Optimal Control of Ordinary Differential Equations and Differential-Algebraic Equations

Matthias Gerdts
Universität der Bundeswehr München
135 PUBLICATIONS **1,399** CITATIONS

MATTHIAS GERDTS

Optimal Control of Ordinary Differential Equations and Differential-Algebraic Equations

Habilitation Thesis
Department of Mathematics
University of Bayreuth
Accepted: December 2006
Revised Version: 23rd January 2007

Address of the Author:

Matthias Gerdts

Schwerpunkt Optimierung und Approximation
Department Mathematik
Universität Hamburg

D-20146 Hamburg

E-Mail: gerdts@math.uni-hamburg.de

WWW: www.math.uni-hamburg.de/home/gerdts/

Preface

This work was mainly developped at the Chair of Applied Mathematics at the Department of Mathematics of the University of Bayreuth. I am indebted to Prof. Dr. Frank Lempio for his ongoing support and many fruitful and valuable discussions. Particularly, I like to thank him for the opportunity of giving a course on nondifferentiable optimization and the close collaboration in the course on discrete dynamic optimization in Borovets/Bulgaria. From both I benefited a lot. In addition, he provided a very good atmosphere and working environment at his chair.

Furthermore, I would like to thank Prof. Dr. Hans Josef Pesch for initiating my interest in optimal control problems, the arrangement of my stay at the University of California, San Diego, his encouragement and the motivation to finish this work.

I thank Prof. Dr. Christof Büskens for the very close and friendly cooperation and for the invitation to the University of Bremen for an interesting research stay.

Last but not least I would like to express my thanks to all members of the Chair of Applied Mathematics and the Chair of Mathematics in the Engineering Sciences at the University of Bayreuth for the very pleasant and enjoyable time.

Finally, I thank my parents for their steady support.

I dedicate this work to Katja and thank her for being there for me!

Contents

Notation

$\mathbb{R}^n$	n-dimensional Euclidian space with norm $\|\cdot\|_2$
0_n	zero in $\mathbb{R}^n$
$[t_0, t_f]$	compact time interval in $\mathbb{R}$ with fixed $t_0 < t_f$
$\mathbb{G}_N$	grid with N subintervals
X, Y, Z	Banach spaces
$\|\cdot\|_X$	norm on a Banach space X
Θ_X	zero of a Banach space X
Θ	generic zero element of some space
$\langle\cdot,\cdot\rangle_X$	scalar product on a pre-Hilbert space X
$U_\varepsilon(x_0) := \{x \in X \mid \|x - x_0\|_X < \varepsilon\}$	open ball around x_0 with radius $\varepsilon > 0$
$\mathrm{im}(T) := \{T(x) \mid x \in X\}$	image of X under the map $T : X \to Y$
$\ker(T) := \{x \in X \mid T(x) = \Theta_X\}$	kernel or null space of a linear map $T : X \to Y$
$T^{-1}(S) := \{x \in X \mid T(x) \in S\}$	preimage of the set $S \subseteq Y$ of a map $T : X \to Y$
X^*	topological dual space of X
$\|f\|_{X^*} = \sup_{\|x\|_X \leq 1} \lvert f(x)\rvert$	norm on X^* defining the strong topology
T^*	adjoint operator of the linear map $T : X \to Y$
$\mathcal{L}(X, Y)$	set of all linear, continuous mappings $f : X \to Y$
X/L	factor space or quotient space
$L^p([t_0, t_f], \mathbb{R}^n)$	space of all mappings $f : [t_0, t_f] \to \mathbb{R}^n$ that are bounded in the norm $\|\cdot\|_p$
$W^{q,p}([t_0, t_f], \mathbb{R}^n)$	Sobolev space of all absolutely continuous functions $f : [t_0, t_f] \to \mathbb{R}^n$ that are bounded in the norm $\|\cdot\|_{q,p}$
$C([t_0, t_f], \mathbb{R}^n)$	space of continuous functions $f : [t_0, t_f] \to \mathbb{R}^n$
$AC([t_0, t_f], \mathbb{R}^n)$	space of absolutely continuous functions $f : [t_0, t_f] \to \mathbb{R}^n$
$BV([t_0, t_f], \mathbb{R}^n)$	functions of bounded variation
$NBV([t_0, t_f], \mathbb{R}^n)$	normalized functions of bounded variation
$\int_{t_0}^{t_f} f(t) d\mu(t)$	Riemann-Stieltjes integral
$\mathrm{conv}(S)$	convex hull of the set S
$\mathrm{relint}(S)$	relative interior of the set S
$\mathrm{cone}(S, x)$	conical hull of the set $S - \{x\}$
$\mathrm{cl}(S)$	closure of S
$\mathrm{int}(S)$	topological interior of S
S^+, S^-	positive and negative dual cone of S
$F'(x)(d)$	Fréchet derivative of F at x in direction d
$\nabla F(x) = F'(x)^\top$	gradient of $F : \mathbb{R}^n \to \mathbb{R}$
$F'_x(x, y)$	partial Fréchet derivative of F at (x, y)
$F'(x; d)$	directional derivative of F at x in direction d
$\delta F(x)(d)$	Gateaux derivative of F at x in direction d
$\partial F(x)$	Clarke's subdifferential of F at x
$\partial_B F(x)$	Bouligand-differential of F at x

$\operatorname{lev}(f,z) := \{x \in X \mid f(x) \leq z\}$	level set of the function f at level z
$T(\Sigma, x)$	(sequential) tangent cone of Σ at $x \in \Sigma$
$T_{lin}(K, S, x)$	linearizing cone of K and S at x
$T_C(x)$	critical cone at x
$\mathcal{H}$	Hamilton function of an optimal control problem
L	Lagrange function

Chapter 1

Introduction

Historically, optimal control problems evolved from variational problems. Variational problems have been investigated more thoroughly since 1696, although the first variational problems, e.g. queen Dido's problem, were formulated in the ancient world already. In 1696 Johann Bernoulli (1667–1748) posed the Brachistochrone problem to other famous contemporary mathematicians like Sir Isaac Newton (1643–1727), Gottfried Wilhelm Leibniz (1646–1716), Jacob Bernoulli (1654–1705), Guillaume François Antoine Marquis de L'Hôspital (1661–1704), and Ehrenfried Walter von Tschirnhaus (1651–1708). Each of these distinguished mathematicians were able to solve the problem. An interesting description of the Brachistochrone problem with many additional historical remarks as well as the solution approach of Johann Bernoulli exploiting Fermat's principle can be found in Pesch [Pes02].

Optimal control problems generalize variational problems by separating control and state variables and admitting control constraints. Strongly motivated by military applications, optimal control theory as well as solution methods evolve rapidly since 1950. The decisive breakthrough was achieved by the Russian mathematician Lev S. Pontryagin (1908–1988) and his coworkers V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mishchenko in proving the maximum principle, which provides necessary optimality conditions for optimal control problems, cf [PBGM64]. Almost at the same time also Magnus R. Hestenes [Hes66] proved a similar theorem. Since then many contributions in view of necessary conditions, cf., e.g., Jacobsen et al. [JLS71], Girsanov [Gir72], Knobloch [Kno75], Neustadt [Neu76], Kirsch et al. [KWW78], Ioffe and Tihomirov [IT79], Maurer [Mau77, Mau79], Hartl et al. [HSV95], sufficient conditions, cf., e.g., Maurer [Mau81], Zeidan [Zei94], Malanowski [Mal97], Maurer and Pickenhain [MP95b], Maurer and Oberle [MO02], Malanowski et al. [MMP04], sensitivity analysis and real-time optimal control, cf., e.g., Maurer and Augustin [MA01], Augustin and Maurer [AM01], Büskens and Maurer [BM96, BM01a], Malanowski and Maurer [MM96, MM98, MM01], Maurer and Pesch [MP94a, MP94b, MP95a], Pesch [Pes78, Pes79, Pes89a, Pes89b], and numerical treatment, cf., e.g., Bulirsch [Bul71], Deuflhard [Deu74, Deu79], Deuflhard et al. [DPR76], Bryson and Ho [BH75], Diekhoff et al. [DLO$^+$77], Chernousko and Lyubushin [CL82], Bock and Plitt [BP84], Kraft [KE85], Oberle [Obe86, OG01], Hargraves and Paris [HP87], Teo and Goh [TG87], Pesch and Bulirsch [PB94], Betts [Bet90], Hiltmann [HCB93], von Stryk [vS94], Malanowski et al. [MBM97], Büskens [Büs98], Dontchev et al. [DHM00], Dontchev et al. [DHV00] have been released.

Applications of optimal control problems can be found in nearly any discipline, for example in natural sciences, engineering sciences, and economy.

While at the beginning the main focus was on optimal control problems subject to ordinary differential equations (ODEs), in the meanwhile also optimal control problems subject to partial differential equations (PDEs), differential-algebraic equations (DAEs), and stochastic optimal control problems are under investigation.

In this book we will particularly focus on optimal control problems subject to differential-algebraic equations. DAEs are composite systems of differential equations and algebraic equations and often are viewed as differential equations on manifolds. The implicit ODE

$$F(t, z(t), \dot{z}(t), u(t)) = 0_{n_z}, \qquad t \in [t_0, t_f] \tag{1.1}$$

provides an example of DAEs. Herein, $[t_0, t_f] \subset \mathbb{R}$, $t_0 < t_f$ is a given compact time interval and $F : [t_0, t_f] \times \mathbb{R}^{n_z} \times \mathbb{R}^{n_z} \times \mathbb{R}^{n_u} \to \mathbb{R}^{n_z}$ is a sufficiently smooth mapping. If the partial derivative $F'_{\dot{z}}(\cdot)$ in (1.1) happens to be non-singular, then (1.1) is just an ODE in implicit form and the implicit function theorem allows to solve (1.1) for $\dot{z}$ in order to obtain an explicit ODE of type $\dot{z}(t) = f(t, z(t), u(t))$. Hence, explicit or implicit ODEs are special cases of DAEs. The more interesting case occurs if the partial derivative $F'_{\dot{z}}(\cdot)$ is *singular*. In this case (1.1) cannot be solved directly for $\dot{z}$ and (1.1) includes differential equations and algebraic equations at the same time. Equations of such type are discussed intensively since the early 1970ies. Though at a first glance DAEs seem to be very similar to ODEs they possess different solution properties, cf. Petzold [Pet82b]. Particularly, DAEs possess different stability properties compared to ODEs (see the perturbation index below) and initial values $z(t_0) = z_0$ have to be defined properly to guarantee at least locally unique solutions (see consistent initial values below).

