MATTHIAS GERDTS • FRANK LEMPIO

DISCRETE DYNAMIC OPTIMIZATION

Addresses of the Authors:

Frank Lempio

Lehrstuhl für Angewandte Mathematik
an der Universität Bayreuth

D-95440 Bayreuth

E-Mail: Frank.Lempio@uni-bayreuth.de

Matthias Gerdts

Mathematisches Institut
Universität Bayreuth

D-95440 Bayreuth

E-Mail: Matthias.Gerdts@uni-bayreuth.de

Preliminary Version: August 15, 2003

Contents

Chapter 1

Introduction

1.1 What's Discrete Dynamic Optimization ?

The **dynamical behaviour** of the **state** of many technical, economical, and biological problems can be described by (discrete) dynamic equations. For instance we might be interested in the development of the population size of a specific species during a certain time period, or we want to describe the dynamical behaviour of chemical processes or mechanical systems or the development of the profit of a company during the next five years, say.

Usually, the dynamical behaviour of a given system can be influenced by the choice of certain **control variables**. For instance, the breeding of rabbits can be influenced by the incorporation of diseases or natural predators. A car can be controlled by the steering wheel, the accelerator pedal, and the brakes. A chemical process can be controlled, e.g., by increasing or decreasing the temperature. The profit of a company is influenced, e.g., by the prices of its products or the number of employees.

Very often, the state variables and/or the control variables cannot assume any value, but are subject to certain **restrictions**. These restrictions may result from certain safety regulations or physical limitations, e.g. the temperature in a nuclear reactor has to be lower than a specific threshold or the altitude of an airplane should be larger than ground level or the steering angle of a car is limited by a maximum steering angle.

In addition, we are particularly interested in those state and control variables which fulfill all restrictions and furthermore minimize or maximize a given **objective function**. For example, the objective of a company is

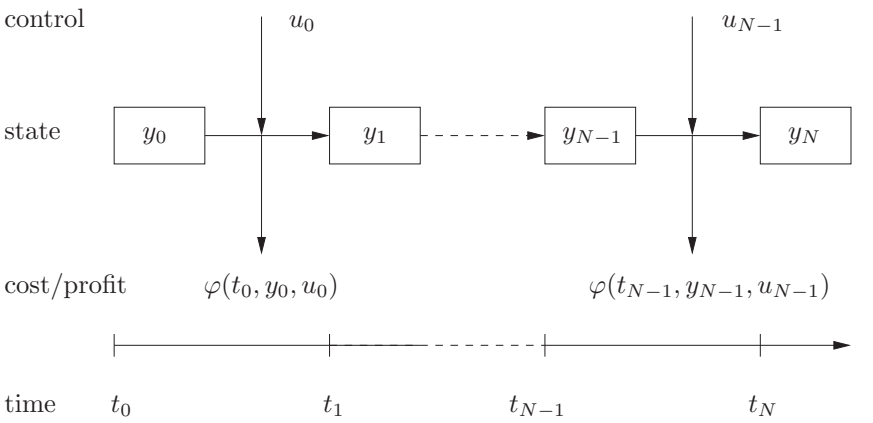


Figure 1.1: Schematic representation of a discrete dynamic optimization problem.

to maximize the profit or to minimize operational costs.

Summarizing, we have the following ingredients for a discrete dynamic optimization problem, cf. Figures 1.1, 1.2:

- The state variables $y(t_j)$ describe the state of a process at certain time points t_j.
- The control variables $u(t_j)$ allow to influence the dynamical behaviour of the state variables at time point t_j.
- The discrete dynamic equation $y(t_{j+1}) = \psi(t_j, y(t_j), u(t_j))$ describes the transition of the state from time point t_j to the subsequent time point t_{j+1} dependent on the current state $y(t_j)$, the control variable $u(t_j)$, and the time point t_j.
- The objective function has to be minimized (or maximized).
- The restrictions on the state and/or control variables have to be fulfilled.

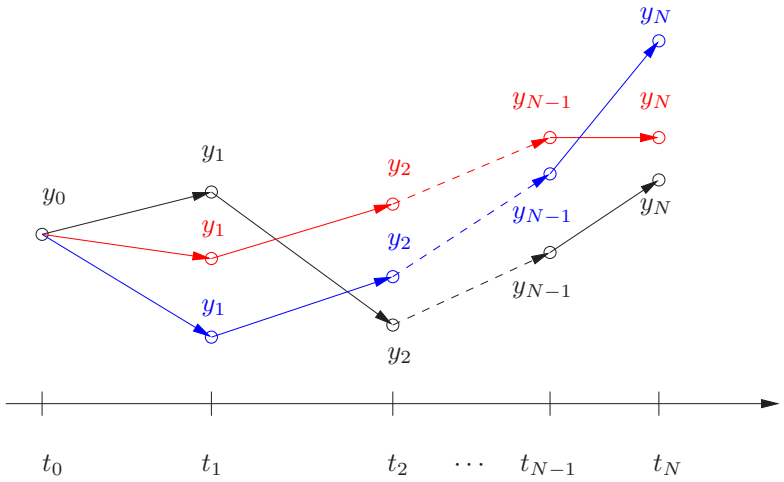


Figure 1.2: Examples of discrete trajectories for different control functions.

This yields the following mathematical formulation of a discrete dynamic optimization problem.

Let

$$\mathbb{G} := \{t_j \mid j = 0, 1, \ldots, N\}$$

denote a **grid** with $N + 1$ fixed (time) points

$$t_0 < t_1 < \ldots < t_N.$$

The task is to determine a **state grid function**

$$y : \mathbb{G} \to \mathbb{R}^n, \qquad t_j \mapsto y(t_j),$$

and a **control grid function**

$$u : \mathbb{G} \to \mathbb{R}^m, \qquad t_j \mapsto u(t_j),$$

such that the objective function

$$f(y, u) := \sum_{j=0}^{N} \varphi(t_j, y(t_j), u(t_j)),$$

with

$$\varphi : \mathbb{G} \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$$

is minimized subject to the **dynamic equations**

$$y(t_{j+1}) = \psi(t_j, y(t_j), u(t_j)), \qquad j = 0, 1, \ldots, N-1,$$

where

$$\psi : \mathbb{G} \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n.$$

In addition, the **state constraints**

$$y(t_j) \in Y(t_j), \qquad j = 0, 1, \ldots, N,$$

with nonempty sets $Y(t_j) \subseteq \mathbb{R}^n$, and the **control constraints**

$$u(t_j) \in U(t_j, y(t_j)), \qquad j = 0, 1, \ldots, N,$$

with nonempty sets $U(t_j, y) \subseteq \mathbb{R}^m$ for $y(t_j) \in Y(t_j)$ and $j = 0, 1, \ldots, N$ have to be fulfilled.

Hence, we arrive at the following

1.1.1. Discrete Dynamic Optimization Problem (DOP).

$$\begin{aligned}
&\textit{Minimize} && \sum_{j=0}^{N} \varphi(t_j, y(t_j), u(t_j)) \\
&\textit{w. r. t.} && y \in \{y \mid y : \mathbb{G} \to \mathbb{R}^n\},\ u \in \{u \mid u : \mathbb{G} \to \mathbb{R}^m\} \\
&\textit{subject to} && y(t_{j+1}) = \psi(t_j, y(t_j), u(t_j)), \quad j = 0, 1, \ldots, N-1, \\
& && y(t_j) \in Y(t_j), \qquad j = 0, 1, \ldots, N, \\
& && u(t_j) \in U(t_j, y(t_j)), \qquad j = 0, 1, \ldots, N.
\end{aligned}$$

■

Remarks :

- Without loss of generality we consider minimization problems. Maximization problems can be easily transformed into equivalent minimization problems by replacing the objective function by its negative.

- Very often, the sets $Y(t_j)$ are expressed in terms of inequality and equality constraints, that is

$$Y(t_j) = \{y \in \mathbb{R}^n \mid g(t_j, y) \leq \Theta,\ h(t_j, y) = \Theta\}.$$

  Similarly, the sets $U(t_j, y)$ often are given by

$$U(t_j, y) = \{u \in \mathbb{R}^m \mid \tilde{g}(t_j, y, u) \leq \Theta,\ \tilde{h}(t_j, y, u) = \Theta\}.$$

  More specifically, the control may be restricted by box constraints only:

$$u(t_j) \in \{v = (v_1, \ldots, v_m)^\top \in \mathbb{R}^m \mid a_j \leq v_j \leq b_j,\ j = 1, \ldots, m\}.$$

1.2 Examples and Applications

We will discuss several areas of applications leading to discrete dynamic optimization problems.

1.2.1 Discretized Optimal Control Problems

We will see that direct discretization methods for infinite dimensional optimal control problems lead to finite dimensional discrete dynamic optimization problems.

Let the following mappings be given:

$$\begin{aligned}
\varphi_a &: \mathbb{R}^n \to \mathbb{R}, \\
\varphi_b &: \mathbb{R}^n \to \mathbb{R}, \\
\varphi &: \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}, \\
\psi &: \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n, \\
\alpha &: \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^{s_a}, \\
\beta &: \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^{s_b}, \\
S &: \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^s
\end{aligned}$$

Let

$$U \subseteq \mathbb{R}^m$$

denote the **control region**.

Then, an optimal control problem on the time interval $[a,b]$, $a < b$ is given by

$$
\begin{aligned}
&\text{Minimize} && \varphi_a(y(a)) + \varphi_b(y(b)) + \int_a^b \varphi(t, y(t), u(t))dt \\
&\text{w. r. t.} && y(\cdot) \in AC([a,b])^n,\ u(\cdot) \in L_\infty([a,b])^m \\
&\text{subject to} && \dot{y}(t) = \psi(t, y(t), u(t)) \qquad \text{for a. a. } t \in [a,b]\ , \\
& && \alpha_i(a, y(a)) \begin{cases} \leq 0, & \text{for } i = 1, \ldots, s'_a\ , \\ = 0, & \text{for } i = s'_a + 1, \ldots, s_a\ , \end{cases} \\
& && \beta_i(b, y(b)) \begin{cases} \leq 0, & \text{for } i = 1, \ldots, s'_b\ , \\ = 0, & \text{for } i = s'_b + 1, \ldots, s_b\ , \end{cases} \\
& && u(t) \in U \qquad \text{for a. a. } t \in [a,b]\ , \\
& && S(t, y(t)) \leq \Theta \qquad \text{for every } t \in [a,b].
\end{aligned}
$$

Direct solution methods for the optimal control problem are based on the discretization of the integral, the differential equation, and the constraints on a chosen grid

$$\mathbb{G} := \{t_i \mid i = 0, 1, \ldots, N\}$$

with

$$a = t_0 < t_1 < \ldots < t_N = b$$

and step sizes $h_j := t_{j+1} - t_j$ for $j = 0, \ldots, N-1$. The differential equation is discretized on $\mathbb{G}$ by a suitable discretization method, e.g. Euler's method or some higher order Runge-Kutta method. For Euler's method we get

$$y_{j+1} = y_j + h_j \psi(t_j, y_j, u_j), \qquad j = 0, 1, \ldots, N-1,$$

where $y_j \approx y(t_j)$, $j = 0, 1, \ldots, N$ and $u_j \approx u(t_j)$, $j = 0, 1, \ldots, N-1$ are approximations of the state function y and the control function u, respec-

tively, on the grid $\mathbb{G}$. The integral in the objective function is approximated by

$$\int_a^b \varphi(t, y(t), u(t))dt \approx \sum_{j=0}^{N-1} h_j \varphi(t_j, y_j, u_j).$$

Finally, the infinitely many state constraints $S(t, y(t)) \leq \Theta$ are replaced by finitely many constraints

$$S(t_j, y_j) \leq \Theta, \qquad j = 0, 1, \ldots, N.$$

Summarizing, we obtain the discretized optimal control problem

$$\begin{aligned}
&\text{Minimize} && \varphi_a(y_0) + \varphi_b(y_N) + \sum_{j=0}^{N-1} h_j \varphi(t_j, y_j, u_j) \\
&\text{w. r. t.} && y_j,\ j = 0, 1, \ldots, N,\ u_j,\ j = 0, 1, \ldots, N-1 \\
&\text{subject to} && y_{j+1} = y_j + h_j \psi(t_j, y_j, u_j), \qquad j = 0, 1, \ldots, N-1, \\
& && \alpha_i(t_0, y_0) \begin{cases} \leq 0, & \text{for } i = 1, \ldots, s'_a\ , \\ = 0, & \text{for } i = s'_a + 1, \ldots, s_a\ , \end{cases} \\
& && \beta_i(t_N, y_N) \begin{cases} \leq 0, & \text{for } i = 1, \ldots, s'_b\ , \\ = 0, & \text{for } i = s'_b + 1, \ldots, s_b\ , \end{cases} \\
& && u_j \in U, \qquad j = 0, 1, \ldots, N-1, \\
& && S(t_j, y_j) \leq \Theta, \qquad j = 0, 1, \ldots, N\ .
\end{aligned}$$

1.2.2 Inventory Management

A company has to determine a minimum cost inventory plan for a fixed number of time periods $t_0 < t_1 < \ldots < t_N$. A product is stored during

Figure 1.3: Inventory Management

these time periods $t_0 < t_1 < \ldots < t_N$ in a store. Let $u_j \geq 0$ be a delivery at time point t_j, $r_j \geq 0$ the demand in $[t_j, t_{j+1})$, and y_j the amount of stored products at time point t_j (just before delivery). Then the balance equations

$$y_{j+1} = y_j + u_j - r_j, \qquad j = 0, 1, \ldots, N-1$$

hold. In addition, we postulate that the demand can be satisfied always, that is $y_{j+1} \geq 0$ for all $j = 0, 1, \ldots, N-1$. Without loss of generality, at the beginning and at the end of the time period the stock level is zero, that is $y_0 = y_N = 0$. The delivery costs at time point t_j are modelled by

$$B(u_j) = \begin{cases} K + cu_j, & \text{if } u_j > 0, \\ 0, & \text{if } u_j = 0, \end{cases}$$

where K denotes the fixed costs and c is the cost per unit. The inventory costs become effective at the end of each time period and are given by hy_{j+1}, $j = 0, 1, \ldots, N-1$. Thus, the total costs are

$$\sum_{j=0}^{N-1} (K\delta(u_j) + cu_j + hy_{j+1}),$$

where

$$\delta(u_j) = \begin{cases} 1, & \text{if } u_j > 0\ , \\ 0, & \text{if } u_j = 0\ . \end{cases}$$

It holds

$$\begin{aligned}\sum_{j=0}^{N-1} u_j &= \sum_{j=0}^{N-1}(y_{j+1}-y_j+r_j) = (y_N-y_0)+\sum_{j=0}^{N-1} r_j = \sum_{j=0}^{N-1} r_j\ ,\\ \sum_{j=0}^{N-1} y_{j+1} &= \sum_{j=0}^{N-1}(y_j+u_j-r_j) = \sum_{j=0}^{N-1} y_j + \sum_{j=0}^{N-1} u_j - \sum_{j=0}^{N-1} r_j = \sum_{j=0}^{N-1} y_j\ .\end{aligned}$$

Taking these relations into account, we can formulate the following inventory problem:

$$\begin{aligned}&\text{Minimize} && \sum_{j=1}^{N} K\delta(u_j)+hy_j && \\ &\text{subject to} && y_{j+1}=y_j+u_j-r_j, && j=0,\dots,N-1,\\ & && y_0=y_N=0, && \\ & && y_j\geq 0, && j=0,\dots,N,\\ & && u_j\geq 0, && j=0,\dots,N.\end{aligned}$$

Example 1.2.1

There are three different methods at different costs available in a production line within one time period. The costs are given by the following table:

Method	I	II	III
stock	0	15	30
costs	0	500	800

The fixed costs for a positive production volume amount to 300 dollar for each time period, the inventory costs are 15 dollar for each item and time period. The demand for each time period is given by 25 items. The task is to determine an optimal production plan for the next three time periods, if the initial inventory is 35 items and the final inventory is restricted to 20 items. It is assumed that the inventory costs are valid at the beginning of each time period. ∎

1.2.3 Knapsack Problem

There is one knapsack together with N items. Item i has weight a_j and value c_j for $j = 1, \ldots, N$. The task is to create a knapsack with maximal value under the restriction that the maximal weight is less or equal A. This leads to the optimization problem:

Maximize

$$\sum_{j=1}^{N} c_j u_j$$

s. t.

$$\sum_{j=1}^{N} a_j u_j \leq A,\ u_j \in \{0,1\},\ j = 1, \ldots, N,$$

where

$$u_j = \begin{cases} 1, & \text{item } j \text{ is put into the knapsack,} \\ 0, & \text{item } j \text{ is not put into the knapsack.} \end{cases}$$

This is a linear optimization problem with integer variables. It can be transformed into an equivalent discrete dynamic optimization problem. Let y_j denote the remaining weight after items $i = 1, \ldots, j-1$ have been included

(or not). Then, the discrete dynamic optimization problem arises:

$$
\begin{aligned}
\textit{Maximize} \quad & \sum_{j=1}^{N} c_j u_j \\
\textit{subject to} \quad & y_{j+1} = y_j - a_j u_j, && j = 1, \ldots, N, \\
& y_1 = A, \\
& 0 \leq y_j, && j = 1, \ldots, N, \\
& u_j \begin{cases} \in \{0,1\}, & \text{if } y_j \geq a_j, \\ = 0, & \text{if } y_j < a_j, \end{cases} && j = 1, \ldots, N.
\end{aligned}
$$

y_{N+1} is the remaining space in the knapsack.

Example 1.2.2

A manager has to choose co-workers for a project. There are four eligible co-workers to choose from. Each of them has a number assigned that indicates the capability of the respective co-worker. The respective numbers are $3, 5, 2$, and 4. The costs for the co-workers are $30, 50, 20$, and 40 thousand dollar, respectively. The manager has a maximal amount of 90 thousand dollars available for the personal expenditure. Which co-workers should the manager choose for the project?

The problem is a knapsack problem. The co-workers 1-4 correspond to $N = 4$ items, which can be put into the knapsack. The personal expenditure $a_j, j = 1, \ldots, N$ corresponds to the weight of the respective item, whereas the capability of the respective co-worker corresponds to the value c_j of item j. The maximal amount A is the maximal weight of the knapsack. The task is to create a knapsack with maximum value subject to the weight restriction.

1.2.4 Assignment Problems

A number A of resources has to be assigned to N projects. Assigning u_j resources to project j results in the profit $\varphi_j(u_j)$. The goal is to maximize the total profit $\sum_{j=1}^{N} \varphi_j(u_j)$. This leads to the optimization problem:

Maximize

$$\sum_{j=1}^{N} \varphi_j(u_j)$$

s. t.

$$\sum_{j=1}^{N} u_j \leq A,\ u_j \geq 0,\ j = 1, \ldots, N.$$

This problem can be transformed into a discrete dynamic optimization problem. For that purpose, we introduce the state y_j, which denotes the remaining resources after assigning resources to the projects $i = 1, \ldots, j-1$. Then the above optimization problem is equivalent with

$$\begin{aligned} \text{Maximize} \quad & \sum_{j=1}^{N} \varphi_j(u_j) \\ \text{subject to} \quad & y_{j+1} = y_j - u_j, && j = 1, \ldots, N, \\ & u_j \in U_j(y_j) = \{0, 1, \ldots, y_j\}, && j = 1, \ldots, N, \\ & y_1 = A. \end{aligned}$$

Example 1.2.3
A company has to assign four sales representatives to four sales regions A,B,C,D. The achievable sales volume depends on the number of assigned representatives according to the following table.

Region/Number	0	1	2	3	4
A	0	25	48	81	90
B	0	35	48	53	65
C	0	41	60	75	92
D	0	52	70	85	95

The company intends to maximize the total sales volume. How does the optimal assignment look like?

1.2.5 Reliability Problems

Let us assume that a computer works if and only if three components A,B and C work properly. In order to increase the reliability of the computer system it is possible to add certain emergency systems to each component.

It costs 100 dollars to add such an emergency system to the first component, 300 dollars for the second component, and 200 dollars for the third component. Furthermore, it is assumed, that at most two emergency systems for each component can be added. The probability, that a component works properly, depends on the number of emergency systems added to the component according to the following table.

Number/System	A	B	C
0	0.85	0.60	0.70
1	0.90	0.85	0.90
2	0.95	0.95	0.98

We are looking for a configuration, that maximizes the reliability of the computer subject to the additional restriction that at most 600 dollars can be spent for additional emergency systems.

Formulation as a dynamic optimization problem:
The components A,B,C are denoted by 1,2,3. Let y_j be the amount remaining after emergency systems have been added to the components $1, \ldots, j-1$.

Let u_j be the number of emergency systems for component j subject to the restriction $u_j \in U = \{0, 1, 2\}$.

Let $p_j(u_j)$ be the probability that component j works properly, if u_j emergency systems have been added. c_j are the costs to add an emergency system to component j, that is $c_1 = 100, c_2 = 300, c_3 = 200$.

Thus, we obtain

$$\textit{Maximize} \quad \prod_{j=1}^{3} p_j(u_j)$$

subject to

$$\begin{aligned}
y_{j+1} &= y_j - c_j \cdot u_j, \quad j = 1, 2, 3 \ , \\
y_1 &= 600 \ , \\
u_j &\in U_j(y_j) = \begin{cases} \{0\}, & \text{if } y_j < c_j \ , \\ \{0, 1\}, & \text{if } c_j \le y_j < 2c_j \ , \\ \{0, 1, 2\}, & \text{if } 2c_j \le y_j \ , \end{cases} \\
y_{j+1} &\in \{0, 100, 200, 300, 400, 500, 600\} \ .
\end{aligned}$$

Unfortunately, the objective function is a product and not a sum. But, we can simply transform the above problem by applying the logarithm to the objective function:

$$\textit{Maximize} \quad \sum_{j=1}^{3} \log p_j(u_j)$$

subject to

$$\begin{aligned} y_{j+1} &= y_j - c_j \cdot u_j, \quad j = 1,2,3 \ , \\ y_1 &= 600 \ , \\ u_j &\in U_j(y_j) = \begin{cases} \{0\}, & \text{if } y_j < c_j \ , \\ \{0,1\}, & \text{if } c_j \le y_j < 2c_j \ , \\ \{0,1,2\}, & \text{if } 2c_j \le y_j \ , \end{cases} \\ y_{j+1} &\in \{0,100,200,300,400,500,600\} \ . \end{aligned}$$

1.3 Overview on Solution Methods

1.3.1 Indirect Solution Methods

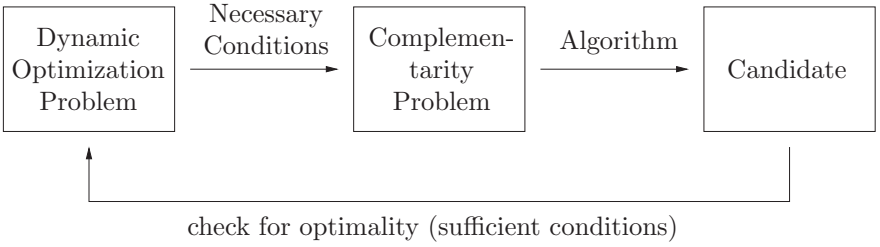


Figure 1.4: Solution process based on indirect methods.

An indirect solution method is based on the evaluation of necessary conditions for the discrete dynamic optimization problem. Necessary conditions are conditions which have to be fulfilled in an optimal solution. We will see later that it is possible to derive necessary conditions in form of a discrete maximum principle for our discrete dynamic optimization problem. These

conditions are based on the well-known Fritz-John necessary conditions for finite dimensional optimization problems.

Under certain circumstances it is possible to solve the resulting conditions numerically or analytically, e. g. by solving a linear or nonlinear system of equations or more generally a complementarity problem. In general, the resulting solutions are only candidates for an optimal solution.

1.3.2 Direct Solution Methods

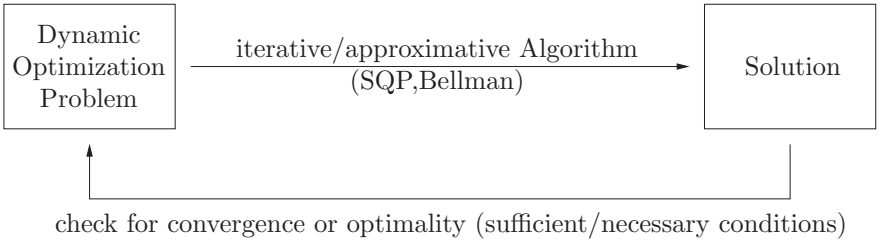


Figure 1.5: Solution process based on direct methods.

Direct solution methods try to solve the discrete dynamic optimization problem directly without solving the necessary conditions explicitly. Very often, direct methods are based on an iterative procedure generating approximations to the optimal solution of the dynamic optimization problem within each iteration step. For instance, the sequential quadratic programming (SQP) method uses quadratic subproblems to approximate a general nonlinear programming problem locally. The important question is whether these iterative respectively approximative algorithms converge to a solution of the original problem or not.

Chapter 2

Dynamic Programming

In this chapter, we discuss one of the earliest methods for the solution of the Discrete Dynamic Optimization Problem 1.1.1 (DOP), the so-called **Dynamic Programming Method**. This method is based on Richard Bellman's **Optimality Principle**.

For a more detailed discussion we refer to the monographs [4], [5], [7], [47]. Many applications and examples can be found in [62].

2.1 Bellman's Optimality Principle

Let $t_k \in \{t_0, t_1, \ldots, t_N\}$ be a fixed time point, $\mathbb{G}_k := \{t_j \mid j = k, k+1, \ldots, N\}$, and $\hat{y} \in Y(t_k)$ an admissible state. Consider the

Discrete Dynamic Optimization Problem $(P(t_k, \hat{y}))$:

$$
\begin{array}{lll}
\textit{Minimize} & \displaystyle\sum_{j=k}^{N} \varphi(t_j, y(t_j), u(t_j)) & \\
\textit{w. r. t.} & y \in \{y \mid y : \mathbb{G}_k \to \mathbb{R}^n\},\ u \in \{u \mid u : \mathbb{G}_k \to \mathbb{R}^m\} & \\
\textit{subject to} & y(t_{j+1}) = \psi(t_j, y(t_j), u(t_j)), & j = k, 1, \ldots, N-1, \\
& y(t_k) = \hat{y}, & \\
& y(t_j) \in Y(t_j), & j = k, k+1, \ldots, N, \\
& u(t_j) \in U(t_j, y(t_j)), & j = k, k+1, \ldots, N.
\end{array}
$$

2.1.1. Definition (Optimal Value Function). *Let $t_k \in \mathbb{G}$. For $\hat{y} \in Y(t_k)$ let $V(t_k, \hat{y})$ denote the optimal value of the problem $(P(t_k, \hat{y}))$. For $\hat{y} \notin Y(t_k)$ we set $V(t_k, \hat{y}) = \infty$. The function $V : \mathbb{G} \times \mathbb{R}^n \to \mathbb{R}$, $(t_k, \hat{y}) \mapsto V(t_k, \hat{y})$ is called* **optimal value function**.

2.1.2. Theorem (Bellman's Optimality Principle). *Let $\hat{y}(\cdot)$ and $\hat{u}(\cdot)$ be an optimal solution of (DOP). Then $\hat{y}|_{\mathbb{G}_k}$ and $\hat{u}|_{\mathbb{G}_k}$ is an optimal solution of $(P(t_k, \hat{y}(t_k)))$.*

Proof. Assume, that $\hat{y}|_{\mathbb{G}_k}$ and $\hat{u}|_{\mathbb{G}_k}$ are not optimal for $(P(t_k, \hat{y}(t_k)))$. Then there exist feasible trajectories $\tilde{y} : \mathbb{G}_k \to \mathbb{R}^n$ and $\tilde{u} : \mathbb{G}_k \to \mathbb{R}^m$ for $(P(t_k, \hat{y}(t_k)))$ with

$$\sum_{j=k}^{N} \varphi(t_j, \tilde{y}(t_j), \tilde{u}(t_j)) < \sum_{j=k}^{N} \varphi(t_j, \hat{y}(t_j), \hat{u}(t_j))$$

and $\tilde{y}(t_k) = \hat{y}(t_k)$. Hence, the trajectories $y : \mathbb{G} \to \mathbb{R}^n$ and $u : \mathbb{G} \to \mathbb{R}^m$

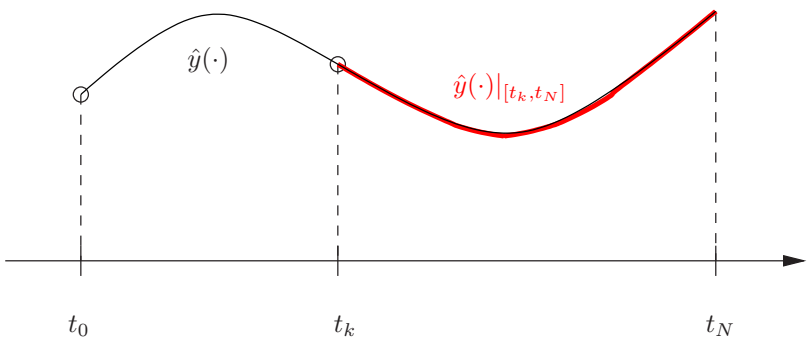


Figure 2.1: Bellman's optimality principle: Rest trajectories of optimal trajectories remain optimal.

with

$$y(t_j) := \begin{cases} \hat{y}(t_j), & \text{for } j = 0, 1, \ldots, k-1, \\ \tilde{y}(t_j), & \text{for } j = k, k+1, \ldots, N, \end{cases}$$

$$u(t_j) := \begin{cases} \hat{u}(t_j), & \text{for } j = 0, 1, \ldots, k-1, \\ \tilde{u}(t_j), & \text{for } j = k, k+1, \ldots, N, \end{cases}$$

are feasible for (DOP) and satisfy

$$\sum_{j=0}^{k-1} \varphi(t_j, \hat{y}(t_j), \hat{u}(t_j)) + \sum_{j=k}^{N} \varphi(t_j, \tilde{y}(t_j), \tilde{u}(t_j)) < \sum_{j=0}^{N} \varphi(t_j, \hat{y}(t_j), \hat{u}(t_j)).$$

This contradicts the optimality of $\hat{y}(\cdot)$ and $\hat{u}(\cdot)$. ∎

The optimality principle states: The decisions in the periods $k, k+1, \ldots, N$ of the Problem (DOP) for a given state y_k are independent of the decisions in the periods $t_0, t_1, \ldots, t_{k-1}$, compare Figure 2.1.

For the validity of the optimality principle it is essential that the discrete dynamic optimization problem can be divided into stages, e.g. the state y

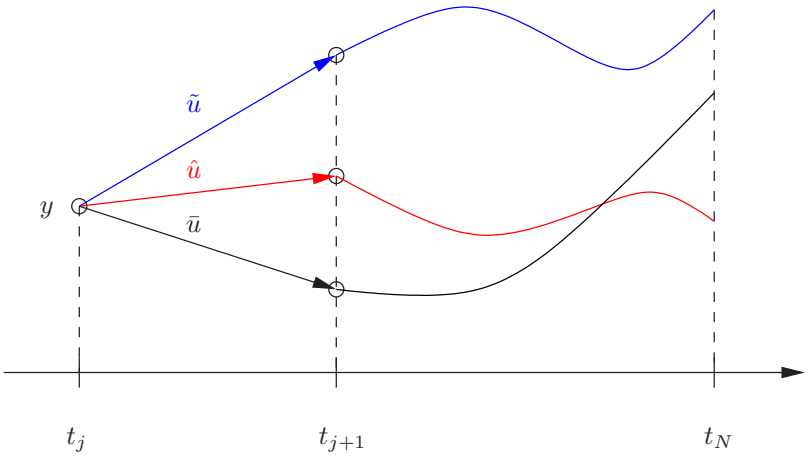


Figure 2.2: Bellman's dynamic programming method: Recursion of the optimal value function.

at t_{j+1} only depends on the values of y and u at the previous stage t_j and not on the respective values at, e.g., t_0 and t_N. Similarly, the objective function is separable and the constraints only restrict y and u at t_j. This allows to apply a stepwise optimization procedure.

2.2 Dynamic Programming Method

The optimality principle allows to derive a recursion for the optimal value function. In the sequel, we use the convention $\varphi(t_j, y, u) = \infty$, if $y \notin Y(t_j)$.

We exploit the fact, that the optimal value for $(P(t_N, y))$ is given by

$$V(t_N, y) = \min_{u \in U(t_N, y)} \varphi(t_N, y, u). \tag{2.2.1}$$

Suppose now, that we already know the optimal value function $V(t_{j+1}, y)$ for any $y \in \mathbb{R}^n$. From the optimality principle we obtain

$$V(t_j, y) = \min_{u \in U(t_j, y)} \left\{ \varphi(t_j, y, u) + V(t_{j+1}, \psi(t_j, y, u)) \right\}, \quad j = 0, 1, \ldots, N-1. \tag{2.2.2}$$

Equations (2.2.1) and (2.2.2) enable us to compute the optimal value function backward in time starting at t_N. The optimal initial state $\hat{y}(t_0)$ of (DOP) is given by

$$\hat{y}(t_0) = \arg \min_{y \in Y(t_0)} V(t_0, y). \tag{2.2.3}$$

Equations (2.2.1)-(2.2.3) form the basis for

Bellman's Dynamic Programming Method I :

(i) **Backward computation**

1. Let $V(t_N, y)$ be given by (2.2.1).
2. For $j = N-1, \ldots, 0$: Calculate $V(t_j, y)$ as in (2.2.2).

(ii) **Forward computation**

1. Let $\hat{y}(t_0)$ be given by (2.2.3).
2. For $j = 0, 1, \ldots, N-1$: Determine

$$\hat{u}(t_j) = \arg \min_{u \in U(t_j, \hat{y}(t_j))} \{\varphi(t_j, \hat{y}(t_j), u) + V(t_{j+1}, \psi(t_j, \hat{y}(t_j), u))\}$$

and set $\hat{y}(t_{j+1}) = \psi(t_j, \hat{y}(t_j), \hat{u}(t_j))$.

