal{J}\left(y(t)\right)$$

subject to

$$\begin{aligned}\frac{\mathrm{d}y}{\mathrm{d}t} &= f(y, \theta_\mathrm{d}, \theta_\mathrm{r}),\ t \in [0, T],\\ y(0) &= y_\mathrm{I},\\ y_{\mathrm{I,min}} \leq y_\mathrm{I} &\leq y_{\mathrm{I,max}}\\ \theta_{\mathrm{d,min}} \leq \theta_\mathrm{d} &\leq \theta_{\mathrm{d,max}},\\ \theta_{\mathrm{r,min}} \leq \theta_\mathrm{r} &\leq \theta_{\mathrm{r,max}},\end{aligned}$$

where $y_{\mathrm{I,min}}, y_{\mathrm{I,max}} \in \mathbb{R}^{10}$, $\theta_{\mathrm{d,min}}, \theta_{\mathrm{d,max}} \in \mathbb{R}^{p_\mathrm{d}}$, $\theta_{\mathrm{r,min}}, \theta_{\mathrm{r,max}} \in \mathbb{R}^{p_\mathrm{r}}$ and $y_{\mathrm{I,min}} \leq y_{\mathrm{I,max}}$, $\theta_{\mathrm{d,min}} \leq \theta_{\mathrm{d,max}}$, $\theta_{\mathrm{r,min}} \leq \theta_{\mathrm{r,max}}$, component wise. The objective function $\mathcal{J}(y(t))$ is defined by

$$\mathcal{J}(y(t)) := \sum_{k=1}^{3} \|y_\mathrm{I} - y(t_k)\|_2^2 \quad \text{with} \quad t_1 = \tau, t_1 = 2\tau, t_3 = 3\tau \tag{8.6}$$

where τ is the desired period of the circadian system. Here, τ is set to 24. For more details on the design problem we refer to [73].
The initial values of y_I and the bounds $y_{\mathrm{I,min}}$, $y_{\mathrm{I,max}}$ are shown in Table 8.5. The initial values of θ_d and the bounds $\theta_{\mathrm{d,min}}$, $\theta_{\mathrm{d,max}}$ are shown in Table 8.6. Those of θ_r and the

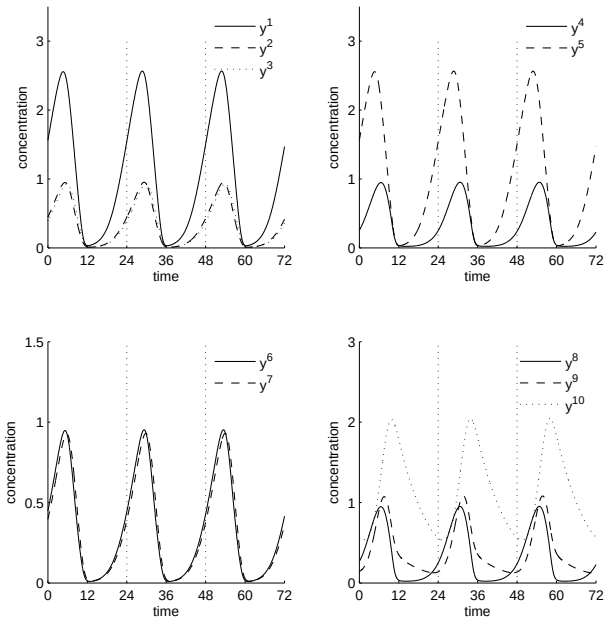


Figure 8.16.: The model states y are shown for the initial design with initial vector θ_{r} for the model of a circadian oscillator.

corresponding bounds $\theta_{\mathrm{r,min}}$, $\theta_{\mathrm{r,max}}$ are shown in Table 8.7. The algorithmic settings are summarized in Table 8.8. The optimal design is calculated without the homotopy strategy of Section 7.4.

The initial design is shown in Figure 8.16. The same design for the worst case realization of θ_{r} within the predefined bounds as given in Table 8.7 is shown in Figure 8.17. The final solution design is shown in Figure 8.18.

To assess the quality of the calculated design, we draw 10000 samples from a uniform distribution over the robustification space $[\theta_{\mathrm{r,min}}, \theta_{\mathrm{r,max}}]$ and for each sample we simulate the circadian model system, given by (8.5), twice.

First, we simulate under the final design conditions, i.e. for the optimal design vectors $\hat{y}_{\mathrm{I}}$ and $\hat{\theta}_{\mathrm{d}}$. Second, we simulate under the initial condition for y_{I} and θ_{d}. Each time, we use the drawn sample to define the global parameter vector θ_{r}.

For each simulation we calculate the objective function in (8.6) and generate histograms of the results for both scenarios, shown in Figure 8.15. The histograms clearly show the increased robustness of the final design in respect of variations in θ_{r}.

i	1	2	3	4	5	6	7	8	9	10
$\theta^i_{\mathrm{d,min}}$	0.4	0.1	0.1	0.8	0.8	0.1	0.1	0.8	0.8	0.1
θ^i_{d}	4.0	1.0	1.0	8.0	8.0	1.0	1.0	8.0	8.0	1.0
$\theta^i_{\mathrm{d,max}}$	40	10	10	80	80	10	10	80	80	10

i	11	12	13	14	15	16	17	18	19	
$\theta^i_{\mathrm{d,min}}$	0.1	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	
θ^i_{d}	1.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	
$\theta^i_{\mathrm{d,max}}$	10	20	20	20	20	20	20	20	20	

Table 8.6.: The initial values of θ_{d} and the bounds $\theta_{\mathrm{d,min}}$, $\theta_{\mathrm{d,max}}$ are shown for the mathematical model of the circadian oscillator in Drosophila.

i	1	2	3	4	5	6	7	8	9	10
$\theta^i_{\mathrm{r,min}}$	0.8	0.8	0.56	0.56	0.16	0.16	0.16	0.16	0.72	0.72
θ^i_{r}	1.0	1.0	0.7	0.7	0.2	0.2	0.2	0.2	0.9	0.9
$\theta^i_{\mathrm{r,max}}$	1.2	1.2	0.84	0.84	0.24	0.24	0.24	0.24	1.08	1.08

i	11	12	13	14	15	16	17	18	19	
$\theta^i_{\mathrm{r,min}}$	1.6	1.6	0.48	0.16	0.96	0.48	0.008	0.008	0.008	
θ^i_{r}	2.0	2.0	0.6	0.2	1.2	0.6	0.01	0.01	0.01	
$\theta^i_{\mathrm{r,max}}$	2.4	2.4	0.72	0.24	1.44	0.72	0.012	0.012	0.012	

Table 8.7.: The initial values of θ_{r} and the bounds $\theta_{\mathrm{r,min}}$, $\theta_{\mathrm{r,max}}$ are shown for the mathematical model of the circadian oscillator in Drosophila.

Optimization settings			Integrator settings	
P	δ	`IPOPT`-tol: Step 1./Step 2.	relTol/absTol	relTolSens/absTolSens
10	10^{-2}	$10^{-8}/10^{-8}$	$10^{-11}/10^{-13}$	$10^{-11}/10^{-13}$

Table 8.8.: On the left hand side the optimization settings are listed comprising the `IPOPT` stopping tolerances for Step 1. and Step 2. of Algorithm 3 and on the right hand side the integration tolerances for the nominal trajectory and the first order sensitivities are listed. We use the `IPOPT` options "`honor_original_bounds=no`" for Step 1. and Step 2. of Algorithm 3 and "`mu_strategy=adaptive`" for Step 2. of Algorithm 3.

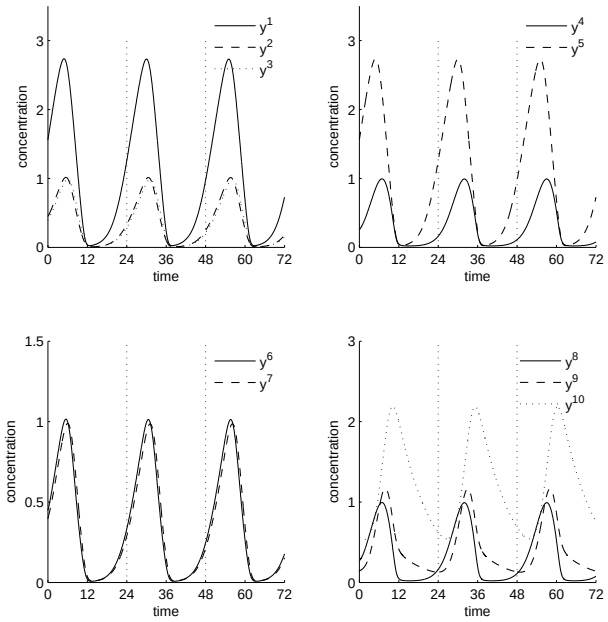


Figure 8.17.: The model states y are shown for the initial design and θ_{r} is set to the worst case value for the model of a circadian oscillator.