It is important to mention, that, presently, (1.1) is too general and therefore too challenging to being tackled theoretically or numerically. Whenever necessary we will restrict the problem to more simple problems. Most often, we will discuss so-called *semi-explicit DAEs*

$$\begin{aligned} \dot{x}(t) &= f(t, x(t), y(t), u(t)), \qquad (1.2)\\ 0_{n_y} &= g(t, x(t), y(t), u(t)), \qquad (1.3) \end{aligned}$$

where the state z in (1.1) is decomposed into components $x : [t_0, t_f] \to \mathbb{R}^{n_x}$ and $y : [t_0, t_f] \to \mathbb{R}^{n_y}$. Herein, $x(\cdot)$ is referred to as *differential variable* and $y(\cdot)$ as *algebraic variable*. Correspondingly, (1.2) is called *differential equation* and (1.3) *algebraic equation*. Actually, the restriction to semi-explicit DAEs is (at least theoretically) not essential, since the implicit DAE (1.1) is equivalent with the semi-explicit DAE

$$\begin{aligned} \dot{z}(t) &= y(t), \\ 0_{n_z} &= F(t, z(t), y(t), u(t)). \end{aligned}$$

The functions $x(\cdot)$ and $y(\cdot)$ define the *state* of the upcoming optimal control problem. The function $u : [t_0, t_f] \to \mathbb{R}^{n_u}$ is considered as an external input and it is referred to as *control variable*. Usually, $u(t)$ is restricted to a set $\mathcal{U} \subseteq \mathbb{R}^{n_u}$, i.e.

$$u(t) \in \mathcal{U}. \tag{1.4}$$

Furthermore, the state and control variables may be restricted by *mixed control-state constraints*

$$c(t, x(t), y(t), u(t)) \leq 0_{n_c}, \tag{1.5}$$

pure state constraints

$$s(t, x(t)) \leq 0_{n_s}, \tag{1.6}$$

and *boundary conditions*

$$\psi(x(t_0), x(t_f)) = 0_{n_\psi}. \tag{1.7}$$

Together with the *objective function*

$$\varphi(x(t_0), x(t_f)) + \int_{t_0}^{t_f} f_0(t, x(t), y(t), u(t))dt \tag{1.8}$$

we will investigate

Problem 1.1 (DAE optimal control problem)
Find functions $x(\cdot)$, $y(\cdot)$, and $u(\cdot)$ such that (1.8) is minimized subject to the constraints (1.2)–(1.7).

Herein, $c : [t_0, t_f] \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_u} \to \mathbb{R}^{n_c}$, $s : [t_0, t_f] \times \mathbb{R}^{n_x} \to \mathbb{R}^{n_s}$, $\psi : \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \to \mathbb{R}^{n_\psi}$, $\varphi : \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \to \mathbb{R}$, and $f_0 : [t_0, t_f] \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_u} \to \mathbb{R}$ are sufficiently smooth functions. Interestingly, the evaluation of the local minimum principle for optimal control problems subject to explicit ODEs in absence of state and control constraints, i.e. $n_c = n_s = n_y = 0$, $\mathcal{U} = \mathbb{R}^{n_u}$, leads to a semi-explicit DAE itself, cf. Ioffe and Tihomirov [IT79]:

$$\begin{aligned} \dot{x}(t) &= f(t, x(t), u(t)), \\ \dot{\lambda}(t) &= -\mathcal{H}'_x(t, x(t), u(t), \lambda(t), l_0)^\top, \\ 0_{n_u} &= \mathcal{H}'_u(t, x(t), u(t), \lambda(t), l_0)^\top. \end{aligned}$$

Herein, $\mathcal{H}(t, x, u, \lambda, l_0) := l_0 f_0(t, x, u) + \lambda^\top f(t, x, u)$ denotes the Hamilton function. The state x and the adjoint (co-state) λ are differential variables while u is an algebraic variable.
For (1.2)-(1.3) it is not clear why u is viewed as a control variable while y should be a component of the state. In fact, this assignment is somehow arbitrary from the optimal control theoretic point of view, since both, u and y, can be viewed as control variables. This becomes clearer in the chapter on local minimum principles for DAE optimal control problems, where it can be seen, that both functions have very similar properties. Actually, this is the reason why the functions φ, ψ, and s only depend on x and not on y and u. But, recall that the DAE in reality is a model for, e.g., a robot, a car, an electric circuit, or a power plant. Hence, the DAE has a meaning for itself independent of whether it occurs in an optimal control problem or not. In this context, it is necessary to distinguish between control u and algebraic variable y. The essential difference is, that an operator can choose the control u, whereas the algebraic variable y cannot be controlled directly since it results from the state component x and the input u. For instance, in the context of mechanical multi-body systems the algebraic variable y corresponds physically to a constraint force.
The fact that y is defined by x and u, is mathematically taken into account for by imposing additional regularity assumptions. The degree of regularity of the DAE (1.2)-(1.3) is measured by a quantity called *index*. For specially structured DAEs, e.g. Hessenberg DAEs, the index definitions coincide, whereas for general systems there are several different index definitions, cf. Duff and Gear [DG86a], Gear [Gea88, Gea90], Campbell and Gear [CG95], Hairer et al. [HLR89], and März [Mär95, Mär98b, Mär98a].

The perturbation index indicates the influence of perturbations and their derivatives on the solution and therefore addresses the *stability of DAEs*.

Definition 1.2 (Perturbation Index, Hairer et al. [HLR89])
The DAE (1.1) has perturbation index $p \in \mathbb{N}$ *along a solution z on $[t_0, t_f]$, if $p \in \mathbb{N}$ is the smallest number such that for all functions $\tilde{z}$ satisfying the perturbed DAE*

$$F(t, \tilde{z}(t), \dot{\tilde{z}}(t), u(t)) = \delta(t), \tag{1.9}$$

there exists a constant S depending on F and $t_f - t_0$ with

$$\|z(t) - \tilde{z}(t)\| \leq S \left(\|z(t_0) - \tilde{z}(t_0)\| + \max_{t_0 \leq \tau \leq t} \|\delta(\tau)\| + \ldots + \max_{0 \leq \tau \leq t} \|\delta^{(p-1)}(\tau)\| \right) \tag{1.10}$$

for all $t \in [t_0, t_f]$ whenever the expression on the right is less than or equal to a given bound.

The perturbation index *is $p = 0$, if the estimate*

$$\|z(t) - \tilde{z}(t)\| \leq S \left(\|z(t_0) - \tilde{z}(t_0)\| + \max_{t_0 \leq \tau \leq t_f} \left\| \int_{t_0}^{\tau} \delta(s) ds \right\| \right) \tag{1.11}$$

holds.

In the sequel we will make use of the subsequent lemma.

Lemma 1.3 (Gronwall) *Let $w, z : [t_0, t_f] \to \mathbb{R}$ be integrable functions with*

$$w(t) \leq L \int_{t_0}^{t} w(\tau) d\tau + z(t)$$

for a.e. $t \in [t_0, t_f]$ with some constant $L \geq 0$. Then it holds

$$w(t) \leq z(t) + L \int_{t_0}^{t} \exp\left(L(t - \tau)\right) z(\tau) d\tau$$

for a.e. $t \in [t_0, t_f]$. If z in addition belongs to $L^\infty([t_0, t_f], \mathbb{R})$ then it holds

$$w(t) \leq \|z(\cdot)\|_\infty \exp\left(L(t - t_0)\right)$$

for a.e. $t \in [t_0, t_f]$.

Proof. According to the assumption we may write

$$w(t) = a(t) + z(t) + \delta(t)$$

with the absolutely continuous function

$$a(t) := L \int_{t_0}^{t} w(\tau) d\tau$$

and a non-negative function $\delta(\cdot) \in L^1([t_0, t_f], \mathbb{R})$. Introducing the expression for w in a yields

$$a(t) = L \int_{t_0}^{t} a(\tau) d\tau + L \int_{t_0}^{t} \left(z(\tau) + \delta(\tau)\right) d\tau.$$

Hence, a solves the inhomogeneous linear differential equation

$$a'(t) = La(t) + L\left(z(t) + \delta(t)\right)$$

for almost every $t \in [t_0, t_f]$ with initial value $a(t_0) = 0$. The well-known solution formula for linear differential equations yields

$$a(t) = L \int_{t_0}^{t} \exp\left(L(t - \tau)\right) \left(z(\tau) + \delta(\tau)\right) d\tau$$

respectively

$$w(t) = L \int_{t_0}^{t} \exp\left(L(t - \tau)\right) \left(z(\tau) + \delta(\tau)\right) d\tau + z(t) + \delta(t).$$

Since $\delta(t) \geq 0$ the first assertion holds. If z is even essentially bounded we find

$$w(t) \leq \|z(\cdot)\|_\infty \left(1 + L \int_{t_0}^{t} \exp\left(L(t - \tau)\right) d\tau\right) = \|z(\cdot)\|_\infty \exp\left(L(t - t_0)\right).$$

■

Remark 1.4 *Obviously, the last assertion of Gronwall's lemma remains true if we apply the norm $\|\cdot\|_\infty$ only to the interval $[t_0,t]$ instead of the whole interval $[t_0,t_f]$.*

The ODE case:

To illustrate the perturbation index, we start with the initial value problem (IVP)

$$\dot{x}(t) = f(t,x(t),u(t)), \qquad x(t_0) = x_0,$$

where f is assumed to be Lipschitz continuous w.r.t. x uniformly with respect to t.
The (absolutely continuous) solution x can be written in integral form:

$$x(t) = x_0 + \int_{t_0}^{t} f(\tau,x(\tau),u(\tau))d\tau.$$

Now, consider the perturbed IVP

$$\dot{\tilde{x}}(t) = f(t,\tilde{x}(t),u(t)) + \delta(t), \qquad \tilde{x}(t_0) = \tilde{x}_0$$

with an integrable perturbation $\delta : [t_0,t_f] \to \mathbb{R}^{n_x}$ and its solution

$$\tilde{x}(t) = \tilde{x}_0 + \int_{t_0}^{t} \left(f(\tau,\tilde{x}(\tau),u(\tau)) + \delta(\tau)\right) d\tau.$$

It holds

$$\begin{aligned} \|x(t)-\tilde{x}(t)\| &\leq \|x_0-\tilde{x}_0\| + \int_{t_0}^{t} \|f(\tau,x(\tau),u(\tau)) - f(\tau,\tilde{x}(\tau),u(\tau))\|d\tau + \left\|\int_{t_0}^{t}\delta(\tau)d\tau\right\| \\ &\leq \|x_0-\tilde{x}_0\| + L\int_{t_0}^{t}\|x(\tau)-\tilde{x}(\tau)\|d\tau + \left\|\int_{t_0}^{t}\delta(\tau)d\tau\right\|. \end{aligned}$$

Application of Gronwall's lemma yields

$$\begin{aligned} \|x(t)-\tilde{x}(t)\| &\leq \left(\|x_0-\tilde{x}_0\| + \max_{t_0\leq\tau\leq t}\left\|\int_{t_0}^{\tau}\delta(s)ds\right\|\right)\exp\left(L(t-t_0)\right) \\ &\leq \left(\|x_0-\tilde{x}_0\| + \max_{t_0\leq\tau\leq t_f}\left\|\int_{t_0}^{\tau}\delta(s)ds\right\|\right)\exp\left(L(t_f-t_0)\right). \end{aligned}$$

Hence, the ODE has perturbation index $p=0$.