3. Determine $\hat{u}(t_N) = \arg \min_{u \in U(t_N, \hat{y}(t_N))} \varphi(t_N, \hat{y}(t_N), u)$.

Bellman's Dynamic Programming Method II :

(i) **Backward computation**

1. Let $V(t_N, y)$ be given by (2.2.1) and $u^*(t_N, y)$ the corresponding optimal control.
2. For $j = N-1, \ldots, 0$: Calculate $V(t_j, y)$ as in (2.2.2).
   Let $u^*(t_j, y)$ denote the optimal control at t_j for y (feedback control).

(ii) **Forward computation**

1. Let $\hat{y}(t_0)$ be given by (2.2.3).
2. For $j = 0, 1, \ldots, N-1$: Determine $\hat{u}(t_j) = u^*(t_j, \hat{y}(t_j))$ and
$$\hat{y}(t_{j+1}) = \psi(t_j, \hat{y}(t_j), \hat{u}(t_j)).$$
3. Determine $\hat{u}(t_N) = u^*(t_N, \hat{y}(t_N))$.

Remark. Both versions of Bellman's dynamic programming method yield an optimal solution of the discrete dynamic programming problem (DOP). Version II is preferable for hand calculations because it provides an optimal feedback control u^* as a function of time and state. Version I is more convenient for computer implementations, since it does not require to store the feedback control u^* for each t_j and y and thus saves memory space. It only computes the optimal trajectories for y and u as functions of time.

Example 2.2.1
Solve the discrete dynamic optimization problem

$$\begin{aligned} \textit{Minimize} \quad & -\sum_{j=0}^{N-1} c(1-u_j)y(j) \\ \textit{s. t.} \quad & y(j+1) = y(j)(0.9+0.6u(j)), \quad j=0,1,\ldots,N-1 \\ & y(0) = k>0, \\ & 0 \le u(j) \le 1, \quad j=0,1,\ldots,N-1 \end{aligned}$$

for $k>0, c>0, b=0.6$ and $N=5$ by the dynamic programming method.

Since $k>0$ and $u(j)\ge 0$ it holds $y(j)>0$ for all j. In the sequel we use the abbreviation $y_j := y(j)$.

The recursive equations (2.2.1) and (2.2.2) for $N=5$ are given by $V(5,y_5)=0$ and

$$V(j,y_j) = \min_{0\le u_j\le 1}\{-cy_j(1-u_j)+V(j+1,y_j(0.9+0.6u_j))\}, \quad 0\le j\le N-1.$$

Evaluation of the recursion and observation of $c > 0, y_j > 0$ yield

$$
\begin{aligned}
V(4,y_4) &= \min_{0\le u_4\le 1}\{-cy_4(1-u_4) + \underbrace{V(5, y_4(0.9+0.6u_4)}_{=0}\} = -cy_4, \qquad \hat{u}_4 = 0,\\
V(3,y_3) &= \min_{0\le u_3\le 1}\{-cy_3(1-u_3) + V(4,(y_3(0.9+0.6u_3))\}\\
&= \min_{0\le u_3\le 1}\{-cy_3(1-u_3) - cy_3(0.9+0.6u_3)\}\\
&= cy_3 \min_{0\le u_3\le 1}\{-1.9+0.4u_3\} = -1.9cy_3, \qquad \hat{u}_3 = 0,\\
V(2,y_2) &= \min_{0\le u_2\le 1}\{-cy_2(1-u_2) + V(3, y_2(0.9+0.6u_2)\}\\
&= \min_{0\le u_2\le 1}\{-cy_2(1-u_2) - 1.9cy_2(0.9+0.6u_2)\}\\
&= cy_2 \min_{0\le u_2\le 1}\{-2.71-0.14u_2\} = -2.85cy_2, \qquad \hat{u}_2 = 1,\\
V(1,y_1) &= \min_{0\le u_1\le 1}\{-cy_1(1-u_1) + V(2, y_1(0.9+0.6u_1)\}\\
&= \min_{0\le u_1\le 1}\{-cy_1(1-u_1) - 2.85cy_1(0.9+0.6u_1)\}\\
&= cy_1 \min_{0\le u_1\le 1}\{-3.565-0.71u_1\} = -4.275cy_1, \qquad \hat{u}_1 = 1,\\
V(0,y_0) &= \min_{0\le u_0\le 1}\{-cy_0(1-u_0) + V(1, y_0(0.9+0.6u_0)\}\\
&= \min_{0\le u_0\le 1}\{-cy_0(1-u_0) - 4.275cy_0(0.9+0.6u_0)\}\\
&= cy_0 \min_{0\le u_0\le 1}\{-4.8475-1.565u_0\} = -6.4125cy_0, \qquad \hat{u}_0 = 1,
\end{aligned}
$$

Hence, the optimal control is $\hat{u}_0 = \hat{u}_1 = \hat{u}_2 = 1, \hat{u}_3 = \hat{u}_4 = 0$. Forward evaluation leads to $\hat{y}_0 = k, \hat{y}_1 = 1.5 \cdot k, \hat{y}_2 = 2.25 \cdot k, \hat{y}_3 = 3.375 \cdot k, \hat{y}_4 = 3.0375 \cdot k, \hat{y}_5 = 2.73375 \cdot k$. The optimal objective value is $-c\hat{y}_3 - c\hat{y}_4 = -c(\hat{y}_3 + \hat{y}_4) = -6.4125 \cdot c \cdot k$.

2.3 Implementation

We discuss an implementable algorithm for the special discrete dynamic optimization problem

$$
\begin{aligned}
&\textit{Minimize} \quad \sum_{j=0}^{N} \varphi(t_j, y_j, u_j) \\
&\textit{s. t.} \\
&\qquad y_{j+1} = \psi(t_j, y_j, u_j) \ , \qquad j = 0, \ldots, N-1, \\
&\qquad y_0 = y_a \ , \\
&\qquad y_j \in [y_l, y_u] \ , \qquad j = 1, \ldots, N \ , \\
&\qquad u_j \in [u_l(t_j, y_j), u_u(t_j, y_j)] \ , \quad j = 0, \ldots, N.
\end{aligned}
$$

The interval $[y_l, y_u]$ is divided into M equidistant sections of length $h = (y_u - y_l)/M$:

$$Y = \{y_l + i \cdot h \mid i = 0, \ldots, M\}.$$

Y denotes the feasible region for the state variables. Similarly, the control region $[u_l(t_j, y_j), u_u(t_j, y_j)]$ is divided into M_j equidistant sections of length $h_j = (u_u(t_j, y_j) - u_l(t_j, y_j))/M_j$:

$$U(t_j, y_j) = \{u_l(t_j, y_j) + i \cdot h_j \mid i = 0, \ldots, M_j\}, \quad j = 0, \ldots, N.$$

$U(t_j, y_j)$ denotes the feasible region for the control variables in step j.

Algorithm: Bellman's Dynamic Programming Method

(i) **Backward computation**

1. For all $y \in Y$ let

$$V(t_N, y) = \min_{u \in U(t_N, y)} \varphi(t_N, y, u).$$

2. For $j = N-1, \ldots, 0$: For all $y \in Y$ determine

$$V(t_j, y) = \min_{\substack{u \in U(t_j, y) \\ \psi(t_j, y, u) \in [y_l, y_u]}} \{\varphi(t_j, y, u) + V(t_{j+1}, \psi(t_j, y, u))\}. \tag{2.3.1}$$

(ii) **Forward computation**

1. Let $\hat{y}_1 = y_a$.
2. For $j = 0, 1, \ldots, N-1$: Determine

$$\hat{u}_j = \arg \min_{\substack{u \in U(t_j, \hat{y}_j) \\ \psi(t_j, \hat{y}_j, u) \in [y_l, y_u]}} \{\varphi(t_j, \hat{y}_j, u) + V(t_{j+1}, \psi(t_j, \hat{y}_j, u))\} \tag{2.3.2}$$

and set $\hat{y}_{j+1} = \psi(t_j, \hat{y}_j, \hat{u}_j)$.

3. Determine $\hat{u}_N = \arg\min_{u \in U(t_N, \hat{y}_N)} \varphi(t_N, \hat{y}_N, u)$.

Remark. The evaluation of (2.3.1) and (2.3.2), respectively, requires the values $V(t_{j+1}, y_{j+1})$ with $y_{j+1} = \psi(t_j, y, u) \in [y_l, y_u]$. It may happen, that y_{j+1} is not a grid point in Y such that $\bar{y} := y_l + i \cdot h < y_{j+1} < y_l + (i+1) \cdot h = \bar{y} + h$ holds for some index i. Then, the value of the optimal value function at y_{j+1} is determined by linear interpolation of the values $V(t_{j+1}, \bar{y})$ and $V(t_{j+1}, \bar{y} + h)$:

$$V(t_{j+1}, y_{j+1}) \approx V(t_{j+1}, \bar{y}) + \frac{y_{j+1} - \bar{y}}{h} \left(V(t_{j+1}, \bar{y} + h) - V(t_{j+1}, \bar{y})\right).$$

Example 2.3.2 **(Test example)**
Knapsack problem ($M = 1000$, $M_j = 1$, $j = 1, \ldots, N$)

No.	Item	weight	value
1.	knapsack	1400 g	1.00
2.	tent	2600 g	0.88
3.	camping mat	1200 g	0.92
4.	sleeping bag	1500 g	0.94
5.	stove	1600 g	0.79
6.	4 soups	each 30 g	0.79
7.	bottle of water	1150 g	0.98
8.	clothes	800 g	0.71
9.	wash bag	300 g	0.74
10.	towel	350 g	0.81
11.	handy	550 g	0.5
12.	wallet	500 g	0.99
13.	pencils	300 g	0.52
14.	map of trails	80 g	0.98
15.	city guide	200 g	0.58
16.	bar of chocolate	100 g	0.98

The maximum weight is $A = 10$ $[kg]$.

Remark. The main drawback of Bellman's dynamic programming method is the so called 'curse of dimension'. As it can be seen from formula (2.3.1) the method requires to compute and to store the values $V(t_j, y)$ for each $j = N, N-1, \ldots, 0$ and each $y \in Y$. Depending on the value N this can become a really huge number. In the worst case each discrete trajectory emanating from y_a has to be considered. Nevertheless, for certain applications, e.g. assignment problems, knapsack problems, and inventory problems with integral data, the dynamic programming method performs quite well.

```
      PROGRAM BELLMAN
      IMPLICIT NONE
      INTEGER N,NDIS
      DOUBLEPRECISION XL,XU,XA
      PARAMETER (N=16,NDIS=1000,XL=0.0D0,XU=10000.0D0)
      INTEGER M(0:N),I

      DOUBLEPRECISION V(0:NDIS+1,0:N),U(0:N),X(0:N),C(N+1),W(N+1),OBJ
      COMMON /NUTZEN/ C
      COMMON /GEWICHT/ W

      C(1) = -1.0D0
      C(2) = -0.88D0
      C(3) = -0.92D0
      C(4) = -0.94D0
      C(5) = -0.79D0
      C(6) = -0.79D0
      C(7) = -0.98D0
      C(8) = -0.71D0
      C(9) = -0.74D0
      C(10)= -0.81D0
      C(11)= -0.5D0
      C(12)= -0.99D0
      C(13)= -0.52D0
      C(14)= -0.98D0
      C(15)= -0.58D0
      C(16)= -0.98D0
      C(17)=  0.0D0

      W(1) = 1400.0D0
      W(2) = 2600.0D0
      W(3) = 1200.0D0
      W(4) = 1500.0D0
      W(5) = 1600.0D0
      W(6) = 120.0D0
      W(7) = 1150.0D0
      W(8) = 800.0D0
      W(9) = 300.0D0
      W(10)= 350.0D0
      W(11)= 550.0D0
      W(12)= 500.0D0
      W(13)= 300.0D0
      W(14)= 80.0D0
      W(15)= 200.0D0
      W(16)= 100.0D0
      W(17)= 0.0D0
      XA = 10000.0D0
      DO 10,I=0,N
         M(I)=1
 10   CONTINUE
      CALL BFG(N,NDIS,XA,XL,XU,M,V,U,X,OBJ)
      END


      SUBROUTINE BFG(N,NDIS,XA,XL,XU,M,V,U,X,OBJ)
      IMPLICIT NONE
      INTEGER N,NDIS,M(0:N)
      DOUBLEPRECISION XA,XL,XU,V(0:NDIS+1,0:N),U(0:N),X(0:N),OBJ
C
      INTEGER I,J,K,L,IRUN,IGIT,NDISTMP
      DOUBLEPRECISION H,HU,XHELP,UHELP,RESG,RESF,INFTY,UL,UU,TAU,
     +    VINT,VHELP,XLTMP,TMP
      LOGICAL B
C     INIT
      H       = (XU-XL)/DBLE(NDIS)
      INFTY   = 1.0D+20
      NDISTMP = NDIS
      XLTMP   = XL
      DO 10,I=0,NDIS
         V(I,N)=INFTY
         XHELP=XLTMP+DBLE(I)*H
         CALL UBOUNDS(N,XHELP,UL,UU)
         HU=(UU-UL)/DBLE(M(N))
         DO 15,J=0,M(N)
            UHELP=UL+DBLE(J)*HU
            CALL PHI(N,XHELP,UHELP,RESG)
            IF (RESG.LT.V(I,N)) V(I,N)=RESG
 15      CONTINUE
 10   CONTINUE
C
C     BACKWARD
      DO 100,J=N-1,0,-1
         B=.FALSE.
         DO 110,I=0,NDISTMP
            V(I,J)=INFTY
            XHELP=XLTMP+DBLE(I)*H
            CALL UBOUNDS(J,XHELP,UL,UU)
            HU=(UU-UL)/DBLE(M(J))
            DO 120,K=0,M(J)
               UHELP=UL+DBLE(K)*HU
               CALL PHI(J,XHELP,UHELP,RESG)
               CALL PSI(J,XHELP,UHELP,RESF)
               IF (RESF.GE.XLTMP .AND. RESF.LE.XU) THEN
                  TAU = (RESF-XLTMP)/H
                  IGIT= INT(TAU)
                  IF (IGIT.EQ.NDIS) THEN
                     VINT=V(IGIT,J+1)
                  ELSE
                     VINT= V(IGIT,J+1)+(TAU-DBLE(IGIT))*
     +                    (V(IGIT+1,J+1)-V(IGIT,J+1))
                  ENDIF
                  IF (RESG+VINT.LT.V(I,J)) V(I,J)=RESG+VINT
               ENDIF
               IF (V(I,J).LT.INFTY) B=.TRUE.
 120        CONTINUE
 110     CONTINUE
         IF (.NOT.B) GOTO 1000
         IF (J.EQ.1) THEN
            NDISTMP=0
            XLTMP=XA
         ENDIF
 100  CONTINUE
C
C     FORWARD
      NDISTMP=NDIS
      XLTMP=XL
      X(0)=XA
      OBJ=0.0D0
      DO 200,J=0,N-1
         VHELP=INFTY
         V(NDISTMP+1,J+1)=0.0D0
         IGIT=INT((X(J)-XLTMP)/H)
         CALL UBOUNDS(J,X(J),UL,UU)
         HU=(UU-UL)/M(J)
         DO 210,K=0,M(J)
            UHELP=UL+DBLE(K)*HU
            CALL PHI(J,X(J),UHELP,RESG)
            CALL PSI(J,X(J),UHELP,RESF)
            IF (RESF.GE.XLTMP .AND. RESF.LE.XU) THEN
               TAU = (RESF-XLTMP)/H
               IGIT= INT(TAU)
               IF (IGIT.EQ.NDIS) THEN
                  VINT=V(IGIT,J+1)
               ELSE
                  VINT= V(IGIT,J+1)+(TAU-DBLE(IGIT))*
     +                 (V(IGIT+1,J+1)-V(IGIT,J+1))
               ENDIF
               IF (RESG+VINT.LT.VHELP) THEN
                  VHELP=RESG+VINT
                  U(J)=UHELP
               ENDIF
            ENDIF
 210     CONTINUE
         IF (VHELP.EQ.INFTY) GOTO 2000
         CALL PHI(J,X(J),U(J),RESG)
         CALL PSI(J,X(J),U(J),RESF)
         X(J+1)=RESF
         OBJ=OBJ+RESG
 200  CONTINUE
      CALL UBOUNDS(N,X(N),UL,UU)
      HU=(UU-UL)/DBLE(M(N))
      TMP=INFTY
      DO 215,J=0,M(N)
         UHELP=UL+DBLE(J)*HU
         CALL PHI(N,X(N),UHELP,RESG)
         IF (RESG.LT.TMP) THEN
            TMP=RESG
            U(N)=UHELP
         ENDIF
 215  CONTINUE
      OBJ=OBJ+TMP

      PRINT*,'SOLUTION FOUND:  '
      PRINT*,'OPTIMAL CONTROL: ',(U(I),I=0,N)
      PRINT*,'OPTIMAL STATE:   ',(X(I),I=0,N)
      PRINT*,'OPTIMAL VALUE:   ',OBJ
      RETURN

C     FEHLER: KEINE ZULAESSIGE LOESUNG
 1000 CONTINUE
      PRINT*,'NO FEASIBLE SOLUTION FOR THIS DISCRETIZATION: YOU HAVE TO
     +REFINE THE DISCRETIZATION!'
      RETURN
 2000 CONTINUE
      PRINT*,'NO FEASIBLE SOLUTION IN FORWARD CALCULATION: YOU HAVE TO R
     +EFINE THE CONTROL REGION!'
      RETURN
      END

      SUBROUTINE PHI(J,X,U,RES)
      IMPLICIT NONE
      INTEGER J
      DOUBLEPRECISION X,U,RES
      DOUBLEPRECISION C(1)
      COMMON /NUTZEN/ C
      RES = C(J+1)*U
      RETURN
      END

      SUBROUTINE PSI(J,X,U,RES)
      IMPLICIT NONE
      INTEGER J
      DOUBLEPRECISION X,U,RES
      DOUBLEPRECISION W(1)
```

```
      COMMON /GEWICHT/ W
      RES = X-W(J+1)*U
      RETURN
      END

      SUBROUTINE UBOUNDS(J,X,UL,UU)
      IMPLICIT NONE
      INTEGER J
      DOUBLEPRECISION X,UL,UU
      UL = 0.0D0
      UU = 1.0D0
      RETURN
      END
```

Chapter 3

Maximum Principle

We prove the maximum principle for discrete dynamic optimization problems by a special version of the Fritz John conditions. More general results on the basis of so-called smooth-convex optimization problems can be found in [33].

Notice the analogy to the maximum principle for continuous optimal control problems, compare the standard textbooks [50, 30, 26, 33].

3.1 Discrete Maximum Principle

As a motivation for discrete dynamic optimization problems originating from continuous control problems we consider

3.1.1. Optimal Control Problem. Let

$$\begin{aligned}
\varphi_a &: \mathbb{R}^n \longrightarrow \mathbb{R} \,, \\
\varphi_b &: \mathbb{R}^n \longrightarrow \mathbb{R} \,, \\
\varphi &: \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m \longrightarrow \mathbb{R} \,, \\
\psi &: \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m \longrightarrow \mathbb{R}^n \,, \\
\alpha &: \mathbb{R} \times \mathbb{R}^n \longrightarrow \mathbb{R}^{s_a} \,, \\
\beta &: \mathbb{R} \times \mathbb{R}^n \longrightarrow \mathbb{R}^{s_b} \,, \\
S &: \mathbb{R} \times \mathbb{R}^n \longrightarrow \mathbb{R}^s \,,
\end{aligned}$$

be mappings and

$$U \subset \mathbb{R}^m \qquad (a \leq b)$$

a *control region.*

Minimize

$$\varphi_a(y(a)) + \varphi_b(y(b)) + \int_a^b \varphi(t, y(t), u(t))dt$$

subject to

$$\begin{aligned}
&y(\cdot) \in AC([a,b])^n,\ u(\cdot) \in L_\infty([a,b])^m\ ,\\
&\dot{y}(t) = \psi(t, y(t), u(t)) \qquad \text{for a. a. } t \in [a,b]\ ,\\
&\alpha_i(a, y(a)) \begin{cases} \leq 0 & (i = 1, \ldots, s'_a)\\ = 0 & (i = s'_a + 1, \ldots, s'_a) \end{cases},\\
&\beta_i(b, y(b)) \begin{cases} \leq 0 & (i = 1, \ldots, s'_b)\\ = 0 & (i = s'_b + 1, \ldots, s'_b) \end{cases},\\
&u(t) \in U \qquad \text{for a. a. } t \in [a,b]\ ,\\
&S(t, y(t)) \leq \Theta \qquad \text{for all } t \in [a,b]\ .
\end{aligned}$$

■

Discretization by Euler's method yields

3.1.2. Discrete Optimal Control Problem. Choose a grid

$$\mathbb{G}_N = \{t_0, t_1, \ldots, t_N\}$$

with

$$\begin{aligned}
&a = t_0 \leq t_1 \leq \ldots \leq t_N = b\ ,\\
&t_j - t_{j-1} = h_{j-1} \qquad (j = 1, \ldots, N)\ .
\end{aligned}$$

Use the notations

$$y_j = y(t_j),\ u_j = u(t_j) \qquad (j = 0, \ldots, N)\ .$$

Minimize

$$\varphi_a(y_0) + \varphi_b(y_N) + \sum_{j=0}^{N-1} h_j \varphi(t_j, y_j, u_j)$$

subject to

$$
\begin{aligned}
& y_j = y_{j-1} + h_{j-1}\psi(t_{j-1}, y_{j-1}, u_{j-1}) \qquad (j = 1, \ldots, N) \ , \\
& \alpha_i(t_0, y_0) \begin{cases} \leq 0 & (i = 1, \ldots, s'_a) \\ = 0 & (i = s'_a + 1, \ldots, s_a) \end{cases} \ , \\
& \beta_i(t_N, y_N) \begin{cases} \leq 0 & (i = 1, \ldots, s'_b) \\ = 0 & (i = s'_b + 1, \ldots, s'_b) \end{cases} \ , \\
& u_j \in U_j \ , \ S(t_j, y_j) \leq \Theta \qquad (j = 0, \ldots, N)
\end{aligned}
$$

over all grid functions

$$
y : \mathbb{G}_N \longrightarrow \mathbb{R}^n, \ u : \mathbb{G}_N \longrightarrow \mathbb{R}^m \ .
$$

■

In the simplest approach to the so-called discrete maximum principle, this discrete optimal control problem is considered as a smooth nonlinear finite dimensional optimization problem

3.1.3. Discrete Maximum Principle. Choose a *fixed* grid

$$
\mathbb{G}_N = \{t_0, \ldots, t_N\}
$$

with

$$
\begin{aligned}
& a = t_0 \leq t_1 \leq \ldots \leq t_N = b \ , \\
& t_j - t_{j-1} = h_{j-1} \qquad (j = 1, \ldots, N) \ .
\end{aligned}
$$

Let $(\hat{y}, \hat{u})$ be a (local) optimal solution of the Discrete Optimal Control Problem 3.1.2 and $\varphi(t_j, \cdot, \cdot)$, $\psi(t_j, \cdot, \cdot)$ be continuously partially differentiable in a neighbourhood of $(\hat{y}_j, \hat{u}_j) \in \mathbb{R}^n \times \mathbb{R}^m$, $\alpha(t_0, \cdot)$ resp. $\beta(t_N, \cdot)$ resp. $S(t_j, \cdot)$ $(j = 0, \ldots, N)$ be continuously partially differentiable in a neighbourhood of $\hat{y}_0$ resp. $\hat{y}_N$ resp. $\hat{y}_j$ $(j = 0, \ldots, N)$.

Let the control regions U_j be convex and

$$
\operatorname{int}(U_j) \neq \emptyset \qquad (j = 0, \ldots, N) \ .
$$

Then there exist multipliers

$$
\begin{aligned}
& \lambda_0 \in \mathbb{R}, \ \lambda_a \in \mathbb{R}^{s_a}, \ \lambda_b \in \mathbb{R}^{s_b} \ , \\
& \mu_j \in \mathbb{R}^s \qquad (j = 0, \ldots, N) \ , \\
& p_j \in \mathbb{R}^n \qquad (j = 0, \ldots, N)
\end{aligned}
$$

with the following properties:

(i)

$$(\lambda_0, \lambda_a, \lambda_b, \mu, p) \neq \Theta \ ,$$

i.e. the multiplier vector is *nontrivial*.

(ii)

$$\lambda_0 \geq 0 \ ,$$

$$(\lambda_a)_i \begin{cases} \geq 0, & (i = 1, \ldots, s'_a) \\ = 0, & \text{if } \alpha_i(t_0, \hat{y}_0) < 0 \end{cases} ,$$

$$(\lambda_b)_i \begin{cases} \geq 0, & (i = 1, \ldots, s'_b) \\ = 0, & \text{if } \beta_i(t_N, \hat{y}_N) < 0 \end{cases} ,$$

$$(\mu_j)_i \begin{cases} \geq 0, & (i = 1, \ldots, s, j = 0, \ldots, N) \\ = 0, & \text{if } S_i(t_j, \hat{y}_j) < 0 \end{cases} ,$$

i.e. the *complementary slackness condition* holds.

(iii)

$$\begin{aligned} p_{j+1} &= p_j - h_j[\psi^\star_x(t_j, \hat{y}_j, \hat{u}_j)p_{j+1} - \lambda_0\varphi^\star_x(t_j, \hat{y}_j, \hat{u}_j)] + S^\star_x(t_j, \hat{y}_j)\mu_j \\ &\qquad \text{for } j = 0, \ldots, N-1 \ , \\ p_0 &= \lambda_0\varphi^\star_{ax}(\hat{y}_0) + \alpha^\star_x(t_0, \hat{y}_0)\lambda_a \ , \\ p_N &= -\lambda_0\varphi^\star_{bx}(\hat{y}_N) - \beta^\star_x(t_N, \hat{y}_N)\lambda_b - S_x(t_N, \hat{y}_N)^\star\mu_N \ , \end{aligned}$$

i.e. the *adjoint equation* holds together with special *boundary conditions*.

iv)

$$[p^\star_{j+1}\psi_u(t_j, \hat{y}_j, \hat{u}_j) - \lambda_0\varphi_u(t_j, \hat{y}_j, \hat{u}_j)](u_j - \hat{u}_j) \leq 0$$

for all $u_j \in U_j \quad (j = 0, \ldots, N-1)$, i.e. the *local maximum principle* holds.

Proof. We give a general outline of the proof, which is based on a proof of a Fritz John-Condition of smooth nonlinear optimization. The proof is divided into several parts which are instructive by themselves.

I. Linearization. The discrete optimal control problem can be considered as a special case of following *nonlinear programming problem*

(NLP) *Minimize* $F(y,u)$ *subject to*

$$\begin{aligned} G(y,u) &\leq \Theta \ , \\ H(y,u) &= \Theta \ , \\ y \in Y,\ u &\in U \ . \end{aligned}$$

Here we choose appropriately

$$\begin{aligned} Y &= (\mathbb{R}^n)^{N+1} \ , \\ U &= U_0 \times \cdots \times U_N \subset (\mathbb{R}^m)^{N+1} = \mathcal{U} \\ F &: Y \times U \longrightarrow \mathbb{R} \ , \\ G &: Y \times U \longrightarrow \mathbb{R}^{s'_a} \times \mathbb{R}^{s'_b} \times (\mathbb{R}^s)^{N+1} \ , \\ H &: Y \times U \longrightarrow (\mathbb{R}^n)^{N+1} \times \mathbb{R}^{s_a - s'_a} \times \mathbb{R}^{s_b - s'_b} \ . \end{aligned}$$

F, G, H are Fréchet-differentiable in a neighbourhood of $(\hat{y}, \hat{u})$, the Fréchet-derivatives being represented by appropriately chosen functional matrices

$$F_{(y,u)}(\hat{y},\hat{u}),\ G_{(y,u)}(\hat{y},\hat{u}),\ H_{(y,u)}(\hat{y},\hat{u}) \ .$$

The problem (NLP) is linearized locally at $(\hat{y}, \hat{u})$ yielding a linearized programming problem

(LP) *Minimize*

$$F(\hat{y},\hat{u}) + F_{(x,u)}(\hat{y},\hat{u})(y - \hat{x}, u - \hat{u})$$

subject to

$$\begin{aligned} G(\hat{y},\hat{u}) + G_{(y,u)}(\hat{y},\hat{u})(y - \hat{y}, u - \hat{u}) &\leq \Theta \ , \\ H(\hat{y},\hat{u}) + H_{(y,u)}(y,\hat{u})(y - \hat{y}, u - \hat{u}) &= \Theta \ , \\ y \in Y,\ u &\in U \ . \end{aligned}$$

We now show that (except some special situations and some special feasible solutions of (LP)) the local optimal solution $(\hat{y}, \hat{u})$ of (NLP) is optimal for (LP).

Choose a feasible solution (y, u) of (LP) with

$$\begin{aligned} y \in Y,\ u &\in \mathrm{int}(U) \ , \\ G(\hat{y},\hat{u}) + G_{(y,u)}(\hat{y},\hat{u})(y - \hat{y}, u - \hat{u}) &< \Theta \ , \\ H(\hat{y},\hat{u}) + H_{(y,u)}(\hat{y},\hat{u})(y - \hat{y}, u - \hat{u}) &= \Theta \ , \end{aligned}$$

which reduces the linearized objective function

$$F(\hat{y},\hat{u}) + F_{(y,u)}(\hat{y},\hat{u})(y-\hat{y},u-\hat{u}) < F(\hat{y},\hat{u}) \ .$$

II. Treatment of the Equality Constraints. We assume that the functional matrix

$$H_{(y,u)}(\hat{y},\hat{u})$$

is surjective. Later we will see that the results which we want to prove hold trivially if $H_{(y,u)}(\hat{y},\hat{u})$ is not surjective.

Let $\{e^\mu\}_{\mu=1,\dots,k}$ with $k = n(N+1) + (s_a - s_a') + (s_b - s_b')$ be a basis of the image of $H_{(y,u)}(\hat{y},\hat{u})$. Then there are vectors $\{v^\mu\}_{\mu=1,\dots,k} \in Y \times \mathcal{U}$, with

$$H_{(y,u)}(\hat{y},\hat{u})v^\mu = e^\mu \qquad\qquad (\mu = 1,\dots,k) \ .$$

Consider the nonlinear system of equations

$$\tilde{H}(t,r) = H\big((\hat{y},\hat{u}) + t(y-\hat{y},u-\hat{u}) + \sum_{\mu=1}^{k} r_\mu v^\mu\big) = \Theta$$

for the unknowns $t, r_1, \dots, r_k$.

We want to solve it for $r_1,\dots,r_k$ as functions of t locally around $t = 0,\ r = \Theta$. We know

(i)

$$\tilde{H}(0,\Theta) = H(\hat{y},\hat{u}) = \Theta \ ,$$

(ii)

$\tilde{H}$ is continuously partially differentiable in a neighbourhood of $(0,\Theta)$,

(iii)

$$\frac{\partial \tilde{H}}{\partial r}(0,\Theta) = H_{(y,u)}(\hat{y},\hat{u})[v^1,\dots,v^k] = E_k$$

is regular.

Therefore, by the Implicit Function Theorem, there exist $\epsilon > 0$, $\delta > 0$ and functions

$$r_\mu(t) \qquad\qquad (-\epsilon \le t \le \epsilon)$$

with

$$\begin{array}{rcll} \tilde{H}(t,r) & = & \Theta & \\ & \longleftrightarrow & & \\ r & = & r(t) & (|t| \le \epsilon,\ \|r\|_\infty \le \delta)\ . \end{array}$$

With these functions, we have solved

$$H\big((\hat{y},\hat{u}) + t(y-\hat{y},u-\hat{u}) + \sum_{\mu=1}^{k} r_\mu(t) v^\mu\big) = \Theta$$

for all $|t| \le \epsilon$.

Moreover, by differentiating this equation with respect to t at $t = 0$ yields

$$H_{(y,u)}(\hat{y},\hat{u})\left[(y-\hat{y},u-\hat{u}) + \sum_{\mu=1}^{k} \frac{d}{dt} r_\mu(t)\Big|_{t=0} v^\mu\right] = \Theta\ .$$

Since $H_{(y,u)}(\hat{y},\hat{u})(y-\hat{y},u-\hat{u}) = \Theta$, we get

$$\begin{aligned} \Theta &= \sum_{\mu=1}^{k} \frac{d}{dt} r_\mu(t)\Big|_{t=0} H_{(y,u)}(\hat{y},\hat{u}) v^\mu\ , \\ &= \sum_{\mu=1}^{k} \frac{d}{dt} r_\mu(t)\Big|_{t=0} e^\mu\ , \end{aligned}$$

hence

$$\frac{d}{dt} r_\mu(t)\Big|_{t=0} = 0 \qquad (\mu = 1,\dots,k)\ ,$$

i.e. the curve

$$(\hat{y},\hat{u}) + t(y-\hat{x},u-\hat{u}) + \sum_{\mu=1}^{k} r_\mu(t) v^\mu \qquad (|t| \le \epsilon)$$

is differentiable with derivative

$$(y-\hat{y},u-\hat{u})$$

at $t = 0$, and *satisfies the nonlinear equality constraints.*

III. Treatment of the Inequality Constraints. If a scalar inequality constraint

$$G_i(\hat{y}, \hat{u}) < 0$$

is strictly satisfied, then by continuity alone there exists a sufficiently small $\epsilon_i > 0$ with

$$G_i((\hat{y}, \hat{u}) + t(y - \hat{y}, u - \hat{u}) + \sum_{\mu=1}^{k} r_\mu(t) v^\mu) < 0 \qquad (0 \le t \le \epsilon_i) \ .$$

If

$$G_i(\hat{y}, \hat{u}) = 0 \ ,$$

then by the chain rule

$$\begin{aligned} & \lim_{t \to 0+} \frac{1}{t} \Big[G_i\big((\hat{y}, \hat{u}) + t(y - \hat{y}, u - \hat{u}) + \sum_{\mu=1}^{k} r_\mu(t) v^\mu\big) - G_i(\hat{y}, \hat{u}) \Big] \\ = \ & \frac{d}{dt} G_i\big((\hat{y}, \hat{u}) + t(y - \hat{y}, u - \hat{u}) + \sum_{\mu=1}^{k} r_\mu(t) v^\mu\big) \Big|_{t=0} \\ = \ & (G_i)_{(y,u)}(\hat{y}, \hat{u})(y - \hat{y}, u - \hat{u}) < 0 \ . \end{aligned}$$

Hence, there exists $\epsilon_i > 0$ with

$$G_i\big((\hat{y}, \hat{u}) + t(y - \hat{y}, u - \hat{u}) + \sum_{\mu=1}^{k} r_\mu(t) v^\mu\big) < 0 \qquad (0 < t \le \epsilon_i) \ .$$

Therefore, for a sufficiently small $\epsilon > 0$, the curve

$$(\hat{y}, \hat{u}) + t(y - \hat{y}, u - \hat{u}) + \sum_{\mu=1}^{k} r_\mu(t) v^\mu \qquad (0 \le t \le \epsilon)$$

satisfies all equality and inequality constraints.