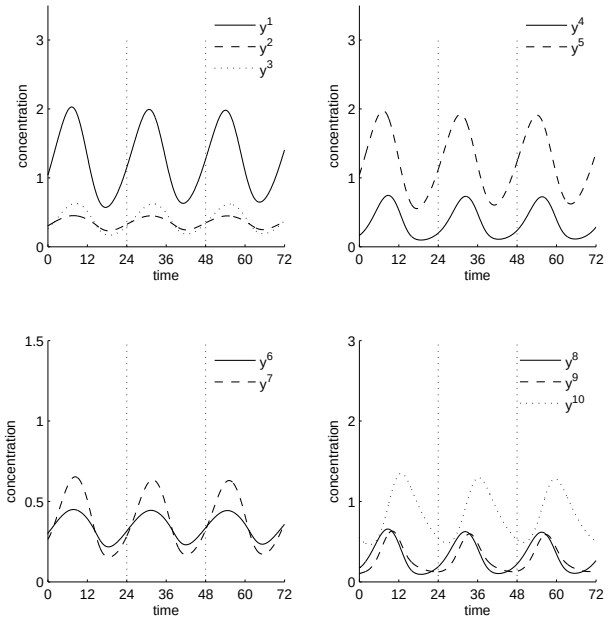


Figure 8.18.: The model states y are shown for the optimized design for the model of a circadian oscillator.

CHAPTER 9

Conclusion and Outlook

In this work, we present a framework for the robust computation of optimal experimental designs for the purpose of model discrimination.
First, motivated by a real experimental setup, which is established by a group of experimental biologists, we derive an extended statistical framework, which explicitly contains the design goals to determine the optimal initial conditions, the number of measurements and the optimal placements of measurements under the constraint that only one measurement at a time point can be performed. We additionally restrict the placement of a measurement such that a next measurement can only be placed after a fixed and constant time span. Further, we allow for system perturbations whereby the placement is also subject to the experimental design. This statistical framework is rigorously derived from the well known Kullback-Leibler divergence and is translated to a discontinous semi-infinite optimization problem.
To tackle this semi-infinite optimization problem we develop a smoothing approach to construct a continuous semi-infinite approximation depending on smoothing parameters, which control the quality of the approximation. The smoothing approach is theoretically validated such that any desired quality of the approximation can be achieved.
We develop an algorithm to numerically calculate such optimal designs by utilization of an *Outer Approximations* scheme to solve the underlying semi-infinite optimization problem. A strategy for the numerical stabilization of the algorithm by use of a homotopy approach is suggested and implemented.
For two relevant biological settings we successfully apply the developed framework and calculate experimental designs for various scenarios. In our examples we clearly find

that the homotopy approach is significantly superior to a cold start of successive design problems $\mathbf{P}_{\Omega_{N+1}}$. For the first test case, the discrimination of two models describing glycolytic oscillations, the *Outer Approximations* scheme completely fails to reach the desired accuracy δ without homotopy strategy. For the second test case, the discrimination of two models describing signal sensing in *dictyostelium discoideum*, the *Outer Approximations* scheme also fails without warm start, however the homotopy strategy works with only two homotopy steps (not presented in this work but the results are essentially the same). We further successfully apply the robust framework to design a Circadian Rhythm to set its period in a robust optimal way.

9.1. Outlook and further work

An extension to the statistical scenario for model discrimination, which is presented in this work, includes the situation to test whether given measurement data can be explained best by one out of a finite set of probability models based on measures P_{1,r_1}, $r_1 \in \{1, ..., M_1\}$, against the hypothesis that the measurement can best be explained by another one out of a second finite set of probability models based on measures P_{2,r_2}, $r_2 \in \{1, ..., M_2\}$. Each probability model P_{j,r_j} might be parametrized by parameters $\theta_{j,r_j} \in \Theta_{j,r_j} \subset \mathbb{R}^{p_{j,r_j}}$, $j \in \{1, 2\}$.
By calculating

$$\hat{\xi} = \arg\max_{\xi \in \Xi} \min_{\substack{r_1 \in \{1,...,M_1\} \\ r_2 \in \{1,...,M_2\}}} \min_{\substack{\theta_{1,r_1} \in \Theta_{1,r_1} \\ \theta_{2,r_2} \in \Theta_{2,r_2}}} \mathcal{I}(P_{2,r_2}(\theta_{2,r_2}) : P_{1,r_1}(\theta_{1,r_1}), \mathcal{O}_1; \xi)$$

we can get a robust worst case estimate of an optimally discriminating design for the case of composite null and alternative hypothesis. This is a more realistic setting in the light of practical applicability.
If the range of each parameter space Θ_{j,r_j}, $r_j \in \{1, ..., M_j\}$, $j \in \{1, 2\}$, is large, then the resulting discriminating design might have poor discriminating power, i.e. each distinct model adapts very well for some region in the permitted parameter space. Therefore, it is important to incorporate appropriate restrictions to these parameter sets. These restrictions should be based on previous knowledge, on the possible and reasonable parameter range and on current measurement information. This implies that for R previous measurement runs the parameter sets Θ_{j,r_j}, $r_j \in \{1, ..., M_j\}$, $j \in \{1, 2\}$, which are considered for the calculation of a robust optimal design, have to be restricted such that each distinct model still fits to previous observations for any parametric realization of its restricted parameter set. One way to incorporate these restrictions is to replace (7.33) (i.e. Step 1. in the *Outer Approximations* scheme for the numerical calculation of

the optimal design problem, which is considered in this work) by

$$(r_1, r_2, \hat{\theta}_{1,r_1,N}, \hat{\theta}_{2,r_2,N}) = \arg \min_{\substack{r_1 \in \{1,\ldots,M_1\} \\ r_2 \in \{1,\ldots,M_2\}}} \min_{\substack{\theta_{1,r_1} \in \Theta_{1,r_1} \\ \theta_{2,r_2} \in \Theta_{2,r_2}}} \widetilde{\mathcal{I}}(2:1, \mathcal{O}_1; \xi_N, \theta_{1,r_1}, \theta_{2,r_2}),$$

subject to

$$\sum_{i=1}^{R} \sum_{j=1}^{N} w_{ij} \left(y_{t_{ij}} - y_{1,r_1}(\theta_{1,r_1}, t_{ij}) \right)^2 < d \quad \text{for} \quad r_1 \in \{1, \ldots, M_1\},$$

$$\sum_{i=1}^{R} \sum_{j=1}^{N} w_{ij} \left(y_{t_{ij}} - y_{2,r_2}(\theta_{2,r_2}, t_{ij}) \right)^2 < d \quad \text{for} \quad r_2 \in \{1, \ldots, M_2\},$$

where w_{ij} are weights associated to the variance of the j-th measurement of measurement run i. The vectors $y_{t_{ij}}$ denote the measured values and $y_{j,r_j}(\theta_{j,r_j}, t_{ij})$ denote the responses of a model of the group of null hypothesis candidate models for $j = 2$ and of the group of alternative hypothesis candidate models for $j = 1$, respectively. The scalar d is a constant defining the allowed degree of lack of fit of the measurement data.
Beside conceptional improvements of the statistical framework, there are plenty possibilities to improve the algorithmic framework, as well. To take a single example, the homotopy approach to stabilize the *Outer Approximations* scheme can be further improved by implementing an effective step size control and continuation strategy.

APPENDIX A

Theoretical background

Definition 23 (Continuous function, Definition 5.1.18 in [91])**.** *Let $\mathcal{V}$ be a real normed space and let S be a convex subset of $\mathcal{V}$.*
(a) A function $f : \mathcal{V} \to \mathbb{R}^m$ is said to be continuous at a point $x \in \mathcal{V}$, if, for every $\delta > 0$, there exists a $\rho > 0$ such that

$$\|f(x') - f(x)\| < \delta, \forall x' \in \mathring{B}(x, \rho).$$

A function $f : \mathcal{V} \to \mathbb{R}^m$ is said to be continuous (continuous on S) if it is continuous at all $x \in \mathcal{V}$ ($x \in S$).
(b) A function $f : \mathcal{V} \to \mathbb{R}^m$ is said to be continuous relative to S (S-continuous), if for every $x \in S$ and for every $\delta > 0$, there exists a $\rho > 0$ such that

$$\|f(x') - f(x)\| < \delta, \forall x' \in \mathring{B}(x, \rho) \cap S.$$

(c) A function $f : \mathcal{V} \to \mathbb{R}$ is said to be upper semicontinuous (u.s.c.) at a point $x \in \mathcal{V}$, if, for every $\delta > 0$, there exists a $\rho > 0$ such that

$$f(x') - f(x) < \delta, \forall x' \in \mathring{B}(x, \rho).$$

(d) A function $f : \mathcal{V} \to \mathbb{R}$ is said to be u.s.c. (u.s.c. on S) if it is u.s.c. at all $x \in \mathcal{V}$ ($x \in S$).
(e) A function $f : \mathcal{V} \to \mathbb{R}$ is said to be upper semicontinuous relative to S ($S - $u.s.c.),

if for every $x \in S$ and for every $\delta > 0$, there exists a $\rho > 0$ such that

$$f(x') - f(x) < \delta, \forall x' \in \mathring{B}(x, \rho) \cap S.$$

(f) A function $f : \mathcal{V} \to \mathbb{R}$ is said to be lower semicontinuous (l.s.c.) at a point $x \in \mathcal{V}$, if, for every $\delta > 0$, there exists a $\rho > 0$ such that

$$f(x') - f(x) > -\delta, \forall x' \in \mathring{B}(x, \rho).$$

(g) A function $f : \mathcal{V} \to \mathbb{R}$ is said to be l.s.c. (l.s.c. on S) if it is l.s.c. at all $x \in \mathcal{V}$ ($x \in S$).
(h) A function $f : \mathcal{V} \to \mathbb{R}$ is said to be lower semicontinuous relative to S ($S - l.s.c.$), if for every $x \in S$ and for every $\delta > 0$, there exists a $\rho > 0$ such that