The index-1 case:

Consider the DAE (1.2)-(1.3) with solution (x,y) and initial value $x(t_0)=x_0$ and the perturbed IVP

$$\begin{aligned} \dot{\tilde{x}}(t) &= f(t,\tilde{x}(t),\tilde{y}(t),u(t)) + \delta_f(t), \qquad \tilde{x}(t_0)=\tilde{x}_0, && (1.12) \\ 0_{n_y} &= g(t,\tilde{x}(t),\tilde{y}(t),u(t)) + \delta_g(t) && (1.13) \end{aligned}$$

in $[t_0,t_f]$. Assume that

(i) f is Lipschitz continuous w.r.t. x and y with Lipschitz constant L_f uniformly with respect to t;

(ii) g is continuously differentiable and $g'_y(t,x,y,u)$ is non-singular and bounded for all $(t,x,y,u) \in [t_0,t_f] \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_y} \times \mathcal{U}$.

According to (ii) the equation

$$0_{n_y} = g(t,x,y,u) + \delta_g$$

can be solved for $y \in \mathbb{R}^{n_y}$ for any $t \in [t_0,t_f]$, $x \in \mathbb{R}^{n_x}$, $u \in \mathcal{U}$ by the implicit function theorem:

$$\tilde{y} = Y(t,x,u,\delta_g), \quad g(t,x,Y(t,x,u,\delta_g),u) + \delta_g = 0_{n_y}.$$

Furthermore, Y is locally Lipschitz continuous w.r.t. t,x,u,δ_g with Lipschitz constant L_Y. Let $(x(\cdot),y(\cdot))$ denote the unperturbed solution of (1.2)-(1.3) resp. of (1.12)-(1.13) for $\delta_f \equiv 0$, $\delta_g \equiv 0$ and $(\tilde{x}(\cdot),\tilde{y}(\cdot))$ the perturbed solution of (1.12)-(1.13). For the algebraic variables we get the estimate

$$\begin{aligned}\|y(t)-\tilde{y}(t)\| &= \|Y(t,x(t),u(t),0_{n_y}) - Y(t,\tilde{x}(t),u(t),\delta_g(t))\| \\ &\leq L_Y\left(\|x(t)-\tilde{x}(t)\| + \|\delta_g(t)\|\right).\end{aligned}$$

With (i) we obtain

$$\begin{aligned}\|x(t)-\tilde{x}(t)\| &\leq \|x_0-\tilde{x}_0\| + \int_{t_0}^{t} \|f(\tau,x(\tau),y(\tau),u(\tau)) - f(\tau,\tilde{x}(\tau),\tilde{y}(\tau),u(\tau))\| d\tau \\ &\quad + \left\|\int_{t_0}^{t} \delta_f(\tau)d\tau\right\| \\ &\leq \|x_0-\tilde{x}_0\| + L_f \int_{t_0}^{t} \|x(\tau)-\tilde{x}(\tau)\| + \|y(\tau)-\tilde{y}(\tau)\| d\tau + \left\|\int_{t_0}^{t} \delta_f(\tau)d\tau\right\| \\ &\leq \|x_0-\tilde{x}_0\| + L_f(1+L_Y)\int_{t_0}^{t} \|x(\tau)-\tilde{x}(\tau)\| d\tau + L_f\int_{t_0}^{t} \|\delta_g(\tau)\| d\tau \\ &\quad + \left\|\int_{t_0}^{t} \delta_f(\tau)d\tau\right\|.\end{aligned}$$

Using the same arguments as in the ODE case we end up in the estimate

$$\begin{aligned}\|x(t)-\tilde{x}(t)\| \leq \left(\|x_0-\tilde{x}_0\| + (t_f-t_0)\left(L_f \max_{t_0\leq\tau\leq t}\|\delta_g(\tau)\| + \max_{t_0\leq\tau\leq t}\|\delta_f(\tau)\|\right)\right)\cdot \\ \cdot \exp\left(L_f(1+L_Y)(t-t_0)\right).\end{aligned}$$

Hence, if the assumptions (i) and (ii) are valid, then the DAE (1.2)-(1.3) has perturbation index $p = 1$.

<u>**The index-k case:**</u>

The above procedure will work for a class of DAEs called Hessenberg DAEs.

Definition 1.5 (Hessenberg DAE)
Let $k \geq 2$. The semi-explicit DAE

$$\begin{array}{rcllllllll}\dot{x}_1(t) &=& f_1(t, & x_0(t), & x_1(t), & x_2(t), & \ldots, & x_{k-2}(t), & x_{k-1}(t), & u(t)), \\ \dot{x}_2(t) &=& f_2(t, & & x_1(t), & x_2(t), & \ldots, & x_{k-2}(t), & x_{k-1}(t), & u(t)), \\ \vdots & & & & & \ddots & & & & \\ \dot{x}_{k-1}(t) &=& f_{k-1}(t, & & & & & x_{k-2}(t), & x_{k-1}(t), & u(t)), \\ 0_{n_y} &=& g(t, & & & & & & x_{k-1}(t), & u(t))\end{array} \qquad (1.14)$$

is called Hessenberg DAE of order k. *Herein, the differential variable is* $x = (x_1, \ldots, x_{k-1})^\top$ *and the algebraic variable is* $y = x_0$.

For $i = 1, \ldots, k-1$ let f_i be i times continuously differentiable and let u be sufficiently smooth. By $(k-1)$-fold differentiation of the algebraic constraint, application of the implicit function theorem and repeating the arguments for ODEs and index-1 DAEs it is easy to see that the Hessenberg DAE of order k has perturbation index k, if the matrix

$$M := g'_{x_{k-1}}(\cdot) \cdot f'_{k-1,x_{k-2}}(\cdot) \cdots f'_{2,x_1}(\cdot) \cdot f'_{1,x_0}(\cdot) \tag{1.15}$$

is non-singular.
Another index concept corresponds to the definition of the order of active state constraints in optimal control problems, cf. Maurer [Mau79]. Notice, that the algebraic constraint (1.3) can be interpreted as a state resp. mixed control-state constraint in an appropriate optimal control problem, which is active on $[t_0, t_f]$. The idea is to differentiate the algebraic constraint (1.3) w.r.t. time and to replace the derivative $\dot{x}$ according to (1.2) until the resulting equations can be solved for the derivative $\dot{y}$ by the implicit function theorem. The smallest number of differentiations needed is called *differential index*, c.f. Brenan et al. [BCP96].
We illustrate the differential index for (1.2)-(1.3) and assume

(i) g and u are continuously differentiable.

(ii) $g'_y(t,x,y,u)$ is non-singular and $g'_y(t,x,y,u)^{-1}$ is bounded for all $(t,x,y,u) \in [t_0,t_f] \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_y} \times \mathcal{U}$.

Differentiation of the algebraic constraint (1.3) w.r.t. time yields

$$\begin{aligned} 0_{n_y} &= g'_t[t] + g'_x[t] \cdot \dot{x}(t) + g'_y[t] \cdot \dot{y}(t) + g'_u[t] \cdot \dot{u}(t) \\ &= g'_t[t] + g'_x[t] \cdot f[t] + g'_y[t] \cdot \dot{y}(t) + g'_u[t] \cdot \dot{u}(t), \end{aligned}$$

where, e.g., $g'_t[t]$ is an abbreviation for $g'_t(t, x(t), y(t), u(t))$. Since $g'_y[t]^{-1}$ exists and is bounded we can solve this equality for $\dot{y}$ and obtain the differential equation

$$\dot{y}(t) = -\left(g'_y[t]\right)^{-1} \left(g'_t[t] + g'_x[t] \cdot f[t] + g'_u[t] \cdot \dot{u}(t)\right). \tag{1.16}$$

Hence, it was necessary to differentiate the algebraic equation once in order to obtain a differential equation for the algebraic variable y. Consequently, the differential index is one. Equations (1.2) and (1.16) are called the *underlying ODE* of the DAE (1.2)-(1.3). Often, ODEs are viewed as index-0 DAEs, since the underlying ODE is identical with the ODE itself and no differentiations are needed.
In the same way it can be shown that the Hessenberg DAE of order $k \geq 2$ has differential index k, if the matrix M in (1.15) is non-singular.
Consequently, the differential index and the perturbation index coincide for Hessenberg DAEs, cf. also Campbell and Gear [CG95]. For more general DAEs the difference between perturbation index and differential index can be arbitrarily high, cf. Hairer and Wanner [HW96], p. 461.