IV. Treatment of the Control Restrictions. We have $\hat{u} \in U$ and assumed

$$u \in \mathrm{int}(U) \ ,$$

consequently, there exists a ball $B(\Theta, \delta)$ in Y around Θ with radius $\delta > 0$ and

$$\hat{u} + t(u + w - \hat{u}) \in U$$

for all $w \in B(\Theta, \delta)$ and all $0 \le t \le 1$,

Since

$$\lim_{t \to 0} \| \frac{1}{t} r(t) \| = 0 \ ,$$

there exists $\epsilon > 0$ with

$$\| \sum_{\mu=1}^{k} \frac{1}{t} r_\mu(t) v^\mu \| \le \delta \qquad (0 < t \le \epsilon) \ .$$

Therefore,

$$\begin{aligned} & \hat{u} + t(u - \hat{u}) + \sum_{\mu=1}^{k} r_\mu(t) v^\mu \\ = \ & \hat{u} + t \Big[u + \sum_{\mu=1}^{k} \frac{1}{t} r_\mu(t) v^\mu - \hat{u} \Big] \in U \end{aligned}$$

for all $0 < t \le \epsilon$.

This means that *all control restrictions are satisfied* for $0 \le t \le \epsilon$.

V. Treatment of the Objective Function. We assumed

$$F(\hat{y}, \hat{u}) + F_{(y,u)}(\hat{y}, \hat{u})(y - \hat{y}, u - \hat{u}) < F(\hat{y}, \hat{u}) \ .$$

Therefore, for the feasible curve

$$\begin{aligned} & \lim_{t \to 0+} \frac{1}{t} \Big[F\big((\hat{y}, \hat{u}) + t(y - \hat{y}, u - \hat{u}) + \sum_{\mu=1}^{k} r_\mu(t) v^\mu\big) - F(\hat{y}, \hat{u}) \Big] \\ = \ & \frac{d}{dt} F\big((\hat{y}, \hat{u}) + t(y - \hat{y}, u - \hat{u}) + \sum_{\mu=1}^{k} r_\mu(t) v^\mu\big) \Big|_{t=0} \\ = \ & F_{(y,u)}(\hat{y}, u)(y - \hat{y}, u - \hat{u}) < 0 \ . \end{aligned}$$

Hence, there exists sufficiently small $\epsilon > 0$ with

$$F\big((\hat{y}, \hat{u}) + t(y - \hat{y}, u - \hat{u}) + \sum_{\mu=1}^{k} r_\mu(t) v^\mu - F(\hat{y}, \hat{u}) < 0$$

for all $0 < t \le \epsilon$.

This is a contradiction to the local optimality of $(\hat{y}, \hat{u})$.

Altogether, we have shown in II–V:

If $H_{(y,u)}(\hat{y}, \hat{u})$ is surjective, then there exists no $(y, u) \in Y \times \mathrm{int}(U)$ with

$$\begin{aligned} G(\hat{y}, \hat{u}) + G_{(y,u)}(\hat{y}, \hat{u})(y - \hat{y}, u - \hat{u}) &< \Theta \ , \\ H(\hat{y}, \hat{u}) + H_{(y,u)}(\hat{y}, \hat{u})(y - \hat{y}, u - \hat{u}) &= \Theta \ , \\ F(\hat{y}, \hat{u}) + F_{(y,u)}(\hat{y}, \hat{u})(y - \hat{y}, u - \hat{u}) &< F(\hat{y}, \hat{u}) \ . \end{aligned}$$

VI. Separation of Convex Sets. Consider the convex sets

$$A = \left\{ \begin{pmatrix} F(\hat{y}, \hat{u}) \\ G(\hat{y}, \hat{u}) \\ H(\hat{y}, \hat{u}) \end{pmatrix} + \begin{pmatrix} F_{(y,u)}(\hat{y}, \hat{u})(y - \hat{x}, u - \hat{u}) \\ G_{(y,u)}(\hat{y}, \hat{u})(y - \hat{y}, u - \hat{u}) \\ H_{(y,u)}(\hat{y}, \hat{u})(y - \hat{x}, u - \hat{u}) \end{pmatrix} : (y, u) \in Y \times U \right\} ,$$

$$B = \left\{ \begin{pmatrix} z_0 \\ z_G \\ z_H \end{pmatrix} : z_0 \leq F(\hat{x}, \hat{u}),\ z_G \leq \Theta,\ z_H = \Theta \right\} .$$

If $H_{(y,u)}(\hat{y}, \hat{u})$ happens not to be surjective, then these sets can be separated by a hyperplane, in fact then both sets are even contained in one hyperplane.

If $H_{(y,u)}(\hat{y}, \hat{u})$ is surjective, then the above considerations in II–V showed that

$$\mathrm{rel\,int}(A) \cap \mathrm{rel\,int}(B) = \emptyset \ .$$

Hence, in any case, both sets can be separated in the image space of the product mapping (F, G, H), which in fact is

$$\mathbb{R} \times \mathbb{R}^{s'_a} \times \mathbb{R}^{s'_b} \times (\mathbb{R}^s)^{N+1} \times \mathbb{R}^{s_a - s'_a} \times \mathbb{R}^{s_b - s'_b} \times (\mathbb{R}^n)^{N+1} \ ,$$

by a hyperplane.

Analytically, this geometric property, has the following meaning.

There exist multiplier vectors of appropriate dimensions

$$\lambda_0 \in \mathbb{R},\ \lambda_G,\ \lambda_H \ ,$$

not all of them being trivial, with

$$\begin{aligned} & \lambda_0 z_0 + \lambda_G^\star z_G + \lambda_H^\star z_H \\ \leq\ & \lambda_0 \big[F(\hat{y},\hat{u}) + F_{(y,u)}(\hat{y},\hat{u})(y-\hat{y},u-\hat{u})\big] \\ & +\lambda_G^\star \big[G(\hat{y},\hat{u}) + G_{(y,u)}(\hat{y},\hat{u})(y-\hat{y},u-\hat{u})\big] \\ & +\lambda_H^\star \big[H(\hat{y},\hat{u}) + H_{(y,u)}(\hat{y},\hat{u})(y-\hat{y},u-\hat{u})\big] \end{aligned}$$

for all $z_0 \leq F(\hat{y},\hat{u}),\ z_G \leq \Theta,\ z_H = \Theta,\ y \in Y,\ u \in U$.

This implies and is in fact equivalent to

$$\begin{aligned} & \lambda_0 \geq 0,\ \lambda_G \geq \Theta \quad \text{(componentwise)}\ , \\ & \lambda_G^\star G(\hat{y},\hat{u}) = 0 \quad \text{(complementary slackness)}\ , \\ & \lambda_0 F_y(\hat{y},\hat{u})(y-\hat{y}) + \lambda_G^\star G_y(\hat{y},\hat{u})(y-\hat{y}) + \lambda_H^\star H_y(\hat{y},\hat{u})(y-\hat{y}) = 0 \\ & \text{for all } y \in Y \quad \text{(adjoint equation)}\ , \\ & \lambda_0 F_u(\hat{y},\hat{u})(u-\hat{u}) + \lambda_G^\star G_u(\hat{y},\hat{u})(u-\hat{u}) + \lambda_H^\star H_u(\hat{y},\hat{u})(u-\hat{u}) \geq 0 \\ & \text{for all } u \in U \quad \text{(local minimum principle)}\ . \end{aligned}$$

Here, we denote by

$$F_y(\hat{y},\hat{u}),\ G_y(\hat{y},\hat{u}),\ H_y(\hat{y},\hat{u})$$

resp.

$$F_u(\hat{y},\hat{u}), G_u(\hat{y},\hat{u}),\ H_u(\hat{y},\hat{u})$$

the partial functional matrices with respect to y resp. u.

This is a variant of the famous *Fritz John-Necessary Optimality Condition* for finite-dimensional smooth nonlinear optimization problems.

Notice, that we needed no Constraint Qualification. Contrary to the Karush-Kuhn-Tucker-Conditions, one gets an additional multiplier $\lambda_0 \geq 0$, corresponding to the objective function, which could turn out to be 0. In that case, this optimality condition would be a very weak one since then the objective function is not involved at all. To avoid that case, one has to assume additional constraint qualification like the local Slater-Condition, which in the control context amounts to controllability conditions and implies additional stability properties of the minimal value function.

VII. Application to the Discrete Control Problem. We specialize the Fritz-John Condition from VI. to the Discrete Optimal Control Problem

3.1.2 and obtain

$$
\begin{aligned}
F(y,u) &= \varphi_a(y_0) + \varphi_b(y_N) + \sum_{j=0}^{N-1} h_j \varphi(t_j, y_j, u_j) \ , \\
G(y,u) &= \begin{pmatrix} \alpha_1(t_0,y_0) \\ \vdots \\ \alpha_{s'_a}(t_0,y_0) \\ \beta_1(t_N,y_N) \\ \vdots \\ \beta'_{s_b}(t_N,y_N) \\ (S(t_j,y_j))_{j=0,\ldots,N} \end{pmatrix} , \\
H(y,u) &= \begin{pmatrix} \alpha_{s'_{a+1}}(t_0,y_0) \\ \vdots \\ \alpha_{s_a}(t_0,y_0) \\ \beta_{s'_{b+1}}(t_N,y_N) \\ \vdots \\ \beta_{s_b}(t_N,y_N) \\ \left[y_j - y_{j-1} - h_{j-1}\psi(t_{j-1},y_{j-1},u_{j-1} \right]_{j=1,\ldots,N} \end{pmatrix} , \\
U &= U_0 \times \ldots \times U_N \ .
\end{aligned}
$$

U is convex and $\mathrm{int}(U) \neq \emptyset$, since we required U_j to be convex and $\mathrm{int}(U_j) \neq \emptyset$ $(j = 0, \ldots, N)$. U_N does not really appear in the problem, therefore one could choose $U_N = R^m$ or omit U_N at all.

The multipliers are denoted by

$$
\begin{aligned}
&\lambda_0 \ , \\
&(\lambda_a)_{1,\ldots,s'_a}, \ (\lambda_b)_{1,\ldots,s'_b}, \ \mu_j \in \mathbb{R}^s \quad (j = 0, \ldots, N) \ , \\
&(\lambda_a)_{s'_a+1,\ldots,s_a}, \ (\lambda_b)_{s'_b+1,\ldots,s_b}, \ p_j \in \mathbb{R}^n \quad (j = 1, \ldots, N) \ .
\end{aligned}
$$

Not all of them are identical 0, therefore property (i) of the Discrete Maximum Principle 3.1.3 is proven.

Property (ii) is equivalent to $\lambda_0 \geq 0, \quad \lambda_G \geq \Theta$ (componentwise), $\lambda_G^\star G(\hat{y}, \hat{u}) = 0$ since $G(\hat{y}, \hat{u}) \leq \Theta$ (componentwise).

The adjoint equations translates to

$$\begin{aligned}
& \lambda_0[\varphi_{ax}(\hat{y}_0)(y_0 - \hat{y}_0) + \varphi_{bx}(\hat{y}_N)(y_N - \hat{y}_N) \\
+ & \sum_{j=0}^{N-1} h_j \varphi_x(t_j, \hat{y}_j, \hat{u}_j)(y_j - \hat{y}_j)] \\
+ & \lambda_a^\star \alpha_x(t_0, \hat{y}_0)(y_0 - \hat{y}_0) + \lambda_b^\star \beta_x(t_N \hat{y}_N)(y_N - \hat{y}_N) \\
+ & \sum_{j=0}^{N} \mu_j^\star S_x(t_j, \hat{y}_j)(y_j - \hat{y}_j) \\
+ & \sum_{j=1}^{N} p_j^\star \big[(y_j - \hat{y}_j) - (y_{j-1} - \hat{y}_{j-1}) \\
& \qquad - h_{j-1} \psi_x(t_{j-1}, \hat{y}_{j-1}, \hat{u}_{j-1})(y_{j-1} - \hat{y}_{j-1})\big] \\
= & \ 0
\end{aligned}$$

for all $y_0, y_1, \ldots, y_N \in \mathbb{R}^n$.

This is a variational equation which, by successive insertion of all unit vectors for the increment $y - \hat{y}$ in $(\mathbb{R}^n)^{N+1}$, turns out to be equivalent to

$$\begin{aligned}
& \lambda_0[\varphi_{ax}(\hat{y}_0) + h_0 \varphi_x(t_0, \hat{y}_0, \hat{u}_0)] \\
& + \lambda_a^\star \alpha_x(t_0, \hat{y}_0) + \mu_0^\star S_x(t_0, \hat{y}_0) \\
& - p_1^\star \big[E_N + h_0 \psi_x(t_0 \hat{y}_0, \hat{u}_0)\big] = 0_{\mathbb{R}^n}^\star \ , \\
& \lambda_0 h_j \varphi_x(t_j, \hat{y}_j, \hat{u}_j) + \mu_j^\star S_x(t_j, \hat{y}_j) \\
& + p_j^\star - p_{j+1}^\star \big[E_N + h_j \psi_x(t_j, \hat{y}_j, \hat{u}_j)\big] = 0_{\mathbb{R}^n}^\star \ , \\
& (j = 1, \ldots, N-1) \ , \\
& \lambda_0 \varphi_{bx}(t_N, \hat{y}_N) + \lambda_b^\star \beta_x(\hat{y}_N) + \mu_N^\star S_x(t_N, \hat{y}_N) \\
& + p_N^\star = 0_{\mathbb{R}^n}^\star \ .
\end{aligned}$$

It is convenient to define in addition

$$p_0^\star = \lambda_a \varphi_{ax}(\hat{y}_0) + \lambda_a^\star \alpha_x(t_0, \hat{y}_0) \ .$$

Then $p_0, p_1, \ldots, p_N$ solve the *adjoint equation*

$$\begin{aligned} p_{j+1} &= p_j - h_j \psi_x^\star(t_j, \hat{y}_j, \hat{u}_j) p_{j+1} + \lambda_0 h_j \varphi_x^\star(t_j, \hat{y}_j, \hat{u}_j) \\ &\quad + S_x^\star(t_j, \hat{y}_j)\mu_j \\ &\quad (j = 0, \ldots, N-1) \end{aligned}$$

together with the *boundary conditions*

$$\begin{aligned} p_0 &= \lambda_0 \varphi_{ax}^\star(\hat{y}_0) + \alpha_x^\star(t_0, \hat{y}_0)\lambda_a \ , \\ p_N &= \lambda_0 \varphi_{bx}^\star(\hat{y}_N) - \beta_x^\star(t_N, \hat{y}_N)\lambda_b - S_x(t_N, \hat{y}_N)^\star \mu_N \ . \end{aligned}$$

The local minimum principle translates to

$$\begin{aligned} &\lambda_0 \sum_{j=0}^{N-1} h_j \varphi_u(t_j, \hat{y}_j, \hat{u}_j)(u_j - \hat{u}_j) \\ &- \sum_{j=1}^{N} p_j^\star h_{j-1} \psi_u(t_{j-1}, \hat{y}_{j-1}, \hat{u}_{j-1})(u_{j-1} - \hat{u}_{j-1}) \geq 0 \end{aligned}$$

for all $u_j \in U_j$ $(j = 0, \ldots, N)$, which is equivalent to the *local maximum principle*

$$\big[p_{j+1}^\star \psi_u(t_j, \hat{y}_j, \hat{u}_j) - \lambda_0 \varphi_u(t_j, \hat{y}_j, \hat{u}_j)\big](u_j - \hat{u}_j) \leq 0$$

for all $u_j \in U_j$ $(j = 0, \ldots, N-1)$. ∎

Concluding Remark. Introduce the *Hamilton-Function*

$$H(t, x, u, p, \lambda_0) = p^\star \psi(t, x, u) - \lambda_0 \varphi(t, x, u) \ ,$$

then the system equation and the adjoint equation can be written as the following discrete Hamiltonean system

$$\begin{aligned} \hat{y}_{j+1} &= \hat{y}_j + h_j \frac{\partial H}{\partial p}(t_j, \hat{y}_j, \hat{u}_j, p_{j+1}, \lambda_0)^\star \ , \\ p_{j+1} &= p_j - h_j \frac{\partial H}{\partial x}(t_j, \hat{y}_j, \hat{u}_j, p_{j+1}, \lambda_0)^\star + S_x^\star(t_j, \hat{y}_j)\mu_j \ , \\ &\quad (j = 0, \ldots, N-1) \ , \end{aligned}$$

which is explicit Euler method for the continuous system equation coupled with implicit Euler method for the continuous adjoint *integral equation*.

Notice that, due to the term $S_x^\star(t_j, \hat{y}_j)\mu_j$, the second difference equation is *not* a discretization of a differential equation. The coupling occurs by the (unknown) local optimal solution $(\hat{y}, \hat{u})$, by the complementary slackness conditions, the boundary conditions, and the local maximum principle

$$\frac{\partial H}{\partial u}(t_j, \hat{y}_j, \hat{u}_j, P_{j+1}, \lambda_0)(u_j - \hat{u}_j) \leq 0$$

for all $u_j \in U_j$ $(j = 0, \ldots, N-1)$, which is a *necessary condition for maximizing*

$$H(t_j, \hat{y}_j, u_j, p_{j+1}, \lambda_0)$$

globally for all $u_j \in U_j$.

Since first order necessary optimality conditions for concave functions are sufficient for a global maximum on U_j, we can conclude: If in addition the Hamilton function is concave with respect to u for each fixed $t_j, \hat{y}_j, p_{j+1}, \lambda_0$, then

$$H(t_j, \hat{y}_j, u_j, p_{j+1}, \lambda_0) \leq H(t_j, \hat{y}_j \hat{u}_j, p_{j+1}, \lambda_0)$$

for all $u_j \in U_j$ $(j = 0, \ldots, N-1)$.

This is the *global discrete maximum principle.*

Using necessary optimality conditions for so-called *differentiable-convex optimization problems*, this result can be slightly weakened.

3.2 Continuous Maximum Principle

Naturally, the question arises what is the continuous analogue of the discrete maximum principle.

We now examine the following continuous optimal control problem more closely.

3.2.1. Optimal Control Problem. Let

$$\begin{aligned}
\varphi_a &: \mathbb{R}^n \longrightarrow \mathbb{R}\ , \\
\varphi_b &: \mathbb{R}^n \longrightarrow \mathbb{R}\ , \\
\varphi &: \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m \longrightarrow \mathbb{R}\ , \\
\psi &: \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m \longrightarrow \mathbb{R}^n\ , \\
\alpha &: \mathbb{R}^n \longrightarrow \mathbb{R}^{s_a}\ , \\
\beta &: \mathbb{R}^n \longrightarrow \mathbb{R}^{s_b}\ , \\
S &: \mathbb{R} \times \mathbb{R}^n \longrightarrow \mathbb{R}^s\ ,
\end{aligned}$$

be mappings and

$$U \subset \mathbb{R}^m$$

a *control region.*

Minimize

$$\varphi_a(y(a)) + \varphi_b(y(b)) + \int_a^b \varphi(t, y(t), u(t))dt$$

subject to

$$\begin{aligned}
&y(\cdot) \in AC([a,b])^n,\ u(\cdot) \in L_\infty([a,b])^m\ ,\\
&\dot{y}(t) = \psi(t, y(t), u(t)) \text{ for a. a. } t \in [a,b]\ ,\\
&\alpha_i(y(a)) \begin{cases} \leq 0 & (i = 1, \ldots, s'_a) \\ = 0 & (i = s'_a, \ldots, s_a) \end{cases}\ ,\\
&\beta_i(y(b)) \begin{cases} \leq 0 & (i = 1, \ldots, s'_b) \\ = 0 & (i = s'_b + 1, \ldots, s_b) \end{cases}\ ,\\
&S(t, y(t)) \leq \Theta \quad \text{for all } t \in [a,b]\ ,\\
&u(t) \in U \qquad\quad \text{for all } t \in [a,b]\ .
\end{aligned}$$

■

This problem can be interpreted as a smooth-convex optimization problem in the sense of [33], for which the following necessary optimality condition holds.

3.2.2. Maximum Principle. Let $(\hat{y}, \hat{u})$ be a (local) optimal solution of 3.2.1. Assume that all mappings

$$\varphi_a, \varphi_b, \varphi, \psi, \alpha, \beta, S$$

are continuous and, *with respect to all state coordinates*, continuously partially differentiable.

Then there exist multipliers

$$\lambda_0 \in \mathbb{R},\ \lambda_a \in \mathbb{R}^{s_a},\ \lambda_b \in \mathbb{R}^{s_b}\ ,$$

functions of bounded variation

$$\nu_i : [a,b] \longrightarrow \mathbb{R} \qquad (i = 1, \ldots, s)\ ,$$

and vector functions

$$p : [a, b] \longrightarrow \mathbb{R}^n$$

with the following properties:

(i) $\lambda_0, \lambda_a, \lambda_b, p(\cdot)$, resp. $\nu(\cdot)$ cannot vanish identically resp. be constant simultaneously.

(ii)

$$\lambda_0 \geq 0 \ ,$$

$$(\lambda_a)_i \begin{cases} \geq 0, & (i = 1, \ldots, s'_a) \\ = 0, & \text{if } \alpha_i(\hat{y}(a)) < 0 \end{cases} ,$$

$$(\lambda_b)_i \begin{cases} \geq 0, & (i = 1, \ldots, s'_b) \\ = 0, & \text{if } \beta_i(\hat{y}(b)) < 0 \end{cases} ,$$

For every $i = 1, \ldots, s$: $\nu_i(\cdot)$ is (weakly) monotonically increasing, continuous from the left, and $\nu_i(t)$ constant on any open subinterval $(a', b') \subset [a, b]$ with $S_i(t, \hat{y}(t)) < 0 \qquad (a' < t < b')$.

(iii)

$$\begin{aligned} p(t) &= -\lambda_0 \varphi^\star_{bx}(\hat{y}(b) - \beta^\star_x(\hat{y}(b))\lambda_b \\ &\quad + \int_t^b [\psi^\star_x(\tau, \hat{y}(\tau), u(\tau), p(\tau) - \lambda_0 \varphi^\star_x(\tau, \hat{y}(\tau), \hat{u}(\tau))] d\tau \\ &\quad - \int_t^b S^\star_x(\tau, \hat{y}(\tau)) d\nu(\tau) \qquad (a \leq t \leq b) \ , \\ p(a) &= \lambda_0 \varphi^\star_{ax}(\hat{y}(a)) + \alpha^\star_x(\hat{y}(a))\lambda_a \ . \end{aligned}$$

iv) For almost all $t \in [a, b]$:

$$\begin{aligned} & p^\star(t)\psi(t, \hat{y}(t), u) - \lambda_0 \varphi(t, \hat{y}(t), u) \\ \leq \ & p^\star(t)\psi(t, \hat{y}(t), \hat{u}(t)) - \lambda_0 \varphi(t, \hat{y}(t), \hat{u}(t)) \qquad (u \in U) \ . \end{aligned}$$

Remarks The condition (i) is again the non-triviality condition for the multiplier (functions). The functions $\nu_i(\cdot)$ define, via Lebesgue-Stieltjes-integrals, regular Borel measures which are identically 0 if $\nu(\cdot)$ is constant.

Condition (ii) is the complementary slackness condition, monotonically increasing functions $\nu_i(\cdot)$ correspond to nonnegative Borel measures. The adjoint equation (iii) is now an integral equation with respect to Lebesgue-Stieltjes integration.

$p(\cdot)$ is of bounded variation and continuous from the left, since the functions $\nu_i(\cdot)$ can be assumed to be continuous from the left. Observe that

$$\begin{aligned}\lim_{t\to b+} p(t) &= -\lambda_0\varphi^\star_{bx}(\hat y(b)) - \beta^\star_x(\hat y(b))\lambda_b \\ &\quad + S^\star_x(b,\hat y(b)) \lim_{t\to b+}[\nu(t)-\nu(b)]\ , \\ \lim_{t\to a+} p(t) &= \lambda_0\varphi^\star_{ax}(\hat y(a)) + \alpha^\star_x(\hat y(a))\lambda_a \\ &\quad + S^\star_x(a,\hat y(a)) \lim_{t\to a+}[\nu(t)-\nu(a)]\ ,\end{aligned}$$

if some of the functions ν_i jump at b resp. a (from the right).

If

$$S_i(a,\hat y(a)) < 0,\ S_i(b,\hat y(b)) < 0 \qquad (i=1,\dots,s)\ ,$$

then $p(\cdot)$ is continuous at a and b and

$$\begin{aligned} p(a) &= \lambda_0\varphi^\star_{ax}(\hat y(a)) + \alpha^\star_x(\hat y(a))\lambda_a\ , \\ p(b) &= -\lambda_0\varphi^\star_{bx}(\hat y(b)) - \beta^\star_x(\hat y(b))\lambda_b\ .\end{aligned}$$

These conditions specialize in case $\varphi_a \equiv 0,\ \varphi_b \equiv 0$ to the *transversality conditions*.

If

$$S_i(t,y(t)) < 0 \qquad (a\le t\le b,\ i=1,\dots,s)\ ,$$

then no state constraints really occur, and the adjoint integral equation reduces to

$$\begin{aligned} p(t) &= -\lambda_0\varphi^\star_{bx}(\hat y(b)) - \beta^\star_x(\hat y(b))\lambda_b \\ &\quad + \int_t^b [\psi^\star_x(\tau,\hat y(\tau),\hat u(\tau))p(\tau) - \lambda_0\varphi^\star_x(\tau,\hat y(\tau),\hat u(\tau))]dt \\ &\quad \text{for } a\le t\le b\ , \\ p(a) &= \lambda_0\varphi^\star_{ax}(\hat y(a)) + \alpha^\star_x(\hat y(a))\lambda_a\ ,\end{aligned}$$

which is equivalent to the boundary value problem for an absolutely continuous n-vector function $p(\cdot)$:

$$\frac{d}{dt}p(t) = -[\psi^\star_x(t,\hat y(t),\hat u(t))p(t) - \lambda_0\varphi^\star_x(t,\hat y(t),\hat u(t))]$$

for almost all $t\in[a,b]$,

$$\begin{aligned} p(a) &= \lambda_0\varphi^\star_{ax}(\hat y(a)) + \alpha^\star_x(\hat y(a))\lambda_a\ , \\ p(b) &= -\lambda_0\varphi^\star_{bx}(\hat y(b)) - \beta^\star_x(\hat y(b))\lambda_b\ .\end{aligned}$$

Introducing again the Hamilton function

$$H(t,x,u,p,\lambda_0) = p^\star \psi(t,x,u) - \lambda_0 \varphi(t,x,u) \ ,$$

the adjoint integral equation gets the form

$$\begin{aligned} p(t) &= -\lambda_0 \varphi_{bx}^\star(\hat{y}(b)) - \beta_x^\star(\hat{y}(b))\lambda_b \\ &\quad + \int_t^b H_x(\tau,\hat{y}(\tau),\hat{u}(\tau),p(\tau),\lambda_0)^\star d\tau \\ &\quad - \int_t^b S_x^\star(\tau,\hat{y}(\tau)) d\nu(\tau) \qquad (a \le t \le b) \ , \\ p(a) &= \lambda_0 \varphi_{ax}^\star(\hat{y}(a)) + \alpha_x^\star(\hat{y}(a))\lambda_0 \ . \end{aligned}$$

If no state constraints are present, this reduces together with the system equation to the Hamiltonean system

$$\begin{aligned} \frac{d}{dt}\hat{y}(t) &= \frac{\partial H}{\partial p}(t,\hat{y}(t),\hat{u}(t),p(t),\lambda_0)^\star \ , \\ \frac{d}{dt}p(t) &= -\frac{\partial H}{\partial x}(t,\hat{y}(t),\hat{u}(t),p(t),\lambda_0)^\star \end{aligned}$$

for almost all $t \in [a,b]$.

The global maximum principle (iv) can be written as

$$\sup_{u \in U} H(t,\hat{y}(t),u,p(t),\lambda_0) = H(t,\hat{y}(t),\hat{u}(t),p(t),\lambda_0)$$

for almost all $t \in [a,b]$.

Proof of the Maximum Principle. The proof of this maximum principle can be found e.g. in [33]. ∎

In the light of these results, the discrete maximum principle can be regarded as an discrete approximation of the continuous maximum principle by explicit Euler method for the system equations, by the implicit Euler method for the adjoint integral resp. differential equation and by satisfaction of the first order necessary optimality condition for the maximization of the Hamiltonean with respect to u in all grid points.

Naturally, there are other possibilities for the discrete approximation of optimal trajectories and controls for the continuous optimal control problem:

- Choose other discretizations of the Hamiltonean system.
- Solve the discrete optimal control problem directly by nonlinear optimization methods.
- Omit (at least for a while) the objective function and analyze the resulting state constrained differential inclusion numerically.

Chapter 4

Direct Methods

We discuss direct methods for the numerical solution of smooth discrete dynamic optimization problems which can be written as a nonlinear programming problem of the form

$$
(NLP) \qquad \begin{array}{lll} \text{Minimize} & f(x) & \\ \text{w. r. t.} & x \in \mathbb{R}^n & \\ \text{subject to} & g_i(x) \leq 0, & i = 1, \ldots, m, \\ & h_j(x) = 0, & j = 1, \ldots, p. \end{array}
$$

We present necessary and sufficient conditions for a local minimum of (NLP). Due to lack of time and space, we analyze only the Lagrange-Newton- resp. SQP-Method as numerical solution methods for (NLP) more closely.

We used the monographs [3], [25], [13], [58], [39], [18], [1], [19].

4.1 Necessary Conditions

Our aim is to derive **necessary and sufficient conditions** for a local minimum of (NLP). Necessary conditions are based on the assumption that a local minimum $\hat{x}$ of (NLP) is known already. Then, conditions are derived

which have to be fulfilled at $\hat{x}$. For example, for unconstrained smooth optimization problems the condition $f'(\hat{x}) = 0$ is necessary for a local minimum. Conversely, given a point which fulfills a necessary condition it is not possible in general to decide whether this point actually is optimal or not. For example, consider $f(x) = -x^2$. The necessary condition $f'(x) = 0$ is fulfilled in $x = 0$, but $x = 0$ is a maximum and not a minimum of f.

We introduce some notations and definitions. The set

$$\Sigma := \{x \in \mathbb{R}^n \mid g_i(x) \leq 0,\ i = 1, \ldots, m,\ h_j(x) = 0,\ j = 1, \ldots, p\}$$

is called **feasible or admissible set of (NLP)**.

4.1.1. Definition (Global Minimum, Local Minimum).

- $\hat{x} \in \Sigma$ *is called a* **global minimum of (NLP)**, *if*

$$f(\hat{x}) \leq f(x) \qquad \forall x \in \Sigma.$$

- $\hat{x} \in \Sigma$ *is called a* **local minimum of (NLP)**, *if there exists an open neighborhood* $U_\varepsilon(\hat{x}) := \{x \in \mathbb{R}^n \mid \|x - \hat{x}\| < \varepsilon\}$ *such that*

$$f(\hat{x}) \leq f(x) \qquad \forall x \in U_\varepsilon(\hat{x}) \cap \Sigma.$$

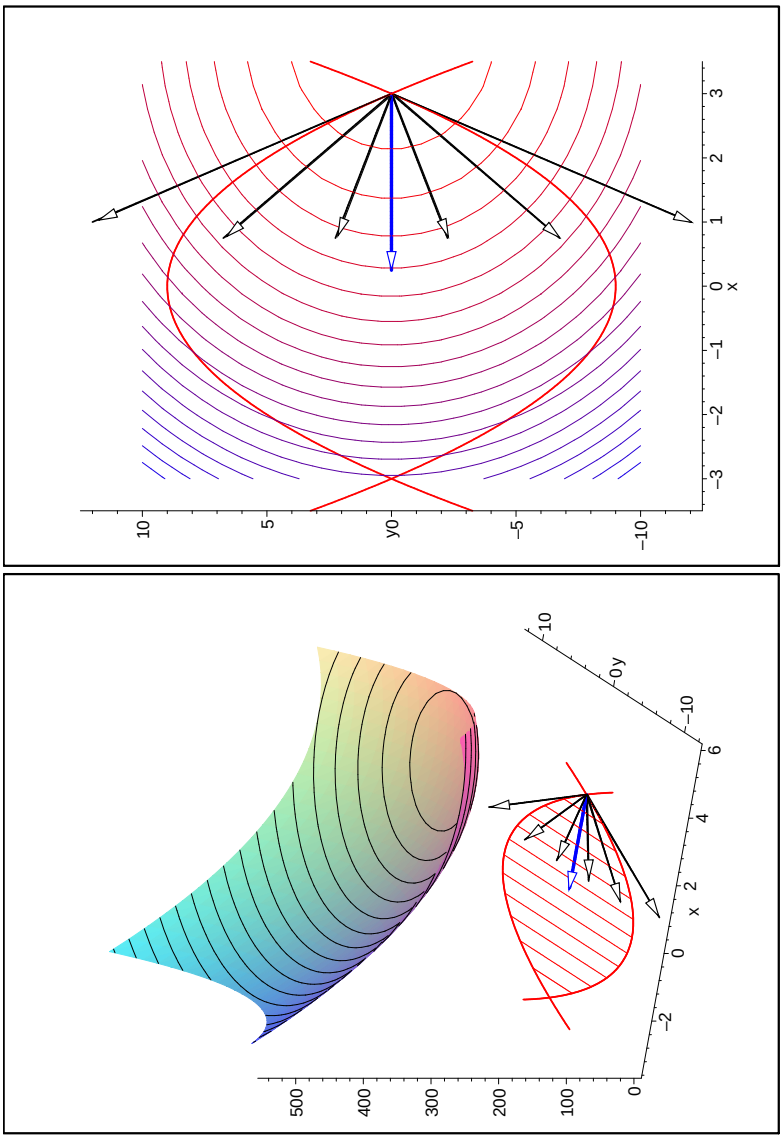
The function

$$L(x, \lambda, \mu) := f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{j=1}^{p} \mu_j h_j(x)$$

is called **Lagrange function of (NLP)**. The vectors $\lambda = (\lambda_1, \ldots, \lambda_m)^\top \in \mathbb{R}^m$ and $\mu = (\mu_1, \ldots, \mu_p)^\top \in \mathbb{R}^p$ respectively their components are called **Lagrange multipliers**.