$$f(x') - f(x) > -\delta, \forall x' \in \mathring{B}(x, \rho) \cap S.$$

Proposition 7 (Proposition 5.1.19 in [91])**.** *Let $\mathcal{V}$ be a real normed space and let S be a convex subset of $\mathcal{V}$.*
(a) A function $f : \mathcal{V} \to \mathbb{R}^m$ is continuous at x^* if and only if, for any sequence $\{x_i\}_{i=0}^{\infty}$ in $\mathcal{V}$ such that $x_i \to x^*$, as $i \to \infty$, $f(x_i) \to f(x^*)$, as $i \to \infty$.
(b) A function $f : \mathcal{V} \to \mathbb{R}^m$ is continuous, relative to S, at $x^* \in S$ if and only if, for any sequence $\{x_i\}_{i=0}^{\infty}$ in S such that $x_i \to x^*$, as $i \to \infty$, $f(x_i) \to f(x^*)$, as $i \to \infty$.
(c) A function $f : \mathcal{V} \to \mathbb{R}$ is u.s.c. at x^* if and only if, for any sequence $\{x_i\}_{i=0}^{\infty}$ in $\mathcal{V}$ such that $x_i \to x^*$, as $i \to \infty$, $\overline{\lim} f(x_i) \leq f(x^*)$.
(d) A function $f : \mathcal{V} \to \mathbb{R}$ is u.s.c., relative to S, at $x^* \in S$ if and only if, for any sequence $\{x_i\}_{i=0}^{\infty}$ in S such that $x_i \to x^*$, as $i \to \infty$, $\overline{\lim} f(x_i) \leq f(x^*)$.
(e) A function $f : \mathcal{V} \to \mathbb{R}$ is l.s.c. at x^* if and only if, for any sequence $\{x_i\}_{i=0}^{\infty}$ in $\mathcal{V}$ such that $x_i \to x^*$, as $i \to \infty$, $\underline{\lim} f(x_i) \geq f(x^*)$.
(f) A function $f : \mathcal{V} \to \mathbb{R}$ is l.s.c., relative to S, at $x^* \in S$ if and only if, for any sequence $\{x_i\}_{i=0}^{\infty}$ in S such that $x_i \to x^*$, as $i \to \infty$, $\underline{\lim} f(x_i) \geq f(x^*)$.

For details we refer to [91].

Definition 24 (Directional derivative, Definition 5.1.30 in [91])**.** *Let $\mathcal{V}$ be a real normed space and suppose that $f : \mathcal{V} \to \mathbb{R}^m$.*
(a) We define the (one-sided) directional derivative of $f(\cdot)$ at a point $x \in \mathcal{V}$ in the direction $h \in \mathcal{V}$ by

$$\mathrm{d}f(x; h) := \lim_{t \downarrow 0} \frac{f(x + th) - f(x)}{t},$$

if this limit exists. Note that $t > 0$ is required.
(b) We say that $f(\cdot)$ is directional differentiable at a point $x^* \in \mathcal{V}$, if the directional derivative $\mathrm{d}f(x^*; h)$ exists for all $h \in \mathcal{V}$.

For details we refer to [91].

Definition 25 (Subgradient, Definition 5.1.31 in [91])**.** *Let $\mathcal{H}$ be a real Hilbert space with inner product $\langle \cdot, \cdot \rangle$.*
Suppose that $f : \mathcal{H} \to \mathbb{R}$ is such that the directional derivative $\mathrm{d}f(x, h)$ exists for all $x, h \in \mathcal{H}$. Then we define the subgradient $\partial f(x) \subset \mathcal{H}$ of $f(\cdot)$ at point $x \in \mathcal{H}$ by

$$\partial f(x) := \{\xi \in \mathcal{H} | \mathrm{d}f(x; h) \geq \langle \xi, h \rangle, \forall h \in \mathcal{H}\}.$$

For details we refer to [91].

Definition 26 (Definition 5.2.1 in [91])**.** *A set $S \subset \mathbb{R}^n$ is said to be convex if for any $x', x'' \in S$ and $\lambda \in [0, 1]$, $[\lambda x' + (1 - \lambda)x''] \in S$.*

For details we refer to [91].

Definition 27 (Definition 5.2.4 in [91])**.** *Let S be a subset of $\mathbb{R}^n$. We say that $\operatorname{conv} S$ is the convex hull of S, if it is the smallest convex set containing S.*

For details we refer to [91].

Definition 28 (Definition 5.2.6 in [91])**.** *Let S_1, S_2 be any two sets in $\mathbb{R}^n$. We say that the hyperplane*

$$H := \{x \in \mathbb{R}^n | \langle x, \nu \rangle = \alpha\},$$

separates S_1 and S_2 if

$$\langle x, \nu \rangle \geq \alpha, \quad \forall x \in S_1$$

and

$$\langle y, \nu \rangle \leq \alpha, \quad \forall y \in S_2.$$

The separation is said to be strict if there exists an $\epsilon > 0$ such that

$$\langle x, \nu \rangle \geq \alpha + \epsilon, \quad \forall x \in S_1$$

and

$$\langle y, \nu \rangle \leq \alpha - \epsilon, \quad \forall y \in S_2.$$

For details we refer to [91].

Theorem 19 (Separation of Convex Sets in $\mathbb{R}^n$, Theorem 5.2.7a in [91]). *Let S_1, S_2 be two nonempty convex sets in $\mathbb{R}^n$ such that $S_1 \cap S_2 = \emptyset$. Then there exists a hyperplane which separates S_1 and S_2. Furthermore, if S_1 and S_2 are closed and either S_1 or S_2 is compact, then the separation can be made strict.*

For details we refer to [91].

Definition 29 (Definition 5.2.8 in [91]). *Suppose that $S \subset \mathbb{R}^n$ is convex. We say that*

$$H = \{x | \langle x - \bar{x}, \nu \rangle = 0\},$$

is a support hyperplane to S through $\bar{x}$ with inward (outward) normal ν if $\bar{x} \in \bar{S}$ (where $\bar{S}$ is the closure of S) and

$$\langle x - \bar{x}, \nu \rangle \geq 0 \ (\leq 0), \ \forall x \in S.$$

For details we refer to [91].

Definition 30 (Definition 5.2.10 in [91]). *A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be convex if, for any $x', x'' \in \mathbb{R}^n$ and $\lambda \in (0, 1)$,*

$$f(\lambda x' + (1 - \lambda)x'') \leq \lambda f(x') + (1 - \lambda) f(x'').$$

A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be strictly convex if, for any $x', x'' \in \mathbb{R}^n$ and $\lambda \in (0, 1)$,

$$f(\lambda x' + (1 - \lambda)x'') < \lambda f(x') + (1 - \lambda) f(x'').$$

A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be concave (strictly concave) if $-f(\cdot)$ is convex (strictly convex).

For details we refer to [91].

Theorem 20 (Theorem 5.2.11 in [91]). *Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is convex. Then $f(\cdot)$ is continuous.*

A proof is given in [91].

Definition 31 (Definition 5.2.17 in [91]). *Let $S \subset \mathbb{R}^n$ be a convex set. We define the support function $\sigma_S : \mathbb{R}^n \to \mathbb{R}$ of S by*

$$\sigma_S(h) := \sup\{\langle h, x \rangle | x \in S\}.$$

Proposition 8 (Proposition 5.2.18 in [91]). *Let $\sigma_S(\cdot)$ be a support function for the convex set $S \subset \mathbb{R}^n$. Then,*
(a) $\sigma_S(\cdot)$ is positively homogeneous, i.e. for all $\lambda \geq 0$,

$$\sigma_S(\lambda h) = \lambda \sigma_S(h);$$

(b) $\sigma_S(\cdot)$ is subadditive, i.e. for all h_1, h_2,

$$\sigma_S(h_1 + h_2) \leq \sigma_S(h_1) + \sigma_S(h_2);$$

(c) $\sigma_S(\cdot)$ is convex; and
(d) if, in addition, S is bounded, then $\sigma_S(\cdot)$ is Lipschitz continuous.

For details we refer to [91].

Proposition 9 (Proposition 5.2.19 in [91]). *Let $S \subset \mathbb{R}^n$ be convex and compact. Suppose that, for a given $h \in \mathbb{R}^n$, $x_h \in S$ is such that $\sigma_S(h) = \langle h, x_h \rangle$. Then*

$$\langle x - x_h, h \rangle \leq 0, \quad \forall x \in S,$$

i.e., $\langle x, h \rangle = \langle x_h, h \rangle$ is a support hyperplane to S with outward normal h.

For details we refer to [91].

Proposition 10 (Proposition 5.2.20 in [91]). *Let $\sigma : \mathbb{R}^n \to \mathbb{R}$ be a Lipschitz continuous, positively homogeneous, subadditive function. Then the set*

$$C := \{x \in \mathbb{R}^n | \langle x, h \rangle \leq \sigma(h), \quad \forall h \in \mathbb{R}^n\},$$

is nonempty, convex, bounded, and closed, and $\sigma(\cdot)$ is the support function for C.

A proof is given in [91].

Proposition 11 (Proposition 5.2.21 in [91]). *Suppose that C and D are two convex and compact subsets of $\mathbb{R}^n$. Then $C \subset D$ if and only if $\sigma_C(h) \leq \sigma_D(h)$ for all $h \in \mathbb{R}^n$*

For details we refer to [91].