Remark 1.6 *Notice, that the index calculations are questionable in the presence of a control variable* u. *We assumed that* u *is sufficiently smooth which is not the case for many optimal control problems. Nevertheless, often the optimal control* u *is at least piecewise smooth and hence the interpretation remains valid locally almost everywhere.*

Consistent initial values

DAEs not only differ in their stability behavior from explicit ODEs. Another difference is that initial values have to be defined properly, that is they have to be consistent. We restrict the discussion to Hessenberg DAEs of order k. A general definition of consistency can be found in Brenan et al. [BCP96].
Define

$$g^{(0)}(t, x_{k-1}(t), u(t)) := g(t, x_{k-1}(t), u(t)). \tag{1.17}$$

Of course, an initial value for x_{k-1} at $t = t_0$ has to satisfy this equality. But this is not sufficient, because also time derivatives of (1.17) impose additional restrictions on initial values. This can be seen as follows. Differentiation of $g^{(0)}$ w.r.t. time and substitution of $\dot{x}_{k-1}(t) = f_{k-1}(t, x_{k-2}(t), x_{k-1}(t), u(t))$ leads to the equation

$$\begin{aligned} 0_{n_y} &= g'_t[t] + g'_{x_{k-1}}[t] \cdot f_{k-1}(t, x_{k-2}(t), x_{k-1}(t), u(t)) + g'_u[t] \cdot \dot{u}(t) \\ &=: g^{(1)}(t, x_{k-2}(t), x_{k-1}(t), u(t), \dot{u}(t)), \end{aligned}$$

which has to be satisfied as well. Recursive application of this differentiation and substitution process leads to the equations

$$0_{n_y} = g^{(j)}(t, x_{k-1-j}(t), \ldots, x_{k-1}(t), u(t), \dot{u}(t), \ldots, u^{(j)}(t)), \qquad j = 1, 2, \ldots, k-2, \tag{1.18}$$

and

$$0_{n_y} = g^{(k-1)}(t, x_0(t), x_1(t), \ldots, x_{k-1}(t), u(t), \dot{u}(t), \ldots, u^{(k-1)}(t)). \tag{1.19}$$

Since the equations (1.18)-(1.19) do not occur explicitly in the original system (1.14), these equations are called *hidden constraints* of the Hessenberg DAE. With this notation consistency is defined as follows.

Definition 1.7 (Consistent Initial Value)
Let $x(t_0) = (x_0(t_0), x_1(t_0), \ldots, x_{k-1}(t_0))^\top$ be given and let u be sufficiently smooth. $x(t_0)$ is called consistent *for the Hessenberg DAE of order k, if $x(t_0)$ satisfies the equations (1.17)-(1.19) at $t = t_0$. The differential component is called consistent if equations (1.17)-(1.18) are satisfied.*

Remark 1.8 *The dependence on derivatives of the control in (1.18) and (1.19) can be avoided by introducing additional state variables ξ_j, $j = 0, \ldots, k-2$, by*

$$\xi_0 := u, \ \xi_1 := \dot{u}, \ \ldots \ \xi_{k-2} := u^{(k-2)}$$

satisfying the differential equations

$$\dot{\xi}_0 = \xi_1, \ \ldots \ \dot{\xi}_{k-3} = \xi_{k-2}, \ \dot{\xi}_{k-2} = \tilde{u}$$

and to consider $\tilde{u} := u^{(k-1)}$ as the new control, cf. Müller [Mül03]. Clearly, this approach is nothing else than constructing a sufficiently smooth control u for the original problem. The resulting problem is not equivalent to the original problem anymore. Nevertheless, this strategy is very useful from a practical point of view.

Index reduction and stabilization

The perturbation index measures the stability of the DAE. DAEs with perturbation index greater than or equal to two are called *higher index DAEs.* With increasing perturbation index the DAE suffers from increasing ill-conditioning, since not only the perturbation δ occurs in (1.10) but also derivatives thereof. Thus, even for small perturbations (in the supremum norm) large deviations in the respective solutions are possible. Higher index DAEs occur, e.g., in mechanics, cf. Führer [Füh88], Simeon [Sim94], Arnold [Arn95], process system engineering, cf. Engl et al. [EKKvS99], and electrical engineering, cf. März and Tischendorf [MT97], Günther [Gün95]. To avoid the problems of severe ill-conditioning it is possible to reduce the index by replacing the algebraic constraint in (1.14) by its derivative $g^{(1)}$. It is easy to see that the resulting DAE has perturbation index $k-1$. More generally, the algebraic constraint $g^{(0)}(\cdot) = 0_{n_y}$ can be replaced by any of the constraints $g^{(j)}(\cdot) = 0_{n_y}$ with $1 \leq j \leq k-1$. Then, the resulting DAE has perturbation index $k-j$. The drawback of this index reduction approach is that numerical integration methods suffer from the so-called *drift-off effect.* Since a numerical integration scheme when applied to an index reduced DAE only obeys hidden constraints up to a certain level, the numerical solution will not satisfy the higher level constraints and it can be observed that the magnitude of violation increases with time t. Hence, the numerical solution drifts off the neglected constraints even when the initial value was consistent. Especially for long time intervals the numerical solution may deviate substantially from the solution of the original DAE. One possibility to avoid this drawback is to perform a projection step onto the neglected constraints after each successful integration step for the index reduced system, cf. Ascher and Petzold [AP91] and Eich [Eic93].

Another approach is to stabilize the index reduced problem by adding the neglected constraints to the index reduced DAE. The resulting system is an overdetermined DAE and numerical integration methods have to be adapted, cf. Führer [Füh88], Führer and Leimkuhler [FL91]. The stabilization approach when applied to mechanical multi-body systems is equivalent with the GGL-stabilization, cf. Gear et al. [GLG85].

The above mentioned definitions are illustrated for a very important field of application – mechanical multi-body systems.

Example 1.9 (Mechanical Multi-body Systems)

Let a system of n rigid bodies be given. Every body of mass m_i has three translational and three rotary degrees of freedom. The position of body i in a fixed reference system is given by its position $r_i = (x_i, y_i, z_i)^\top$. The orientation of the bodies coordinate system with respect to the reference coordinate system is given by the angles α_i, β_i, and γ_i. Hence, body i is characterized by the coordinates

$$q_i = (x_i, y_i, z_i, \alpha_i, \beta_i, \gamma_i)^\top \in \mathbb{R}^6$$

and the whole multi-body system is described by

$$q = (q_1, \ldots, q_n)^\top \in \mathbb{R}^{6n}.$$

In general the motion of the n bodies is restricted by holonomic constraints

$$0_{n_g} = g(q).$$

The kinetic energy of the multi-body system is given by

$$T(q, \dot{q}) = \frac{1}{2} \sum_{i=1}^{n} \left(m_i \cdot \dot{q}_i^\top \dot{q}_i + w_i^\top \cdot I_i \cdot w_i \right),$$

where w_i denotes the angular velocity of body i w.r.t. the reference system and I_i is the moment of inertia of body i. The Euler-Lagrangian equations of the first kind are given by

$$\begin{aligned} \frac{d}{dt}\left(T'_{\dot{q}}(q,\dot{q})\right)^\top - \left(T'_q(q,\dot{q})\right)^\top &= F(q,\dot{q},u) - g'(q)^\top\lambda, \\ 0_{n_g} &= g(q), \end{aligned}$$

where F is the vector of applied forces and moments, which may depend on the control u. Explicit calculation of the derivatives leads to the *descriptor form* of mechanical multi-body systems:

$$\begin{aligned} \dot{q}(t) &= v(t), && (1.20) \\ M(q(t))\dot{v}(t) &= f(q(t),v(t),u(t)) - g'(q(t))^\top\lambda(t), && (1.21) \\ 0_{n_g} &= g(q(t)), && (1.22) \end{aligned}$$

where M is the symmetric and positive definite mass matrix and f includes the applied forces and moments and the Coriolis forces. Multiplication of the second equation with M^{-1} yields a Hessenberg DAE of order 3.

Let g be twice continuously differentiable. The constraint $g(q(t)) = 0_{n_g}$ is called *constraint on position level*. Differentiation w.r.t. time of this algebraic constraint yields the *constraint on velocity level*

$$g'(q(t)) \cdot v(t) = 0_{n_g}$$

and the *constraint on acceleration level*

$$g'(q(t)) \cdot \dot{v}(t) + g''_{qq}(q(t))(v(t),v(t)) = 0_{n_g}.$$

Replacing $\dot{v}$ by

$$\dot{v}(t) = M(q(t))^{-1}\left(f(q(t),v(t),u(t)) - g'(q(t))^\top\lambda(t)\right)$$

yields

$$g'(q(t))M(q(t))^{-1}\left(f(q(t),v(t),u(t)) - g'(q(t))^\top\lambda(t)\right) + g''_{qq}(q(t))(v(t),v(t)) = 0_{n_g}.$$

If $\operatorname{rank}(g'(q)) = n_g$, then the matrix $g'(q)M(q)^{-1}g'(q)^\top$ is non-singular and the latter equation can be solved for the algebraic variable λ. By the same reasoning as before it can be shown that the descriptor form has index 3. Consistent initial values (q_0, v_0, λ_0) satisfy the constraints on position, velocity, and acceleration level.

For numerical methods it is advisable to perform an index reduction, i.e. the constraint on position level is replaced by the constraint on velocity or acceleration level. An even better idea is to use the *GGL-stabilization*

$$\begin{aligned} \dot{q}(t) &= v(t) - g'(q(t))^\top\mu(t), && (1.23) \\ M(q(t))\dot{v}(t) &= f(q(t),v(t),u(t)) - g'(q(t))^\top\lambda(t), && (1.24) \\ 0_{n_g} &= g(q(t)), && (1.25) \\ 0_{n_g} &= g'(q(t)) \cdot v(t). && (1.26) \end{aligned}$$

This DAE is equivalent to an index 2 Hessenberg DAE, if the second equation is multiplied with M^{-1}. Furthermore, differentiation of the first algebraic equation yields

$$0_{n_y} = g'(q(t)) \cdot \left(v(t) - g'(q(t))^\top\mu(t)\right) = -g'(q(t))g'(q(t))^\top\mu(t).$$

If $\text{rank}(g'(q)) = n_g$ then $g'(q)g'(q)^\top$ is non-singular and it follows $\mu \equiv 0_{n_g}$. Hence, the GGL-stabilization (1.23)-(1.26) is equivalent to the overdetermined (stabilized) descriptor form

$$\begin{aligned}
\dot{q}(t) &= v(t), \\
M(q(t))\dot{v}(t) &= f(q(t), v(t), u(t)) - g'(q(t))^\top \lambda(t), \\
0_{n_g} &= g(q(t)), \\
0_{n_g} &= g'(q(t)) \cdot v(t).
\end{aligned}$$

■

After these introductory remarks on DAEs an outline of the book is as follows. In Chapter 2 the functional analytic background needed for the derivation of local minimum principles for DAE optimal control problems is provided. For the readers convenience we included a rather detailed presentation of the background material. Although most of the material is standard, we found that some parts, e.g. the more advanced properties of Stieltjes integrals and the results on variational equalities and inequalities involving functions of bounded variation and Stieltjes integrals, are hard to find in the literature.