4.1.2. Assumption.

- *The functions* $f : \mathbb{R}^n \to \mathbb{R}$, $g_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, \ldots, m$ *and* $h_j : \mathbb{R}^n \to \mathbb{R}$, $i = 1, \ldots, p$ *are continuously differentiable.*
- *The feasible set* Σ *is not empty and closed.*


10
5
y0
−5
−10
3
2
1
0
x
−1
−2
−3
10
0 y
−10
6
4
2
x
0
−2
500
400
300
200
100
0

From the figure we noticed that the directional derivative $f'(\hat{x}; d) = \nabla f(\hat{x})^\top d$ has to be nonnegative in a local minimum $\hat{x}$ for every feasible direction d, that is

$$f'(\hat{x}; d) = \nabla f(\hat{x})^\top d \geq 0 \qquad \forall d \text{ feasible.}$$

'The function f is nondecreasing in every feasible direction.'

We formalize the term 'feasible direction'.

4.1.3. Definition (Tangent Cone). *The set*

$$T(\Sigma, x) = \left\{ d \in \mathbb{R}^n \;\middle|\; \begin{array}{l} \text{there exist sequences } \{\alpha_k\}_{k\in\mathbb{N}}, \alpha_k \downarrow 0 \text{ and} \\ \{x_k\}_{k\in\mathbb{N}}, x_k \in \Sigma \text{ with } \lim_{k\to\infty} x_k = x, \text{ such that} \\ \lim_{k\to\infty} (x_k - x)/\alpha_k = d \text{ holds.} \end{array} \right\}$$

is called **tangent cone to Σ at x.**

What does this definition mean? If there exist a sequence $\{x_k\}$ in Σ, which approaches x arbitrarily close, then the direction

$$d = \lim_{k\to\infty} \frac{x_k - x}{\|x_k - x\|}$$

and all positive multiples belong to the tangent cone. If the limit fails to exist, all convergent subsequences belong to the tangent cone.

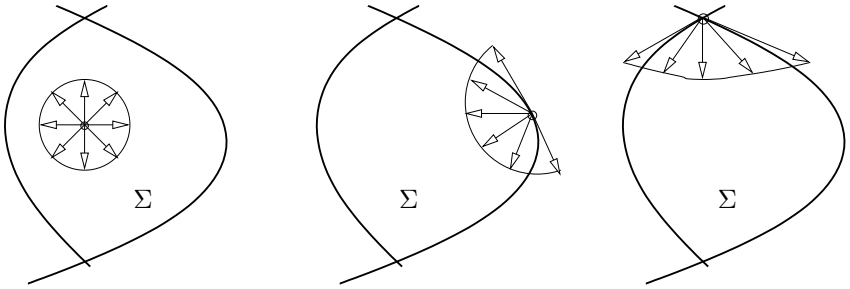


Figure 4.1: Tangent cone at different feasible points of the feasible set Σ.

4.1.4. Remarks. *A set $K \subseteq \mathbb{R}^n$ is called* **cone with apex** 0*, if $d \in K$ implies $\alpha d \in K$ for all $\alpha \geq 0$. It can be shown that the above defined tangent cone is a closed cone with apex zero.*

Now we are ready to formulate a first necessary condition.

4.1.5. Theorem. *Let $\hat{x}$ be a local minimum of (NLP). Then*

$$\nabla f(\hat{x})^\top d \geq 0 \qquad \forall d \in T(\Sigma, \hat{x}).$$

Proof. Let $d \in T(\Sigma, \hat{x})$. Then there exist sequences $\alpha_k \downarrow 0$ and $x_k \in \Sigma$, $x_k \to \hat{x}$ such that $(x_k - \hat{x})/\alpha_k \to d$. Since $\hat{x}$ is a local minimum and f is continuously differentiable, the mean value theorem yields

$$0 \leq f(x^k) - f(\hat{x}) = \nabla f(\hat{x} + \theta_k(x^k - \hat{x}))^\top \cdot (x^k - \hat{x}),$$

for all sufficient large $k \in \mathbb{N}$ and some $0 < \theta_k < 1$. Division by $\alpha_k > 0$ leads to

$$0 \leq \nabla f(\hat{x} + \theta_k(x^k - \hat{x}))^\top \cdot (x^k - \hat{x})/\alpha_k.$$

Since $x^k \to \hat{x}$ and $(x^k - \hat{x})/\alpha_k \to d$ hold for $k \to \infty$ we get

$$0 \leq \nabla f(\hat{x})^\top \cdot d.$$

Since $d \in T(\Sigma, \hat{x})$ was chosen arbitrarily the proof is complete. ∎

Unfortunately, it is hard to check this necessary condition, since the tangent cone is difficult to be handled. Next, we will derive more convenient necessary conditions, which involve the functions g_i, $i = 1, \ldots, m$ and h_j, $j = 1, \ldots, p$.

4.1.6. Definition (Index Set of Active Constraints, Linearizing Cone).

(i) The set $A(x) = \{i \mid g_i(x) = 0,\ 1 \leq i \leq m\}$ is called **index set of active inequality constraints**.

(ii) The set

$$\begin{aligned} T_{lin}(x) \quad = \quad & \{d \in \mathbb{R}^n \mid \nabla g_i(x)^\top d \leq 0,\ i \in A(x), \\ & \nabla h_j(x)^\top d = 0,\ j = 1, \ldots, p\} \end{aligned}$$

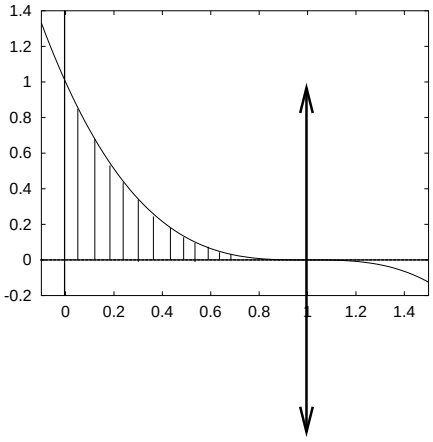
is called **linearizing cone for $g_i(x) \leq 0$, $i = 1, \ldots, m$ and $h_j(x) = 0$, $j = 1, \ldots, p$.**

4.1.7. Remarks. The conjecture $T(\Sigma, x) = T_{lin}(x)$ is **not true**. In general only $T(\Sigma, x) \subseteq T_{lin}(x)$ holds. Counter-example: $n = 2, m = 3, p = 0$ and

$$g(x_1, x_2) = \left(-(1 - x_1)^3 + x_2, -x_1, -x_2\right)^\top. \tag{4.1.1}$$

Let $x = (1, 0)^\top$. It is $T(\Sigma, x) = \{(\alpha, 0)^\top \mid \alpha \leq 0\}$ and $T_{lin}(x) = \{(\alpha, 0)^\top \mid \alpha \in \mathbb{R}\}$.

.



It is important to mention, that the tangent cone $T(\Sigma, x)$ is independent of the representation of the set Σ by inequality and equality constraints, whereas the linearizing cone $T_{lin}(x)$ depends on the functions g_i and h_j describing Σ. For instance, we may add the inequality constraint $g_4(x_1, x_2) = x_1 - 1 \leq 0$ to the constraints in (4.1.1) without changing the set Σ, that is $T(\Sigma, x)$ at $x = (1, 0)^\top$ remains unchanged. But, it turns out, that $T_{lin}(x)$ at $x = (1, 0)^\top$ for the constraints given by (4.1.1) and g_4 equals $T(\Sigma, x)$.

In the proof of the discrete maximum principle in Chapter 3 we already proved the necessary conditions of Fritz-John for a more general optimization problem than (NLP). For (NLP) the Fritz-John necessary conditions are summarized in the following theorem.

4.1.8. Theorem (Necessary Fritz-John Conditions). *Let $\hat{x}$ be a local minimum of (NLP). Then there exist multipliers $\lambda_0 \geq 0$, $\lambda = (\lambda_1, \ldots, \lambda_m)^\top \in$*

$\mathbb{R}^m$ *and* $\mu = (\mu_1, \ldots, \mu_p)^\top \in \mathbb{R}^p$ *not all zero, i.e.* $(\lambda_0, \lambda, \mu) \neq \Theta$, *such that*

$$\lambda_0 \nabla f(\hat{x}) + \sum_{i=1}^{m} \lambda_i \nabla g_i(\hat{x}) + \sum_{j=1}^{p} \mu_j \nabla h_j(\hat{x}) = 0_{\mathbb{R}^n}, \tag{4.1.2}$$

$$g_i(\hat{x}) \leq 0,\ i = 1, \ldots, m, \tag{4.1.3}$$

$$h_j(\hat{x}) = 0,\ j = 1, \ldots, p, \tag{4.1.4}$$

$$\lambda_i g_i(\hat{x}) = 0,\ i = 1, \ldots, m, \tag{4.1.5}$$

$$\lambda_i \geq 0,\ i = 1, \ldots, m. \tag{4.1.6}$$

Every vector $(x, \lambda_0, \lambda, \mu) \in \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^m \times \mathbb{R}^p$ *satisfying the Fritz-John conditions* $\lambda_0 \geq 0$, $(\lambda_0, \lambda, \mu) \neq \Theta$ *and (4.1.2)-(4.1.6) is called* **Fritz-John point** *of (NLP).*

Unfortunately, the case $\lambda_0 = 0$ may occur. In this case, the objective function f does not occur in the Fritz-John conditions. The main statement of the theorem is that there exists a **nontrivial** vector $(\lambda_0, \lambda, \mu) \neq \Theta$.

In the sequel we are interested in conditions such that $\lambda_0 \neq 0$ holds in the Fritz-John conditions. In this case we can chose $\lambda_0 = 1$ without loss of generality since the Lagrange multipliers appear linearly in the Fritz-John conditions.

4.1.9. Definition (Constraint Qualification of Abadie). *The constraints in (NLP) fulfill the* **constraint qualification of Abadie at** $\hat{x}$, *if*

$$T(\Sigma, \hat{x}) = T_{lin}(\hat{x})$$

holds.

It holds

4.1.10. Theorem (Necessary Karush-Kuhn-Tucker Conditions). *Let* $\hat{x}$ *be a local minimum of (NLP) and let the constraint qualification of Abadie be fulfilled at* $\hat{x}$. *Then there exist multipliers* $\lambda = (\lambda_1, \ldots, \lambda_m)^\top \in \mathbb{R}^m$ *and* $\mu = (\mu_1, \ldots, \mu_p)^\top \in \mathbb{R}^p$ *such that*

$$\nabla_x L(\hat{x}, \lambda, \mu) = 0_{\mathbb{R}^n}, \tag{4.1.7}$$

$$g_i(\hat{x}) \leq 0, \quad i = 1, \ldots, m, \tag{4.1.8}$$

$$h_j(\hat{x}) = 0, \quad j = 1, \ldots, p, \tag{4.1.9}$$

$$\lambda_i g_i(\hat{x}) = 0, \quad i = 1, \ldots, m, \tag{4.1.10}$$

$$\lambda_i \geq 0, \quad i = 1, \ldots, m. \tag{4.1.11}$$

Every vector $(x, \lambda, \mu) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p$ *satisfying the KKT conditions (4.1.7)-(4.1.11) is called* **stationary point** *or* **KKT point** *of (NLP).*

Proof. We apply Farkas' Lemma: There exists $z \geq \Theta$ with $Az = b$ if and only if $A^\top d \geq \Theta$ implies $b^\top d \geq 0$.

Since $\hat{x}$ is a local minimum of (NLP) Theorem 4.1.5 together with the constraint qualification of Abadie implies

$$d \in T(\Sigma, \hat{x}) = T_{lin}(\hat{x}) \quad \Rightarrow \quad \nabla f(\hat{x})^\top d \geq 0.$$

Let

$$A^\top := \begin{pmatrix} -\nabla g_i(\hat{x})^\top & (i \in A(\hat{x})) \\ \nabla h_j(\hat{x})^\top & (j = 1, \ldots, p) \\ -\nabla h_j(\hat{x})^\top & (j = 1, \ldots, p) \end{pmatrix}.$$

Then

$$d \in T_{lin}(\hat{x}) \Leftrightarrow A^\top d \geq \Theta \quad \Rightarrow \quad \underbrace{\nabla f(\hat{x})^\top}_{=:b^\top} d \geq 0.$$

Hence, we can apply Farkas' Lemma and the system $Az = b, z \geq \Theta$ has a solution. We denote the components of $z \geq \Theta$ by $\lambda_i \geq 0, i \in A(\hat{x})$ and $\mu_j^+, \mu_j^- \geq 0$, $j = 1, \ldots, p$. Thus, we get

$$\sum_{i \in A(\hat{x})} \lambda_i \left(-\nabla g_i(\hat{x})\right) + \sum_{j=1}^{p} \underbrace{(\mu_j^+ - \mu_j^-)}_{=:-\mu_j} \nabla h_j(\hat{x}) = \nabla f(\hat{x}).$$

With $\mu_j := \mu_j^- - \mu_j^+$ and $\lambda_i = 0$ for $i \notin A(\hat{x})$ it follows

$$\nabla_x L(\hat{x}, \lambda, \mu) = \nabla f(\hat{x}) + \sum_{i=1}^{m} \lambda_i \nabla g_i(\hat{x}) + \sum_{j=1}^{p} \mu_j h_j(\hat{x}) = 0_{\mathbb{R}^n}.$$

The condition $\lambda_i g_i(\hat{x}) = 0$ for $i = 1, \ldots, m$ is fulfilled since $\lambda_i \geq 0, i \in A(\hat{x})$ and $\lambda_i = 0, i \notin A(\hat{x})$ holds. This completes the proof. ∎

4.1.11. Remarks.

- The geometric meaning of the KKT conditions is that the gradient of the objective function f at a local minimum $\hat{x}$ can be expressed as

a nonnegative linear combination of the gradients of the active constraints. If the constraint qualification of Abadie is not fulfilled as in the first part of Remark 4.1.7, it is in general not possible to express the gradient of the objective function as a nonnegative linear combination of the gradients of the active constraints. For instance, consider the optimization problem of minimizing $f(x_1, x_2) = -x_1$ subject to the inequality constraints given by (4.1.1). The minimum is attained at $\hat{x} = (1, 0)^\top$ and we have $T(\Sigma, \hat{x}) \neq T_{lin}(\hat{x})$ and

$$\nabla f(\hat{x}) = (-1, 0)^\top,\ \nabla g_1(\hat{x}) = (0, 1)^\top,\ \nabla g_3(\hat{x}) = (0, -1)^\top.$$

Obviously, $\nabla f(\hat{x})$ is linearly independent of $\nabla g_1(\hat{x})$ and $\nabla g_3(\hat{x})$ and thus cannot be expressed as a linear combination thereof.

- In the special case $m = p = 0$ the KKT conditions reduce to the well known necessary condition $\nabla f(\hat{x}) = 0_{\mathbb{R}^n}$ for unconstrained problems. In the case $m = 0, p > 0$ we derive the multiplier rule of Lagrange

$$\nabla f(\hat{x}) + \sum_{j=1}^{p} \mu_j \nabla h_j(\hat{x}) = 0_{\mathbb{R}^n}.$$

- The KKT conditions are only necessary conditions but not sufficient. A counter-example is given by $f(x_1, x_2) = x_1, g_1(x_1, x_2) = (x_1 - 4)^2 + x_2^2 - 16, h_1(x_1, x_2) = (x_1 - 3)^2 + (x_2 - 2)^2 - 13, m = p = 1$. In this example there are three KKT points, one of which is a local minimum, one is the global minimum, and one is the global maximum.

In practice, the Abadie constraint qualification turns out to be difficult to be checked. Fortunately, there are alternative constraint qualifications. All of the following conditions imply the constraint qualification of Abadie.

(i) **Constraint qualification of Mangasarian-Fromowitz:**

(a) The gradients $\nabla h_j(\hat{x})$, $j = 1, \ldots, p$ are linearly independent;

(b) There exists a vector $d \in \mathbb{R}^n$ with

$$\nabla g_i(\hat{x})^\top d < 0 \text{ for } i \in A(\hat{x}) \text{ and } \nabla h_j(\hat{x})^\top d = 0 \text{ for } j = 1, \ldots, p.$$

It is comparatively easy to see that the Fritz-John conditions cannot hold with $\lambda_0 = 0$ if the constraint qualification of Mangasarian-Fromowitz holds: Assume, that the constraint qualification of

Mangasarian-Fromowitz holds and that the Fritz-John conditions hold at $\hat{x}$ with $\lambda_0 = 0$, that is, there exist multipliers $\lambda_i \geq 0$, $i = 1, \ldots, m$, μ_j, $j = 1, \ldots, p$ not all zero with $\lambda_i g_i(\hat{x}) = 0$ for all $i = 1, \ldots, m$ and

$$\begin{aligned} 0_{\mathbb{R}^n} &= \underbrace{\lambda_0}_{=0} \nabla f(\hat{x}) + \sum_{i=1}^{m} \lambda_i \nabla g_i(\hat{x}) + \sum_{j=1}^{p} \mu_j \nabla h_j(\hat{x}) \\ &= \sum_{i=1}^{m} \lambda_i \nabla g_i(\hat{x}) + \sum_{j=1}^{p} \mu_j \nabla h_j(\hat{x}). \end{aligned} \tag{4.1.12}$$

Multiplying the latter equation with d from (b) yields

$$0 = \sum_{i=1}^{m} \underbrace{\lambda_i}_{\substack{\geq 0, i \in A(\hat{x}) \\ = 0, i \notin A(\hat{x})}} \underbrace{d^\top \nabla g_i(\hat{x})}_{< 0, i \in A(\hat{x})} + \sum_{j=1}^{p} \mu_j \underbrace{d^\top \nabla h_j(\hat{x})}_{=0} = \sum_{i \in A(\hat{x})} \underbrace{\lambda_i}_{\geq 0} \underbrace{d^\top \nabla g_i(\hat{x})}_{< 0}$$

Hence, $\lambda_i = 0$ for all $i \in A(\hat{x})$. Since $\lambda_i = 0$ for $i \notin A(\hat{x})$ we have $\lambda_i = 0$ for all $i = 1, \ldots, m$. From (4.1.12) it follows

$$0_{\mathbb{R}^n} = \sum_{j=1}^{p} \mu_j \nabla h_j(\hat{x}).$$

According to (a) the vectors $\nabla h_j(\hat{x})$ are linearly independent. It follows $\mu_j = 0$ for all $j = 1, \ldots, m$. Hence, we have $(\lambda_0, \lambda, \mu) = \Theta$, which is a contradiction to the statement that not all multipliers are zero.

(ii) **Linear independence constraint qualification**:
The gradients $\nabla g_i(\hat{x})$, $i \in A(\hat{x})$ and $\nabla h_j(\hat{x})$, $j = 1, \ldots, p$ are linearly independent.
In this case the Lagrange multipliers are unique.

Again, it is easy to see that the Fritz-John conditions cannot hold with $\lambda_0 = 0$, if the above gradients are linearly independent. The linear independence together with Equation (4.1.12) implies $\lambda_i = \mu_j = 0$ for all $i = 1, \ldots, m$, $j = 1, \ldots, p$. Again, this is a contradiction to $(\lambda_0, \lambda, \mu) \neq \Theta$.

The uniqueness of the Lagrange multipliers follows from the following considerations. Assume, that there are Lagrange multipliers λ_i, $i = 1, \ldots, m$, μ_j, $j = 1, \ldots, p$ and $\tilde{\lambda}_i$, $i = 1, \ldots, m$, $\tilde{\mu}_j$, $j = 1, \ldots, p$

satisfying the KKT conditions. In particular we have

$$0_{\mathbb{R}^n} = \nabla f(\hat{x}) + \sum_{i=1}^{m} \lambda_i \nabla g_i(\hat{x}) + \sum_{j=1}^{p} \mu_j \nabla h_j(\hat{x}),$$

$$0_{\mathbb{R}^n} = \nabla f(\hat{x}) + \sum_{i=1}^{m} \tilde{\lambda}_i \nabla g_i(\hat{x}) + \sum_{j=1}^{p} \tilde{\mu}_j \nabla h_j(\hat{x}).$$

Subtracting these equations leads to

$$0_{\mathbb{R}^n} = \sum_{i=1}^{m} (\lambda_i - \tilde{\lambda}_i) \nabla g_i(\hat{x}) + \sum_{j=1}^{p} (\mu_j - \tilde{\mu}_j) \nabla h_j(\hat{x}).$$

For inactive inequality constraints we have $\lambda_i = \tilde{\lambda}_i = 0$, $i \notin A(\hat{x})$. Since the gradients of the active constraints are assumed to be linearly independent, it follows $0 = \lambda_i - \tilde{\lambda}_i$, $i \in A(\hat{x})$ and $0 = \mu_j - \tilde{\mu}_j$, $j = 1, \ldots, p$. Hence, the Lagrange multipliers are unique.

(iii) **Constraint qualification of Kuhn-Tucker**:
For all $d \in T_{lin}(\hat{x})$ there exists a continuously differentiable curve $\varphi : [0, 1] \to \Sigma$ emanating from $\varphi(0) = \hat{x}$ with direction $\varphi'(0) = d$.

The constraint qualification of Kuhn-Tucker implies the constraint qualification of Abadie. We show: If the constraint qualification of Abadie does not hold, then the constraint qualification of Kuhn-Tucker is not fulfilled.

So, let the constraint qualification of Abadie not be fulfilled. Since, $T(\Sigma, \hat{x}) \subseteq T_{lin}(\hat{x})$ there exists a direction $d \in T_{lin}(\hat{x})$ with $d \notin T(\Sigma, \hat{x})$. It follows that for all sequences $\alpha_k \downarrow 0$ and all sequences $x^k \to \hat{x}$, $x^k \in \Sigma$ it holds $\lim_{k\to\infty}(x^k - \hat{x})/\alpha_k \neq d$. Assume, that the constraint qualification of Kuhn-Tucker is fulfilled. Then there is a curve $\varphi : [0, 1] \to \Sigma$ with $\varphi(0) = \hat{x}$ and $\varphi'(0) = d$. Let $\alpha_k \downarrow 0$, $\alpha_k \leq 1$ be an arbitrary sequence and $x^k := \varphi(\alpha_k)$. It holds $x^k \in \Sigma$ and, since φ is continuous, $x^k = \varphi(\alpha_k) \to \varphi(0) = \hat{x}$. Furthermore, since φ is continuously differentiable, we have

$$\frac{x^k - \hat{x}}{\alpha_k} = \frac{\varphi(\alpha_k) - \varphi(0)}{\alpha_k} \to \varphi'(0) = d.$$

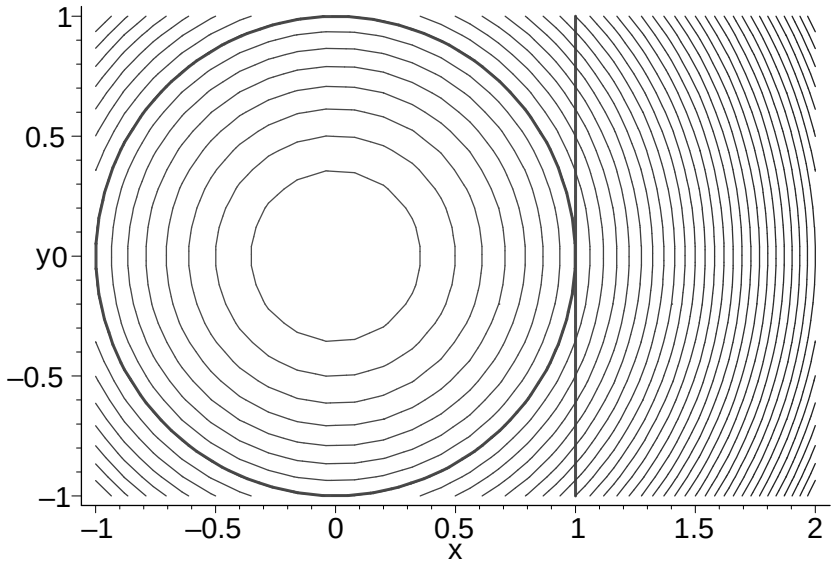
But this is a contradiction to $d \notin T(\Sigma, \hat{x})$, since $\lim_{k\to\infty}(x^k - \hat{x})/\alpha_k \neq d$ for all sequences $\alpha_k \downarrow 0$ and all sequences $x^k \to \hat{x}$, $x^k \in \Sigma$.

Notice, that the Fritz-John conditions with $\lambda_0 = 1$ are the KKT conditions. If any of the above constraint qualifications holds we may set – without loss of generality – $\lambda_0 = 1$ in the Fritz-John conditions.

Example 4.1.1
Consider the nonlinear programming problem

$$\text{Minimize } x_1^2 + x_2^2 \text{ s. t. } 1 - x_1 \leq 0,\ 1 - x_1^2 - x_2^2 \leq 0.$$

The feasible set is given by $\Sigma = \{(x_1, x_2)^\top \in \mathbb{R}^2 \mid x_1 \geq 1\}$. From the graphics it is easy to see that the minimum is attained at $\hat{x} = (1, 0)^\top$.



Lagrange function:

$$L(x_1, x_2, \lambda_1, \lambda_2) = x_1^2 + x_2^2 + \lambda_1(1 - x_1) + \lambda_2(1 - x_1^2 - x_2^2).$$

KKT conditions:

$$\begin{aligned} 2x_1 - \lambda_1 - 2\lambda_2 x_1 &= 0, \\ 2x_2 - 2\lambda_2 x_2 &= 0, \\ 1 - x_1 &\leq 0, \\ 1 - x_1^2 - x_2^2 &\leq 0, \\ \lambda_1(1 - x_1) &= 0, \\ \lambda_2(1 - x_1^2 - x_2^2) &= 0, \\ \lambda_1, \lambda_2 &\geq 0. \end{aligned}$$

KKT points:

$$(\hat{x}_1, \hat{x}_2, \hat{\lambda}_1, \hat{\lambda}_2) = (1, 0, 2(1-\alpha), \alpha), \qquad 0 \leq \alpha \leq 1.$$

The optimal solution $\hat{x} = (1,0)^\top$ is unique, whereas the Lagrange multipliers are not uniquely determined.

Which constraint qualifications are fulfilled at $\hat{x}$?

- **Abadie:**
  Tangent cone:

  $$T(\Sigma, \hat{x}) = \{(d_1, d_2)^\top \in \mathbb{R}^2 \mid d_1 \geq 0\}.$$

  The index set of active inequality constraints is $A(\hat{x}) = \{1, 2\}$ and

  $$\nabla f(x_1, x_2) = \begin{pmatrix} 2 \\ 0 \end{pmatrix}, \ \nabla g_1(x^*) = \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \ \nabla g_2(x^*) = \begin{pmatrix} -2 \\ 0 \end{pmatrix}.$$

  Linearizing cone:

  $$T_{lin}(x^*) = \{(d_1, d_2)^\top \in \mathbb{R}^2 \mid d_1 \geq 0\}.$$

  Thus, $T_{lin}(\hat{x}) = T(\Sigma, \hat{x})$ and the constraint qualification of Abadie is fulfilled.

- **Mangasarian-Fromowitz:**
  For the direction $d = (1,0)^\top$ we have $\nabla g_1(\hat{x})^\top d = -1 < 0$ and $\nabla g_2(\hat{x})^\top d = -2 < 0$. Thus, the Mangasarian-Fromowitz condition is fulfilled.

- **Kuhn-Tucker:**
  Let $d = (d_1, d_2)^\top \in T_{lin}(\hat{x})$. The line

$$\varphi(t) = \begin{pmatrix} 1 \\ 0 \end{pmatrix} + t \cdot \begin{pmatrix} d_1 \\ d_2 \end{pmatrix}, \quad 0 \le t \le 1$$

  is a continuously differentiable curve with $\varphi(0) = \hat{x}$ and $\varphi'(0) = d$. With $t \ge 0$ and $d_1 \ge 0$ it follows $1+td_1 \ge 1$ and $(1+td_1)^2+(td_2)^2 \ge 1$, i.e. $\varphi(t) \in \Sigma$. Hence, the constraint qualification of Kuhn-Tucker is fulfilled.

- **Linear independence:**
  The condition is not fulfilled since $\nabla g_1(\hat{x})$ and $\nabla g_2(\hat{x})$ are not linearly independent.

Finally, we investigate a second order necessary condition and assume that $(\hat{x}, \hat{\lambda}, \hat{\mu})$ is a KKT point of (NLP). We need the cone

$$\begin{aligned} T_2(\hat{x}) \quad := \quad & \{d \in \mathbb{R}^n \mid \nabla g_i(\hat{x})^\top d \le 0,\ i \in A(\hat{x}), \hat{\lambda}_i = 0, \\ & \nabla g_i(\hat{x})^\top d = 0,\ i \in A(\hat{x}), \hat{\lambda}_i > 0, \\ & \nabla h_j(\hat{x})^\top d = 0,\ j = 1, \dots, p\}. \end{aligned}$$

4.1.12. Theorem (Second Order Necessary Condition). *Let $f, g_i, i = 1, \dots, m$, and h_j, $j = 1, \dots, p$ be twice continuously differentiable. Let $\hat{x}$ be a local minimum of (NLP) and let the gradients $\nabla g_i(\hat{x})$, $i \in A(\hat{x})$ and $\nabla h_j(\hat{x})$, $j = 1, \dots, p$ be linearly independent. Then it holds*

$$d^\top L''_{xx}(\hat{x}, \hat{\lambda}, \hat{\mu}) d \ge 0 \qquad \forall d \in T_2(\hat{x}),$$

where $(\hat{x}, \hat{\lambda}, \hat{\mu})$ denotes a KKT point of (NLP).

Proof. Let $d \in T_2(\hat{x})$, $d \ne 0$. Let

$$\begin{aligned} A_0(\hat{x}) &:= \{i \in A(\hat{x}) \mid \hat{\lambda}_i = 0\}, \\ A_>(\hat{x}) &:= \{i \in A(\hat{x}) \mid \hat{\lambda}_i > 0\}, \\ A_0^<(\hat{x}) &:= \{i \in A_0(\hat{x}) \mid \nabla g_i(\hat{x})^\top d < 0\}, \\ A_0^=(\hat{x}) &:= \{i \in A_0(\hat{x}) \mid \nabla g_i(\hat{x})^\top d = 0\}. \end{aligned}$$

Since the vectors $\nabla g_i(\hat{x})$, $i \in A_>(\hat{x}) \cup A_0^=(\hat{x})$ and $\nabla h_j(\hat{x})$, $j = 1, \ldots, p$ are linearly independent, it can be shown similarly to the proof of the Fritz-John conditions, that there exist a $\varepsilon > 0$ and a twice continuously differentiable curve $x : (-\varepsilon, \varepsilon) \to \mathbb{R}^n$ with $x(0) = \hat{x}$, $x'(0) = d$, $g_i(x(t)) = 0$, $i \in A_>(\hat{x}) \cup A_0^=(\hat{x})$, $h_j(x(t)) = 0$, $j = 1, \ldots, p$ for all $t \in (-\varepsilon, \varepsilon)$, and $x(t) \in \Sigma$ for all $t \in [0, \varepsilon)$. Let

$$\varphi(t) := L(x(t), \hat{\lambda}, \hat{\mu}), \qquad t \in (-\varepsilon, \varepsilon).$$

φ is twice continuously differentiable with

$$\varphi'(t) = x'(t)^\top L'_x(x(t), \hat{\lambda}, \hat{\mu})$$

and

$$\varphi''(t) = x''(t)^\top L'_x(x(t), \hat{\lambda}, \hat{\mu}) + x'(t)^\top L''_{xx}(x(t), \hat{\lambda}, \hat{\mu}) x'(t).$$

Since $(\hat{x}, \hat{\lambda}, \hat{\mu})$ is a KKT point it follows

$$\varphi'(0) = d^\top L'_x(\hat{x}, \hat{\lambda}, \hat{\mu}) = 0$$

and

$$\varphi''(0) = d^\top L''_{xx}(\hat{x}, \hat{\lambda}, \hat{\mu}) d.$$

Assume now, that $\varphi''(0) < 0$. Then, the continuity of φ'' implies $\varphi''(t) < 0$ for all sufficiently small $t \in (-\varepsilon, \varepsilon)$. Taylor expansion of φ at $t = 0$ yields

$$\varphi(t) = \varphi(0) + t\varphi'(0) + \frac{t^2}{2}\varphi''(\xi_t)$$

for all $t \in (-\varepsilon, \varepsilon)$ and a point ξ_t between 0 and t. From $\varphi'(0) = 0$ and $\varphi''(\xi_t) < 0$ for all sufficient small $t \in (-\varepsilon, \varepsilon)$ we have $\varphi(t) < \varphi(0)$. Since

$$\varphi(0) = L(\hat{x}, \hat{\lambda}, \hat{\mu}) = f(\hat{x})$$

and

$$\varphi(t) = L(x(t), \hat{\lambda}, \hat{\mu}) = f(x(t))$$

(to see this, exploit the properties of $x(t)$ and that $(\hat{x}, \hat{\lambda}, \hat{\mu})$ is a KKT point), it follows $f(x(t)) < f(\hat{x})$ for all sufficiently small $t \in [0, \varepsilon)$. Since $x(t) \in \Sigma$ this contradicts the minimality of $\hat{x}$. ∎

4.2 Sufficient Conditions

To decide, whether a given point that fulfills the necessary conditions is optimal we need **sufficient conditions**. For unconstrained smooth optimization problems, the condition $f''(x) > 0$ where $f''(x)$ denotes the Hessian of f at x is a sufficient condition. But it is not necessary as the example $f(x) = x^4$ shows.

A sufficient condition in terms of the tangent cone is given by the following theorem.