Definition 32 (Outer semicontinuous, Definition 5.3.1 in [91]). *A set-valued function (map) $f : \mathbb{R}^n \to 2^{\mathbb{R}^m}$ is said to be outer semicontinuous (o.s.c.) at $\hat{x}$, if $f(\hat{x})$ is closed and, for every compact set S such that $f(\hat{x}) \cap S = \emptyset$, there exists $\hat{\rho} > 0$ such that $f(x) \cap S = \emptyset$ for all $x \in B(\hat{x}, \hat{\rho})$.*
A set valued function $f : \mathbb{R}^n \to 2^{\mathbb{R}^m}$ is o.s.c. if it is o.s.c. at every $x \in \mathbb{R}^n$.

Definition 33 (Inner semicontinuous, Definition 5.3.2 in [91]). *A set-valued function (map) $f : \mathbb{R}^n \to 2^{\mathbb{R}^m}$ is said to be inner semicontinuous (i.s.c.) at $\hat{x}$, if for every open set G such that $f(\hat{x}) \cap G \neq \emptyset$, there exists $\hat{\rho} > 0$ such that $f(x) \cap G \neq \emptyset$ for all $x \in B(\hat{x}, \hat{\rho})$. A set valued function $f : \mathbb{R}^n \to 2^{\mathbb{R}^m}$ is i.s.c. if it is i.s.c. at every $x \in \mathbb{R}^n$.*

Definition 34 (Definition 5.3.3 in [91]). *A set-valued function $f : \mathbb{R}^n \to 2^{\mathbb{R}^m}$ is said to be continuous if it is both o.s.c. and i.s.c..*

For details we refer to [91].

Definition 35 (Definition 5.3.6 in [91]). *Consider a sequence of sets $\{A_i\}_{i=0}^{\infty}$ in $\mathbb{R}^n$.*
(a) The point $\hat{x}$ is said to be a limit point of $\{A_i\}_{i=0}^{\infty}$ if $d(\hat{x}, A_i) \to 0$ as $i \to \infty$, where

$$d(\hat{x}, A_i) := \inf\{\|x - \hat{x}\| | x \in A_i\},$$

i.e., if there exist $x_i \in A_i$ for all $i \in \mathbb{N}$, such that $x_i \to \hat{x}$, as $i \to \infty$.
(b) The point $\hat{x}$ is a cluster point of $\{A_i\}_{i=0}^{\infty}$ if it is a limit point of a subsequence of $\{A_i\}_{i=0}^{\infty}$.
(c) We denote the set of limit points of $\{A_i\}_{i=0}^{\infty}$ by $\underline{\text{Lim}} A_i$ and call it the inner limit, and we denote the set of cluster points of $\{A_i\}_{i=0}^{\infty}$ by $\overline{\text{Lim}} A_i$ and call it the outer limit.
(d) We will say that the sets A_i converge to the set A if $\underline{\text{Lim}} A_i = \overline{\text{Lim}} A_i = A$. which we denote either by $A_i \to A$, as $i \to \infty$, or by $\text{Lim}\, A_i = A$.

For details we refer to [91].

Theorem 21 (Theorem 5.3.7 in [91]). *(a) A function $f : \mathbb{R}^n \to 2^{\mathbb{R}^m}$ is o.s.c. at $\hat{x}$ if and only if , for any sequence $\{x_i\}_{i=0}^{\infty}$ such that $x_i \to \hat{x}$, as $i \to \infty$, $\overline{\text{Lim}} f(x_i) \subset f(\hat{x})$. Morever, $f(\cdot)$ is o.s.c. if and only if its graph $G(f) := \{(x, y) | y \in f(x)\}$ is closed.*
(b) Suppose that $f : \mathbb{R}^n \to 2^{\mathbb{R}^m}$ is such that $f(x)$ is compact for all $x \in \mathbb{R}^n$ and bounded on bounded sets. Then $f(\cdot)$ is o.s.c. at $\hat{x}$ if and only if, for every open set G such that $f(\hat{x}) \subset G$, there exists a $\hat{\rho} > 0$ such that $f(x) \subset G$ for all $x \in B(\hat{x}, \hat{\rho})$.
(c) A function $f : \mathbb{R}^n \to 2^{\mathbb{R}^m}$ is i.s.c. at $\hat{x}$ if and only if, for any sequence $\{x_i\}_i^{\infty}$ such that $x_i \to \hat{x}$, as $i \to \infty$, $\underline{\text{Lim}} f(x_i) \supset f(\hat{x})$.

A proof is given in [91].

Theorem 22 (Theorem 5.4.3 in [91]). *Consider the function*

$$\psi(x) := \max_{y \in Y(x)} \phi(x, y),$$

with $\phi : \mathbb{R}^n \times \mathbb{R}^m$ continuous and $Y : \mathbb{R}^n \to 2^{\mathbb{R}^m}$ continuous and compact-valued. Let

$$\hat{Y}(x) := \{y \in Y(x) | \psi(x) = \phi(x, y)\}.$$

Then $\hat{Y}(\cdot)$ is o.s.c and compact-valued. Furthermore, if $\hat{Y}(x) = \{\hat{y}(x)\}$, a singleton, then $\hat{y}(\cdot)$ is continuous at x.

A proof is given in [91].

Theorem 23 (Theorem 5.3.8 in [91])**.** *Suppose that $g : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ is continuous and that $Y : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ is o.s.c. and bounded on bounded sets. Then the set-valued function $G : \mathbb{R}^n \to 2^{\mathbb{R}^n}$, defined by*

$$G(x) = \operatorname{conv}\left(\bigcup_{y \in Y(x)} \{g(x,y)\}\right), \tag{A.1}$$

is o.s.c.. Furthermore, the map $G(\cdot)$ is bounded on bounded sets.

A proof is given in [91].

Corollary 5 (Corollary 5.3.9 in [91])**.** *Suppose that $g : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ is continuous and that $Y : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ is continuous. Then the set-valued function $G : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ defined by (A.1) is continuous.*

For details we refer to [91].

Theorem 24 (Corollary 5.4.6 in [91])**.** *Consider the function $\psi(x) = \max_{j \in \mathbf{q}} f^j(x)$, with $f^j : \mathbb{R}^n \to \mathbb{R}$, $j \in \mathbf{q} := \{1, \ldots, q\}$, continuously differentiable. Then,*
(a) The directional derivative $\mathrm{d}\psi(x;h)$ exists for all x, $h \in \mathbb{R}^n$ and is given by

$$\mathrm{d}\psi(x;h) = \max_{j \in \hat{\mathbf{q}}(x)} \langle \nabla f^j(x), h \rangle$$

where

$$\hat{\mathbf{q}}(x) := \{j \in \mathbf{q} | f^j(x) = \psi(x)\};$$

(b) the directional derivative $\mathrm{d}\psi(x;h)$ is upper semicontinuous, and, for every $x \in \mathbb{R}^n$, the function is positively homogeneous, subadditive, and Lipschitz continuous;
(c) the subgradient $\partial\psi(x)$ of $\psi(\cdot)$ at $x \in \mathbb{R}^n$ is given by

$$\partial\psi(x) = C := \operatorname{conv}\left(\bigcup_{j \in \hat{\mathbf{q}}(x)} \{\nabla f^j(x)\}\right),$$

and

$$\mathrm{d}\psi(x;h) = \max_{\xi \in \partial\psi(x)} \langle \xi, h \rangle.$$

Furthermore, $\partial\psi(x)$ is o.s.c..

A proof is given in [91].

Theorem 25 (Theorem 5.4.8 in [91])**.** *Consider the function $\psi : \mathbb{R}^n \to \mathbb{R}$ defined by*

$$\psi(x) := \max_{j \in \mathbf{q}} \psi^j(x),$$

where, for $j \in \mathbf{q} := \{1, \ldots, q\}$,

$$\psi^j(x) := \max_{y \in Y_j} \phi^j(x, y).$$

Suppose that, for all $j \in \mathbf{q}$,
(i) the functions $\phi^j : \mathbb{R}^n \times \mathbb{R}^{m_j} \to \mathbb{R}$ are continuous and the sets $Y_j \subset \mathbb{R}^{m_j}$ are compact;
(ii) the gradients $\nabla_x \phi^j(\cdot, \cdot)$ exist and are continuous.
Then,
(a) The directional derivative $\mathrm{d}\psi(x; h)$ exists for all $x, h \in \mathbb{R}^n$, and is given by

$$\mathrm{d}\psi(x; h) = \max_{j \in \hat{\mathbf{q}}(x)} \left(\max_{y \in \hat{Y}_j(x)} \langle \nabla_x \phi^j(x, y), h \rangle \right) = \max_{j \in \hat{\mathbf{q}}(x)} \mathrm{d}\psi^j(x; h),$$

where for $j \in \mathbf{q}$,

$$\hat{Y}_j(x) := \{y \in Y_j | \phi^j(x, y) = \psi^j(x)\},$$

and

$$\hat{\mathbf{q}}(x) := \{j \in \mathbf{q} | \psi^j(x) = \psi(x)\};$$

(b) The directional derivative $\mathrm{d}\psi(\cdot; \cdot)$ is upper semicontinuous, and for every $x \in \mathbb{R}^n$, $\mathrm{d}\psi(x; \cdot)$ is positively homogeneous, subadditive, and Lipschitz continuous;
(c) The subgradient $\partial\psi(x)$ is given by

$$\partial\psi(x) = C := \mathrm{conv} \left(\bigcup_{j \in \hat{\mathbf{q}}(x)} \partial\psi^j(x) \right) = \mathrm{conv} \left[\bigcup_{j \in \hat{\mathbf{q}}(x)} \mathrm{conv} \left(\bigcup_{y \in \hat{Y}_j(x)} \{\nabla_x \phi^j(x, y)\} \right) \right] \tag{A.2}$$

and

$$\mathrm{d}\psi(x; h) = \max_{\xi \in \partial\psi(x)} \langle \xi, h \rangle; \tag{A.3}$$

(d) The subgradient $\partial\psi(\cdot)$ is o.s.c.