Chapter 3 summarizes results on finite and infinite optimization problems. Similar as in Chapter 2, for the sake of completeness, we provide a detailed presentation of the results including most of the proofs. We found it useful to include, e.g., the proof of the first order necessary Fritz-John conditions since it sheds light on many difficulties arising in infinite dimensions. Furthermore, the exposition attempts to combine several versions of necessary conditions from the literature, that differ sligthly in their assumptions or the problem classes under consideration, and to adapt them to our purposes. Readers who are familiar with infinite and finite dimensional optimization theory may skip large parts of this chapter. However, Chapter 3 is of central importance for all upcoming theoretical and numerical approaches towards optimal control problems.

Since most current numerical integration methods are only capable of solving nonlinear index-1 and index-2 Hessenberg DAEs, we restrict the discussion to optimal control problems subject to these two classes of DAEs. Of course, explicit ODEs are included in either of such DAEs as subclasses. The optimal control problem is considered as an infinite dimensional optimization problem of type

$$F(x, y, u) \to \min \qquad \text{s.t.} \qquad G(x, y, u) \in K,\ H(x, y, u) = \Theta,\ u \in \mathcal{U},$$

where F, G, H are mappings between appropriate Banach spaces, K is a cone, and $\mathcal{U}$ is a convex set with non-empty interior. The proof of the local minimum principles in Chapter 4 exploits first order necessary conditions for such general infinite dimensional optimization problems. An additional regularity assumption similar to the Mangasarian-Fromowitz condition in finite dimensional optimization provides a constraint qualification for the local solution. This chapter is a major contribution of the author to the theory of optimal control processes with DAE constraints. For such problems only very few theoretical results are known up to now.

The local minimum principle usually leads to a boundary value problem and thus provides the basis of the so-called indirect solution approach for optimal control problems. The indirect method attempts to solve the minimum principle resp. the boundary value problem numerically. In Chapter 3 we summarize also well-known necessary and sufficient conditions for finite dimensional optimization problems. These conditions are the basis for the direct solution approach for optimal control problems. This direct approach is based on a discretization of the optimal control problem and leads to a finite dimensional optimization problem, which is solved numerically by the sequential quadratic programming (SQP) method. Direct discretization methods for optimal control problems are discussed in Chapter 6. These methods need integration methods for DAEs that are briefly summarized in Chapter 5. In particular, the class of linearized Runge-Kutta methods is introduced and analysed. This new class turns out to be efficient for discretized optimal control problems. Application of the necessary finite dimensional Fritz-John conditions yields a discrete version of the local minimum principle and allows to compute approximations for the adjoints (co-states). On the other hand, the adjoint equation can be used for the calculation of gradients within the SQP method. An alternative method for calculating gradients is the sensitivity equation approach, which is preferable if the number of variables is small compared to the number of constraints. Finally, selected applications from real-time optimization, parameter identification, and mixed-integer optimal control are discussed in Chapter 7. The methods and results presented in chapters 6 and 7 include contributions of the author to the numerical aspects of DAE optimal control problems, e.g. sensitivity analysis, gradient calculation, determination of consistent initial values and mixed-integer optimal control.

Chapter 2

Basics from Functional Analysis

2.1 Vector Spaces

Let $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$ and X be a set. Let two algebraic operations $+ : X \times X \to X$ (addition) and $\cdot : \mathbb{K} \times X \to X$ (scalar multiplication) be defined. Recall, that $(X, +)$ is called an *Abelian group*, if

(i) $(x + y) + z = x + (y + z)$ holds for all $x, y, z \in X$ (associative law);

(ii) there exists an element $\Theta_X \in X$ with $\Theta_X + x = x$ for all $x \in X$ (existence of null-element);

(iii) for every $x \in X$ there exists $x' \in X$ with $x + x' = \Theta_X$ (existence of inverse);

(iv) $x + y = y + x$ holds for all $x, y \in X$ (commutative law).

Θ_X is the null-element of $(X, +)$. If no confusion is possible, we will use simply Θ. In the special cases $X = \mathbb{R}^n$ and $X = \mathbb{R}$ the null-element is denoted by 0_n and 0, respectively.
$(X, +, \cdot)$ is called a *vector space* or *linear space* over $\mathbb{K}$, if $(X, +)$ is an Abelian group and if the following computational rules are satisfied:

(i) $(s \cdot t) \cdot x = s \cdot (t \cdot x)$ for all $s, t \in \mathbb{K}\ x \in X$;

(ii) $s \cdot (x + y) = s \cdot x + s \cdot y$ for all $s \in \mathbb{K},\ x, y \in X$;

(iii) $(s + t) \cdot x = s \cdot x + t \cdot x$ for all $s, t \in \mathbb{K},\ x \in X$;

(iv) $1 \cdot x = x$ for all $x \in X$.

A *topology* τ on a set X is a subset of the power set 2^X of X with the following properties:

(i) $\emptyset, X$ belong to τ.

(ii) The union of arbitrary many elements of τ belongs to τ.

(iii) The intersection of finitely many elements of τ belongs to τ.

The elements of a topology τ are called *open sets*. The complement of an open set is a *closed set*. The tuple (X, τ) is called a *topological vector space* over $\mathbb{K}$, if

(i) X is a vector space over $\mathbb{K}$,

(ii) X is endowed with a topology τ,

(iii) addition and multiplication by scalars are continuous functions in the given topology.

Let X and Y be topological vector spaces and $f : X \to Y$ a mapping. f is called *continuous at* $x \in X$, if for all open (closed) sets V containing $y = f(x) \in Y$ there exists an open (closed) set U in X with $x \in U$ and $f(U) \subseteq V$. f is called *continuous on* X, if f is continuous at every $x \in X$.

In order to define a topology in a vector space it is sufficient to specify a basis $\mathcal{A}$ of open sets around zero. A basis around zero is characterized by the property that for any neighborhood V of zero there exists $U \in \mathcal{A}$ such that $U \subseteq V$. Herein, every open set containing x is called a *neighborhood of* x. Then, a set S is open, if and only if for every $x \in S$ there exists an element $U \in \mathcal{A}$ such that $x + U \subseteq S$. Furthermore, $x \in S$ is called an *interior point of* S, if there is a neighborhood U of x with $U \subseteq S$. The set of all interior points of S is denoted by $\text{int}(S)$. The *closure* $\text{cl}(S)$ *of* S is the set of all points x satisfying $U \cap S \neq \emptyset$ for all neighborhoods U of x. x is a *boundary point of* S, if $x \in \text{cl}(S)$ and $x \notin \text{int}(S)$. We have that S is open, if and only if $S = \text{int}(S)$. Furthermore, S is closed, if and only if $S = \text{cl}(S)$.

A *metric space* is a tupel (X, d), where X is a set and $d : X \times X \to \mathbb{R}$ is a mapping such that for every $x, y, z \in X$ it holds

(i) $d(x, y) \geq 0$ and $d(x, y) = 0$, if and only if $x = y$;

(ii) $d(x, y) = d(y, x)$;

(iii) $d(x, y) \leq d(x, z) + d(z, y)$.

d is called a *metric* on X.

In a metric space the sequence $\{x_n\}$ is said to *converge* to $x \in X$, that is $x_n \to x$, if $d(x_n, x) \to 0$. If a sequence converges, its limit is unique. Recall that a metric space X is called *complete*, if every Cauchy sequence from X has a limit in X. A sequence $\{x_n\}$ in X is a *Cauchy sequence*, if for every $\varepsilon > 0$ there is a $N(\varepsilon) \in \mathbb{N}$ such that $d(x_n, x_m) < \varepsilon$ for all $n, m > N(\varepsilon)$. In a metric space every convergent sequence is a Cauchy sequence.

A metric space (X, d) is called *compact*, if for every sequence $\{x_n\}$ in X there exists a subsequence $\{x_{n_k}\}$ converging to an element $x \in X$. Every compact metric space is complete. In finite dimensional spaces, compactness is equivalent with closedness and boundedness. In infinite dimensional spaces this is not true. For instance, it is shown in Luenberger [Lue69], p. 40, that the unit sphere in general is not compact. A subset S of a metric space (X, d) is called *dense in* X, if $cl(S) = X$ holds.

Let X be a vector space over $\mathbb{K}$. The tupel $(X, \|\cdot\|_X)$ is called a *normed vector space*, if $\|\cdot\|_X : X \to \mathbb{R}$ is a mapping such that for every $x, y \in X$ and every $\lambda \in \mathbb{K}$ it holds

(i) $\|x\|_X \geq 0$ and $\|x\|_X = 0$ if and only if $x = \Theta_X$;

(ii) $\|\lambda x\|_X = |\lambda| \cdot \|x\|_X$;

(iii) $\|x + y\|_X \leq \|x\|_X + \|y\|_X$.

The mapping $\|\cdot\|_X$ is called a *norm on* X. Since every norm defines a metric by $d(x, y) := \|x - y\|_X$ the terminologies 'convergence, complete, Cauchy sequence,...' can be translated directly to normed spaces $(X, \|\cdot\|_X)$.

Definition 2.1.1 (Banach space)
A complete normed vector space is called Banach space*.*

In a Banach space for $r > 0$ let

$$\begin{aligned} U_r(x) &:= \{y \in X \mid \|y - x\|_X < r\}, \\ \overline{U_r(x)} &:= \{y \in X \mid \|y - x\|_X \leq r\} \end{aligned}$$

denote the open and closed balls around x with radius r, respectively. Then $\mathcal{A} = \{U_r(\Theta_X) \mid r > 0\}$ defines a basis of open sets about zero. We say, the norm $\|\cdot\|_X$ induces the *strong topology on* X.

Finally, we introduce Hilbert spaces.

Definition 2.1.2 (Inner Product, Scalar Product)
Let X be a vector space over $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$. The mapping $\langle \cdot, \cdot \rangle_X : X \times X \to \mathbb{K}$ is called an inner product *or* scalar product*, if the following conditions hold for all $x, y, z \in X$ and all $\lambda \in \mathbb{K}$.*

(i) $\langle x, y \rangle_X = \overline{\langle y, x \rangle_X}$;

(ii) $\langle x + y, z \rangle_X = \langle x, z \rangle_X + \langle y, z \rangle_X$;

(iii) $\langle x, \lambda y \rangle_X = \lambda \langle x, y \rangle_X$;

(iv) $\langle x, x \rangle_X \geq 0$ *and* $\langle x, x \rangle_X = 0$*, if and only if* $x = \Theta_X$.