4.2.1. Theorem. *Let $\hat{x} \in \Sigma$ and f be continuously differentiable at $\hat{x}$. Let*

$$\nabla f(\hat{x})^\top d > 0 \qquad \forall d \in T(\Sigma, \hat{x}) \setminus \{0_{\mathbb{R}^n}\}.$$

Then there exists a neighborhood U of $\hat{x}$ and some $\alpha > 0$ such that

$$f(x) \geq f(\hat{x}) + \alpha \|x - \hat{x}\| \qquad \forall x \in \Sigma \cap U.$$

Proof. Let us assume that the statement is wrong. Then for any open ball centered at $\hat{x}$ with radius $1/i$ there exists a point $x^i \in \Sigma$ with

$$f(x^i) - f(\hat{x}) < \frac{1}{i}\|x^i - \hat{x}\|, \quad \|x^i - \hat{x}\| < \frac{1}{i}, \quad \forall i \in \mathbb{N}. \tag{4.2.1}$$

Since the unit ball w. r. t. $\|\cdot\|$ is compact in $\mathbb{R}^n$, there exists a convergent subsequence of $\{x^i\}$ with

$$\lim_{k\to\infty} \frac{x^{i_k} - \hat{x}}{\|x^{i_k} - \hat{x}\|} = \hat{d}, \quad \lim_{k\to\infty} \|x^{i_k} - \hat{x}\| = 0,$$

that is, $\hat{d} \in T(\Sigma, \hat{x}) \setminus \{0\}$. Taking the limit in (4.2.1) yields

$$\nabla f(\hat{x})^\top \hat{d} = \lim_{k\to\infty} \frac{f(x^{i_k}) - f(\hat{x})}{\|x^{i_k} - \hat{x}\|} \leq 0,$$

which contradicts the assumption. ∎

A second order sufficient condition is given by

4.2.2. Theorem (Second Order Sufficient Condition). *Let f, g_i, $i = 1,\dots,m$, and h_j, $j = 1,\dots,p$ be twice continuously differentiable. Let $(\hat{x},\hat{\lambda},\hat{\mu})$ be a KKT point of (NLP) with*

$$d^\top L''_{xx}(\hat{x},\hat{\lambda},\hat{\mu})d > 0 \qquad \forall d \in T_2(\hat{x}),\ d \neq 0_{\mathbb{R}^n}. \tag{4.2.2}$$

Then there exists a neighborhood U of $\hat{x}$ and some $\alpha > 0$ such that

$$f(x) \geq f(\hat{x}) + \alpha\|x-\hat{x}\|^2 \qquad \forall x \in \Sigma \cap U.$$

Proof. **I**. Let $d \in T(\Sigma,\hat{x})$, $d \neq 0_{\mathbb{R}^n}$. Then there exist sequences $x^k \in \Sigma$, $x^k \to \hat{x}$ and $\alpha_k \downarrow 0$ with

$$\lim_{k\to\infty} \frac{x^k - \hat{x}}{\alpha_k} = d.$$

For $i \in A(\hat{x})$ we have

$$0 \geq \frac{g_i(x^k) - g_i(\hat{x})}{\alpha_k} = g'(\xi_k)\frac{x^k-\hat{x}}{\alpha_k} \to g'(\hat{x})d$$

by the mean-value theorem. Similarly, we show $h'_j(\hat{x})d = 0$ for $j = 1,\dots,p$. Since $(\hat{x},\hat{\lambda},\hat{\mu})$ is a KKT point with $\hat{\lambda}_i = 0$, if $g_i(\hat{x}) < 0$, we obtain

$$f'(\hat{x})d = -\sum_{i=1}^{m} \hat{\lambda}_i g'_i(\hat{x})d - \sum_{j=1}^{p} \hat{\mu}_j h'_j(\hat{x})d \geq 0\ .$$

Hence, $\hat{x}$ fulfills the first order necessary condition $f'(\hat{x})d \geq 0$ for all $d \in T(\Sigma,\hat{x})$.

II. Assume, that the statement of the theorem is wrong. Then for any ball around $\hat{x}$ with radius $1/k$ there exist a point $x^k \in \Sigma$ with

$$f(x^k) - f(\hat{x}) < \frac{1}{k}\|x^k-\hat{x}\|^2\ , \quad \|x^k-\hat{x}\| \leq \frac{1}{k} \qquad \forall k \in \mathbb{N}\ . \tag{4.2.3}$$

Since the unit ball w. r. t. $\|\cdot\|$ is compact in $\mathbb{R}^n$, there exists a convergent subsequence of $\{x^k\}$ with

$$\lim_{k\to\infty} \frac{x^k-\hat{x}}{\|x^k-\hat{x}\|} = d, \quad \lim_{k\to\infty} \|x^k-\hat{x}\| = 0.$$

Notice, that we identified the convergent subsequence with the sequence $\{x^k\}$ for notational convenience. Hence, $d \in T(\Sigma,\hat{x}) \setminus \{0_{\mathbb{R}^n}\}$. Taking the limit in (4.2.3) yields

$$f'(\hat{x})d = \lim_{k\to\infty} \frac{f(x^k)-f(\hat{x})}{\|x^k-\hat{x}\|} \leq 0.$$

Together with I. we have

$$f'(\hat{x})d = 0 \ .$$

III. Since $\hat{x}$ is a KKT point, it follows

$$f'(\hat{x})d = -\sum_{i\in I(\hat{x})} \underbrace{\hat{\lambda}_i}_{\geq 0} \underbrace{g_i'(\hat{x})d}_{\leq 0} - \sum_{j=1}^{p} \hat{\mu}_j \underbrace{h_j'(\hat{x})d}_{=0} = 0 \ .$$

Thus, it is $g_i'(\hat{x})d = 0$, if $\hat{\lambda}_i > 0$. Hence, $d \in T_2(\hat{x})$.

According to (4.2.3) it holds

$$\lim_{k\to\infty} \frac{f(x^k) - f(\hat{x})}{\|x^k - \hat{x}\|^2} \leq \lim_{k\to\infty} \frac{1}{k} = 0 \tag{4.2.4}$$

for the direction d. Furthermore, it is

$$\begin{aligned}
L(x^k, \hat{\lambda}, \hat{\mu}) &= f(x^k) + \sum_{i=1}^{m} \hat{\lambda}_i g_i(x^k) + \sum_{j=1}^{p} \hat{\mu}_j h_j(x^k) \leq f(x^k) \ , \\
L(\hat{x}, \hat{\lambda}, \hat{\mu}) &= f(\hat{x}) + \sum_{i=1}^{m} \hat{\lambda}_i g_i(\hat{x}) + \sum_{j=1}^{p} \hat{\mu}_j h_j(\hat{x}) = f(\hat{x}) \ , \\
L_x'(\hat{x}, \hat{\lambda}, \hat{\mu}) &= f'(\hat{x}) + \sum_{i=1}^{m} \hat{\lambda}_i g_i'(\hat{x}) + \sum_{j=1}^{p} \hat{\mu}_j h_j'(\hat{x}) = 0_{\mathbb{R}^n}^{\top} \ .
\end{aligned}$$

Taylor expansion of L w. r. t. to x at $\hat{x}$ yields

$$\begin{aligned}
f(x^k) \geq L(x^k, \hat{\lambda}, \hat{\mu}) &= L(\hat{x}, \hat{\lambda}, \hat{\mu}) + L_x'(\hat{x}, \hat{\lambda}, \hat{\mu})(x^k - \hat{x}) \\
&\quad + \frac{1}{2}(x^k - \hat{x})^{\top} L_{xx}''(\xi^k, \hat{\lambda}, \hat{\mu})(x^k - \hat{x}) \\
&= f(\hat{x}) + \frac{1}{2}(x^k - \hat{x})^{\top} L_{xx}''(\xi^k, \hat{\lambda}, \hat{\mu})(x^k - \hat{x}),
\end{aligned}$$

where ξ^k is some point between $\hat{x}$ and x^k. Division by $\|x^k - \hat{x}\|^2$ and taking the limit, yields together with (4.2.4)

$$0 \geq \frac{1}{2} d^{\top} L_{xx}''(\hat{x}, \hat{\lambda}, \hat{\mu}) d.$$

This contradicts the assumption $d^{\top} L_{xx}''(\hat{x}, \hat{\lambda}, \hat{\mu}) d > 0$ for all $d \in T_2(\hat{x})$. ∎

4.3 Perturbed Nonlinear Optimization Problems and Sensitivity

In view of the upcoming SQP method we need results about the sensitivity of solutions under perturbations. We investigate parametric optimization problems

$$
(NLP(w)) \qquad \begin{array}{lll} \text{Minimize} & f(x,w) & \\ \text{w. r. t.} & x \in \mathbb{R}^n & \\ \text{subject to} & g_i(x,w) \leq 0, & i = 1,\dots,m, \\ & h_j(x,w) = 0, & j = 1,\dots,p. \end{array}
$$

Herein, $f, g_1, \dots, g_m, h_1, \dots, h_p : \mathbb{R}^n \times \mathbb{R}^q \to \mathbb{R}$ are sufficiently smooth functions and $w \in \mathbb{R}^q$ denotes a parameter. Let $\hat{w}$ denote a fixed nominal parameter. We are interested in the behavior of the optimal solutions $\hat{x}(w)$ as functions of w in a neighborhood of the nominal parameter $\hat{w}$. The following result is based on [12], [13], [58]. The admissible set of (NLP(w)) is defined by

$$\Sigma(w) := \{x \in \mathbb{R}^n \mid g_i(x,w) \leq 0,\ i = 1,\dots,m,\ h_j(x,w) = 0,\ j = 1,\dots,p\}.$$

The index set of active inequality constraints is given by

$$A(x,w) = \{i \mid g_i(x,w) = 0,\ 1 \leq i \leq m\}.$$

4.3.1. Definition (Strongly Regular Local Solution). *A local minimum $\hat{x}$ of $(NLP(w))$ is called* **strongly regular** *if the following properties hold:*

- *$\hat{x}$ is admissible, i.e. $\hat{x} \in \Sigma(w)$.*
- *$\hat{x}$ fulfills the linear independence constraint qualification, i.e. the gradients $\nabla_x g_i(\hat{x})$, $i \in A(\hat{x},w)$, $\nabla_x h_j(\hat{x},w)$, $j = 1,\dots,p$ are linearly independent.*
- *The KKT conditions hold at $(\hat{x}, \hat{\lambda}, \hat{\mu})$.*
- *The strict complementarity condition holds, i.e. $\hat{\lambda}_i - g_i(\hat{x},w) > 0$ for all $i = 1,\dots,m$.*

- *The second order sufficient condition (4.2.2) holds.*

4.3.2. Theorem (Sensitivity Theorem). *Let $f, g_1, \ldots, g_m, h_1, \ldots, h_p : \mathbb{R}^n \times \mathbb{R}^q \to \mathbb{R}$ be twice continuously differentiable and $\hat{w}$ a nominal parameter. Let $\hat{x}$ be a strongly regular local minimum of $(NLP(\hat{w}))$, $\hat{\lambda}$, $\hat{\mu}$ denote the corresponding Lagrange multipliers. Then there exist neighborhoods $V_\epsilon(\hat{w})$ and $U_\delta(\hat{x}, \hat{\lambda}, \hat{\mu})$, such that $(NLP(w))$ has a unique strongly regular local minimum $(x(w), \lambda(w), \mu(w)) \in U_\delta(\hat{x}, \hat{\lambda}, \hat{\mu})$ for each $w \in V_\epsilon(\hat{w})$. Furthermore, it holds $A(\hat{x}, \hat{w}) = A(x(w), w)$. In addition, $(x(w), \lambda(w), \mu(w))$ is continuously differentiable w. r. t. w with*

$$\begin{pmatrix} \frac{dx}{dw}(\hat{w}) \\ \frac{d\lambda}{dw}(\hat{w}) \\ \frac{d\mu}{dw}(\hat{w}) \end{pmatrix} = - \begin{pmatrix} L''_{xx} & (g'_x)^\top & (h'_x)^\top \\ \hat{\Lambda} \cdot g'_x & \hat{\Gamma} & \Theta \\ h'_x & \Theta & \Theta \end{pmatrix}^{-1} \cdot \begin{pmatrix} L''_{xw} \\ \hat{\Lambda} \cdot g'_w \\ h'_w \end{pmatrix} \tag{4.3.1}$$

where $\hat{\Lambda} = \operatorname{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_m)$, $\hat{\Gamma} = \operatorname{diag}(g_1, \ldots, g_m)$. All functions and their derivatives are evaluated at $(\hat{x}, \hat{\lambda}, \hat{\mu}, \hat{w})$.

Proof. Consider the nonlinear equation

$$F(x, \lambda, \mu, w) := \begin{pmatrix} L'_x(x, \lambda, \mu, w)^\top \\ \Lambda \cdot g(x, w) \\ h(x, w) \end{pmatrix} = 0_{\mathbb{R}^{n+m+p}}, \tag{4.3.2}$$

where $\Lambda := \operatorname{diag}(\lambda_1, \ldots, \lambda_m)$. F is continuously differentiable and it holds $F(\hat{x}, \hat{\lambda}, \hat{\mu}, \hat{w}) = 0_{\mathbb{R}^{n+m+p}}$. We intend to apply the implicit function theorem. Hence, we have to show the non-singularity of

$$F'_x(\hat{x}, \hat{\lambda}, \hat{\mu}, \hat{w}) = \begin{pmatrix} L''_{xx}(\hat{x}, \hat{\lambda}, \hat{\mu}, \hat{w}) & (g'_x(\hat{x}, \hat{w}))^\top & (h'_x(\hat{x}, \hat{w}))^\top \\ \hat{\Lambda} \cdot g'_x(\hat{x}, \hat{w}) & \hat{\Gamma} & \Theta \\ h'_x(\hat{x}, \hat{w}) & \Theta & \Theta \end{pmatrix}.$$

In order to show this, we assume without loss of generality, that the index set of active inequality constraints is given by $A(\hat{x}, \hat{w}) = \{l + 1, \ldots, m\}$,

where l denotes the number of inactive inequality constraints. Then, the strict complementarity condition implies

$$\hat{\Lambda} = \begin{pmatrix} \Theta & \Theta \\ \Theta & \hat{\Lambda}_2 \end{pmatrix}, \qquad \hat{\Gamma} = \begin{pmatrix} \hat{\Gamma}_1 & \Theta \\ \Theta & \Theta \end{pmatrix},$$

with non-singular matrices $\hat{\Lambda}_2 := \operatorname{diag}(\hat{\lambda}_{l+1}, \ldots, \hat{\lambda}_m)$ and $\hat{\Gamma}_1 := \operatorname{diag}(g_1(\hat{x}, \hat{w}), \ldots, g_l(\hat{x}, \hat{w}))$. Consider the linear equation system

$$\begin{pmatrix} L''_{xx}(\hat{x}, \hat{\lambda}, \hat{\mu}, \hat{w}) & (g'_x(\hat{x}, \hat{w}))^\top & (h'_x(\hat{x}, \hat{w}))^\top \\ \hat{\Lambda} \cdot g'_x(\hat{x}, \hat{w}) & \hat{\Gamma} & \Theta \\ h'_x(\hat{x}, \hat{w}) & \Theta & \Theta \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = \begin{pmatrix} 0_{\mathbb{R}^n} \\ 0_{\mathbb{R}^m} \\ 0_{\mathbb{R}^p} \end{pmatrix}$$

for $v_1 \in \mathbb{R}^n$, $v_2 = (v_{21}, v_{22})^\top \in \mathbb{R}^{l+(m-l)}$, and $v_3 \in \mathbb{R}^p$. Exploitation of the special structure of $\hat{\Lambda}$ and $\hat{\Gamma}$ yields $\hat{\Gamma}_1 v_{21} = 0_{\mathbb{R}^l}$ and since $\hat{\Gamma}_1$ is non-singular it follows $v_{21} = 0$. With this, it remains to investigate the reduced system

$$\begin{pmatrix} A & B^\top & C^\top \\ B & \Theta & \Theta \\ C & \Theta & \Theta \end{pmatrix} \begin{pmatrix} v_1 \\ v_{22} \\ v_3 \end{pmatrix} = \begin{pmatrix} 0_{\mathbb{R}^n} \\ 0_{\mathbb{R}^{m-l}} \\ 0_{\mathbb{R}^p} \end{pmatrix}$$

with $A := L''_{xx}(\hat{x}, \hat{\lambda}, \hat{\mu}, \hat{w})$, $B := (g'_{x,i}(\hat{x}, \hat{w}))_{i=l+1,\ldots,m}$, and $C := h'_x(\hat{x}, \hat{w})$. Notice, that the second block equation has been multiplied with $\hat{\Lambda}_2^{-1}$. The last two block equations yield $Bv_1 = 0_{\mathbb{R}^{m-l}}$ and $Cv_1 = 0_{\mathbb{R}^p}$. Multiplication of the first block equation from the left with $v_1^\top$ yields

$$0 = v_1^\top A v_1 + (Bv_1)^\top v_{22} + (Cv_1)^\top v_3 = v_1^\top A v_1.$$

Since A is positive definite on $T_2(\hat{x}) \setminus \{0_{\mathbb{R}^n}\}$, i.e. it holds $d^\top A d > 0$ for all $d \neq 0_{\mathbb{R}^n}$ with $Bd = 0_{\mathbb{R}^{m-l}}$ and $Cd = 0_{\mathbb{R}^p}$, it follows $v_1 = 0_{\mathbb{R}^n}$. Taking this property into account, the first block equation reduces to $B^\top v_{22} + C^\top v_3 = 0_{\mathbb{R}^n}$. By the linear independence of the gradients $\nabla g_i(\hat{x}, \hat{w})$, $i \in A(\hat{x}, \hat{w})$ and $\nabla h_j(\hat{x}, \hat{w})$, $j = 1, \ldots, p$ we obtain $v_{22} = 0_{\mathbb{R}^{m-l}}$, $v_3 = 0_{\mathbb{R}^p}$. Putting all together, the above linear equation system has the unique solution $v_1 = 0_{\mathbb{R}^n}$,

$v_2 = 0_{\mathbb{R}^m}$, $v_3 = 0_{\mathbb{R}^p}$, which means that the matrix F'_x is non-singular and the implicit function theorem is applicable.

By the implicit function theorem there exist neighborhoods $V_\epsilon(\hat{w})$ and $U_\delta(\hat{x}, \hat{\lambda}, \hat{\mu})$, and uniquely defined functions

$$(x(\cdot), \lambda(\cdot), \mu(\cdot)) : V_\varepsilon(\hat{w}) \to U_\delta(\hat{x}, \hat{\lambda}, \hat{\mu})$$

satisfying

$$F(x(w), \lambda(w), \mu(w), w) = 0_{\mathbb{R}^{n+m+p}} \tag{4.3.3}$$

for all $w \in V_\epsilon(\hat{w})$. Furthermore, these functions are continuously differentiable and (4.3.1) arises by differentiation of the identity (4.3.3) w.r.t. w.

It remains to verify, that $x(w)$ actually is a strongly regular local minimum of $(NLP(w))$. The continuity of the functions $x(w), \lambda(w)$ and g together with $\lambda_i(\hat{w}) = \hat{\lambda}_i > 0$, $i = l+1, \ldots, m$ and $g_i(x(\hat{w}), \hat{w}) = g_i(\hat{x}, \hat{w}) < 0$, $i = 1, \ldots, l$ guarantees $\lambda_i(w) > 0$, $i = l+1, \ldots, m$ and $g_i(x(w), w) < 0$, $i = 1, \ldots, l$ for w sufficiently close to $\hat{w}$. From (4.3.3) it follows $g_i(x(w), w) = 0$, $i = l+1, \ldots, m$, and $h_j(x(w), w) = 0$, $j = 1, \ldots, p$. Thus, $x(w) \in \Sigma(w)$ and the KKT conditions are satisfied. Furthermore, the index set $A(x(w), w) = A(\hat{x}, \hat{w})$ remains unchanged in a neighborhood of $\hat{w}$. In addition, due to the continuity of the first and second derivatives, the gradients of the active constraints remain linearly independent and L''_{xx} remains positive definite on $T_2(x(w))$ for w sufficiently close to $\hat{w}$. ∎

4.4 Numerical Methods

We analyze the Lagrange-Newton-Method and the SQP-Method more closely. The SQP-Method is discussed in, e.g., [29], [51], [25], [60], [53], [54], [1], [19].

The SQP-Method exists in several implementations, e.g. [55], [35], [24], [23].

Special adaptions of the SQP method to discretized optimal control problems are described in [22], [56], [59], [6].

4.4.1 Lagrange-Newton-Method

In this section we restrict the discussion to the equality constrained nonlinear optimization problem

$(NLPEQ)$	Minimize	$f(x)$
	w. r. t.	$x \in \mathbb{R}^n$
	subject to	$h_j(x) = 0, \quad j = 1, \ldots, p.$

4.4.1. Assumption. *The functions $f : \mathbb{R}^n \to \mathbb{R}$ and $h_j : \mathbb{R}^n \to \mathbb{R}$, $j = 1, \ldots, p$ are twice continuously differentiable.*

Let $\hat{x}$ be a local minimum of (NLPEQ) and let the gradients $\nabla h_j(\hat{x})$, $j = 1, \ldots, p$ be linearly independent. Then the first order necessary conditions (KKT conditions) are valid: There exist multipliers $\hat{\mu} = (\hat{\mu}_1, \ldots, \hat{\mu}_p)^\top \in \mathbb{R}^p$ such that

$$\begin{aligned} \nabla_x L(\hat{x}, \hat{\mu}) = \nabla f(\hat{x}) + \sum_{j=1}^{p} \hat{\mu}_j \nabla h_j(\hat{x}) &= 0_{\mathbb{R}^n}, \\ h_j(\hat{x}) &= 0, \quad j = 1, \ldots, p. \end{aligned}$$

This is a nonlinear equation system for $\hat{x}$ and $\hat{\mu}$ and we can rewrite it in the form

$$F(\hat{x}, \hat{\mu}) = 0_{\mathbb{R}^{n+p}}, \tag{4.4.1}$$

where $F : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}^{n+p}$ and

$$F(x, \mu) := \begin{pmatrix} \nabla_x L(x, \mu) \\ h(x) \end{pmatrix}, \quad h(x) = \begin{pmatrix} h_1(x) \\ \vdots \\ h_p(x) \end{pmatrix}.$$

The Lagrange-Newton is based on the application of Newton's method to solve the necessary conditions (4.4.1). This leads to the following algorithm:

Algorithm: Lagrange-Newton-Method

(i) Choose $x^{(0)} \in \mathbb{R}^n$ and $\mu^{(0)} \in \mathbb{R}^p$ and set $k = 0$.

(ii) If $F(x^{(k)}, \mu^{(k)}) = 0_{\mathbb{R}^{n+p}}$, STOP.

(iii) Solve the linear system of equations

$$\begin{pmatrix} L''_{xx}(x^{(k)}, \mu^{(k)}) & h'(x^{(k)})^\top \\ h'(x^{(k)}) & \Theta \end{pmatrix} \cdot \begin{pmatrix} d \\ v \end{pmatrix} = - \begin{pmatrix} \nabla_x L(x^{(k)}, \mu^{(k)}) \\ h(x^{(k)}) \end{pmatrix} \tag{4.4.2}$$

and set

$$x^{(k+1)} = x^{(k)} + d, \qquad \mu^{(k+1)} = \mu^{(k)} + v. \tag{4.4.3}$$

(iv) Set $k := k + 1$ and go to (ii).

Exploitation and adaptation of the well-known convergence results of Newton's method to this particular situation leads to the following convergence result.

4.4.2. Theorem (Local Convergence).

(i) Let $(\hat{x}, \hat{\mu})$ be a KKT point.

(ii) Let f, h_j, $j = 1, \ldots, p$ be twice continously differentiable with Lipschitz-continuous second derivatives f'' and h''_j, $j = 1, \ldots, p$.

(iii) Let the matrix

$$\begin{pmatrix} L''_{xx}(\hat{x}, \hat{\mu}) & h'(\hat{x})^\top \\ h'(\hat{x}) & \Theta \end{pmatrix} \tag{4.4.4}$$

be nonsingular.

Then there exists $\varepsilon > 0$ such that the Lagrange-Newton-Method converges for all $(x^{(0)}, \mu^{(0)}) \in U_\varepsilon(\hat{x}, \hat{\mu})$ (local convergence). Furthermore, the convergence is quadratic, i.e. there exists a constant $C \geq 0$ such that

$$\|(x^{(k+1)}, \mu^{(k+1)}) - (\hat{x}, \hat{\mu})\| \leq C \|(x^{(k)}, \mu^{(k)}) - (\hat{x}, \hat{\mu})\|^2$$

for all sufficiently large k.

4.4.3. Remarks.

- The matrix in (4.4.4) is called Kuhn-Tucker-matrix (KT-matrix). The KT-matrix is nonsingular, if the following conditions are satisfied:

  (i) the gradients $\nabla h_j(\hat{x})$, $j = 1, \ldots, p$ are linearly independent;

  (ii) it holds
  $$v^\top L''_{xx}(\hat{x}, \hat{\mu})v > 0$$
  for all $0_{\mathbb{R}^n} \neq v \in \mathbb{R}^n$ with
  $$h'(\hat{x}) \cdot v = 0_{\mathbb{R}^p}.$$

- The convergence is super-linear, if the second derivatives of f and h_j, $j = 1, \ldots, p$ are not Lipschitz-continuous: There exists a sequence $\{C_k\}$ with $\lim_{k\to\infty} C_k = 0$ such that
  $$\|(x^{(k+1)}, \mu^{(k+1)}) - (\hat{x}, \hat{\mu})\| \leq C_k \|(x^{(k)}, \mu^{(k)}) - (\hat{x}, \hat{\mu})\|$$
  for sufficiently large k.

4.4.2 Sequential Quadratic Programming (SQP)

The linear system (4.4.2) of equations in item (iii) of the Lagrange-Newton method can be obtained in a different way.

Let us again consider the equality constrained problem (NLPEQ). We assume, that the problem can be approximated locally at some point $(x^{(k)}, \mu^{(k)})$ by the quadratic optimization problem

$$\begin{aligned} \min_{d\in\mathbb{R}^n} \quad & \frac{1}{2} d^\top L''_{xx}(x^{(k)}, \mu^{(k)})d + \nabla f(x^{(k)})^\top d \\ \text{s. t.} \quad & h(x^{(k)}) + h'(x^{(k)})d \;=\; 0_{\mathbb{R}^p}. \end{aligned}$$

The Lagrange function for the quadratic problem is given by

$$\bar{L}(d, \eta) := \frac{1}{2} d^\top L''_{xx}(x^{(k)}, \mu^{(k)})d + \nabla f(x^{(k)})^\top d + \eta^\top \left(h(x^{(k)}) + h'(x^{(k)})d \right).$$

The evaluation of the first order necessary conditions leads to

$$\begin{aligned} L''_{xx}(x^{(k)}, \mu^{(k)})d + \nabla f(x^{(k)}) + h'(x^{(k)})^\top \eta &= 0_{\mathbb{R}^n}, \\ h(x^{(k)}) + h'(x^{(k)})d &= 0_{\mathbb{R}^p}, \end{aligned}$$

respectively

$$\begin{pmatrix} L''_{xx}(x^{(k)}, \mu^{(k)}) & h'(x^{(k)})^\top \\ h'(x^{(k)}) & \Theta \end{pmatrix} \cdot \begin{pmatrix} d \\ \eta \end{pmatrix} = - \begin{pmatrix} \nabla f(x^{(k)}) \\ h(x^{(k)}) \end{pmatrix}. \quad (4.4.5)$$

If we subtract $h'(x^{(k)})^\top \mu^{(k)}$ on both sides of the first equation in (4.4.5) we get the linear equation system

$$\begin{pmatrix} L''_{xx}(x^{(k)}, \mu^{(k)}) & h'(x^{(k)})^\top \\ h'(x^{(k)}) & \Theta \end{pmatrix} \cdot \begin{pmatrix} d \\ \eta - \mu^{(k)} \end{pmatrix} = - \begin{pmatrix} \nabla_x L(x^{(k)}, \mu^{(k)}) \\ h(x^{(k)}) \end{pmatrix}. \quad (4.4.6)$$

A comparison of (4.4.6) with (4.4.2) reveals that these two linear equation systems are identical, if we set $v := \eta - \mu^{(k)}$. According to (4.4.3) it follows that the new iterates are given by

$$x^{(k+1)} = x^{(k)} + d, \qquad \mu^{(k+1)} = \mu^{(k)} + v = \eta.$$

Summary:
For equality constrained optimization problems, the Lagrange-Newton method is identical with the above depicted successive quadratic programming method, if we use the Lagrange multiplier η of the QP subproblem as the new approximation for the multiplier μ.

This observation motivates the following extension of the quadratic optimization problem for the inequality constrained optimization problem (NLP).

QP Problem $(QP(x^{(k)}, \lambda^{(k)}, \mu^{(k)}))$**:**

$$\begin{aligned} \min_{d \in \mathbb{R}^n} \quad & \frac{1}{2} d^\top L''_{xx}(x^{(k)}, \lambda^{(k)}, \mu^{(k)}) d + f'(x^{(k)}) d \\ \text{s.t.} \quad & g_i(x^{(k)}) + g'_i(x^{(k)}) d \leq 0, \quad i = 1, \ldots, m, \\ & h_j(x^{(k)}) + h'_j(x^{(k)}) d = 0, \quad j = 1, \ldots, p. \end{aligned}$$

Algorithm:
Local sequential quadratic programming (SQP) Method

(i) Choose $(x^{(0)}, \lambda^{(0)}, \mu^{(0)}) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p$ and set $k = 0$.

(ii) If $(x^{(k)}, \lambda^{(k)}, \mu^{(k)})$ is a KKT point of (NLP), STOP.

(iii) Compute a KKT point $(d^{(k)}, \lambda^{(k+1)}, \mu^{(k+1)}) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p$ of the the quadratic programming problem $(QP(x^{(k)}, \lambda^{(k)}, \mu^{(k)})$.

(iv) Set $x^{(k+1)} = x^{(k)} + d^{(k)}$, $k := k + 1$ and go to (ii).

4.4.4. Remarks.

- *It is not necessary to know the index set $A(\hat{x})$ of active inequality constraints in advance.*
- *The iterates $x^{(k)}$ are not necessarily admissible, i.e. it may happen that $x^{(k)} \notin \Sigma$ holds.*
- *There are powerful algorithms for the numerical solution of quadratic optimization problems, compare e.g. [20], [27], [21], [58].*

The local convergence of the SQP method is established in the following theorem.

4.4.5. Theorem (Local Convergence of SQP Method).

(i) Let $\hat{x}$ be a local minimum of (NLP).

(ii) *Let the functions* f, g_i, $i = 1, \ldots, m$, *and* h_j, $j = 1, \ldots, p$ *be twice continuously differentiable with Lipschitz continuous second derivatives* f'', g_i'', $i = 1, \ldots, m$, *and* h_j'', $j = 1, \ldots, p$.

(iii) *Let the gradients* $\nabla g_i(\hat{x})$, $i \in A(\hat{x})$, *and* $\nabla h_j(\hat{x})$, $j = 1, \ldots, p$ *be linearly independent.*

(Then $\hat{x}$ *fulfills the KKT conditions with unique multipliers* $\hat{\lambda}_i \geq 0$, $i = 1, \ldots, m$, *and* $\hat{\mu}_j$, $j = 1, \ldots, p$.*)*

(iv) *Let the strict complementarity condition* $\hat{\lambda}_i - g_i(\hat{x}) > 0$ *for all* $i \in A(\hat{x})$ *hold.*

(v) *Let*

$$d^\top L''_{xx}(\hat{x}, \hat{\lambda}, \hat{\mu}) d > 0$$

hold for all $0_{\mathbb{R}^n} \neq d \in \mathbb{R}^n$ *with*

$$g_i'(\hat{x})d = 0, \quad i \in A(\hat{x}), \quad h_j'(\hat{x})d = 0, \quad j = 1, \ldots, p.$$

Then there exists $\varepsilon > 0$ *such that for arbitrary starting values* $(x^{(0)}, \lambda^{(0)}, \mu^{(0)}) \in U_\varepsilon(\hat{x}, \hat{\lambda}, \hat{\mu})$ *all QP problems* $(QP(x^{(k)}, \lambda^{(k)}, \mu^{(k)}))$ *possess a locally unique solution* $d^{(k)}$ *with unique multipliers* $\lambda^{(k+1)}$ *and* $\mu^{(k+1)}$. *Furthermore, the sequence* $\{(x^{(k)}, \lambda^{(k)}, \mu^{(k)})\}$ *converges quadratically to* $(\hat{x}, \hat{\lambda}, \hat{\mu})$.

Proof. The proof exploits the Sensitivity Theorem 4.3.2 for parametric optimization problems to show that the index set of active constraints remains unchanged within a certain neighborhood of the solution. Then, it is possible to show that the SQP method locally coincides with the Lagrange-Newton-Method.

1. We consider $(QP(\hat{x}, \hat{\lambda}, \hat{\mu}))$ as the unperturbed quadratic problem with nominal parameter $\hat{w} = (\hat{x}, \hat{\lambda}, \hat{\mu})$. Furthermore, we notice, that the KKT conditions for $(QP(\hat{w}))$ and (NLP) coincide for $\hat{d} = 0_{\mathbb{R}^n}$. Hence, $(0_{\mathbb{R}^n}, \hat{\lambda}, \hat{\mu})$ is a KKT point of $(QP(\hat{w}))$. In addition, assumptions (iii)-(v) guarantee that $\hat{d}$ is a strongly regular local minimum of $(QP(\hat{w}))$.

   Hence, we may apply Theorem 4.3.2: There exist a neighborhood $V_\varepsilon(\hat{w})$ such that $(QP(w))$ has a unique strongly regular local minimum $(d(w), \lambda(w), \mu(w))$ for each $w \in V_\varepsilon(\hat{w})$. Herein, $(d(w), \lambda(w), \mu(w))$ is continuously differentiable w. r. t. w.

Furthermore, the index set of active constraints remains unchanged in that neighborhood: $A(\hat{x}) = A_{QP}(\hat{d}, \hat{w}) = A_{QP}(d(w), w)$. $A_{QP}(d, w)$ denotes the index set of active inequality constraints of $(QP(w))$ at d.