Proof. (a) First we show for $x, h \in \mathbb{R}^n$ that

$$\left| \varliminf_{t \downarrow 0} \frac{\psi(x + th) - \psi(x)}{t} \right| < \infty$$

and

$$\left|\overline{\lim_{t\downarrow 0}}\,\frac{\psi(x+th)-\psi(x)}{t}\right|<\infty.$$

Since for $j\in\mathbf{q}$, $\phi^j(\cdot,y)$ is local Lipschitz continuous in x since the gradients $\nabla_x\phi^j(\cdot,y)$ exist and are continuous, and $\nabla_x\phi^j(x,\cdot)$ is continuous with respect to $y\in Y_j$ with Y_j compact, there exists a $\rho,L>0$ with

$$|\phi^j(x',y)-\phi^j(x'',y)|<L\|x'-x''\|,\quad \forall x',x''\in B(x,\rho),\ \forall y\in Y_j,\ \forall j\in\mathbf{q}.$$

Hence for $x',x''\in B(x,\rho)$,

$$\begin{aligned}\psi(x')-\psi(x'') &= \phi^{j'}(x',y')-\phi^{j''}(x'',y'')\\ &= \left[\phi^{j'}(x',y')-\phi^{j'}(x'',y')\right]+\underbrace{\left[\phi^{j'}(x'',y')-\phi^{j''}(x'',y'')\right]}_{\leq 0}\\ &\leq \phi^{j'}(x',y')-\phi^{j'}(x'',y')\leq L\|x'-x''\|,\end{aligned}$$

where $j'\in\hat{\mathbf{q}}(x')$ $j''\in\hat{\mathbf{q}}(x'')$ and $y'\in\hat{Y}_{j'}(x')$ $y''\in\hat{Y}_{j''}(x'')$. Interchanging x' and x'' above, we conclude that $\psi(\cdot)$ is local Lipschitz continuous. Hence, for $x,h\in\mathbb{R}$,

$$-L\|h\|\leq\underline{\lim_{t\downarrow 0}}\,\frac{\psi(x+th)-\psi(x)}{t}\leq\overline{\lim_{t\downarrow 0}}\,\frac{\psi(x+th)-\psi(x)}{t}\leq L\|h\|.$$

Next, since $\phi^j(x,y)\leq\psi(x)$ for all $j\in\mathbf{q}$, $y\in Y_j$, we obtain that for $t>0$,

$$\begin{aligned}\frac{\psi(x+th)-\psi(x)}{t} &= \max_{j\in\mathbf{q}}\left(\max_{y\in Y_j}\frac{\phi^j(x+th,y)-\psi(x)}{t}\right)\\ &= \max_{j\in\hat{\mathbf{q}}(x+th)}\left(\max_{y\in\hat{Y}_j(x+th)}\frac{\phi^j(x+th,y)-\psi(x)}{t}\right)\\ &\leq \max_{j\in\hat{\mathbf{q}}(x+th)}\left(\max_{y\in\hat{Y}_j(x+th)}\frac{\phi^j(x+th,y)-\phi^j(x,y)}{t}\right)\end{aligned}$$

Now the functions $g^j(t,y):=[\phi^j(x+th,y)-\phi^j(x,y)]/t$ are continuous, provided we define $g^j(0,y)=\mathrm{d}_x\phi(x,y;h)$, for all $j\in\mathbf{q}$. Since $\hat{Y}_j(\cdot)$ are *o.s.c* by Theorem 22, it follows from Theorem 6 that the max functions $\tilde{g}^j(t)$,

$$\tilde{g}^j(t):=\max_{y\in\hat{Y}_j(x+th)}g^j(t,y)=\begin{cases}\displaystyle\max_{y\in\hat{Y}_j(x+th)}\frac{\phi^j(x+th,y)-\phi^j(x,y)}{t}, & t>0,\\ \displaystyle\max_{y\in\hat{Y}_j(x)}\mathrm{d}_x\phi(x,y;h), & t=0,\end{cases}$$

for $h \in \mathbb{R}^n$ and $j \in \mathbf{q}$ are *u.s.c.*. Hence

$$\overline{\lim_{t\downarrow 0}} \frac{\psi(x+th)-\psi(x)}{t} \leq \max_{j\in\hat{\mathbf{q}}(x)} \left(\max_{y\in\hat{Y}_j(x)} \langle \nabla_x \phi^j(x,y), h\rangle \right). \tag{A.4}$$

Second,

$$\begin{aligned}\frac{\psi(x+th)-\psi(x)}{t} &= \max_{j\in\mathbf{q}} \left(\max_{y\in Y_j} \frac{\phi^j(x+th,y)-\psi(x)}{t} \right) \\ &\geq \max_{j\in\hat{\mathbf{q}}(x)} \left(\max_{y\in\hat{Y}_j(x)} \frac{\phi^j(x+th,y)-\phi^j(x,y)}{t} \right),\end{aligned}$$

because $\psi(x) = \phi^j(x,y)$ for all $j \in \hat{\mathbf{q}}(x)$ with $y \in \hat{Y}_j(x)$, and $j \in \hat{\mathbf{q}}(x) \subset \mathbf{q}$ with $y \in \hat{Y}_j(x) \subset Y$. Hence (by the same arguments as before), we must have that

$$\underline{\lim_{t\downarrow 0}} \frac{\psi(x+th)-\psi(x)}{t} \geq \max_{j\in\hat{\mathbf{q}}(x)} \left(\max_{y\in\hat{Y}_j(x)} \langle \nabla_x \phi^j(x,y), h\rangle \right). \tag{A.5}$$

Combining (A.4) and (A.5), we conclude that

$$\mathrm{d}\psi(x;h) = \lim_{t\downarrow 0} \frac{\psi(x+th)-\psi(x)}{t} = \max_{j\in\hat{\mathbf{q}}(x)} \left(\max_{y\in\hat{Y}_j(x)} \langle \nabla_x \phi^j(x,y), h\rangle \right).$$

(b) Since for $j \in \mathbf{q}$, $\hat{Y}^j(x)$ are *o.s.c* and bounded and $\langle \nabla \phi^j(x,y), h\rangle$ are continuous in (x,y,h), it follows from Theorem 6 that $\mathrm{d}\psi(\cdot;\cdot)$ is *u.s.c.*.
To establish that $\mathrm{d}\psi(x;\cdot)$ is Lipschitz continuous, we note that for any $h', h'' \in \mathbb{R}^n$,

$$\begin{aligned}\mathrm{d}\psi(x;h') - \mathrm{d}\psi(x;h'') &\leq \max_{j\in\hat{\mathbf{q}}(x)} \left(\max_{y\in\hat{Y}_j(x)} \langle \nabla_x \phi^j(x,y), h'-h''\rangle \right) \\ &\leq \max_{j\in\hat{\mathbf{q}}(x)} \left(\max_{y\in\hat{Y}_j(x)} \|\nabla_x \phi^j(x,y)\| \, \|h'-h''\| \right).\end{aligned}$$

Reversing h' and h'', we see that $\mathrm{d}\psi(x;.)$ is Lipschitz continuous with Lipschitz constant $L = \max_{j\in\hat{\mathbf{q}}(x)} \left(\max_{y\in\hat{Y}_j(x)} \|\nabla_x \phi^j(x,y)\| \right)$.
$\mathrm{d}\psi(x;\cdot)$ is subadditive and positively homogeneous, by inspection.
(c) By Definition 25, the subgradient $\partial\psi(x)$ is defined by

$$\partial\psi(x) := \{\xi \in \mathbb{R}^n | \langle \xi, h\rangle \leq \mathrm{d}\psi(x;h), \forall h \in \mathbb{R}^n\}.$$

Since $\mathrm{d}\psi(x;\cdot)$ is subadditive, positively homogeneous and Lipschitz continuous, the ex-

pression (A.3) now follows from Proposition 10. Next because

$$\max_{j\in\hat{\mathbf{q}}(x)}\left(\max_{y\in\hat{Y}_j(x)}\langle\nabla_x\phi^j(x,y),h\rangle\right)=\max_{\xi\in C}\langle\xi,h\rangle,$$

with C defined in (A.2), wee see that the equality in (A.2) follows from Proposition 11. (d) It follows directly from Theorem 23 that $\partial\psi(\cdot)$ is *o.s.c.*. □

Theorem 26 (Implicit Function Theorem, Theorem 5.1.33 in [91]). *Suppose that $\mathcal{V}$ is a real normed space and that $g:\mathbb{R}^l\times\mathcal{V}\to\mathbb{R}^l$ is $k\geq 1$ times continuously differentiable. If $x^*\in\mathbb{R}^l$ and $y^*\in\mathcal{V}$ are such that $g(x^*,y^*)=0$ and the Jacobian $g_x(x^*,y^*)$ is nonsingular, then there exist $\rho_x,\rho_y>0$ and a k-times continuously differentiable funtion $\Phi:B(y^*,\rho_y)\to B(x^*,\rho_x)$ such that $\Phi(y^*)=x^*$,*

$$\Phi_y(y^*)=-g_x(x^*,y^*)^{-1}g_y(x^*,y^*),$$

and

$$g(\Phi(y),y)=0,\quad\forall y\in B(y^*,\rho_y).$$

For details we refer to [91].