Definition 2.1.3 (pre-Hilbert Space, Hilbert Space)
A pre-Hilbert space *is a vector space together with an inner product. A complete pre-Hilbert space is called* Hilbert space.

Notice, that $\|x\|_X := \sqrt{\langle x, x \rangle_X}$ defines a norm on a pre-Hilbert space X.

2.2 Mappings, Dual Spaces, and Properties

From now on, unless otherwise specified, $(X, \|\cdot\|_X), (Y, \|\cdot\|_Y), (Z, \|\cdot\|_Z), \ldots$ denote real Banach spaces over $\mathbb{K} = \mathbb{R}$ with norms $\|\cdot\|_X, \|\cdot\|_Y, \|\cdot\|_Z, \ldots$. If no confusion is possible, we omit the norms and call $X, Y, Z, \ldots$ Banach spaces and assume that appropriate norms are defined.

Definition 2.2.1
Let $T : X \to Y$ be a mapping from a Banach space $(X, \|\cdot\|_X)$ into a Banach space $(Y, \|\cdot\|_Y)$.

(i) The image of T *is given by*

$$im(T) := \{T(x) \mid x \in X\}.$$

The kernel of T *is given by*

$$ker(T) := \{x \in X \mid T(x) = \Theta_Y\}.$$

Given a set $S \subseteq Y$, the preimage of S under T *is given by*

$$T^{-1}(S) := \{x \in X \mid T(x) \in S\}.$$

(ii) T is called linear*, if*

$$T(x_1 + x_2) = T(x_1) + T(x_2), \quad T(\lambda x_1) = \lambda T(x_1) \qquad \forall x_1, x_2 \in X, \lambda \in \mathbb{R}.$$

(iii) T is called continuous at $x \in X$*, if for every sequence $x_i \to x$ it holds $T(x_i) \to T(x)$. T is called continuous on X, if T is continuous at all $x \in X$.*

(iv) Let $D \subseteq X$ be open. A function $f : D \to Y$ is called locally Lipschitz continuous at $x \in D$ with constant L*, if there exists a $\varepsilon > 0$ such that*

$$\|f(y) - f(z)\|_Y \leq L\|y - z\|_X \qquad \forall y, z \in U_\varepsilon(x).$$

(v) T is called bounded*, if $\|T(x)\|_Y \leq C \cdot \|x\|_X$ holds for all $x \in X$ and some constant $C \geq 0$.*

The operator norm *is defined by*

$$\|T\|_{X,Y} = \sup_{x \neq \Theta_X} \frac{\|T(x)\|_Y}{\|x\|_X} = \sup_{\|x\|_X \leq 1} \|T(x)\|_Y = \sup_{\|x\|_X = 1} \|T(x)\|_Y.$$

(vi) *If* $Y = \mathbb{R}$, *then* T *is called a* functional*.*

(vii) *A functional* $f : X \to \mathbb{R}$ *is called* upper semicontinuous at x*, if for every sequence* $\{x_i\}$ *with* $x_i \to x$ *it holds*

$$\limsup_{i\to\infty} f(x_i) \leq f(x).$$

A functional $f : X \to \mathbb{R}$ *is called* lower semicontinuous at x*, if for every sequence* $\{x_i\}$ *with* $x_i \to x$ *it holds*

$$f(x) \leq \liminf_{i\to\infty} f(x_i).$$

(viii) *The set of all linear continuous functionals on* X *endowed with the norm*

$$\|f\|_{X^*} = \sup_{\|x\|_X \leq 1} |f(x)|$$

is called dual space of X *and is denoted by* X^**. This norm defines the* strong topology on X^**.* X *is called* reflexive*, if* $(X^*)^* = X$ *holds with respect to the strong topology.*

(ix) *Let* $T : X \to Y$ *be linear. The* adjoint operator $T^* : Y^* \to X^*$ *is a linear operator defined by*

$$T^*(y^*)(x) = y^*(T(x)) \qquad \forall y^* \in Y^*,\ x \in X.$$

A linear operator is continuous, if and only if it is bounded. A linear operator is continuous, if and only if it is continuous at zero. If T is continuous, then so is T^*. The set $\mathcal{L}(X, Y)$ of all linear continuous mappings from X into Y endowed with the operator norm $\|\cdot\|_{X,Y}$ is a Banach space. The dual space of a Banach space is a Banach space.

Theorem 2.2.2 (cf. Ljusternik and Sobolew [LS76], p. 109, Th. 4)
Let X, Y *be Banach spaces and* $T : X \to Y$ *a continuous, linear, surjective, and injective operator. Then the inverse operator* T^{-1} *exists and it is linear and continuous.*

It is well-known, that given n Banach spaces $X_1, X_2, \ldots, X_n$ with norms $\|\cdot\|_1, \|\cdot\|_2, \ldots, \|\cdot\|_n$, the product space

$$X = X_1 \times X_2 \times \cdots \times X_n$$

equipped with one of the norms

$$\|x\|_X = \max_{1\leq i\leq n} \|x_i\|_{X_i}, \qquad \|x\|_X = \left(\sum_{i=1}^n \|x_i\|_{X_i}^p\right)^{1/p}, \qquad p \in \mathbb{N},$$

is also a Banach space. The dual space of X is

$$X^* = \{x^* = (x_1^*, x_2^*, \ldots, x_n^*) \mid x_i^* \in X_i^*,\ i = 1, \ldots, n\}, \qquad x^*(x) = \sum_{i=1}^n x_i^*(x_i).$$

The following theorem can be found in Werner [Wer95], p. 135, Th. IV.3.3.

Theorem 2.2.3 (Open Mapping Theorem)
Let $T : X \to Y$ *be a linear, continuous, and surjective operator. Let* $S \subseteq X$ *be open. Then* $T(S) \subseteq Y$ *is open.*

Let $T : X \to Y$ be linear and continuous. Then, the following statements are equivalent, cf. Werner [Wer95], p. 143, Th. IV.5.1:

(i) $\mathrm{im}(T)$ is closed.

(ii) $\mathrm{im}(T) = (\ker(T^*))^{\perp} := \{x \in X \mid x^*(x) = 0 \ \forall x^* \in X^*\}$.

(iii) $\mathrm{im}(T^*)$ is closed.

(iv) $\mathrm{im}(T^*) = (\ker(T))^{\perp} := \{x^* \in X^* \mid x^*(x) = 0 \ \forall x \in X\}$.

Let $T : X \to Y$ be a linear and continuous mapping. Then, $\ker(T)$ is a closed subspace of X.

Definition 2.2.4 (Factor space, Quotient space, codimension)
Let X be a vector space and $L \subseteq X$ a subspace. The factor space *or* quotient space *X/L consists of all sets $[x]$, $x \in X$, where $[x] := x + L$. Vectors $x, y \in X$ are called* L-equivalent, *if $x - y \in L$. The dimension of X/L is called* codimension of L *or* defect of L.

If X is finite dimensional, then $\dim(X/L) = \dim(X) - \dim(L)$, cf. Kowalsky [Kow69], p. 214.
Let L be a linear subspace of the real vector space X and $x \in X$. The set $x + L$ is called a *linear manifold*. Let $\{e_i\}_{i \in I}$ be a basis of X and $\{e_j\}_{j \in J}$, $J \subseteq I$ a basis of L. The *dimension* of the linear manifold $x + L$ is defined as $|J|$ and the *codimension* or *defect* of $x + L$ is defined as $|I \setminus J|$. Hyperplanes are linear manifolds having defect one. Furthermore, H is a hyperplane, if and only if it can be represented as $H = \{x \in X \mid f(x) = \gamma\}$, where f is a non-zero linear functional $f : X \to \mathbb{R}$ and $\gamma \in \mathbb{R}$. Given a set S in a real vector space X, the codimension of S is defined as the codimension of the linear subspace parallel to the affine hull of S.
If L is a closed subspace of the Banach space X then

$$\|[x]\|_{X/L} := \inf\{\|y - x\|_X \mid y \in L\}$$

defines a norm on X/L and X/L endowed with the norm $\|\cdot\|_{X/L}$ is a Banach space.
The canonical mapping $w : X \to X/L$, $x \mapsto [x]$ is surjective and $\ker(w) = L$, cf. Kowalsky [Kow69], p. 213.

Theorem 2.2.5 (cf. Kowalsky [Kow69], p. 214)
Let $T : X \to Y$ be a linear mapping from the vector space X into the vector space Y and $L \subseteq ker(T)$ a subspace of X. Then T can be factorized uniquely as $T = \hat{T} \circ w$ with a linear mapping $\hat{T} : X/L \to Y$. Furthermore, $\hat{T}$ is injective, if and only if $ker(T) = L$. $\hat{T}$ is surjective, if and only if T is surjective.

As a direct consequence of the theorem we find

Corollary 2.2.6 *Let $T : X \to Y$ be a linear and surjective mapping from the vector space X onto the vector space Y. Then there exists a linear, surjective and injective mapping $\hat{T} : X/ker(T) \to Y$.*

A linear continuous operator $T : X \to Y$ is called an *isomorphism*, if T is surjective and injective and T^{-1} is linear and continuous. Two normed spaces X and Y are called *isomorph*, if there exists an isomorphism between X and Y.