2. Due to the continuity of the constraints, we may neglect the inactive constraints at $\hat{x}$ and obtain the (locally) equivalent optimization problem

$$\min f(x) \quad \text{s.t.} \quad g_i(x) = 0,\ i \in A(\hat{x}),\ h_j(x) = 0,\ j = 1, \ldots, p.$$

We can apply the Lagrange-Newton-Method and under the assumptions (i)-(v) Theorem 4.4.2 yields the local quadratic convergence

$$(x^{(k)}, \lambda^{(k)}_{A(\hat{x})}, \mu^{(k)}) \to (\hat{x}, \hat{\lambda}_{A(\hat{x})}, \hat{\mu}).$$

Notice, that the multipliers are unique according to (iii). We may add $\lambda_i^{(k)} = 0$ for $i \notin A(\hat{x})$ to obtain

$$w^{(k)} := (x^{(k)}, \lambda^{(k)}, \mu^{(k)}) \to \hat{w} = (\hat{x}, \hat{\lambda}, \hat{\mu}).$$

3. Let δ denote the radius of convergence of the Lagrange-Newton-Method. Let $r := \min\{\varepsilon, \delta\}$. For $w^{(0)} \in U_r(\hat{w})$ all subsequent iterates $w^{(k)}$ of the Lagrange-Newton-Method remain in that neighborhood. Furthermore, $(d^{(k)}, \lambda^{(k+1)}, \mu^{(k+1)})$ with $d^{(k)} = x^{(k+1)} - x^{(k)}$ fulfills the necessary conditions of $(QP(w^{(k)}))$, cf. (4.4.5) and (4.4.6). According to 1., the solution $(d(w^{(k)}), \lambda(w^{(k)}), \mu(w^{(k)}))$ of $(QP(w^{(k)}))$ is unique. Hence, the SQP iteration coincides with the Lagrange-Newton iteration.

∎

4.4.6. Remarks (Approximation of Hessian). The use of the exact Hessian L''_{xx} of the Lagrange function in the QP problem has two drawbacks from numerical point of view:

- In most practical applications the Hessian is not known explicitly. The numerical approximation of the Hessian by finite differences is very expensive.
- The Hessian may be indefinite. This makes the numerical solution of the QP problem more difficult. It is desirable to have a positive definite matrix in the QP problem.

In practice, the Hessian of the Lagrange function in iteration k is replaced by a suitable matrix B_k. Powell [51] suggested to use the modified BFGS-update formula

$$B_{k+1} = B_k + \frac{q^{(k)}(q^{(k)})^\top}{(q^{(k)})^\top s^{(k)}} - \frac{B_k s^{(k)}(s^{(k)})^\top B_k}{(s^{(k)})^\top B_k s^{(k)}}, \tag{4.4.7}$$

where

$$\begin{aligned}
s^{(k)} &= x^{(k+1)} - x^{(k)}, \\
q^{(k)} &= \theta_k \eta^{(k)} + (1-\theta_k) B_k s^{(k)}, \\
\eta^{(k)} &= \nabla_x L(x^{(k+1)}, \lambda^{(k)}, \mu^{(k)}) - \nabla_x L(x^{(k)}, \lambda^{(k)}, \mu^{(k)}), \\
\theta_k &= \begin{cases} 1, & \text{if } (s^{(k)})^\top \eta^{(k)} \geq 0.2 (s^{(k)})^\top B_k s^{(k)}, \\ \frac{0.8 (s^{(k)})^\top B_k s^{(k)}}{(s^{(k)})^\top B_k s^{(k)} - (s^{(k)})^\top \eta^{(k)}}, & \text{otherwise.} \end{cases}
\end{aligned}$$

This update formula guarantees that B_{k+1} remains symmetric and positive definite if B_k was symmetric and positive definite. For $\theta_k = 1$ we get the well known BFGS update formula, which is used in variable metric methods (or quasi Newton methods) for unconstrained optimization.

If the exact Hessian is replaced by the modified BFGS update formula, the convergence of the resulting SQP method is only super-linear and not quadratic.

4.4.3 Globalization of the Local SQP Method

The convergence result shows that the SQP method converges for all starting values which are within some neighborhood of a local minimum of (NLP). Unfortunately, in practice this neighborhood is not known and it cannot be guaranteed, that the starting values are within this neighborhood. Fortunately, the SQP method can be globalized in the sense that it converges for arbitrary starting values (under suitable conditions). The idea is to determine the new iterate $x^{(k+1)}$ according to the formula

$$x^{(k+1)} = x^{(k)} + t_k d^{(k)}$$

with a step length $t_k > 0$. The step length t_k is obtained by performing a so-called **line search** in the direction $d^{(k)}$ for a suitable **penalty function**.

The penalty function allows to decide whether the new iterate $x^{(k+1)}$ is in some sense 'better' than the old iterate $x^{(k)}$. The new iterate will be better than the old one, if either a sufficient decrease in the objective function f or an improvement of the total constraint violations is achieved while the respective other value is not substantially declined.

Example 4.4.2 (Penalty Functions)
A typical penalty function for (NLP) based on the 1-norm is given by the l_1-penalty function

$$l_1(x;\alpha) := f(x) + \alpha \sum_{i=1}^{m} \max\{0, g_i(x)\} + \alpha \sum_{j=1}^{p} |h_j(x)|.$$

Notice, that the l_1-penalty function penalizes infeasible points $x \notin \Sigma$.

More generally, penalty functions based on q-norms are given by

$$l_q(x;\alpha) = f(x) + \alpha \left(\sum_{i=1}^{m} (\max\{0, g_i(x)\})^q + \sum_{j=1}^{p} h_j(x)^q \right)^{1/q}, \ 1 \le q < \infty,$$

and

$$l_\infty(x;\alpha) = f(x) + \alpha \max\{0, g_1(x), \ldots, g_m(x), |h_1(x)|, \ldots, |h_p(x)|\}.$$

A general class of penalty functions is defined by

$$P_r(x;\alpha) := f(x) + \alpha \cdot r(x), \tag{4.4.8}$$

where $\alpha > 0$ denotes a penalty parameter and $r : \mathbb{R}^n \to [0, \infty)$ is a function with the property

$$r(x) = 0 \qquad \Leftrightarrow \qquad x \in \Sigma.$$

4.4.7. Definition (Exact Penalty Function). *A penalty function (4.4.8) is called* **exact in a local minimum $\hat{x}$ of (NLP)**, *if there exists a finite parameter $\hat{\alpha} > 0$, such that $\hat{x}$ is a local minimum of $P_r(\cdot;\alpha)$ for all $\alpha \ge \hat{\alpha}$.*

Example 4.4.3

Figure 4.2 shows the l_1-penalty function for the problem

$$\begin{aligned} f(x,y) &= (x-2)^2+(y-3)^2, \\ h(x,y) &= y+\frac{x}{2}-\frac{1}{2}, \\ g_1(x,y) &= y+2x^2-2, \\ g_2(x,y) &= x^2-y-1, \end{aligned}$$

for different values of α. The optimal solution is given by $\hat{x}=(3/5,1/5)^\top$, $\hat{\lambda}=(0,0)^\top$, and $\hat{\mu}=28/5$. The constraints g_1 and g_2 are not active at $\hat{x}$.

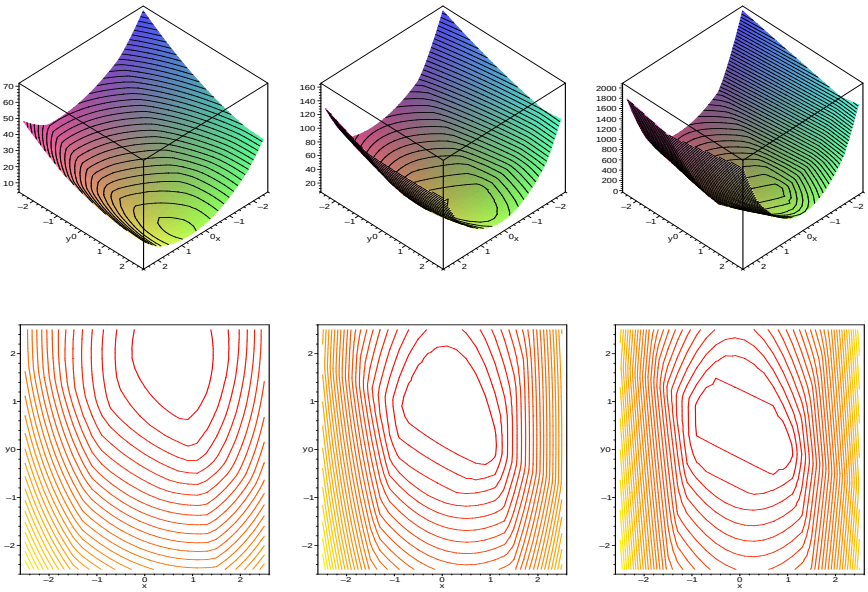


Figure 4.2: 3D-representation (top) and contours (bottom) of the l_1-penalty function for $\alpha=1$ (left), $\alpha=28/5$ (middle), and $\alpha=100$ (right).

Unfortunately, it can be shown that the penalty function (4.4.8) is not differentiable in a local minimum $\hat{x}$, if it is exact and $\nabla f(\hat{x}) \neq 0_{\mathbb{R}^n}$ holds. Please notice, that the condition $\nabla f(\hat{x}) \neq 0_{\mathbb{R}^n}$ usually holds for constrained nonlinear programming problems.

The following theorem states, that the penalty functions l_q for $1 \leq q \leq \infty$ are exact, if in addition a constraint qualification holds.

4.4.8. Theorem. *Let $\hat{x} \in \Sigma$ be an isolated local minimum of (NLP). Let the constraint qualification of Mangasarian-Fromowitz be fulfilled at $\hat{x}$. Then l_q is exact for $1 \leq q \leq \infty$.*

Proof. cf. Geiger and Kanzow [19], p. 225 ∎

Theorem 4.4.8 suggests to replace the constrained optimization problem (NLP) by the unconstrained minimization problem

$$\min_{x \in \mathbb{R}^n} l_q(x; \alpha)$$

for a sufficiently large $\alpha > 0$. We will exploit this relationship to determine a step length t by a one-dimensional line search. For the ease of representation we restrict the following discussion to the l_1-penalty function.

Assume, that we are in iteration k of the SQP method and $d^{(k)}$ is the optimal solution of the QP subproblem. It can be shown that the directional derivative

$$l_1'(x^{(k)}; d^{(k)}; \alpha) := \lim_{h \downarrow 0} \frac{l_1(x^{(k)} + hd^{(k)}; \alpha) - l_1(x^{(k)}; \alpha)}{h}$$

of l_1 at $x^{(k)}$ in direction $d^{(k)}$ exists. Furthermore, let $d^{(k)}$ be a **descent direction** of the penalty function, i.e. let the condition

$$l_1'(x^{(k)}; d^{(k)}; \alpha) < 0$$

be fulfilled (in fact, it turns out that under suitable assumptions $d^{(k)} \neq 0$ actually is a descent direction of l_1). Then, we consider the real-valued function

$$\varphi(t) := l_1(x^{(k)} + td^{(k)}; \alpha), \qquad t \geq 0,$$

cf. Figure 4.3.

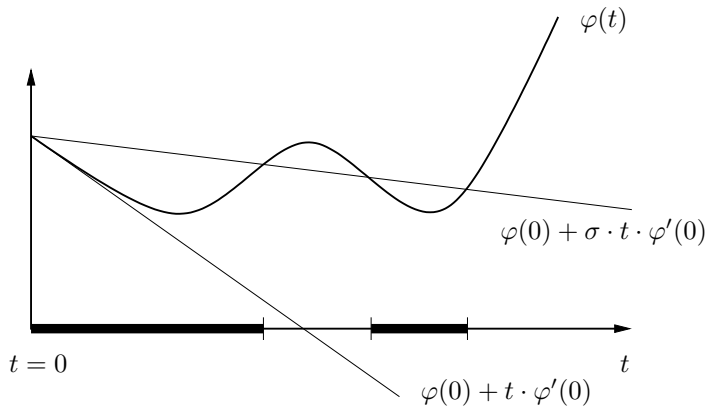


Figure 4.3: One dimensional line search by Armijo's method.

The step length t_k is now determined such that $\varphi(t_k) \approx \min_{t>0} \varphi(t)$ holds. Actually, it suffices to obtain a sufficient decrease in the penalty function, i.e. the condition

$$l_1(x^{(k)} + t_k d^{(k)}; \alpha) \leq l_1(x^{(k)}; \alpha) + \sigma t_k l_1'(x^{(k)}; d^{(k)}; \alpha) \tag{4.4.9}$$

with a constant $\sigma \in (0, 1)$ has to be fulfilled. A commonly used line search method which guarantees (4.4.9) is Armijo's method.

Algorithm: Armijo Line Search

(i) Choose parameters $\beta \in (0, 1)$ and $\sigma \in (0, 1)$ and set $t := 1$.

(ii) If the condition

$$l_1(x^{(k)} + t d^{(k)}; \alpha) \leq l_1(x^{(k)}; \alpha) + \sigma t l_1'(x^{(k)}; d^{(k)}; \alpha).$$

is fulfilled, set $t_k := t$ and STOP. Otherwise go to (iii).

(iii) Set $t := \beta \cdot t$ and go to (ii).

The bold intervals in Figure 4.3 represent the valid step lengths.

Incorporation of all modifications of the local SQP method yields a globalized version of the SQP method.

Algorithm:
Globalized sequential quadratic programming (SQP) Method

(i) Choose $(x^{(0)}, \lambda^{(0)}, \mu^{(0)}) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p$, $B_0 \in \mathbb{R}^{n\times n}$ symmetric and positive definite, $\alpha > 0$, $\beta \in (0,1)$, $\sigma \in (0,1)$, and set $k = 0$.

(ii) If $(x^{(k)}, \lambda^{(k)}, \mu^{(k)})$ is a KKT point of (NLP), STOP.

(iii) Compute a KKT point $(d^{(k)}, \lambda^{(k+1)}, \mu^{(k+1)}) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p$ of the quadratic programming problem $(QP(x^{(k)}, \lambda^{(k)}, \mu^{(k)}))$ with L''_{xx} replaced by B_k.

(iv) Determine a step size $t_k = \max\{\beta^j \mid j = 0, 1, 2, \ldots\}$ such that

$$l_1(x^{(k)} + t_k d^{(k)}; \alpha) \leq l_1(x^{(k)}; \alpha) + \sigma t_k l_1'(x^{(k)}; d^{(k)}; \alpha).$$

(v) Compute B_{k+1} according to (4.4.7) and set $x^{(k+1)} := x^{(k)} + t_k d^{(k)}$.

(vi) Set $k := k + 1$ and go to (ii).

It is important to mention, that there exist also differentiable exact penalty functions. But these penalty functions are not of the form (4.4.8). A commonly used differentiable exact penalty function for (NLP) is the **augmented Lagrange function**

$$\begin{aligned} L_a(x, \lambda, \mu; \alpha) &= f(x) + \mu^\top h(x) + \frac{\alpha}{2}\|h(x)\|^2 \\ &\quad + \frac{1}{2\alpha}\sum_{i=1}^{m}\left(\left(\max\{0, \lambda_i + \alpha g_i(x)\}\right)^2 - \lambda_i^2\right) \\ &= f(x) + \sum_{j=1}^{p}\left(\mu_j h_j(x) + \frac{\alpha}{2}h_j(x)^2\right) \qquad (4.4.10) \\ &\quad + \sum_{i=1}^{m}\begin{cases} \lambda_i g_i(x) + \frac{\alpha}{2}g_i(x)^2, & \text{if } \lambda_i + \alpha g_i(x) \geq 0, \\ -\frac{\lambda_i^2}{2\alpha}, & \text{otherwise.} \end{cases} \end{aligned}$$

A SQP method employing the augmented Lagrange function is discussed in

[53], [54].

The augmented Lagrange function is motivated by the subsequent considerations. (NLP) can be transformed into an equivalent equality constrained nonlinear optimization problem by introducing slack variables $s = (s_1, \ldots, s_m)^\top \in \mathbb{R}^m$:

$$(NLP') \qquad \begin{aligned} \min_{x,s} \quad & f(x) \\ \text{s.t.} \quad & h_j(x) = 0, \quad j = 1, \ldots, p, \\ & g_i(x) + s_i^2 = 0, \quad i = 1, \ldots, m. \end{aligned}$$

Let $\hat{x}$ be a local minimum of (NLP') and $\alpha > 0$. Then, $\hat{x}$ is also a local minimum of

$$(NLP'') \qquad \begin{aligned} \min_{x,s} \quad & f(x) + \tfrac{\alpha}{2} \left(\|h(x)\|^2 + \|g(x) + s\|^2 \right) \\ \text{s.t.} \quad & h_j(x) = 0, \quad j = 1, \ldots, p, \\ & g_i(x) + s_i^2 = 0, \quad i = 1, \ldots, m. \end{aligned}$$

The Lagrange function for (NLP") is given by

$$\begin{aligned} \bar{L}(x, s, \lambda, \mu; \alpha) = \; & f(x) + \frac{\alpha}{2} \left(\|h(x)\|^2 + \|g(x) + s\|^2 \right) \\ & + \sum_{j=1}^{p} \mu_j h_j(x) + \sum_{i=1}^{m} \lambda_i \left(g_i(x) + s_i^2 \right). \end{aligned}$$

If $(\hat{x}, \hat{s}, \hat{\lambda}, \hat{\mu})$ is a KKT point of (NLP') and if the second order sufficient condition holds then it turns out that $\bar{L}(\cdot, \cdot, \hat{\lambda}, \hat{\mu}; \alpha)$ is exact in $\hat{x}$ and $\hat{s}$, i.e. $\bar{L}(\cdot, \cdot, \hat{\lambda}, \hat{\mu}; \alpha)$ has a local minimum at $\hat{x}$ and $\hat{s}$ for sufficiently large $\alpha > 0$. Thus, we have to minimize $\bar{L}(x, s, \hat{\lambda}, \hat{\mu}; \alpha)$ w. r. t. to x and s. For fixed x it is possible to calculate the minimizing $\hat{s}$ explicitly:

$$\hat{s}_i = \left(\max \left\{ 0, -\left(\frac{\lambda_i}{\alpha} + g_i(x) \right) \right\} \right)^{1/2}, \quad i = 1, \ldots, m.$$

Introducing these values in $\bar{L}$ yields the augmented Lagrange function (4.4.10).

4.4.9. Remarks. In practical applications a suitable value for the penalty parameter α is not known a priori. Strategies for adapting α iteratively and individually for each constraint can be found in [54] and [51].

4.4.4 Inconsistent QP Problem

So far, we always assumed that the QP problem has a solution. The following example shows that this assumption is not always justified, even if the original problem is feasible.

Example 4.4.4
Consider the constraint

$$g(x) = 1 - x^2 \leq 0$$

and $x^{(0)} = 0$. *In the QP problem we get the constraint*

$$g(x^{(0)}) + g'(x^{(0)}) \cdot d = 1 \leq 0.$$

Obviously, this constraint cannot be fulfilled.

Powell [51] suggested to relax the constraints of the QP problem in such a way, that the relaxed QP problem possesses admissible points. The original QP problem is then replaced by

Relaxed QP Problem $(QP'(x^{(k)}, \lambda^{(k)}, \mu^{(k)}))$**:**

$$\begin{aligned} \min_{d \in \mathbb{R}^n} \quad & \frac{1}{2} d^\top B_k d + f'(x^{(k)})d + \frac{\alpha}{2}\delta^2 \\ \text{s.t.} \quad & g_i(x^{(k)})(1 - \sigma_i \delta) + g_i'(x^{(k)})d \leq 0, \quad i = 1, \ldots, m, \\ & h_j(x^{(k)})(1 - \delta) + h_j'(x^{(k)})d = 0, \quad j = 1, \ldots, p. \end{aligned}$$

where

$$\sigma_i = \begin{cases} 0, & \text{if } g_i(x^{(k)}) < 0, \\ 1, & \text{otherwise,} \end{cases} \qquad i = 1, \ldots, m.$$

The point $d = 0$ and $\delta = 1$ is always admissible for $(QP'(x^{(k)}, \lambda^{(k)}, \mu^{(k)}))$. If the optimal solution (d, δ) of $(QP'(x^{(k)}, \lambda^{(k)}, \mu^{(k)}))$ fulfills $\delta = 0$, then d is optimal for $(QP(x^{(k)}, \lambda^{(k)}, \mu^{(k)}))$ as well.

4.4.5 An Active Set Method for the Solution of QP Problems

We consider the quadratic optimization problem

$$\min f(x) := \frac{1}{2}x^\top Qx + c^\top x \quad \text{s.t.} \quad l \le Ax \le u \tag{4.4.11}$$

with a symmetric and positive definite matrix $Q \in \mathbb{R}^{n\times n}$, a matrix $A \in \mathbb{R}^{m\times n}$, and vectors $c \in \mathbb{R}^n$, $l, u \in \mathbb{R}^m$. Let $a_i^\top$ denote the i^{th} row of A.

The Lagrange function is given by

$$L(x, \lambda^u, \lambda^l) = \frac{1}{2}x^\top Qx + c^\top x + (\lambda^u)^\top(Ax - u) + (\lambda^l)^\top(l - Ax)$$

The KKT conditions are

$$\begin{aligned}
0_{\mathbb{R}^n} &= \nabla_x L(x, \lambda^u, \lambda^l) = Qx + c + A^\top(\lambda^u - \lambda^l),\\
\lambda^u &\ge 0_{\mathbb{R}^m},\\
\lambda^l &\ge 0_{\mathbb{R}^m},\\
a_i^\top x < u_i &\Rightarrow \lambda_i^u = 0, \quad i = 1, \ldots, m,\\
l_i < a_i^\top x &\Rightarrow \lambda_i^l = 0, \quad i = 1, \ldots, m,\\
l &\le Ax \le u.
\end{aligned}$$

The KKT conditions can be formulated equivalently by use of the modified Lagrange function

$$\tilde{L}(x, \lambda) = \frac{1}{2}x^\top Qx + c^\top x + \lambda^\top Ax$$

where $\lambda := \lambda^u - \lambda^l$. Then, the KKT conditions are given by

$$\begin{aligned}
0_{\mathbb{R}^n} &= \nabla_x \tilde{L}(x, \lambda) = Qx + c + A^\top\lambda,\\
l &\le Ax \le u
\end{aligned}$$

and the complementarity conditions

$$\begin{aligned}
l_i = a_i^\top x = u_i &\Rightarrow \lambda_i \text{ arbitrary},\\
l_i < a_i^\top x < u_i &\Rightarrow \lambda_i = 0,\\
l_i = a_i^\top x < u_i &\Rightarrow \lambda_i \le 0,\\
l_i < a_i^\top x = u_i &\Rightarrow \lambda_i \ge 0.
\end{aligned}$$

Let

$$E := \{i \mid l_i = u_i,\ 1 \le i \le m\}$$

denote the index set of equality constraints and

$$I := \{1, 2, \ldots, m\} \backslash E$$

the index set of inequality constraints. The idea of the subsequent active set strategy is to choose an approximation of the optimal active set (which is not known in advance) and to solve a sequence of equality constrained quadratic problems where inactive constraints at the current iterate are neglected. If it turns out that the current iterate respectively the current index set is not yet optimal, the index set of active constraints is adapted and another iteration is performed.

Algorithm: Active Set Strategy

(0) Let $x^{(0)}$ be feasible and $\lambda^{(0)}$ be given. Set $k = 0$. Determine the index set of active inequality constraints $I_0 := I_0^u \cup I_0^l$ with $I_0^u := \{i \in I \mid a_i^\top x^{(0)} = u_i\}$ and $I_0^l := \{i \in I \mid a_i^\top x^{(0)} = l_i\}$.

(1) If $(x^{(k)}, \lambda^{(k)})$ is a KKT point of (4.4.11), STOP.

(2) Set $\lambda_i^{(k+1)} = 0$ for $i \in I \backslash I_k$. Compute a solution $\Delta x^{(k)}, \lambda_{E \cup I_k}^{(k+1)}$ of the linear equation system

$$\begin{pmatrix} Q & A_{E \cup I_k}^\top \\ A_{E \cup I_k} & \Theta \end{pmatrix} \begin{pmatrix} \Delta x \\ \lambda_{E \cup I_k} \end{pmatrix} = \begin{pmatrix} -\nabla f(x^{(k)}) \\ \Theta \end{pmatrix}. \tag{4.4.12}$$

(3) The following cases may occur:

(a) If $\Delta x^{(k)} = 0$ and

$$\lambda_i^{(k+1)} \begin{cases} \ge 0, & \text{if } i \in I_k^u, \\ \le 0, & \text{if } i \in I_k^l, \end{cases} \tag{4.4.13}$$

STOP. The complementarity conditions are fulfilled and $(x^{(k)}, \lambda^{(k+1)})$ is a KKT point.

(b) If $\Delta x^{(k)} = 0$ and (4.4.13) is not satisfied, then determine the index set

$$J = \{i \in I_k^u \mid \lambda_i^{(k+1)} < 0\} \cup \{i \in I_k^l \mid \lambda_i^{(k+1)} > 0\}$$

of violated complementarity conditions and determine an index $q \in J$ with

$$|\lambda_q^{(k+1)}| = \max\{|\lambda_i^{(k+1)}| \mid i \in J\}.$$

Set $x^{(k+1)} = x^{(k)}$, $I_{k+1} = I_k \backslash \{q\}$, $k = k+1$ and go to (1).

(c) If $\Delta x^{(k)} \neq 0$ and $x^{(k)} + \Delta x^{(k)}$ is feasible, then set $x^{(k+1)} = x^{(k)} + \Delta x^{(k)}$, $I_{k+1}^u = I_k^u$, $I_{k+1}^l = I_k^l$, $I_{k+1} = I_k$, $k = k+1$ and go to (1).

(d) If $\Delta x^{(k)} \neq 0$ and $x^{(k)} + \Delta x^{(k)}$ is infeasible, then determine a step length $t_k > 0$, such that $x^{(k)} + t_k \Delta x^{(k)}$ is feasible, that is

$$l_i \leq a_i^\top (x^{(k)} + t_k \Delta x^{(k)}) \leq u_i, \qquad i \in I \backslash I_k.$$

(Notice, that the indices $i \in E \cup I_k$ satisfy $a_i^\top \Delta x^{(k)} = 0$. Since $x^{(k)}$ is feasible, the corresponding constraints are fulfilled for arbitrary t_k.)

Two cases may occur:

* If $a_i^\top \Delta x^{(k)} > 0$, then determine an index q with

$$\begin{aligned} t_u &:= \frac{u_q - a_q^\top x^{(k)}}{a_q^\top \Delta x^{(k)}} \\ &= \min\left\{ \frac{u_i - a_i^\top x^{(k)}}{a_i^\top \Delta x^{(k)}} \;\middle|\; a_i^\top \Delta x^{(k)} > 0,\ i \in I \backslash I_k \right\}. \end{aligned}$$

(The constraint q becomes active at the upper bound u_q.)

* If $a_i^\top \Delta x^{(k)} < 0$, then determine an index r with

$$\begin{aligned} t_l &:= \frac{l_r - a_r^\top x^{(k)}}{a_r^\top \Delta x^{(k)}} \\ &= \min\left\{ \frac{l_i - a_i^\top x^{(k)}}{a_i^\top \Delta x^{(k)}} \;\middle|\; a_i^\top \Delta x^{(k)} < 0,\ i \in I \backslash I_k \right\}. \end{aligned}$$

(The constraint r becomes active at the lower bound l_r.)

Set $t_k = \min\{t_l, t_u\}$, $x^{(k+1)} = x^{(k)} + t_k \Delta x^{(k)}$ and

$$\begin{cases} I_{k+1}^u = I_k^u \cup \{q\}, I_{k+1}^l = I_k^l, & \text{if } t_k = t_u, \text{ or} \\ I_{k+1}^u = I_k^u, I_{k+1}^l = I_k^l \cup \{r\}, & \text{if } t_k = t_l. \end{cases}$$

Set $I_{k+1} = I_{k+1}^u \cup I_{k+1}^l$ and go to (1).

4.4.5.1. Remarks.

- An initial feasible point for the active set strategy can be computed similar to phase one of the simplex method known from linear programming.

- The structure of the linear equation system (4.4.12) can be exploited by reduction methods (range space or null space methods).

4.5 Numerical Example: Emergency Landing Manoeuvre

During the ascent phase of a winged two-stage hypersonic flight system some malfunction necessitates to abort the ascent shortly after separation. The upper stage of the flight system is still able to manoeuvre although the propulsion system is damaged, cf. [40], [9]. For security reasons an emergency landing trajectory with maximum range has to be found. This leads to the following optimal control problem for $t \in [0, t_f]$:

Minimize

$$-\left(\frac{\Lambda(t_f) - \Lambda(0)}{\Lambda(0)}\right)^2 - \left(\frac{\Theta(t_f) - \Theta(0)}{\Theta(0)}\right)^2 \tag{4.5.1}$$

subject to the ordinary differential equation (ODE) system for the velocity v, the inclination γ, the azimuth angle χ, the altitude h, the latitude Λ, and

the longitude Θ

$$\begin{aligned}
\dot{v} &= -D(v,h;C_L)\frac{1}{m} - g(h)\sin\gamma + \\
&\quad + \omega^2\cos\Lambda(\sin\gamma\cos\Lambda - \cos\gamma\sin\chi\sin\Lambda)R(h), \\
\dot{\gamma} &= L(v,h;C_L)\frac{\cos\mu}{mv} - \left(\frac{g(h)}{v} - \frac{v}{R(h)}\right)\cos\gamma + \\
&\quad + 2\omega\cos\chi\cos\Lambda + \omega^2\cos\Lambda(\sin\gamma\sin\chi\sin\Lambda + \cos\gamma\cos\Lambda)\frac{R(h)}{v}, \\
\dot{\chi} &= L(v,h;C_L)\frac{\sin\mu}{mv\cos\gamma} - \cos\gamma\cos\chi\tan\Lambda\frac{v}{R(h)} + \\
&\quad + 2\omega(\sin\chi\cos\Lambda\tan\gamma - \sin\Lambda) - \omega^2\cos\Lambda\sin\Lambda\cos\chi\frac{R(h)}{v\cos\gamma}, \\
\dot{h} &= v\sin\gamma, \\
\dot{\Lambda} &= \cos\gamma\sin\chi\frac{v}{R(h)}, \\
\dot{\Theta} &= \cos\gamma\cos\chi\frac{v}{R(h)\cos\Lambda},
\end{aligned}$$

with functions

$$\begin{aligned}
L(v,h,C_L) &= q(v,h)\ F\ C_L, & \rho(h) &= \rho_0\exp(-\beta h), \\
D(v,h,C_L) &= q(v,h)\ F\ C_D(C_L), & R(h) &= r_0 + h, \\
C_D(C_L) &= C_{D_0} + k\ {C_L}^2, & g(h) &= g_0(r_0/R(h))^2, \\
q(v,h) &= \tfrac{1}{2}\rho(h)v^2
\end{aligned}$$

and constants

$$\begin{aligned}
F &= 305, & r_0 &= 6.371\cdot 10^6, & C_{D_0} &= 0.017, \\
k &= 2, & \rho_0 &= 1.249512, & \beta &= 1/6900, \\
g_0 &= 9.80665, & \omega &= 7.27\cdot 10^{-5}, & m &= 115000.
\end{aligned}$$

Since the propulsion system is damaged, the mass m remains constant.

Box constraints for the two control functions C_L and μ are given by

$$\begin{array}{rcccl} 0.01 & \leq & C_L & \leq & 0.18326, \\ -\dfrac{\pi}{2} & \leq & \mu & \leq & \dfrac{\pi}{2}. \end{array}$$

The initial values for the state correspond to a starting position above Bayreuth/Germany

$$\begin{pmatrix} v(0) \\ \gamma(0) \\ \chi(0) \\ h(0) \\ \Lambda(0) \\ \Theta(0) \end{pmatrix} = \begin{pmatrix} 2150.5452900 \\ 0.1520181770 \\ 2.2689279889 \\ 33900.000000 \\ 0.8651597102 \\ 0.1980948701 \end{pmatrix}.$$

An additional restriction is given by the terminal condition

$$h(t_f) = 500.$$

The dynamic pressure constraint

$$q(h) \leq q_{max}, \qquad q_{max} = 60000$$

is an additional nonlinear state constraint. The final time t_f is assumed to be free and thus t_f is an additional optimization variable.

The infinite dimensional optimal control problem is discretized similar to the method depicted in the introduction. Instead of a simple Euler approximation for the differential equation we used a higher order Runge-Kutta scheme. The control is approximated by a continuous and piecewise linear function.

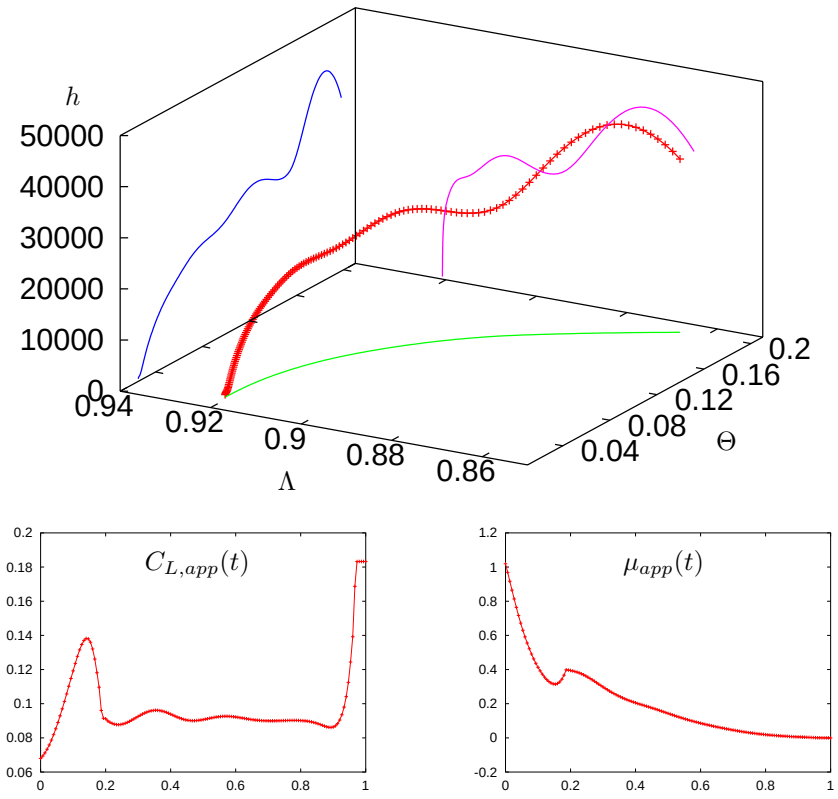


Figure 4.4: Numerical solution: 3D plot of the state (left) and approximate optimal controls $C_{L,app}$ and μ_{app} (right, normalized time scale) for 151 grid points.