Corollary 6 (Corollary 5.1.34 in [91]). *Suppose that $g:\mathbb{R}^n\to\mathbb{R}^l$ is $k\geq 1$ times continuously differentiable and that $x^*\in\mathbb{R}^n$ is such that $g(x^*)=0$ and $g_x(x^*)$ has row rank l. Then, given any $h\neq 0$ in $\mathbb{R}^n$ such that $g_x(x^*)h=0$, there exists a $t_h>0$ and a k times continuously differentiable function $s:[0,t_h]\to\mathbb{R}^n$ such that (i) $s(0)=x^*$, (ii) $s'(0)=h$, and (iii) $g(s(t))=0$ for all $t\in[0,t_h]$.*

A proof is given in [91].

Theorem 27 (Chain Rule Theorem, Theorem 5.4.12 in [91]). *Suppose that $\psi(x)=\max_{j\in\mathbf{q}}f^j(x)$ with $f^j:\mathbb{R}^n\to\mathbb{R}$ continuously differentiable, or that*

$$\psi(x)=\max_{j\in\mathbf{q}}\max_{y\in Y_j}\phi^j(x,y),$$

with $\phi^j:\mathbb{R}^n\times\mathbb{R}^{m_j}\to\mathbb{R}$ continuous, $\nabla_x\phi^j(\cdot,\cdot)$ continuous, and $Y_j\subset\mathbb{R}^{m_j}$, $j\in\mathbf{q}$ compact. If $t'>0$ and $s:[0,t')\to\mathbb{R}^n$ is a continuously differentiable function such that $s(0)=\hat{x}$ and $\dot{s}(0)=h$, and the function $\sigma:[0,t')\to\mathbb{R}$ is defined by $\sigma(t):=\psi(s(t))$, then

$$\mathrm{d}\sigma(0;1)=\mathrm{d}\psi(\hat{x};h)$$

and

$$\partial\sigma(0) = \operatorname{conv}\left(\bigcup_{\xi\in\partial\psi(\hat{x})}\{\langle\xi, h\rangle\}\right). \tag{A.6}$$

A proof is given in [91].

Theorem 28 (von Neumann Theorem, bounded version, Corollary 5.5.6 in [91])**.** *Let $\phi : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ be a continuous function such that $\phi(\cdot, y)$ is convex for all $y \in \mathbb{R}^m$ and $\phi(x, \cdot)$ is concave for all $x \in \mathbb{R}^n$, and let Y be a compact, convex subset of $\mathbb{R}^m$. Suppose that $\phi(x, y) \to \infty$, as $\|x\| \to \infty$, uniformly in $y \in Y$. Then*

$$\min_{x\in\mathbb{R}^n}\max_{y\in Y}\phi(x, y) = \max_{y\in Y}\min_{x\in\mathbb{R}^n}\phi(x, y).$$

Morever there exist vectors $x_0 \in \mathbb{R}^n$ and $y_0 \in Y$ such that

$$\max_{y\in Y}\phi(x_0, y) = \phi(x_0, y_0) = \min_{x\in\mathbb{R}^n}\phi(x, y_0).$$

A proof is given in [91].

Bibliography

[1] J. Albersmeyer. *Adjoint-based algorithms and numerical methods for sensitivity generation and optimization of large scale dynamic systems*. PhD thesis, University of Heidelberg, Heidelberg, December 2010.

[2] J. Albersmeyer and H. G. Bock. Sensitivity generation in an adaptive BDF-method. In *Modeling, Simulation and Optimization of Complex Processes: Proceedings of the Third International Conference on High Performance Scientific Computing*. Springer, 2008.

[3] P. R. Amestoy, I. S. Duff, J. Koster, and J.-Y. L'Excellent. A fully asynchronous multifrontal solver using distributed dynamic scheduling. *SIAM Journal on Matrix Analysis and Applications*, 23(1):15–41, 2001.

[4] P. R. Amestoy, A. Guermouche, J.-Y. L'Excellent, and S. Pralet. Hybrid scheduling for the parallel solution of linear systems. *Parallel Computing*, 32(2):136–156, 2006.

[5] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen. *LAPACK Users' Guide*. Society for Industrial and Applied Mathematics, Philadelphia, PA, third edition, 1999.

[6] E. Anderson, Z. Bai, J. Dongarra, A. Greenbaum, A. McKenney, J. Du Croz, S. Hammerling, J. Demmel, C. Bischof, and D. Sorensen. Lapack: a portable linear algebra library for high-performance computers. In *Proceedings of the 1990 ACM/IEEE conference on Supercomputing*, Supercomputing '90, pages 2–11, Los Alamitos, CA, USA, 1990. IEEE Computer Society Press.

[7] J. F. Apgar, J. E. Toettcher, D. Endy, F. M. White, and B. Tidor. Stimulus design for model selection and validation in cell signaling. *PLoS Computational Biology*, 4(2):e30, 02 2008.

[8] J. S. Arora, O. A. Elwakeil, A. I. Chahande, and C. C. Hsieh. Global optimization methods for engineering applications: A review. *Structural and Multidisciplinary Optimization*, 9:137–159, 1995. 10.1007/BF01743964.

[9] U. Ascher and L. Petzold. *Computer methods for ordinary differential equations and differential-algebraic equations*. SIAM, Philadelphia, 1998.

[10] A. Atkinson and A. Donev. *Optimum Experimental Designs*. Oxford University Press, 1992.

[11] A. C. Atkinson and V. V. Fedorov. The design of experiments for discriminating between two rival models. *Biometrika*, 62(1):57–70, 1975.

[12] A. C. Atkinson and V. V. Fedorov. Optimal design: Experiments for discriminating between several models. *Biometrika*, 62(2):289–303, 1975.

[13] E. Balsa-Canto, A. A. Alonso, and J. R. Banga. Computational procedures for optimal experimental design in biological systems. *IET Systems Biology*, 2(4):163–172, July 2008.

[14] I. Bauer. *Numerische Verfahren zur Lösung von Anfangswertaufgaben und zur Generierung von ersten und zweiten Ableitungen mit Anwendungen bei Optimierungsaufgaben in Chemie und Verfahrenstechnik*. PhD thesis, University of Heidelberg, 1999.

[15] I. Bauer, H. G. Bock, S. Körkel, and J. P. Schlöder. Numerical methods for optimum experimental design in DAE systems. *Journal of Computational and Applied Mathematics*, 120:1–25, 2000.

[16] I. Bauer, H. G. Bock, and J. P. Schlöder. DAESOL – a BDF-code for the numerical solution of differential algebraic equations. Technical report, University of Heidelberg, 1998.

[17] V. Becker, M. Schilling, J. Bachmann, U. Baumann, A. Raue, T. Maiwald, J. Timmer, and U. Klingmuller. Covering a broad dynamic range: Information processing at the erythropoietin receptor. *Science*, 328(5984):1404–1408, 2010, http://www.sciencemag.org/cgi/reprint/328/5984/1404.pdf.

[18] B. M. Bell. Automatic differentiation software cppad., 2010.

[19] B. M. Bell and J. V. Burke. Algorithmic differentiation of implicit functions and optimal values. In C. H. Bischof, H. M. Bücker, P. D. Hovland, U. Naumann, and J. Utke, editors, *Advances in Automatic Differentiation*, pages 67–77. Springer, 2008.

[20] J. P. Bernacki and R. M. Murphy. Model discrimination and mechanistic interpretation of kinetic data in protein aggregation studies. *Biophysical Journal*, 96:2871–2887, 2009.

[21] M. Berz. Algorithms for higher order automatic differentiation in many variables with applications to beam physics, 1991.

[22] P. Billingsley. *Probability and Measure*. John Wiley & Sons Inc, 1986.

[23] G. Blanch. On modified divided differences I. *Mathematical Tables and Other Aids to Computation*, 8(45):pp. 1–11, 1954.

[24] G. Blanch. On modified divided differences II. *Mathematical Tables and Other Aids to Computation*, 8(46):pp. 67–75, 1954.

[25] H. G. Bock. Numerical treatment of inverse problems in chemical reaction kinetics. In K. H. Ebert, P. Deuflhard, and W. Jäger, editors, *Modelling of Chemical Reaction Systems*, volume 18 of *Springer Series in Chemical Physics*, pages 102–125. Springer, Heidelberg, 1981.

[26] H. G. Bock. Recent advances in parameter identification techniques for ODE. In P. Deuflhard and E. Hairer, editors, *Numerical Treatment of Inverse Problems in Differential and Integral Equations*, pages 95–121. Birkhäuser, Boston, 1983.

[27] H. G. Bock. Randwertproblemmethoden zur Parameteridentifizierung in Systemen nichtlinearer Differentialgleichungen. In *Bonner Mathematische Schriften*, volume 183. University of Bonn, 1987.

[28] H. G. Bock and K. J. Plitt. A multiple shooting algorithm for direct solution of optimal control problems. In *Proceedings of the Ninth IFAC World Congress, Budapest*. Pergamon, Oxford, 1984.

[29] P. N. Brown, G. D. Byrne, and A. C. Hindmarsh. Vode: A variable-coefficient ode solver. 10(5):1038–1051, 1989.

[30] K. P. Burnham and D. R. Anderson. *Model Selection and Multimodel inference: A practical information-theoretic approach*. Springer, 2002.

[31] R. H. Byrd, J. Nocedal, and R. A. Waltz. Knitro: An integrated package for nonlinear optimization. In *Large Scale Nonlinear Optimization, 35–59, 2006*, pages 35–59. Springer Verlag, 2006.