Theorem 2.2.7 *Let X, Y be Banach spaces and $T : X \to Y$ a linear, continuous, and surjective operator. Then there exists a linear, continuous, surjective, and injective mapping $\hat{T} : X/ker(T) \to Y$ such that $\hat{T}^{-1}$ exists and it is linear and continuous (that is, $X/ker(T)$ and Y are isomorphic).*

Proof. According to Corollary 2.2.6 it remains to show that $\hat{T}$ is continuous and that $\hat{T}^{-1}$ exists and is continuous. The continuity follows from $\hat{T}([x]) = T(x)$ and the continuity of T. Since X, Y are Banach spaces, the existence of the inverse operator $\hat{T}^{-1}$ and its continuity follows from Theorem 2.2.2. ■

Theorem 2.2.8 *Let $F : X \to Y \times \mathbb{R}^n$ be defined by $F(x) = (G(x), H(x))$, where $G : X \to Y$ is a linear, continuous, and surjective operator and $H : X \to \mathbb{R}^n$ is linear and continuous. Then $im(F) = F(X)$ is closed in $Y \times \mathbb{R}^n$.*

Proof. Suppose that $\mathrm{im}(F)$ is not closed. Then there exists a sequence $(y_i, z_i) \in \mathrm{im}(F)$ with $\lim_{i\to\infty}(y_i, z_i) = (y_0, z_0) \notin \mathrm{im}(F)$. Then, there exists a (algebraic) hyperplane, which contains $\mathrm{im}(F)$ but not (y_0, z_0). Hence, there exist a linear functional $l_y \in Y'$ and a continuous linear functional $l_z \in (\mathbb{R}^n)^*$, not both zero, with

$$\begin{aligned} l_y(G(x)) + l_z(H(x)) &= 0, \qquad (x \in X) \\ l_y(y_0) + l_z(z_0) &\neq 0. \end{aligned}$$

Notice, that l_y is not necessarily continuous. In fact, we will show that l_y is actually continuous. Let $\eta_i \in Y$ be an arbitrary sequence converging to zero in Y. Unfortunately, there may be many points in the preimage of η_i under the mapping G, i.e. G^{-1} is in general no mapping. To circumvent this problem, instead we consider the factor space $X/\ker(G)$ whose elements $[x]$ are defined by $[x] := x + \ker(G)$ for $x \in X$. The space $X/\ker(G)$ endowed with the norm $\|[x]\|_{X/\ker(G)} := \inf\{\|y - x\|_X \mid y \in \ker(G)\}$ is a Banach space. Notice that $\ker(G)$ is a closed subspace of X.

If we consider G as a mapping from the factor space $X/\ker(G)$ onto Y, then the inverse operator of this mapping is continuous, because G is surjective and X and Y are Banach spaces. Hence, to each $\eta_i \in Y$ there corresponds an element W_i of $X/\ker(G)$. Since the inverse operator is continuous and η_i is a null-sequence, the sequence W_i converges to zero. Furthermore, we can choose a representative $\xi_i \in W_i$ such that

$$\begin{aligned} \|\xi_i\|_X &\leq 2\|W_i\|_{X/\ker(G)}, \\ G(\xi_i) &= \eta_i \end{aligned}$$

hold for every $i \in \mathbb{N}$. Since W_i is a null-sequence the same holds for ξ_i by the first inequality. This yields

$$\begin{aligned} \lim_{i\to\infty} l_y(\eta_i) &= \lim_{i\to\infty} l_y(G(\xi_i)) \\ &= \lim_{i\to\infty} -l_z(H(\xi_i)) \\ &= 0, \end{aligned}$$

since $\xi_i \to \Theta_X$ and H and l_z are continuous. This shows, that l_y is actually continuous in Θ_X and hence on X.

In particular, we obtain

$$\begin{aligned} 0 &= \lim_{i\to\infty} (l_y(y_i) + l_z(z_i)) \\ &= l_y\left(\lim_{i\to\infty} y_i\right) + l_z\left(\lim_{i\to\infty} z_i\right) \\ &= l_y(y_0) + l_z(z_0) \end{aligned}$$

in contradiction to $0 \neq l_y(y_0) + l_z(z_0)$. ■

2.3 Function Spaces

Some important real Banach and Hilbert spaces are summarized in the sequel.

2.3.1 L^p-Spaces

Let $1 \leq p < \infty$. The space $L^p([a,b],\mathbb{R})$ consists of all measurable functions $f : [a,b] \to \mathbb{R}$ with

$$\int_a^b |f(t)|^p dt < \infty,$$

where the integral denotes the Lebesgue integral. Notice, that functions that differ only on a set of measure zero, i.e. functions that are equal almost everywhere, are considered to be the same. In this sense, the elements of L^p are equivalence classes. The space $L^\infty([a,b],\mathbb{R})$ consists of all measurable functions $f : [a,b] \to \mathbb{R}$ which are essentially bounded, i.e.

$$\operatorname*{ess\ sup}_{a\leq t\leq b} |f(t)| := \inf_{\substack{N\subset[a,b]\\ \mu(N)=0}} \sup_{t\in[a,b]\setminus N} |f(t)| < \infty.$$

The spaces $L^p([a,b],\mathbb{R})$, $1 \leq p < \infty$ endowed with the norm

$$\|f\|_p := \left(\int_a^b |f(t)|^p dt\right)^{1/p}$$

and the space $L^\infty([a,b],\mathbb{R})$ endowed with the norm

$$\|f\|_\infty := \operatorname*{ess\ sup}_{a\leq t\leq b} |f(t)|$$

are Banach spaces, cf. Kufner et al. [KOS77], Ths. 2.8.2, 2.11.7. For $1 < p < \infty$ the dual space of $L^p([a,b],\mathbb{R})$ is given by $L^q([a,b],\mathbb{R})$ where $1/p + 1/q = 1$, i.e. for $f^* \in (L^p([a,b],\mathbb{R}))^*$ there exists a unique function $g \in L^q([a,b],\mathbb{R})$ such that

$$f^*(f) = \int_a^b f(t)g(t)dt, \qquad \forall f \in L^p([a,b],\mathbb{R}),$$

and $\|f^*\| = \|g\|_q$, cf. Kufner et al. [KOS77], Th. 2.9.5. For $1 < p < \infty$ the spaces L^p are reflexive, cf. Kufner et al. [KOS77], Th. 2.10.1.

The dual space of $L^1([a,b],\mathbb{R})$ is given by $L^\infty([a,b],\mathbb{R})$, i.e. for $f^* \in (L^1([a,b],\mathbb{R}))^*$ there exists a unique function $g \in L^\infty([a,b],\mathbb{R})$ such that

$$f^*(f) = \int_a^b f(t)g(t)dt, \qquad \forall f \in L^1([a,b],\mathbb{R}),$$

cf. Kufner et al. [KOS77], Th. 2.11.8. The spaces $L^1([a,b],\mathbb{R})$ and $L^\infty([a,b],\mathbb{R})$ are not reflexive, cf. Kufner et al. [KOS77], Ths. 2.11.10, 2.11.11.

The dual space of $L^\infty([a,b],\mathbb{R})$ does not have a nice structure. According to Kufner et al. [KOS77], Rem. 2.17.2, the space $L^\infty([a,b],\mathbb{R})$ is isometrically isomorph with the space of all finitely additive measures on the family of measurable subsets of $[a,b]$ which are absolutely continuous w.r.t. the Lebesgue measure.

$L^2([a,b],\mathbb{R})$ is a Hilbert space with the inner product

$$\langle f, g\rangle_{L^2} := \int_a^b f(t)g(t)dt.$$

For $1 \leq p \leq \infty$ the space $L^p([a,b],\mathbb{R}^n)$ is defined as the product space

$$L^p([a,b],\mathbb{R}^n) := L^p([a,b],\mathbb{R}) \times \cdots \times L^p([a,b],\mathbb{R}),$$

where each element f of $L^p([a,b],\mathbb{R}^n)$ is a mapping from $[a,b]$ in $\mathbb{R}^n$.

2.3.2 Absolutely Continuous Functions

A function $f : [a, b] \to \mathbb{R}$ is said to be *absolutely continuous*, if for every $\varepsilon > 0$ there exists a $\delta(\varepsilon) > 0$ such that

$$\sum_{i=1}^{m} |b_i - a_i| < \delta(\varepsilon) \qquad \Rightarrow \qquad \sum_{i=1}^{m} |f(b_i) - f(a_i)| < \varepsilon,$$

where $m \in \mathbb{N}$ is arbitrary and $(a_i, b_i) \subseteq [a, b]$, $i = 1, \dots, m$ are disjoint intervals. The set of all absolutely continuous functions is called $AC([a, b], \mathbb{R})$. The following properties of absolutely continuous functions can be found in Natanson [Nat75].

If $f \in L^1([a, b], \mathbb{R})$ then the undetermined integral

$$F(t) := C + \int_a^t f(\tau)d\tau$$

is absolutely continuous and it holds

$$F'(t) = f(t) \qquad \text{a.e. in } [a, b].$$

On the other hand, if $f' \in L^1([a, b], \mathbb{R})$ exists everywhere and is finite, then it holds

$$f(t) = f(a) + \int_a^t f'(\tau)d\tau. \tag{2.3.1}$$

Furthermore, the *partial integration*

$$\int_a^b f(t)g'(t)dt + \int_a^b g(t)f'(t)dt = [f(t)g(t)]_a^b \, .$$

holds for absolutely continuous functions f and g on $[a, b]$.

In addition, an absolutely continuous function on $[a, b]$ is continuous, of bounded variation, f' exists almost everywhere in $[a, b]$ and satisfies (2.3.1) for $a \le t \le b$. If f' is zero almost everywhere in $[a, b]$, then f is constant. If the derivatives f' and g' of absolutely continuous functions f and g are equal almost everywhere, then the difference $f - g$ is constant.

2.3.3 $W^{q,p}$-Spaces

Let $1 \le q, p \le \infty$. The space $W^{q,p}([a, b], \mathbb{R})$ consists of all absolutely continuous functions $f : [a, b] \to \mathbb{R}$ with absolutely continuous derivatives up to order $q - 1$ and

$$\|f\|_{q,p} < \infty,$$

where the norm is given by

$$\begin{aligned} \|f\|_{q,p} &:= \left(\sum_{i=0}^{q} \|f^{(i)}\|_p^p \right)^{1/p}, \qquad 1 \le p < \infty, \\ \|f\|_{q,\infty} &:= \max_{0 \le i \le q} \|f^{(i)}\|_\infty. \end{aligned}$$

The spaces $W^{q,p}([a, b], \mathbb{R})$, $1 \le q, p \le \infty$ endowed with the norm $\| \cdot \|_{q,p}$ are Banach spaces. The spaces $W^{q,2}([a, b], \mathbb{R})$ are Hilbert spaces with the inner product

$$\langle f, g \rangle_{W^{q,2}} := \sum_{i=0}^{q} \int_a^b f^{(i)}(t)g^{(i)}(t)dt.$$

2.3.4 Functions of Bounded Variation

A function $f : [a, b] \to \mathbb{R}$ is said to be of *bounded variation*, if there exists a constant K such that for any partition

$$\mathbb{G}_m := \{a = t_0 < t_1 < \ldots < t_m = b\}$$

of $[a, b]$ it holds

$$\sum_{i=1}^{m} |f(t_i) - f(t_{i-1})| \leq K.$$

The *total variation of f* is

$$TV(f, a, b) := \sup_{\text{all } \mathbb{G}_m} \sum_{i=1}^{m} |f(t_i) - f(t_{i-1})|.$$

The space $BV([a, b], \mathbb{R})$ consists of all functions of bounded variation on $[a, b]$.
We summarize some facts about functions of bounded variation, cf. Natanson [Nat75].