Figure 4.4 shows the numerical solution for 151 discretization points. The approximate optimal final time for this highly nonlinear optimization problem is calculated to $t_f = 727.106$ seconds and the approximate objective function value is -0.7649265. Note that the control C_L has a boundary arc at the end of the time interval $[0, t_f]$. The state constraint is active in the time interval $[130.77, 135.73]$. The CPU time needed to compute the numerical solution on a Pentium III processor with 750 MHz is 3 minutes and 25 seconds.

Output from SQP method:

```
                    NPSOL  ---  Version 5.0-2      Sept 1995
                    ====================================


 Majr Minr    Step  Fun  Merit function Norm gZ  Violtn   nZ Penalty Conv
    0   14 0.0E+00    1 -2.12604554E-02 1.6E-05 5.7E+01  296 1.1E-03 F FF
    1   12 1.0E+00    2 -7.35703021E-01 4.8E-02 1.5E+02  289 1.1E-03 F FF
    2   46 6.4E-02    5 -7.40216486E-01 2.5E-02 1.5E+02  284 1.1E-03 T FF
    3   65 9.7E-02    8 -7.43820660E-01 4.8E-02 1.5E+02  294 1.1E-03 T FF
    4   12 1.2E-01   11 -7.47107812E-01 4.3E-02 1.4E+02  295 1.1E-03 T FF


 Majr Minr    Step  Fun  Merit function Norm gZ  Violtn   nZ Penalty Conv
    5    3 3.5E-01   13 -7.54207110E-01 8.6E-02 1.2E+02  295 1.1E-03 T FF
    6    1 3.1E-01   15 -7.56906775E-01 6.3E-02 1.0E+02  295 1.1E-03 T FF
    7    2 6.9E-02   18 -7.58036347E-01 4.6E-02 1.0E+02  294 1.1E-03 T FF
    8    9 8.1E-02   21 -7.58933278E-01 3.6E-02 1.0E+02  292 1.1E-03 T FF
    9    5 5.4E-02   24 -7.59674831E-01 3.4E-02 9.6E+01  295 1.1E-03 T FF


 Majr Minr    Step  Fun  Merit function Norm gZ  Violtn   nZ Penalty Conv
   10    1 7.8E-02   27 -7.60170761E-01 3.2E-02 9.1E+01  295 1.1E-03 T FF
   11    3 3.4E-01   29 -7.61649322E-01 3.4E-02 7.2E+01  295 1.1E-03 T FF
   12    2 3.6E-01   31 -7.62812035E-01 2.9E-02 5.6E+01  294 1.1E-03 T FF
   13    4 1.0E-01   34 -7.63108761E-01 1.9E-02 5.2E+01  295 1.1E-03 T FF
   14    1 2.2E-01   36 -7.63240180E-01 2.9E-02 4.8E+01  295 1.1E-03 T FF


 Majr Minr    Step  Fun  Merit function Norm gZ  Violtn   nZ Penalty Conv
   15    1 2.7E-01   38 -7.63558575E-01 2.3E-02 4.2E+01  295 1.1E-03 T FF
   16    1 2.4E-01   40 -7.63801119E-01 1.6E-02 3.7E+01  295 1.1E-03 T FF
   17    2 2.5E-01   42 -7.63987546E-01 1.4E-02 3.3E+01  296 1.1E-03 T FF
   18    3 1.0E-01   45 -7.64093590E-01 1.1E-02 3.1E+01  294 1.1E-03 T FF
   19    2 2.6E-01   47 -7.64238416E-01 1.4E-02 2.6E+01  295 1.1E-03 T FF


 Majr Minr    Step  Fun  Merit function Norm gZ  Violtn   nZ Penalty Conv
   20    1 3.3E-01   49 -7.64397491E-01 1.2E-02 2.0E+01  295 1.1E-03 T FF
   21    3 2.2E-01   51 -7.64467957E-01 9.0E-03 1.8E+01  295 1.1E-03 T FF
   22    1 2.2E-01   53 -7.64530145E-01 9.2E-03 1.6E+01  295 1.1E-03 T FF
   23    2 2.6E-01   55 -7.64587833E-01 9.9E-03 1.4E+01  296 1.1E-03 T FF
   24    2 2.3E-01   57 -7.64626394E-01 9.5E-03 1.2E+01  295 1.1E-03 T FF


 Majr Minr    Step  Fun  Merit function Norm gZ  Violtn   nZ Penalty Conv
   25    1 2.6E-01   59 -7.64666079E-01 7.9E-03 9.8E+00  295 1.1E-03 T FF
   26    1 2.4E-01   61 -7.64697801E-01 7.0E-03 8.3E+00  295 1.1E-03 T FF
   27    2 2.1E-01   63 -7.64723606E-01 7.6E-03 7.4E+00  294 1.1E-03 T FF
   28    1 2.0E-01   65 -7.64741153E-01 7.9E-03 6.5E+00  294 1.1E-03 T FF
   29    1 2.6E-01   67 -7.64760232E-01 7.5E-03 5.5E+00  294 1.1E-03 T FF

....


 Majr Minr    Step  Fun  Merit function Norm gZ  Violtn   nZ Penalty Conv
  200    1 1.0E+00  356 -7.64926474E-01 1.2E-07 4.0E-10  295 1.1E-03 T TF
  201    1 1.0E+00  357 -7.64926474E-01 1.3E-07 8.4E-10  295 1.1E-03 T TF
  202    1 1.0E+00  358 -7.64926474E-01 9.9E-08 7.0E-10  295 1.1E-03 T TF
  203    1 1.0E+00  359 -7.64926474E-01 6.9E-08 4.0E-10  295 1.1E-03 T TF
  204    1 1.0E+00  360 -7.64926474E-01 6.5E-08 1.7E-10  295 1.1E-03 T TF


 Majr Minr    Step  Fun  Merit function Norm gZ  Violtn   nZ Penalty Conv
  205    1 1.0E+00  361 -7.64926474E-01 7.4E-08 1.9E-10  295 1.1E-03 T TF
  206    1 1.0E+00  362 -7.64926474E-01 6.9E-08 1.8E-10  295 1.1E-03 T TF
  207    1 1.0E+00  363 -7.64926474E-01 3.8E-08 1.7E-10  295 1.1E-03 T TF
  208    1 1.0E+00  364 -7.64926474E-01 2.2E-08 7.1E-11  295 1.1E-03 T TF
  209    1 1.0E+00  365 -7.64926474E-01 2.2E-08 1.0E-11  295 1.1E-03 T TF


 Majr Minr    Step  Fun  Merit function Norm gZ  Violtn   nZ Penalty Conv
  210    1 1.0E+00  366 -7.64926474E-01 3.1E-08 3.6E-11  295 1.1E-03 T TF
  211    1 1.0E+00  367 -7.64926474E-01 3.0E-08 5.7E-11  295 1.1E-03 T TF
  212    1 1.0E+00  368 -7.64926474E-01 1.5E-08 2.6E-11  295 1.1E-03 T TF
  213    1 1.0E+00  369 -7.64926474E-01 1.4E-08 4.1E-11  295 1.1E-03 T TF
  214    1 1.0E+00  370 -7.64926474E-01 1.7E-08 8.7E-12  295 1.1E-03 T TT

 Exit NPSOL - Optimal solution found.

 Final nonlinear objective value =  -0.7649265
```

Chapter 5

State-Constrained Differential Inclusions

5.1 Preliminaries

We summarize some fields of applications from natural sciences, engineering sciences and operations research where differential inclusions arise as a natural tool of mathematical modelling:

- differential equations with discontinuities with respect to the state variables
- non-smooth optimization problems
- control problems

We consider two special applications a little closer.

5.1.1. Non-Smooth Hamiltonean Systems. Let

$$\begin{aligned} \varphi &: \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m \longrightarrow \mathbb{R} \ , \\ \psi &: \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m \longrightarrow \mathbb{R}^n \ , \\ \alpha &: \mathbb{R}^n \longrightarrow \mathbb{R}^{s_a} \ , \\ \beta &: \mathbb{R}^n \longrightarrow \mathbb{R}^{s_b} \ , \end{aligned}$$

be mappings and $U \subset \mathbb{R}^m$ be a control region.

Minimize

$$\int_a^b \varphi(t, y(t), u(t))dt$$

subject to $y(\cdot) \in AC([a,b])^n,\ u(\cdot) \in L_\infty([a,b])^m,$

$$\dot{y}(t) = \psi(t, y(t), u(t)) \qquad \text{for a. a. } t \in [a,b]\ ,$$

$$\alpha_i(y(a)) \begin{cases} \leq 0\ (i = 1, \ldots, s'_a) \\ = 0\ (i = s'_a, \ldots, s_a) \end{cases},\quad \beta_i(y(b)) \begin{cases} \leq 0\ (i = 1, \ldots, s'_b) \\ = 0\ (i = s'_b, \ldots, s_b) \end{cases},$$

$$u(t) \in U \text{ for all } t \in [a,b]\ .$$

Let $\hat{y}(\cdot), \hat{u}(\cdot)$ be a (local) optimal solution of this control problem and assume that all mappings $\varphi, \psi, \alpha, \beta$ are continuous and, with respect to all state coordinates, continuously partially differentiable. Then the following version of Pontryagin's Maximum Principle holds:

There exist multipliers

$$\lambda_0 \in \mathbb{R},\ \lambda_a \in \mathbb{R}^{s_a},\ \lambda_b \in \mathbb{R}^{s_b}$$

and a function

$$p(\cdot) \in AC([a,b])^n$$

with the following properties

(i) $\lambda_0, \lambda_a, \lambda_b, p(\cdot)$ do not vanish (identically) simultaneously.

(ii)

$$\lambda_0 \geq 0\ ,$$

$$(\lambda_a)_i \begin{cases} \geq 0, & (i = 1, \ldots, s'_a), \\ = 0, & \text{if } \alpha_i(\hat{y}(a)) < 0, \end{cases}$$

$$(\lambda_b)_i \begin{cases} \geq 0, & (i = 1, \ldots, s'_b), \\ = 0, & \text{if } \beta_i(\hat{y}(b)) < 0\ , \end{cases}$$

(*complementary slackness condition*).

(iii)

$$\dot{p}(t) = -\big[\psi_x^\star(t, \hat{y}(t), \hat{u}(t))p(t) - \lambda_0 \varphi_x^\star(t, \hat{y}(t), \hat{u}(t))\big]$$

for almost all $t \in [a,b]$,

$$\begin{aligned} p(a) &= \alpha_x(\hat{y}(a))^\star \lambda_a\ , \\ p(b) &= -\beta_x(\hat{y}(b))^\star \lambda_b \end{aligned}$$

(*adjoint system equation*, together with (ii) the *transversality conditions* follow).

(iv) For almost all $t \in [a,b]$:

$$\begin{aligned} & p^\star(t)\psi(t,\hat{y}(t),u) - \lambda_0\varphi(t,\hat{y}(t),u) \\ \leq\ & p^\star(t)\psi(t,\hat{y}(t),\hat{u}(t)) - \lambda_0\varphi(t,\hat{y}(t),\hat{u}(t)) \qquad (u \in U) \end{aligned}$$

(*maximum principle*).

Introducing the Hamiltonean function

$$H(t,x,u,p,\lambda_0) = p^\star\psi(t,x,u) - \lambda_0\varphi(t,x,u)\ ,$$

the system equation and the adjoint system equation can be written as

$$\begin{aligned} \frac{d}{dt}\hat{y}(t) &= \frac{\partial H}{\partial p}(t,\hat{y}(t),\hat{u}(t),p(t),\lambda_0)^\star\ , \\ \frac{d}{dt}\hat{p}(t) &= -\frac{\partial H}{\partial x}(t,\hat{y}(t),\hat{u}(t),p(t),\lambda_0)^\star \end{aligned}$$

for almost all $t \in [a,b]$.

The maximum principle (iv) reads

$$\sup_{u\in U} H(t,\hat{y}(t),u,p(t),\lambda_0) = H(t,\hat{y}(t),\hat{u},p(t),\lambda_0)\ .$$

We denote by $\hat{U}(t,\hat{y}(t),p(t),\lambda_0)$ the set of maximal solutions of this auxiliary maximization problem, and arrive at a boundary value problem for a Hamiltonean differential inclusion

$$\begin{pmatrix} \frac{d}{dt}\hat{y}(t) \\ \frac{d}{dt}p(t) \end{pmatrix} \in \left\{ \begin{pmatrix} \frac{\partial H}{\partial p}(t,\hat{y}(t),u,p(t),\lambda_0) \\ -\frac{\partial H}{\partial x}(t,\hat{y}(t),u,p(t),\lambda_0) \end{pmatrix} : u \in \hat{U}(t,\hat{y}(t),p(t),\lambda_0) \right\}\ .$$

■

The set-valued mapping $\hat{U}(t,x,p,\lambda_0)$ reflects the structure of the optimal control, especially its jump behaviour, depending on the values x and p of the unknown trajectory resp. adjoint trajectory.

Sometimes, especially, if the control region U is simple or even a set of finitely many discrete control vectors, the set-valued mapping

$$\hat{U}(t,x,p,\lambda_0)$$

can be computed directly or numerically.

But in realistic applications, additional state constraints occur, and the adjoint system equation has to be replaced by an integral equation with respect to Lebesgue-Stieltjes integration. In that case, it is nearly hopeless to solve the resulting set-valued integral equations numerically.

But it could be possible to attack a state-constrained optimal control problem by direct methods. If one is interested in approximations of the reachable sets one could even omit for a while the objective function. This is even obligatory, if the objective function is not known in advance or if, for political reasons, there is no objective function which is generally accepted.

5.1.2. Control Problem with State Constraints. *Compute*

$$y(\cdot) \in AC([a,b])^n,\ u(\cdot) \in L_\infty([a,b])^m$$

with

$$\begin{aligned}
&\dot{y}(t) = \psi(t, y(t), u(t)) \qquad \text{for a. a. } t \in [a,b]\ , \\
&y(a) \in Y_a\ , \\
&S_i(t, y(t)) \le 0 \qquad (i = 1, \dots, s) \\
&u(t) \in U \qquad (a \le t \le b)\ .
\end{aligned}$$

This problem is related to the following *state constrained differential inclusion*

$$\begin{aligned}
&\dot{y}(t) \in \psi(t, y(t), U)\ , \\
&y(a) \in Y_a\ , \\
&S_i(t, y(t)) \le \Theta \qquad (i = 1, \dots, s)\ .
\end{aligned}$$

Interesting sets to be analyzed analytically or approximated numerically are the *set of all feasible* ("viable") *solutions*, resp. the *reachable set* at time b, resp. the *stability behaviour* of all solutions for $t \longrightarrow \infty$, resp. the set-valued limit, if it exists at all in an appropriate sense. ∎

All problem classes described up to now can be modelled by the following

5.1.3. State-Constrained Differential Inclusion. Let $Y_a \subset \mathbb{R}^n$ be a nonempty subset of $\mathbb{R}^n$, $F : \mathbb{R} \times \mathbb{R}^n \Longrightarrow \mathbb{R}^n$ a set-valued mapping, $\Theta : \mathbb{R} \Longrightarrow \mathbb{R}^n$ a set-valued mapping.

Compute $y(\cdot) \in AC([a,b])^n$ *with*

$$\begin{aligned} \dot{y}(t) &\in F(t,y(t)) \qquad \text{for a. a. } t \in [a,b] \ , \\ y(a) &\in Y_a \ , \\ y(t) &\in \Theta(t) \qquad (a \le t \le b) \ . \end{aligned}$$

■

5.1.4. Definitions. We will use the following notations:

$Y[b;a,Y_a]$ denotes the *set of all solutions* of the differential inclusion

$$\begin{aligned} \dot{y}(t) &\in F(t,y(t)) \qquad \text{for a.a. } t \in [a,b] \ , \\ y(t) &\in Y_a \end{aligned}$$

without state constraints.

$Y_\Theta[b;a,Y_a]$ denotes the *set of all solutions of State-Constrained Differential Inclusion* 5.1.3.

As long as initial time a, initial set Y_a, and the final time b are clear from the context, we will suppress them in the notation,

$$\begin{aligned} Y &= Y(b) = Y(b;a,Y_a) \ , \\ Y_\Theta &= Y_\Theta(b) = Y_\Theta(b;a,Y_a) \ . \end{aligned}$$

Analogously, by

$$X(t;a,Y_a) = \{x \in \mathbb{R}^n : \text{there exists } y(\cdot) \in Y(b;a,Y_a) \text{ with } y(t) = x\}$$

resp.

$$X_\Theta(t;a,Y_a) = \left\{x \in \mathbb{R}^n : \text{there exists } y(\cdot) \in Y^\Theta(b;a,Y_a) \text{ with } y(t) = x\right\}$$

we denote the *reachable set at time* $t \in [a,b]$ *with* resp. *without state constraints.*

Remark. Naturally, in general

$$X_\Theta(t;a,Y_a) \underset{\neq}{\subset} X(t;a,Y_a) \cap \Theta(t) \ .$$

State constraints complicate the structure of solution sets and reachable sets considerably, even for relative simple differential inclusions. ■

For convergence results for subsets in $\mathbb{R}^n$ resp. subsets of the space of trajectories we need some measure for the distance of sets. We will use systematically Hausdorff-distance in $\mathbb{R}^n$ with respect to some norm $\|\cdot\|$ in $\mathbb{R}^n$ resp. Hausdorff-distance in the space of trajectories with respect to some function space norm $\|\cdot\|$. Whereas in $\mathbb{R}^n$ any norm could be used, the choice of the norm in function spaces is crucial.

5.1.5. Hausdorff-Distance. Let $A \subset X$ be a nonempty subset of a normed space X with norm $\|\cdot\|$ and $x \in X$.

$$\operatorname{dist}(x, A) = \inf_{a \in A} \|x - a\|$$

is called *distance* of x from A.

Let A , B be two nonempty sets, then

$$\operatorname{dist}(B, A) = \sup_{b \in B} \operatorname{dist}(x, A)$$

is called *one-sided distance* of B from A, and

$$\operatorname{haus}(B, A) = \max\{\operatorname{dist}(B, A), \operatorname{dist}(A, B)\}$$

is called **Hausdorff-distance** between A and B.

Remark. Conceptually, to prove that

$$\operatorname{haus}(A, B) \leq \epsilon \ ,$$

instead of exploiting the definition directly, it is often easier to prove:

For each $a \in A$ there is $b \in B$ with

$$\|a - b\| \leq \epsilon$$

and for each $b \in B$ there is $a \in A$ with

$$\|b - a\| \leq \epsilon \ .$$

Especially in the space of trajectories, where we usually use supremum norm $\|\cdot\|_\infty$ with respect to some norm $\|\cdot\|$ in $\mathbb{R}^n$,

$$\|y(\cdot) - z(\cdot)\|_\infty \leq \epsilon$$

is proved by showing

$$\|y(t) - z(t)\| \leq \epsilon \qquad (a \leq t \leq b) \ .$$

■

5.2 Discrete Approximations

To approximate the set

$$Y(b;a,y_a)$$

of all solutions of the differential inclusion

$$\begin{aligned} \dot{y}(t) &\in F(t,y(t)) \qquad \text{for a. a. } t \in [a,b] \ , \\ y(a) &\in y_a \end{aligned}$$

without state constraints, there are some discretization methods available which reach from nothing more than mere constructive existence proofs, methods of broken order of convergence equal to 1/2, first order of convergence, second order of convergence, merely conceptual higher order of convergence methods to methods of arbitrary order of convergence for linear differential inclusions with additional smoothness properties. We need for our purposes only one exemplary method, and choose the simplest one, the following

5.2.1. Set-Valued Euler Method. Choose $N \in \mathbb{N}$ and an equidistant grid

$$\mathbb{G}_N = \{t_0, \ldots, t_N\}$$

with

$$\begin{aligned} t_j &= a + jh \qquad (j = 0, \ldots, N) \ , \\ h &= \frac{b-a}{N} \ . \end{aligned}$$

Compute all sequences

$$\eta^N = \left(\eta_0^N, \eta_1^N, \ldots, \eta_N^N\right) \in (\mathbb{R}^n)^{N+1}$$

with

$$\begin{aligned} \eta_{j+1}^N &\in \eta_j^N + hF(t_j, \eta_j^N) \qquad (j = 0, \ldots, N-1) \ , \\ \eta_0 &= y_a \ , \\ \eta_j &\in \Theta \qquad (j = 0, \ldots, N) \ . \end{aligned}$$

By

$$Y^N = Y^N(b) = Y^N(b;a,y_a)$$

resp.

$$Y_\Theta^N = Y_\Theta^N(b) = Y_\Theta^N(t; a, y_a)$$

we denote the set of all discrete solutions, without state constraints (i.e. $\Theta = \mathbb{R}^n$) resp. with state constraints.

These sequences could be equally well be understood as grid functions

$$\eta^N : \mathbb{G}_N \longrightarrow (\mathbb{R}^n)^{N+1} \text{ resp. } \eta_\Theta^N : \mathbb{G}_N \longrightarrow (\mathbb{R}^n)^{N+1} \ .$$

■

The distance between the set Y^N of all discrete solutions and the set Y of all continuous solutions is measured by discrete Hausdorff-distance:

$$\text{haus}(Y^N, Y) = \text{haus}(Y^N, Y|_{\mathbb{G}_N})$$

where Hausdorff-distance in the discrete spaces is to be understood with respect to the discrete supremum norm

$$\|y^N(\cdot) - \eta^N(\cdot)\|_\infty = \sup_{j=0,\dots,N} \|y_j^N - \eta_j^N\| \ .$$

The following result is well-known from [11].

5.2.2. Theorem. *Let $F : [a,b] \times \mathbb{R}^n \Longrightarrow \mathbb{R}^n$ be nonempty, convex, and compact valued, Lipschitz continuous, satisfying the linear growth condition, i.e. there exists a constant $c_1 \geq 0$ with*

$$\|\xi\| \leq c_1(\|x\| + 1)$$

for all $\xi \in F(t,x),\ t \in [a,b],\ x \in \mathbb{R}^n$. Then there exists a constant $c_2 \geq 0$ with

$$\text{haus}(Y^N, Y) \leq c_2 h \ .$$

■

This result has to be extended to differential inclusions with state constraints. This extension needs two stability properties, indeed variants of the Gronwall-Filippov-Ważewski-Lemma for state-constrained continuous resp. discrete differential inclusions.

The continuous case has been proved in [16], [17], some ideas of the proofs being based on [57].

5.2.3. Theorem. *In addition to the assumptions of Theorem 5.2.2, let $\Theta(t) \equiv \Theta$ be a nonempty compact subset of $\mathbb{R}^n$, and let there exist positive constants μ, r, q and a continuous selection*

$$f : [a,b] \times [\Theta \cap B(\partial\Theta, \mu)] \longrightarrow \mathbb{R}^n \ ,$$
$$f(t,x) \in F(t,x) \qquad \text{for all } t \in [a,b],\ x \in \Theta \cap B(\partial\Theta, \mu) \ ,$$

such that

$$B(x + hf(t,x), hr) \subset \Theta$$

for all $t \in [a,b],\ x \in \Theta \cap B(\partial\Theta, \mu),\ h \in (0,q]$. Here, $B(A,\epsilon)$ denotes the ϵ-neighbourhood of a set $A \subset \mathbb{R}^n$

$$B(A,\epsilon) = \{x \in \mathbb{R}^n : \exists\ z \in A \text{ with } \|x - z\| \le \epsilon\} \ .$$

Then for every compact subset $Y_a \subset \Theta$ there exists a constant c_3 with the following property:

For every $y_a \in Y_a$ and every $y(\cdot) \in Y$ there exists $y_\Theta(\cdot) \in Y_\Theta$ with

$$\sup_{a \le t \le b} \|y(t) - y_\Theta(t)\| \le c_3 \sup_{a \le t \le b} \operatorname{dist}(y(t), \Theta) \ .$$

■

The discrete analogue of this result was proved in [10].

5.2.4. Theorem. *Let all assumptions of Theorem 5.2.3 be satisfied.*

Then there exists $N_0 \in N$ such that for every $c_0 > 0$ there is a constant c_4 with the following properties:

For all $N \in \mathbb{N},\ N \ge N_0$ and all $\eta^N \in Y^N$ with

$$\sup_{j=0,\dots,N} \operatorname{dist}(\eta_j^N, \Theta) \le c_0 h$$

there exists $\eta_\Theta^N \in Y_\Theta^N$ with

$$\sup_{j=0,\dots,N} \|\eta_j^N - \eta_{\Theta j}^N\| \le c_4 h \ .$$

■

Before we combine Theorems 5.2.2, 5.2.3, and 5.2.4 to a first order of convergence result for discrete approximations of state-constrained differential inclusions, we have a closer look to the decisive assumption:

$$B(x + hf(t,x), hr) \subset \Theta$$

for all $t \in [a,b],\ x \in \Theta \cap B(\partial\Theta,\mu),\ h \in (0,q]$

Intuitively, for points $x \in \Theta$ near the boundary $\partial\Theta$ of Θ it should be possible to choose a selection $f(t,x) \in F(t,x)$ which transfers x into the interior of Θ.

There is another interesting connection with **viability**. Choose $t \in [a,b]$ and $x \in \partial\Theta$ arbitrarily, and a sequence $(h_\nu)_{\nu\in\mathbb{N}}$ of positive numbers with

$$\lim_{\nu\longrightarrow\infty} h_\nu = 0 \ .$$

Then we have

$$\begin{aligned} &x + h_\nu f(t,x) \in \Theta \qquad (\nu \in \mathbb{N}) \ , \\ &\lim_{\nu\longrightarrow\infty} x + h_\nu f(t,x) = x \in \Theta \ , \\ &\lim_{\nu\longrightarrow\infty} \frac{1}{h_\nu}[x + h_\nu f(t,x) - x] = f(t,x) \in F(t,x) \ , \end{aligned}$$

i.e. $f(t,x)$ belongs to the so-called tangential cone $T(\Theta,x)$ of the set Θ at the point $x \in \partial\Theta$. If $x \in \mathrm{int}(\Theta)$, then $T(\Theta,x) = \mathbb{R}^n$, hence it follows

$$F(t,x) \cap\ T(\Theta,x) \neq \emptyset \qquad (t \in [a,b],\ x \in \Theta) \ .$$

This is the famous **viability condition**, which is (at least in the autonomous case) necessary and sufficient for the existence of a solution to the state-constrained differential inclusion.

Consequently, the essential assumption, which we need, is a strengthened version of the viability condition. It implies the stability properties described in Theorems 5.2.3 and 5.2.4.

We now combine the results of Theorems 5.2.2, 5.2.3, and 5.2.4. Typically, since Hausdorff-distances are involved, the proof consists of two parts.
I. Choose

$$\eta_\Theta^N \in Y_\Theta^N$$

arbitrarily, then naturally

$$\eta_\Theta^N \in Y^N \ .$$

By Theorem 5.2.2 there exists $c_2 \geq 0$ and $y(\cdot) \in Y$ with

$$\max_{j=0,\ldots,N} \|\eta_{\Theta j}^N - y(t_j)\| \leq c_2 h \ .$$

Since all solutions $y \in Y$ are uniformly Lipschitz due to the growth condition with some Lipschitz constant L, for points t between t_j and t_{j+1} it follows

$$\begin{aligned} \|\eta^N_{\Theta j} - y(t)\| &\leq \|\eta^N_{\Theta j} - y(t_j)\| \\ &\quad + \|y(t_j) - y(t)\| \\ &\leq c_2 h + Lh \end{aligned}$$

for $t_j \leq t \leq t_{j+1}$ and $j = 0, \ldots, N-1$.

Hence,

$$\sup_{a \leq t \leq b} \operatorname{dist}(y(t), \Theta) \leq c_2 h + Lh \ .$$

Due to Theorem 5.2.3, there exists $y_\Theta \in Y_\Theta$ with

$$\sup_{a \leq t \leq b} \|y(t) - y_\Theta(t)\| \leq c_3(c_2 + L)h \ .$$

Altogether, we have

$$\max_{j=0,\ldots,N} \|\eta^N_{\Theta j} - y_\Theta(t_j)\| \leq C_2 h + c_3(c_2 + L)h \ .$$

II. Choose

$$y_\Theta \in Y_\Theta$$

arbitrarily, then naturally

$$y_\Theta \in Y \ .$$

By Theorem 5.2.2 there exists $c_2 \geq 0$ and $\eta^N \in Y^N$ with

$$\max_{j=0,\ldots,N} \|y_\Theta(t_j) - \eta^N_j\| \leq c_2 h \ ,$$

hence

$$\max_{j=0,\ldots,N} \operatorname{dist}(\eta^N_j, \Theta) \leq c_2 h \ .$$

Due to Theorem 5.2.4, there exists $\eta^N_\Theta \in y^N_\Theta$ with

$$\max_{j=0,\ldots,N} \|\eta^N_j - \eta^N_{\Theta j}\| \leq c_4 h \ ,$$

at least for all $N \geq N_0$.

Altogether it follows

$$\max_{j=0,\ldots,N} \|y_\Theta(t_j) - \eta^N_{\Theta j}\| \leq (c_2 + c_4)h \ .$$

With I. and II., we have proved

5.2.5. Theorem ([10]). *Assume all assumptions of Theorems 5.2.2 and 5.2.3 to be satisfied. Then there exist a positive constant C and $N_0 \in \mathbb{N}$ such that for all $N \in \mathbb{N}$, $N \geq N_0$,*

$$\operatorname{haus}(y_\Theta^N, y_\Theta) \leq Ch \ .$$

■

5.3 Linear Differential Inclusions

All numerical results where obtained in the Numerical Analysis, Optimization, and Optimal Control Group of the Chair of Applied Mathematics at the University of Bayreuth, by [2] (approximation of reachable sets of linear differential inclusions without state constraints) and by [10] (approximation of solution sets and reachable sets of nonlinear differential inclusions with state constraints). The following results are cited from [10].

First, we apply the set-valued Euler-method to a simple control problem without state constraints. For this more or less academic problem, the reachable sets can be approximated very closely by set-valued integration methods, compare [2], i. e. a very good reference solution is available, and the order of convergence with respect to Hausdorff-distance can be checked.

5.3.1. Example.
Consider the linear differential inclusion

$$\dot{y}(t) \in F(t, y(t)) = Ay(t) + B[-1, 1] \qquad (\text{for a. a. } t \in [0, 2\pi]),$$
$$y(0) = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

where

$$A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

The approximation of the reachable set at time $t = 2\pi$ by Euler's method has first order of convergence with respect to the step size h_N.

Figure 5.1 shows the approximations of the reachable set for $N = 20$ (left upper picture), 40 (right upper one), 80 (left lower one) and 150 (right lower one) and, indeed, gives a geometrical impression of first order convergence.

Table 5.1 gives estimates of the order of convergence of the approximated reachable set for time $t = 2\pi$, which corresponds to Euler's method. These estimates are determined with the help of a reference set (the smallest one with dark colour in Figure 5.1), which is computed by using a second order combination method (iterated set-valued trapezoidal rule with the Euler-Cauchy method as an ODE solver; see [2]) for the linear differential inclusion with $N_{ref} = 100\ 000$.

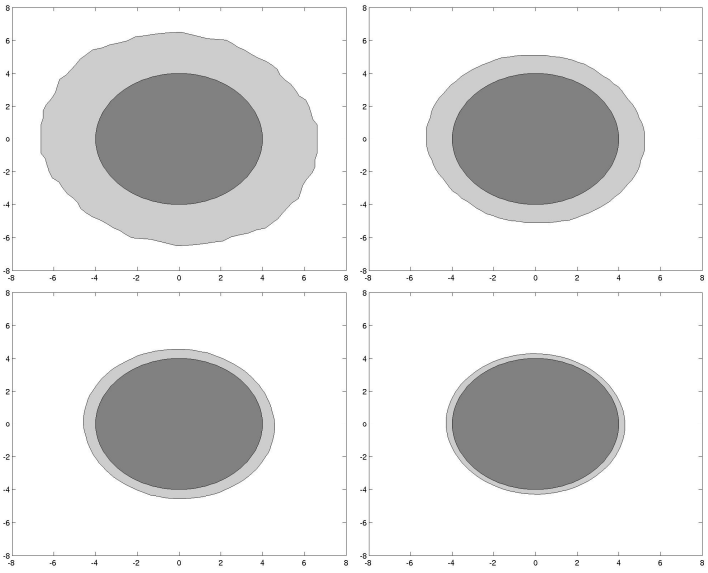


Figure 5.1: Euler method for linear DI without state constraints

N	Hausdorff distance to the reference set	estimated order of convergence
10	9.27119812	——
16	4.34742923	1.611
20	2.84531090	1.704
32	1.36388572	1.672
40	1.29393261	1.136
64	0.80931448	0.753
80	0.63130263	1.035
150	0.32569214	1.052

Table 5.1: Estimated order of convergence (Euler's method)

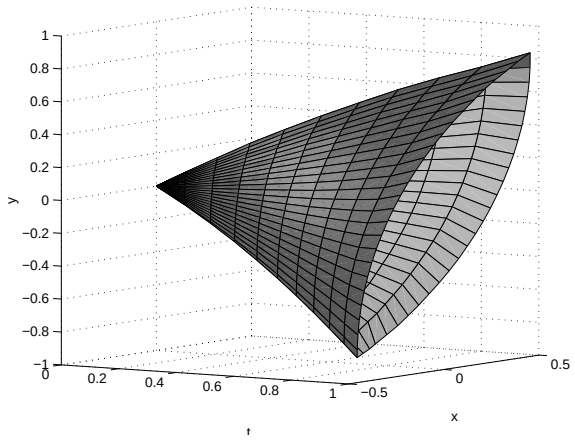


Figure 5.2: Integral funnel without state constraints (Euler's method)

Figure 5.2 shows the integral funnel without state constraints, which is calculated by Euler's method for $N = 60$ on the interval $[0, 1]$. It represents the evolution of the reachable set in time.