[32] G. D. Byrne and A. C. Hindmarsh. A polyalgorithm for the numerical solution of ordinary differential equations. *ACM Transactions on Mathematical Software*, 1(1):71–96, 1975.

[33] M. Calvo, T. Grande, and R. D. Grigorieff. On the zero stability of the variable order variable stepsize bdf-formulas. *Numerische Mathematik*, 57:39–50, 1990. 10.1007/BF01386395.

[34] M. Calvo, F. Lisbona, and J. Montijano. On the stability of variable-stepsize nordsieck bdf methods. *SIAM Journal on Numerical Analysis*, 24(4):pp. 844–854, 1987.

[35] M. Calvo, J. Montijano, and L. Rández. A0-stability of variable stepsize bdf methods. *Journal of Computational and Applied Mathematics*, 45(1-2):29–39, 1993.

[36] M. Calvo, J. I. Montijano, and L. Rández. On the change of step size in multistep codes. *Numerical Algorithms*, 4:283–304, 1993.

[37] H. Chernoff. Large-sample theory: Parametric case. *The Annals of Mathematical Statistics*, 27(1):pp. 1–22, 1956.

[38] B. Christianson. Reverse accumulation and accurate rounding error erstimates for taylor series coefficient. *Optimization Methods and Software*, 1(1):81–94, 1992.

[39] W. J. Cody and J. T. Coonen. Algorithm 722: Functions to support the ieee standard for binary floating-point arithmetic. *ACM Trans. Math. Softw.*, 19:443–451, December 1993.

[40] M. J. Cooney and K. A. McDonald. Optimal dynamic experiments for bioreactor model discrimination. *Applied Microbiology and Biotechnology*, 43:826–837, 1995.

[41] C. F. Curtiss and J. O. Hirschfelder. Integration of stiff equations. *Proceedings of the National Academy of Sciences of the United States of America*, 38:235–243, 1952.

[42] P. Deuflhard and F. Bornemann. *Numerische Mathematik II*. Walter de Gruyter, New York, second edition, 2002.

[43] V. Fedorov and P. Hackl. *Model-Oriented Design of Experiments*. Springer, 1997.

[44] W. F. Feehery, J. E. Tolsma, and P. I. Barton. Efficient sensitivity analysis of large-scale differential-algebraic systems. *Applied Numerical Mathematics*, 25(1):41–54, 1997.

[45] A. Fiacco and G. McCormick. *Nonlinear programming: sequential unconstrained minimization techniques*. Classics in applied mathematics. Society for Industrial and Applied Mathematics, 1990.

[46] R. Fletcher and S. Leyffer. Nonlinear programming without a penalty function. *Mathematical Programming*, 91:239–269, 2002.

[47] A. Forsgren, P. E. Gill, and M. H. Wright. Interior methods for nonlinear optimization. *SIAM Review*, 44(4):525–597, 2002.

[48] N. S. Foundation and D. of Energy. BLAS. http://www.netlib.org/blas/, 2010.

[49] C. Gear. *Asymptotic estimation of errors and derivatives for the numerical solution of ordinary differential equations*. Number Nr. 595-600 in Report (University of Illinois at Urbana-Champaign. Dept. of Computer Science). Dept. of Computer Science, University of Illinois at Urbana-Champaign, 1973.

[50] C. W. Gear and K. W. Tu. The effect of variable mesh size on the stability of multistep methods. *SIAM Journal on Numerical Analysis*, 11(5):pp. 1025–1043, 1974.

[51] J. C. Gilbert. Automatic differentiation and iterative processes. *Optimization Methods and Software*, 1(1):13–21, 1992.

[52] A. Goldbeter. *Biochemical oscillations and cellular rhythms: The molecular bases of periodic and chaotic behaviour*. Cambridge University Press, 1996.

[53] A. Griewank. *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*. Number 19 in Frontiers in Appl. Math. SIAM, Philadelphia, PA, 2000.

[54] A. Griewank, C. Bischof, G. Corliss, A. Carle, and K. Williamson. Derivative convergence for iterative equation solvers, 1993.

[55] A. Griewank, J. Utke, and A. Walther. Evaluating higher derivative tensors by forward propagation of univariate taylor series. *Mathematics of Computation*, 69(231):pp. 1117–1130, 2000.

[56] R. D. Grigorieff. Stability of multistep-methods on variable grids. *Numerische Mathematik*, 42:359–377, 1983. 10.1007/BF01389580.

[57] J. Großmann, Ch.; Terno. *Numerik der Optimierung*. B. G. Teubner, 1994.

[58] E. Hairer, S. P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations I: Nonstiff Problems*. Number 8 in Springer Series in Computational Mathematics. Springer, Berlin, second edition, 2000.

[59] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems*. Number 14 in Springer Series in Computational Mathematics. Springer, New York, first edition, 1991.

[60] R. Hettich and K. O. Kortanek. Semi-infinite programming: Theory, methods, and applications. *SIAM Review*, 35(3):pp. 380–429, 1993.

[61] N. J. Higham. *Accuracy and Stability of Numerical Algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, second edition, 2002.

[62] A. C. Hindmarsh, P. N. Brown, K. E. Grant, S. L. Lee, R. Serban, D. E. Shumaker, and C. S. Woodward. Sundials: Suite of nonlinear and differential/algebraic equation solvers. *ACM Trans. Math. Softw.*, 31:363–396, September 2005.

[63] R. Horn. Statistical methods for model discrimination. applications to gating kinetics and permeation of the acetylcholine receptor channel. *Biophysical Journal*, 51:255–263, 1987.

[64] HSL. A collection of fortran codes for large-scale scientific computation. See http://www.hsl.rl.ac.uk, 2007.

[65] W. G. Hunter and A. M. Reiner. Designs for discriminating between two rival models. *Technometrics*, 7(3):307–323, 1965.

[66] R. Jain, A. L. Knorr, J. Bernacki, and R. Srivastava. Investigation of bacteriophage ms2 viral dynamics using model discrimination analysis and the implications for phage therapy. *Biotechnology Progress*, 22(6):1650–1658, 2006.

[67] S. Körkel, I. Bauer, H. G. Bock, and J. P. Schlöder. A sequential approach for nonlinear optimum experimental design in DAE systems. In F. Keil, W. Mackens, H. Voss, , and J. Werther, editors, *Scientific Computing in Chemical Engineering II*, volume 2. Springer Verlag, Berlin, 1999.

[68] A. Kremling, S. Fischer, K. Gadkar, F. J. Doyle, T. Sauter, E. Bullinger, F. Allgöwer, and E. D. Gilles. A benchmark for methods in reverse engineering and model discrimination: problem formulation and solutions. *Genome Research*, 14(9):1773–1785, September 2004.

[69] C. Kreutz and J. Timmer. Systems biology: experimental design. *FEBS Journal*, 276:923–942, 2009.

[70] S. Kullback. *Information Theory and Statistics*. Dover Publications Inc., 1997.

[71] L. Lacey and A. Dunne. The design of pharmacokinetic experiments for model discrimination. *Journal of Pharmacokinetics and Pharmacodynamics*, 12:351–365, 1984.

[72] C. L. Lawson, R. J. Hanson, D. R. Kincaid, and F. T. Krogh. Basic linear algebra subprograms for fortran usage. *ACM Trans. Math. Softw.*, 5:308–323, September 1979.

[73] D. Lebiedz, M. Rehberg, and D. Skanda. Robust optimal design of synthetic biological networks. In W. Weber and M. Fussenegger, editors, *Synthetic Gene Networks*, volume 813 of *Methods in Molecular Biology*, pages 45–55. Humana Press, 2012. 10.1007/978-1-61779-412-4_3.

[74] D. B. Leineweber. Analyse und restrukturierung eines verfahrens zur direkten lösung von optimal-steuerungsproblemen. Diploma thesis, University of Heidelberg, 1995.

[75] J. C. Leloup and A. Goldbeter. A model for circadian rhythms in drosophila incorporating the formation of a complex between the per and tim proteins. *Journal of Biological Rhythms*, 13(1):70–87, Feb 1998.

[76] A. Levchenko and P. Iglesias. Models of eukaryotic gradient sensing: Application to chemotaxis of amoebae and neutrophils. *Biophysical Journal*, 82:50–63, 2002.

[77] J. López-Fidalgo, C. Tommasi, and P. C. Trandafir. An optimal experimental design criterion for discriminating between non-normal models. *Journal of the Royal Statistical Society Series B*, 69(2):231–242, 2007.

[78] R. M. M. Mattheij and J. Molenaar. *Ordinary differential equations in theory and practice*. SIAM, 2002.

[79] B. Melykuti, E. August, A. Papachristodoulou, and H. El-Samad. Discriminating between rival biochemical network models: three approaches to optimal experiment design. *BMC Systems Biology*, 4(1):38, 2010.

[80] H. D. Minh. *Numerical Methods for Simulation and Optimization of Chemically Reacting Flows in Catalytic Monoliths*. PhD thesis, University of Heidelberg, 2005.

[81] J. I. Myung and M. A. Pitt. Optimal experimental design for model discrimination. *Psychological review*, 116(3):499–518, July 2009.

[82] R. D. Neidinger. An efficient method for the numerical evaluation of partial derivatives of arbitrary order. *ACM Trans. Math. Softw.*, 18:159–173, June 1992.

[83] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer, New York, second edition, 2006.

[84] A. Nordsieck. On numerical integration of ordinary differential equations. *Mathematics of Computation*, 16:22–49, 1962.