Now we apply Euler's method to the same problem with state-constraints. With the help of the finest approximation as reference set, again the order of convergence can be checked.

5.3.2. Example.
Consider the linear differential inclusion in Example 5.3.1 on the interval $[0,1]$ *with the following state constraints:*

$$\Theta := \{y = (y_1, y_2)^t \in S : \ g(y) := y_2 - a \leq 0\},$$

where $a = 0.1$ *and* $S \subset \mathbb{R}^2$ *is a compact box containing the values of all trajectories on* $[0,1]$ *(e.g.* $S = [-0.5, 0.5] \times [-1, 1]$*, see Figure 5.2).*

Figure 5.3 shows the approximations of the reachable set with state constraints at time $t = 1$ for $N = 5$ (left upper picture), 10 (right upper one), 20 (left lower one), 40 (right lower one) by Euler's method, which exhibits first order of convergence, since all required assumptions are fulfilled.

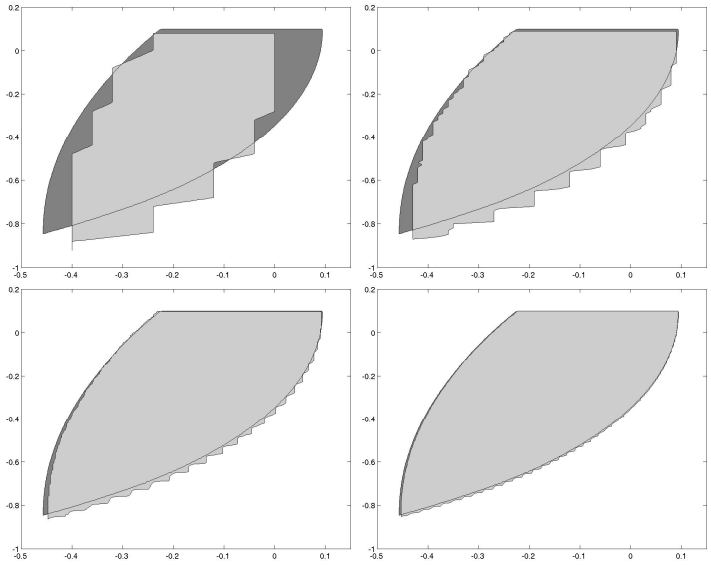


Figure 5.3: Euler method for linear DI with state constraints

Table 5.2 gives estimates of the order of convergence of the approximated reachable set with state constraints.

N	Hausdorff distance to the reference set	estimated order of convergence
2	0.24254329	——
4	0.17968973	0.432
5	0.11978088	0.769
8	0.09532583	0.673
10	0.06202789	0.847
16	0.04049424	0.860
20	0.02764814	0.943
32	0.01612128	0.977
40	0.01129338	1.023
60	0.00583071	1.096

Table 5.2: estimated order of convergence (Euler's method with state constraints)

Figure 5.4 shows the integral funnel with state constraints for the time interval $[0, 1]$, which is calculated by Euler's method for $N = 60$. It represents the evolution of the reachable set for the constrained problem in time. Only for optical reasons, we chose in this figure the value $a = 0.4$.

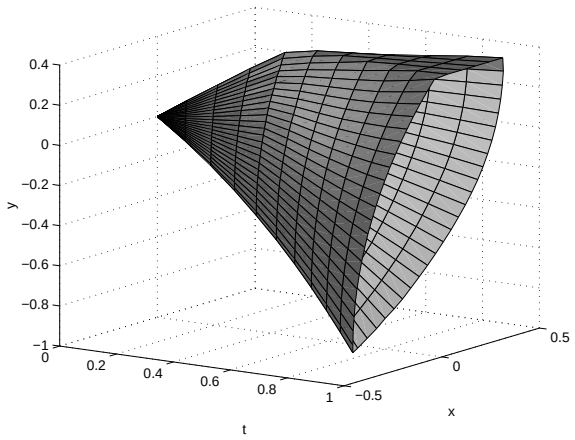


Figure 5.4: Integral funnel with state constraints (Euler's method)

In Figure 5.5 we see that the reachable set with state constraints (dotted line) represents only a subset of the intersection of the state constraints Θ and the reachable set without state constraints (the solid line).

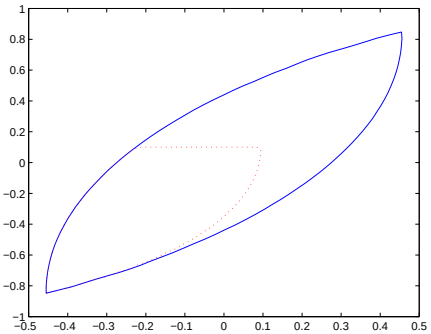


Figure 5.5: Reachable sets with/without state constraints (Euler's method N=60)

5.4 Climate Impact Research

Particularly intriguing applications of dynamical systems with uncertainties arise in the field of climate impact research. As an example, we cite a simple model from [49] which is based on a WBGU scenario for climate change assessment (Wissenschaftlicher Beirat der Bundesrepublik Deutschland für globale Umweltveränderungen, German Advisory Council on Global Change).

5.4.1. Climate Change Model (4D). Let

- $F(\cdot)$ be the **cumulated anthropogenic CO_2-emission**,
- $C(\cdot)$ the **atmospheric carbon concentration**,
- $T(\cdot)$ the **global annual mean temperature**,
- $E(\cdot)$ the **anthropogenic CO_2-emission**,
- $u(\cdot)$ the **relative rate of anthropogenic CO_2-emission change**.

The **control region** is an interval

$$U = [u_{min}, u_{max}] \ .$$

In addition, restrictions are required for the global mean temperature,

$$T_{\min} \ \leq \ T(t) \ \leq T_{\max},$$

and damage costs, essentially due to the relative rate of change of global mean temperature, are required to be bounded via the constraint

$$S(T, \dot{T}) \ \leq \ S_{\max} \ .$$

The function S could be modelled in different ways, one proposal is the following

$$S(T,\dot{T}) = \begin{cases} S_{\max}(\frac{\dot{T}}{\dot{T}_{\max}})^2 (T - T_{\min})^{-1} & \text{for } T_{\min} \leq T \leq T_{\min} + 1, \\ S_{\max}(\frac{\dot{T}}{\dot{T}_{\max}})^2 & \text{for } T_{\min} + 1 \leq T \leq T_{\max} - 1, \\ S_{\max}(\frac{\dot{T}}{\dot{T}_{\max}})^2 (T_{\max} - T)^{-1} & \text{for } T_{\max} - 1 \leq T \leq T_{\max} \ . \end{cases}$$

Roughly, this yields the following **tolerable window** for the rate of change of global mean temperature $\dot{T}$ versus global mean temperature T.

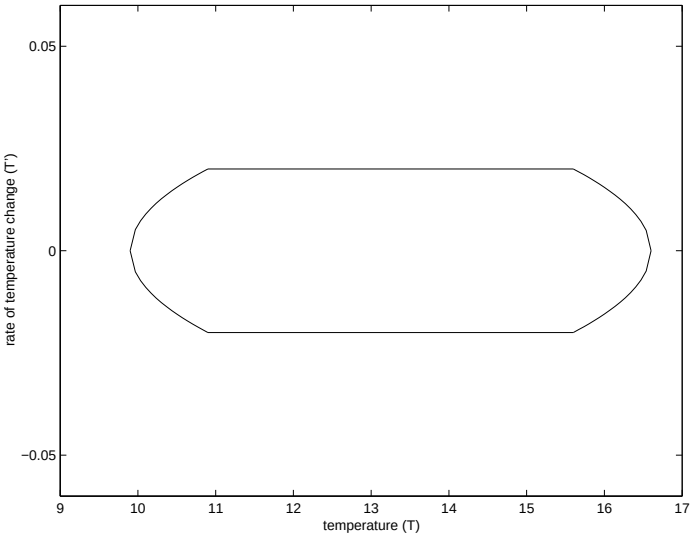


Figure 5.6: Boundary of the state constraints in terms of T and $\dot{T}$

The **systems equations** are

$$\begin{aligned}
\dot{E}(t) &= E(t)u(t) \ , \\
\dot{F}(t) &= E(t) \ , \\
\dot{C}(t) &= B \cdot F(t) + \beta \cdot E(t) - \sigma \cdot (C(t) - C_1) \ , \\
\dot{T}(t) &= \mu \cdot \ln\left(\frac{C(t)}{C_1}\right) - \alpha \cdot (T(t) - T_1)
\end{aligned}$$

for $t \in [0, t_f]$, where B, β, σ, μ, α, T_1 are constants.
Hence, the mixed phase constraints for T and $\dot{T}$ can be expressed as

$$S\left(T, \mu \cdot \ln\left(\frac{C(t)}{C_1}\right) - \alpha \cdot (T(t) - T_1)\right) \ \leq \ S_{\max} \ .$$

This window is roughly of the following form

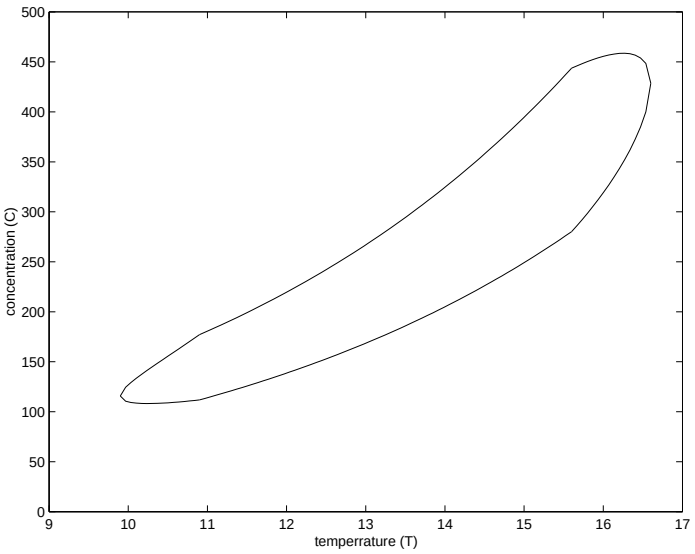


Figure 5.7: Boundary of the state constraints in terms of T and C

Within the context of global warming, this tolerable window is reduced further to

$$\begin{aligned} T_1 \leq\ & T(t) \leq T_{max}, \\ 0 \leq\ & \dot{T}(t) \leq \dot{T}_{crit}(T(t)) \end{aligned}$$

where $T_1 = T_{preindustrial} = 14.6\ ^oC$ is the preindustrial global mean temperature and

$$\dot{T}_{crit}(T(t)) = \begin{cases} \dot{T}_{\max} & \text{if } T_1 \leq T(t) \leq T_{\max} - 1, \\ \dot{T}_{\max}\sqrt{T_{\max} - T(t)} & \text{if } T_{\max} - 1 \leq T(t) \leq T_{\max}\ . \end{cases}$$

Choosing the set $\Theta \subset \mathbb{R}^2$ appropriately, we arrive at last at the following state-constrained control problem:

Determine **states**

$$F(\cdot), C(\cdot), T(\cdot), E(\cdot) \in \mathrm{AC}([0, t_f])$$

and **controls**

$$u(\cdot) \in \mathrm{L}_\infty([0, t_f])$$

with

$$\begin{aligned}
\dot{E}(t) &= E(t)u(t) \ , \\
\dot{F}(t) &= E(t) \ , \\
\dot{C}(t) &= B \cdot F(t) + \beta \cdot E(t) - \sigma \cdot (C(t) - C_1) \ , \\
\dot{T}(t) &= \mu \cdot \ln\left(\frac{C(t)}{C_1}\right) - \alpha \cdot (T(t) - T_1)
\end{aligned}$$

for almost all $t \in [0, t_f]$,

$$(C(t), T(t)) \in \Theta \qquad (0 \le t \le t_f) \ ,$$

and

$$u(t) \in [u_{min}, u_{max}] \qquad (0 \le t \le t_f) \ .$$

■

In connection with such models, the most important question is:

Is it possible to control the emission profile $E(\cdot)$ in such away that the system state stays feasible ('viable') on a (long) time interval $[0, t_f]$?

The answer to this question amounts to the knowledge of the **set of all feasible solutions** on some time interval $[0, t_f]$ resp. of the time development of the **reachable sets**.

For the above four-dimensional climate change model, the set of all feasible solutions was approximated by the set-valued Euler method for state-constrained differential inclusions in [10].

In the following, we study a reduced version with state space dimension 3 where instead of relative CO_2-emission $u(\cdot)$ the emission profile $E(\cdot)$ itself is used as control.

All numerical results, presented in the following, have been obtained by I. A. Chahma [10], some auxiliary tools have been contributed by R. Baier [2], and C. Büskens [8], all of them members of the Research Group on Numerical Analysis, Optimization and Optimal Control at the University of Bayreuth.

5.4.2. Climate Change Model (3D).
Determine **states**

$$F(\cdot), C(\cdot), T(\cdot) \in \mathrm{AC}([0, t_f])$$

and **controls**

$$E(\cdot) \in \mathrm{L}_\infty([0, t_f])$$

with

$$\dot{F}(t) = E(t) \ , \tag{5.4.1}$$

$$\dot{C}(t) = B \cdot F(t) + \beta \cdot E(t) - \sigma \cdot (C(t) - C_1) \ , \tag{5.4.2}$$

$$\dot{T}(t) = \mu \cdot \ln\left(\frac{C(t)}{C_1}\right) - \alpha \cdot (T(t) - T_1) \tag{5.4.3}$$

for almost all $t \in [0, t_f]$,

$$(C(t), T(t)) \in \Theta \qquad (0 \le t \le t_f) \ ,$$

and

$$E_{min}(t) \le E(t) \le E_{max}(t) \qquad (0 \le t \le t_f) \ .$$

■

Using some insight into feasible emission profiles of the 4-dimensional model, the lower and upper bounds of the control region are modelled as

$$E_{\min}(t) = E_0 e^{-u_{\max} t}$$

and

$$E_{\max}(t) = \min\left\{E_0 e^{u_{\max} t}, \hat{E}_{max}\right\}$$

with

$$E_0 = 7.9, \ u_{max} = 0.02, \ \hat{E}_{max} = 13.2307 \ ,$$

cp. the following figure.

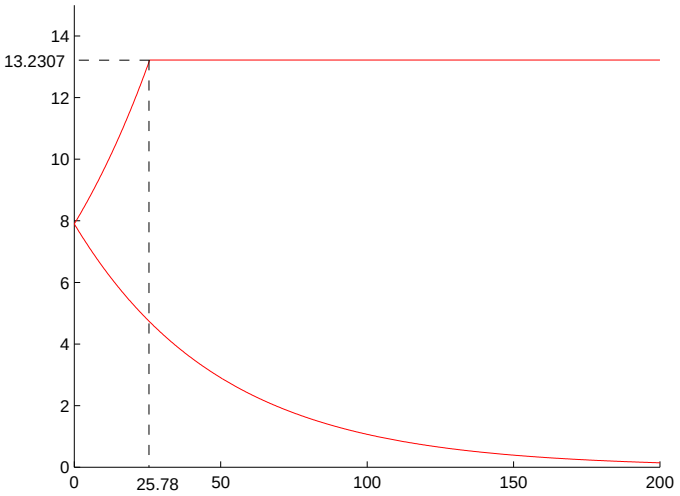


Figure 5.8: Bounds for $E(\cdot)$

Hitherto, only special feasible trajectories could be computed, e. g. trajectories corresponding to specially parameterized families of controls.

Another approach would consist in the maximization of one component of the trajectory, i. e.

$$F(t_f), \text{ resp. } C(t_f), \text{ resp. } T(t_f),$$

which requires numerical methods for optimal control problems with state constraints.

A refinement would consist in successive variation of an (arbitrarily chosen, e. g. affine) objective function and numerical solution of the corresponding family of optimal control problems with state constraints to approximate the reachable set at time t_f at least locally.

Instead, set-valued Euler method yields the following global insight into the time development of the reachable set on the whole time interval.

Figure 5.9 shows the reachable set after 30 years, which is calculated by Euler's method for $N = 30$.

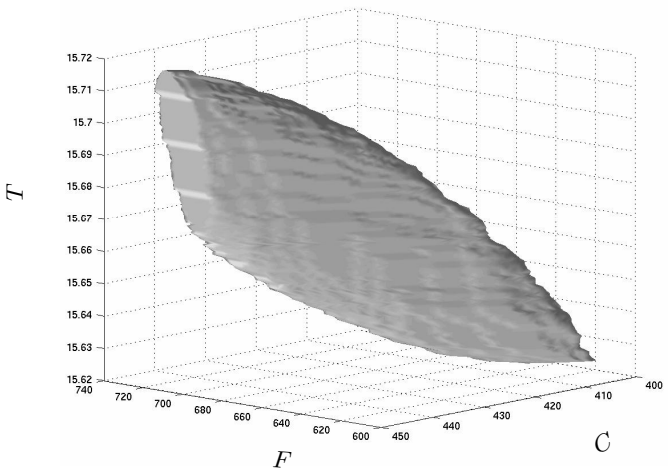


Figure 5.9: Reachable set after 30 years in F-C-T axes

Figures 5.10-5.12 show the 2D-projections corresponding to C-F, C-T and F-T axes. All these pictures correspond to Euler's method for $N = 30$ and end time $t_f = 30$.

Figure 5.13 shows the reachable set after 200 years, which corresponds to Euler's method for $N = 100$.

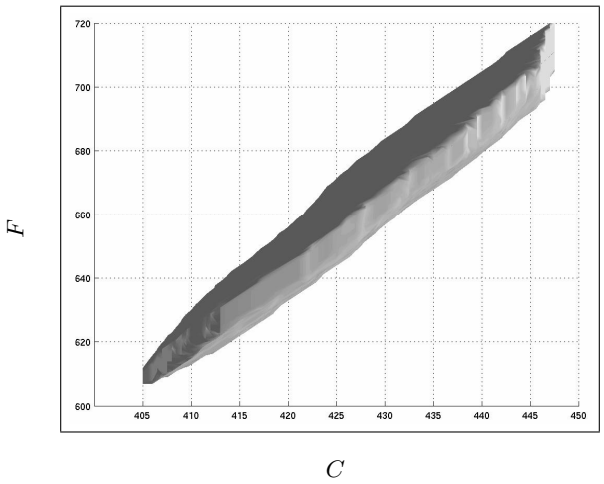


Figure 5.10: 2D-projection of the reachable set in C-F axis

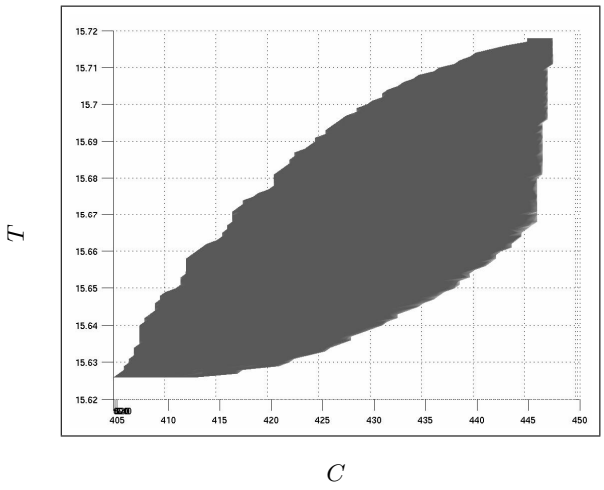


Figure 5.11: 2D-projection of the reachable set in C-T axis

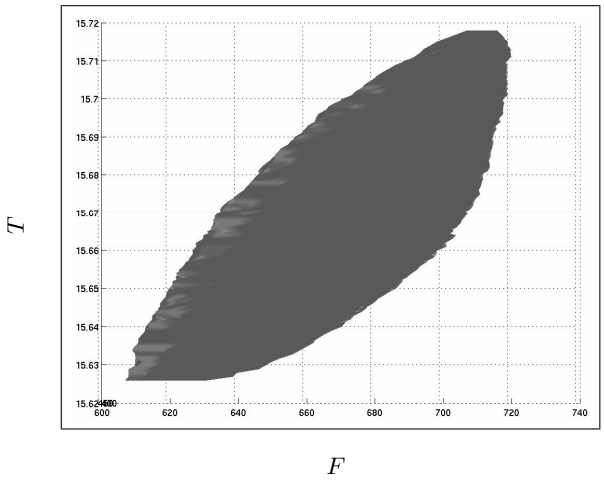


Figure 5.12: 2D-projection of the reachable set in F-T axis

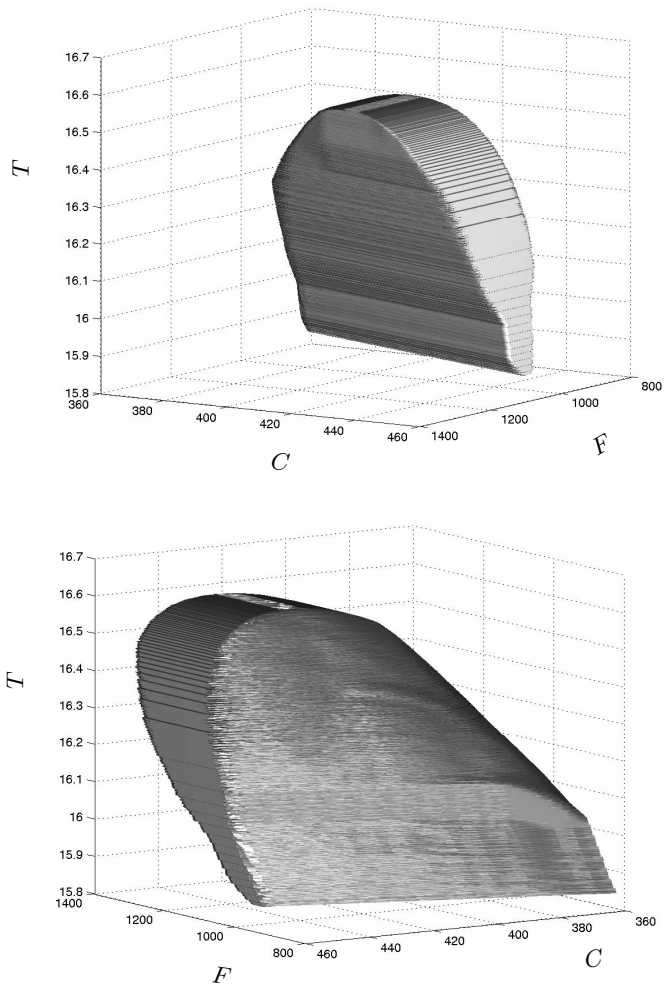


Figure 5.13: Reachable set after 200 years in F-C-T axis

Bibliography

[1] Walter Alt. *Nichtlineare Optimierung: Eine Einführung in Theorie, Verfahren und Anwendungen*. Vieweg, 2002. 51, 72

[2] R. Baier. Mengenwertige Integration und die diskrete Approximation erreichbarer Mengen. *Bayreuther Mathematische Schriften*, 50:1–248, 1995. 108, 109, 118

[3] M. S. Bazaraa and C. M. Shetty. *Nonlinear Programming: Theory and Algorithms*. John Wiley & Sons, 1979. 51

[4] Richard E. Bellman. *Dynamic Programming*. University Press, 1957. 17

[5] Richard E. Bellman and Stuart E. Dreyfus. *Applied Dynamic Programming*. University Press, 1971. 17

[6] John T. Betts and W. P. Huffman. Exploiting Sparsity in the Direct Transcription Method for Optimal Control. *Computational Optimization and Applications*, 14(2):179–201, 1999. 72

[7] Immanuel M. Bomze and W. Grossmann. *Optimierung - Theorie und Algorithmen*. BI-Wissenschaftsverlag, Mannheim, 1993. 17

[8] C. Büskens. *Optimierungmethoden Sensitivitätsanalyse für optimale Steuerprozesse mit Steuer– und Zustands–Beschränkungen*. PhD thesis, University of Münster, Münster, 1998. 118

[9] Christof Büskens and Matthias Gerdts. Numerical Solution of Optimal Control Problems with DAE Systems of Higher Index. In *Optimalsteuerungsprobleme in der Luft- und Raumfahrt, Workshop in Greifswald des Sonderforschungsbereichs 255: Transatmospärische Flugsysteme*, pages 27–38, München, 2000. 91

[10] I. A. Chahma. Set-valued discrete approximation of state-constrained differential inclusions. *Bayreuther Mathematische Schriften*, 67:3–162, 2003. 105, 108, 118

[11] Asen L. Dontchev and Elza M. Farkhi. Error Estimates for Discretized Differential Inclusions. *Computing*, 41:349–358, 1989. 104

[12] A. V. Fiacco. *Introduction to Sensitivity and Stability Analysis in Nonlinear Programming*, volume 165 of *Mathematics in Science and Engineering*. Academic Press, New-York, 1983. 69

[13] Anthony V. Fiacco and Garth P. McCormick. *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, volume 4 of *Classics In Applied Mathematics*. SIAM, Philadelphia, 1990. 51, 69

[14] R. Fletcher. *Practical Methods of Optimization, Volume 1, Unconstrained Optimization*. John Wiley & Sons, Chichester–New York–Brisbane–Toronto, 1980.

[15] R. Fletcher. *Practical Methods of Optimization, Volume 2, Constrained Optimization*. John Wiley & Sons, Chichester–New York–Brisbane–Toronto, 1981.

[16] F. Forcellini and F. Rampazzo. On nonconvex differential inclusions whose state is constrained in the closure of an open set. *DIE*, 12(4):471–497, 1999. 104

[17] H. Frankowska and F. Rampazzo. Filippov's and Filippov-Wazewski's theorems on closed domains. *J. Diff. Equa.*, 161(2):449–478, 2000. 104

[18] Carl Geiger and Christian Kanzow. *Numerische Verfahren zur Lösung unrestringierter Optimierungsaufgaben*. Springer, 1999. 51

[19] Carl Geiger and Christian Kanzow. *Theorie und Numerik restringierter Optimierungsaufgaben*. Springer, 2002. 51, 72, 83

[20] P. E. Gill and W. Murray. Numerically stable methods for quadratic programming. *Mathematical Programming*, 14:349–372, 1978. 77

[21] P. E. Gill, W. Murray, M. A. Saunders, and M. H. Wright. Inertia-controlling methods for general quadratic programming. *SIAM Review*, 33(1):1–36, 1991. 77

[22] Philip E. Gill, Walter Murray, and Michael A. Saunders. *Large-scale SQP Methods and their Application in Trajectory Optimization*, volume 115 of *International Series of Numerical Mathematics*, pages 29–42. Birkhäuser, Basel, 1994. 72

[23] Philip E. Gill, Walter Murray, and Michael A. Saunders. Snopt: An SQP algorithm for large-scale constrained optimization. *SIAM Journal on Optimization*, 12:979–1006, 2002. 72

[24] Philip E. Gill, Walter Murray, Michael A. Saunders, and Margaret H. Wright. User's guide for NPSOL 5.0: A FORTRAN package for nonlinear programming. *Technical Report NA 98-2, Department of Mathematics, University of California, San Diego,California*, 1998. 72

[25] Philip E. Gill, Walter Murray, and Margaret H. Wright. *Practical Optimization*. Academic Press, London, 1981. 51, 72

[26] I. V. Girsanov. *Lectures on Mathematical Theory of Extremum Problems*. Springer-Verlag, Berlin–Heidelberg–New York, 1972. 31

[27] D. Goldfarb and A. Idnani. A numerically stable dual method for solving strictly convex quadratic programs. *Mathematical Programming*, 27:1–33, 1983. 77

[28] G. Grosche, V. Ziegler, D. Ziegler, and E. Zeidler. *Teubner-Taschenbuch der Mathematik, Teil II, 7. Auflage*. B. G. Teubner-Verlag, Stuttgart–Leipzig, 1995.

[29] S. P. Han. A Globally Convergent Method for Nonlinear Programming. *Journal of Optimization Theory and Applications*, 22(3):297–309, 1977. 72

[30] M. R. Hestenes. *Calculus of Variations and Optimal Control Theory*. Applied Mathematics Series. John Wiley and Sons, Inc., New York–London–Sydney, 1966. 31

[31] M. R. Hestenes. *Optimization Theory—The Finite Dimensional Case*. John Wiley and Sons, New York-London-Sydney-Toronto, 1975.

[32] A. D. Ioffe and V. M. Tichomirov. *Theorie der Extremalaufgaben*. VEB Deutscher Verlag der Wissenschaften, Berlin, 1979.

[33] A. D. Ioffe and V. M. Tihomirov. *Theory of Extremal Problems*, volume 6 of *Studies in Mathematics and Applications*. North-Holland Publishing Company, Amsterdam–New York–Oxford, 1979. 31, 46, 49

[34] D. Klatte and B. Kummer. *Nonsmooth Equations in Optimization*. Kluwer Academic Publishers, Dordrecht, 2002.

[35] Dieter Kraft. A Software Package for Sequential Quadratic Programming. *DFVLR-FB-88-28, Oberpfaffenhofen*, 1988. 72

[36] P.-J. Laurent. *Approximation et Optimisation*. Hermann, Paris, 1972.

[37] E. S. Levitin. *Perturbation Theory in Mathematical Programming and Its Applications*. John Wiley and Sons, Chichester, 1994.

[38] L. A. Ljusternik and W. L. Sobolew. *Elemente der Funktionalanalysis*. Verlag Harri Deutsch, Zürich–Frankfurt–Thun, 1979.

[39] Olvi L. Mangasarian. *Nonlinear Programming*, volume 10 of *Classics In Applied Mathematics*. SIAM, Philadelphia, 1994. 51

[40] M. Mayrhofer and G. Sachs. Notflugbahnen eines zweistufigen Hyperschall-Flugsystems ausgehend vom Trennmanöver. *Seminarbericht des Sonderforschungsbereichs 255: Transatmosphärische Flugsysteme, TU München*, pages 109–118, 1996. 91

[41] I. P. Natanson. *Theorie der Funktionen einer reellen Veränderlichen*. Verlag Harri Deutsch, Zürich–Frankfurt–Thun, 1977.

[42] G. L. Nemhauser. *Einführung in die Praxis der dynamischen Programmierung*. R. Oldenbourg Verlag, München–Wien, 1969.

[43] G. L. Nemhauser and L. A. Wolsey. *Integer and Combinatorial Optimization*. J. Wiley & Sons, New York–Chichester, 1988.

[44] K. Neumann. *Operations Research Verfahren, Band I*. Theorie und Praxis des Operations Research. Carl Hanser Verlag, München–Wien, 1975.

[45] K. Neumann. *Operations Research Verfahren, Band III*. Theorie und Praxis des Operations Research. Carl Hanser Verlag, München–Wien, 1975.

[46] K. Neumann. *Operations Research Verfahren, Band II*. Theorie und Praxis des Operations Research. Carl Hanser Verlag, München–Wien, 1977.

[47] Klaus Neumann and Martin Morlock. *Operations Research*. Carl Hanser Verlag, 2002. 17

[48] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer-Verlag, New York, 1999.

[49] G. Petschel-Held, H. Schellnhuber, T. Bruckner, F. Toth, and K. Hasselmann. The tolerable windows approach: Theoretical and methodological foundations. *Climatic Change*, 41:303–331, 1999. 115

[50] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mishchenko. *The Mathematical Theory of Optimal Processes*. Interscience Publ., New York, 1962. 31

[51] M. J. D. Powell. A fast algorithm for nonlinearily constrained optimization calculation. In G.A. Watson, editor, *Numerical Analysis*, volume 630 of *Lecture Notes in Mathematics*, Berlin-Heidelberg-New-York, 1978. Springer. 72, 80, 86, 87

[52] R. T. Rockafellar. *Convex Analysis*, volume 28 of *Princeton Math. Series*. Princeton Univ. Press, Princeton, 1970.

[53] Klaus Schittkowski. The Nonlinear Programming Method of Wilson, Han, and Powell with an Augmented Lagrangian Type Line Search Function. Part 1: Convergence Analysis, Part 2: An Efficient Implementation with Linear Least Squares Subproblems. *Numerische Mathematik*, 383:83–114, 115–127, 1981. 72, 86

[54] Klaus Schittkowski. On the Convergence of a Sequential Quadratic Programming Method with an Augmented Lagrangian Line Search Function. *Mathematische Operationsforschung und Statistik, Series Optimization*, 14(2):197–216, 1983. 72, 86

[55] Klaus Schittkowski. NLPQL: A Fortran subroutine for solving constrained nonlinear programming problems. *Annals of Operations Research*, 5:484–500, 1985. 72

[56] Volker H. Schulz. *Reduced SQP Methods for Large-Scale Optimal Control Problems in DAE with Application to Path Planning Problems for*

Satellite Mounted Robots. PhD thesis, Interdisziplinäres Zentrum für Wissenschaftliches Rechnen, Universität Heidelberg, 1996. 72

[57] H. M. Soner. Optimal control with state constraints. *SIAM J. Cont. Optim.*, 24(3):552–561, 1986. 104

[58] Peter Spellucci. *Numerische Verfahren der nichtlinearen Optimierung*. Birkhäuser, Basel, 1993. 51, 69, 77

[59] Marc C. Steinbach. *Fast Recursive SQP Methods for Large-Scale Optimal Control Problems*. PhD thesis, Interdisziplinäres Zentrum für Wissenschaftliches Rechnen, Universität Heidelberg, 1995. 72

[60] J. Stoer. Principles of sequential quadratic programming methods for solving nonlinear programs. In K Schittkowski, editor, *Computational Mathematical Programming*, volume F15 of *NATO ASI Series*, pages 165–207, Berlin-Heidelberg-New-York, 1985. Springer. 72

[61] H. A. Taha. *Operations Research — An Introduction*. Prentice Hall, Upper Saddle River, New Jersey, 6th edition, 1997.

[62] Wayne L. Winston. *Operations Research: Applications and Algorithms*. Brooks/Cole–Thomson Learning, 4 edition, 2004. 17

[63] E. Zeidler. *Teubner-Taschenbuch der Mathematik*. B. G. Teubner-Verlag, Stuttgart–Leipzig, 1996.