[85] M. R. Osborne. On Nordsieck's method for the numerical solution of ordinary differential equations. *BIT Numerical Mathematics*, 6:51–57, 1966. 10.1007/BF01939549.

[86] L. R. Petzold. A description of DASSL: a differential/algebraic system solver. In *Scientific computing (Montreal, Quebec, 1982)*, pages 65–68. IMACS, New Brunswick, NJ, 1983.

[87] B. Philippe. Stabilité de la méthode des différentiations rétrogrades à pas et ordre variables (méthode de Gear). *C. R. Acad. Sc. Paris*, 294(13):435–437, 1982.

[88] E. Polak. On the convergence of optimization algorithms. *Rev. Française Informat. Recherche Opérationnelle*, 3(16):17–34, 1969.

[89] E. Polak. On the mathematical foundations of nondifferentiable optimization in engineering design. *SIAM Review*, 29(1):pp. 21–89, 1987.

[90] E. Polak. On the use of consistent approximations in the solution of semi-infinite optimization and optimal control problems. *Mathematical Programming*, 62:385–414, 1993. 10.1007/BF01585175.

[91] E. Polak. *Optimization: Algorithms and Consistent Approximations*. Springer, 1997.

[92] L. Pronzato and E. Walter. Robust experiment design via maximin optimization. *Mathematical Biosciences*, 89(2):161–176, 1988.

[93] V. Pérez, J. Renaud, and L. Watson. Homotopy curve tracking in approximate interior point optimization. *Optimization and Engineering*, 10:91–108, 2009. 10.1007/s11081-008-9042-6.

[94] K. Radhakrishnan and A. C. Hindmarsh. Description and use of LSODE, the Livermore Solver for Ordinary Differential Equations. Technical report, Lawrence Livermore National Laboratory, 1993.

[95] D. Salmon. Minimax controller design. *Automatic Control, IEEE Transactions on*, 13(4):369–376, 8 1968.

[96] O. Schenk, M. Bollhöfer, and R. A. Römer. On large-scale diagonalization techniques for the anderson model of localization. *SIAM Review*, 50(1):91–112, 2008.

[97] O. Schenk and K. Gärtner. Solving unsymmetric sparse systems of linear equations with pardiso. *Future Gener. Comput. Syst.*, 20:475–487, April 2004.

[98] O. Schenk and K. Gärtner. On fast factorization pivoting methods for symmetric indefinite systems. *Elec. Trans. Numer. Anal.*, 23:158–179, 2006.

[99] O. Schenk, A. Wächter, and M. Hagemann. Matching-based preprocessing algorithms to the solution of saddle-point problems in large-scale nonconvex interior-point optimization. *Comput. Optim. Appl.*, 36:321–341, April 2007.

[100] M. Schilling, T. Maiwald, S. Hengl, D. Winter, C. Kreutz, W. Kolch, W. D. Lehmann, J. Timmer, and U. Klingmuller. Theoretical and experimental analysis links isoform- specific erk signalling to cell fate decisions. *Molecular Systems Biology*, 5, December 2009.

[101] L. F. Shampine. Limiting precision in differential equation solvers. *Mathematics of Computation*, 28(125):pp. 141–144, 1974.

[102] L. F. Shampine and P. Bogacki. The effect of changing the stepsize in linear multistep codes. *SIAM J. Sci. Stat. Comput.*, 10:1010–1023, September 1989.

[103] K. Shimizu and E. Aiyoshi. Necessary conditions for min-max problems and algorithms by a relaxation procedure. *IEEE Transactions on Automatic Control*, 25(1):62–66, 1980.

[104] D. Skanda and D. Lebiedz. A robust optimization approach to experimental design for model discrimination of dynamical systems. *Mathematical Programming*, pages 1–29. 10.1007/s10107-012-0532-0.

[105] D. Skanda and D. Lebiedz. An optimal experimental design approach to model discrimination in dynamic biochemical systems. *Bioinformatics*, 26:939–45, 2010.

[106] D. Skanda and D. Lebiedz. A robust optimization approach to experimental design for model discrimination of dynamical systems. arXiv, Jan 2011, 1101.3663v1.

[107] D. Skanda and D. Lebiedz. A robust optimization approach to experimental design for model discrimination of dynamical systems. arXiv, Feb 2012, 1101.3663v2.

[108] R. D. Skeel. Equivalent forms of multistep formulas. *Mathematics of Computation*, 33(148):pp. 1229–1250, 1979.

[109] R. D. Skeel and L. W. Jackson. The Stability of Variable-Stepsize Nordsieck Methods. *SIAM Journal on Numerical Analysis*, 20(4):pp. 840–853, 1983.

[110] J. Stelling, E. D. Gilles, and F. J. Doyle. Robustness properties of circadian clock architectures. *Proceedings of the National Academy of Sciences of the United States of America*, 101(36):13210–13215, Sep 2004.

[111] H. J. Stetter. Asymptotic expansions for the error of discretization algorithms for non-linear functional equations. *Numerische Mathematik*, 7:18–31, 1965. 10.1007/BF01397970.

[112] W. E. Stewart, Y. Shon, and G. E. P. Box. Discrimination and goodness of fit of multiresponse mechanistic models. *AIChE Journal*, 44(6):1404–1412, 1998.

[113] J. Stoer and R. Bulirsch. *Introduction to Numerical Analysis*. Number 12 in Texts in Applied Mathematics. Springer, New York, third edition, 2002.

[114] C. Stricker, S. Redman, and D. Daley. Statistical analysis of synaptic transmission: model discrimination and confidence limits. *Biophysical Journal Of The Royal Statistical Society Series B*, 67:532–547, 1994.

[115] I. Swameye, T. G. Müller, J. Timmer, O. Sandra, and U. Klingmüller. Identification of nucleocytoplasmic cycling as a remote sensor in cellular signaling by data-based modeling. *PNAS*, 100(3):1028–1033, February 2003.

[116] R. Takors, W. Wiechert, and D. Weuster-Botz. Experimental design for the identification of macrokinetic models and model discrimination. *Biotechnol Bioeng*, 56(5):564–576, Dec 1997.

[117] J. Timmer, T. G. Müller, I. Swameye, O. Sandra, and U. Klingmüller. Modeling the nonlinear dynamics of cellular signal transduction. *International Journal of Bifurcation and Chaos*, 14(6):2069–2079, 2004.

[118] A. Tishler and I. Zang. A switching regression method using inequality conditions. *Journal of Econometrics*, 11(2-3):259 – 274, 1979.

[119] D. Uciński and B. Bogacka. T-optimum designs for multiresponse dynamic heteroscedastic models. In A. D. Bucchianico and H. Lauter, editors, *Proc. of the 7th International Workshop on Model-Oriented Design and Analysis*, pages 191–199. Physica Verlag, 2004.

[120] D. Uciński and B. Bogacka. T-optimum designs for discrimination between two multiresponse dynamic models. *Journal of the Royal Statistical Society Series B*, 67(1):3–18, 2005.

[121] A. Van Der Sluis. Condition numbers and equilibration of matrices. *Numerische Mathematik*, 14(1):14–23, 1969.

[122] R. J. Vanderbei. LOQO: An Interior Point Code For Quadratic Programming. Technical report, Optimization Methods and Software, 1998.

[123] M. C. Villalobos, M. C. Villalobos, R. A. Tapia, and Y. Zhang. The sphere of convergence of newton's method on two equivalent systems from nonlinear programming. Technical report, 1999.

[124] A. Wächter. *An Interior Point Algorithm for Large-Scale Nonlinear Optimization with Applications in Process Engineering*. PhD thesis, Carnegie Mellon University, 2002.

[125] A. Wächter and L. T. Biegler. Line search filter methods for nonlinear programming: Local convergence. *SIAM Journal on Optimization*, 16(1):32–48, 2005.

[126] A. Wächter and L. T. Biegler. Line search filter methods for nonlinear programming: Motivation and global convergence. *SIAM Journal on Optimization*, 16(1):1–31, 2005.

[127] A. Wächter and L. T. Biegler. On the implementation of a primal-dual interior point filter line search algorithm for large-scale nonlinear programming. *Mathematical Programming*, 106(1):25–57, 2006.

[128] R. Welsch, J. Arango, C. Bär, B. Salazar, S. Al-Babili, J. Beltrán, P. Chavarriaga, H. Ceballos, J. Tohme, and P. Beyer. Provitamin a accumulation in cassava (manihot esculenta) roots driven by a single nucleotide polymorphism in a phytoene synthase gene. *The Plant Cell Online*, 2010.

[129] R. Welsch, F. Wüst, C. Bär, S. Al-Babili, and P. Beyer. A third phytoene synthase is devoted to abiotic stress-induced abscisic acid formation in rice and defines functional diversification of phytoene synthase genes. *Plant Physiology*, 147(1):367–380, 2008, http://www.plantphysiol.org/content/147/1/367.full.pdf+html.

[130] R. C. Whaley and J. Dongarra. Automatically Tuned Linear Algebra Software. Technical Report UT-CS-97-366, University of Tennessee, December 1997. URL : `http://www.netlib.org/lapack/lawns/lawn131.ps`.

[131] R. C. Whaley, A. Petitet, and J. J. Dongarra. Automated empirical optimization of software and the ATLAS project. *Parallel Computing*, 27(1–2):3–35, 2001. Also available as University of Tennessee LAPACK Working Note #147, UT-CS-00-448, 2000 (`www.netlib.org/lapack/lawns/lawn147.ps`).

[132] I. Zang. Discontinuous optimization by smoothing. *Mathematics of Operations Research*, 6(1):pp. 140–152, 1981.