- The derivative f' of a function $f \in BV([a, b], \mathbb{R})$ exists almost everywhere in $[a, b]$ and the integral $\int_a^b f'(t)dt$ exists.
- $f \in BV([a, b], \mathbb{R})$ possesses only countably many jumps. For every jump point of f the left and right-sided limits exist.
- Every function f of bounded variation can be represented as

  $$f(t) = g(t) + s(t) + r(t),$$

  where g is absolutely continuous, s is the jump function of f, i.e.

  $$\begin{aligned} s(a) &= 0, \\ s(t) &= (f(a+) - f(a)) + \sum_{t_i < t} (f(t_i+) - f(t_i-)) + (f(t) - f(t-)), \qquad a < t \leq b. \end{aligned}$$

  and r is singular. Herein, r is called *singular*, if it is a non-constant, continuous function of bounded variation, whose derivative is zero almost everywhere. Note, that a singular function is not absolutely continuous, since then it would be constant.

  If f is continuous, then s is zero. If f is continuous and monotonically increasing, then g and r are also monotonically increasing.

2.4 Stieltjes Integral

The Stieltjes integral generalizes the Riemann integral and is needed for the formulation of the minimum principle for state constrained optimal control problems. Let f and μ be two functions defined on the interval $[a, b]$. Let a partition

$$\mathbb{G}_m := \{a = t_0 < t_1 < \ldots < t_m = b\}$$

of $[a, b]$ be given. Let $\xi_i \in [t_{i-1}, t_i]$ be arbitrary and

$$S(f, \mu) := \sum_{i=1}^{m} f(\xi_i) \left(\mu(t_i) - \mu(t_{i-1})\right).$$

If $S(f,\mu)$ converges to a finite value S independently of the points ξ_i as $\max_i\{t_{i+1}-t_i\} \to 0$, then S is called *Stieltjes integral of f w.r.t. μ* and we write

$$\int_a^b f(t)d\mu(t).$$

More precisely, S is called Stieltjes integral of f w.r.t. μ, if for every $\varepsilon > 0$ there exists a $\delta > 0$ such that for every partition with $\max_i\{t_{i+1}-t_i\} < \delta$ it holds $|S(f,\mu) - S| < \varepsilon$.
The existence of the Stieltjes integral is guaranteed, if f is continuous and μ is of bounded variation on $[a,b]$. Notice, that the Riemann integral is a special case of the Stieltjes integral for $\mu(t) = t$. Some properties of the Stieltjes integral are summarized below, cf. Natanson [Nat75] and Widder [Wid46].

- If one of the integrals $\int_a^b f(t)d\mu(t)$ or $\int_a^b \mu(t)df(t)$ exists, then so does the other and it holds
$$\int_a^b f(t)d\mu(t) + \int_a^b \mu(t)df(t) = [f(t)\mu(t)]_a^b .$$
This is the *partial integration rule.*

- It holds
$$\int_a^b d\mu(t) = \mu(b) - \mu(a).$$
If μ is constant then
$$\int_a^b f(t)d\mu(t) = 0.$$

- If f is continuous and μ is of bounded variation in $[a,b]$ then
$$F(t) = \int_a^t f(\tau)d\mu(\tau), \quad a \le t \le b$$
is of bounded variation. Moreover
$$\begin{aligned} F(t+) - F(t) &= f(t)(\mu(t+) - \mu(t)), \qquad a \le t < b, \\ F(t) - F(t-) &= f(t)(\mu(t) - \mu(t-)), \qquad a < t \le b. \end{aligned}$$

- If f is continuous, $g \in L^1([a,b],\mathbb{R})$ and
$$\mu(t) = \int_c^t g(\tau)d\tau, \quad a \le c \le b,\ a \le t \le b,$$
then μ is of bounded variation on (a,b) and
$$\int_a^b f(t)d\mu(t) = \int_a^b f(t)g(t)dt = \int_a^b f(t)\mu'(t)dt.$$
The latter integral is a Lebesgue integral.

  If g is continuous, h is of bounded variation in $[a,b]$ and
$$\mu(t) = \int_c^t g(\tau)dh(\tau), \quad a \le c \le b,\ a \le t \le b,$$
then
$$\int_a^b f(t)d\mu(t) = \int_a^b f(t)g(t)dh(t).$$

- If f is of bounded variation and μ is absolutely continuous on $[a,b]$, then
$$\int_a^b f(t)d\mu(t) = \int_a^b f(t)\mu'(t)dt,$$
where the integral on the right is a Lebesgue integral.

- If f is continuous and μ is monotonically increasing, then there exists a $\xi \in [a,b]$ such that
$$\int_a^b f(t)d\mu(t) = f(\xi)(\mu(b) - \mu(a)).$$
This is a *Mean-value theorem.*

Lemma 2.4.1 *Let f be continuous and μ monotonically increasing on $[a,b]$. Let μ be differentiable at t. Then, it holds*
$$\frac{d}{dt}\int_a^t f(s)d\mu(s) = f(t)\frac{d}{dt}\mu(t).$$

Proof. Let μ be differentiable at t. Define
$$F(t) := \int_a^t f(s)d\mu(s).$$
The Mean-value theorem yields
$$F(t+h) - F(t) = \int_t^{t+h} f(s)d\mu(s) = f(\xi_h)(\mu(t+h) - \mu(t)), \quad t \leq \xi_h \leq t+h.$$
Hence,
$$\lim_{h\to 0}\frac{F(t+h)-F(t)}{h} = \lim_{h\to 0} f(\xi_h)\cdot\frac{\mu(t+h)-\mu(t)}{h} = f(t)\cdot\frac{d}{dt}\mu(t).$$
Observe, that $t \leq \xi_h \leq t+h \to t$ holds. ■

2.4.1 Continuous Functions

The space $C([a,b],\mathbb{R})$ of continuous functions endowed with the norm
$$\|f\|_\infty = \max_{a\leq t\leq b}|f(t)|$$
is a Banach space. We turn our attention to the representation of elements of the dual space of $C([a,b],\mathbb{R})$. Let μ be a function of bounded variation on $[a,b]$. Then, for every continuous function f on $[a,b]$ the number
$$\Phi(f) := \int_a^b f(t)d\mu(t)$$
exists. A closer investigation of the mapping $\Phi : C([a,b],\mathbb{R}) \to \mathbb{R}$ reveals, that Φ is a bounded, linear functional (and hence a continuous functional) on the space of continuous functions $C([a,b],\mathbb{R})$. Furthermore, all continuous linear functionals on $C([a,b],\mathbb{R})$ have the representation Φ, cf. Natanson [Nat75], p. 266. It holds

Theorem 2.4.2 (Riesz, cf. Natanson [Nat75], p. 266)
Let $\Phi : C([a,b],\mathbb{R}) \to \mathbb{R}$ be a linear, continuous functional. Then there exists a function μ of bounded variation on $[a,b]$ such that for every $f \in C([a,b],\mathbb{R})$ it holds
$$\Phi(f) = \int_a^b f(t)d\mu(t).$$

μ is defined a.e. in $[a,b]$ with exception of an additive constant, cf. Göpfert and Riedrich [GR80], p. 59, Luenberger [Lue69], p. 113.
The *normalized space of functions of bounded variations* $NBV([a,b],\mathbb{R})$ consists of all functions μ of bounded variation which are continuous from the right on (a,b) and satisfy $\mu(a)=0$. A norm is given by $\|\mu\| = TV(f,a,b)$. With this definition, the correspondence between the dual of $C([a,b],\mathbb{R})$ and $NBV([a,b],\mathbb{R})$ is unique.

2.5 Set Arithmetic

Let $S,T \subseteq X$ be sets and $\lambda \in \mathbb{R}$. The arithmetic operations '+', '−', and '·' are defined as

$$S \pm T := \{s \pm t \mid s \in S,\ t \in T\}, \qquad \lambda S := \{\lambda s \mid s \in S\}.$$

If S,T are convex, then $S+T$ and λS are also convex.
A set S is called a *cone with vertex* Θ, if $x \in S$ implies $\alpha x \in S$ for all $\alpha \geq 0$. If S is a cone with vertex Θ, then $x_0 + S$ is a cone with vertex at x_0. Let $S \subseteq X$ be a set and $x \in S$. The set

$$\operatorname{cone}(S,x) := \{\alpha(z-x) \mid \alpha \geq 0,\ z \in S\} = \mathbb{R}_+ \cdot (S - \{x\})$$

is called *conical hull of* $S - \{x\}$. If S is a cone with vertex at Θ, the conical hull of $S-\{x\}$ can be written as

$$\operatorname{cone}(S,x) = \{z - \alpha x \mid \alpha \geq 0,\ z \in S\} = S - \{\alpha x \mid \alpha \geq 0\} = S - \mathbb{R}_+ \cdot \{x\}.$$

If S is a convex cone with vertex at Θ then it holds

$$\operatorname{cone}(S,x) = S + \{\alpha x \mid \alpha \in \mathbb{R}\} = S + \mathbb{R} \cdot \{x\}.$$

To show the latter, notice that $\operatorname{cone}(S,x) \subseteq S + \mathbb{R}\cdot\{x\}$. Hence, it suffices to show that given an element $y \in S + \mathbb{R}\cdot\{x\}$ it can be written as $s - \alpha x$ with $s \in S$ and $\alpha \geq 0$. So, let $y = s_1 + \alpha_1 x$, $s_1 \in S$, $\alpha_1 \in \mathbb{R}$. If $\alpha_1 \leq 0$, we are ready. Hence, assume $\alpha_1 > 0$. From

$$y = s_1 + \alpha_1 x + x - x = s_1 + \underbrace{(1+\alpha_1)x}_{=:s_2 \in S} - x = 2\underbrace{\left(\frac{1}{2}s_1 + \frac{1}{2}s_2\right)}_{\in S} - x$$

it follows $y \in \operatorname{cone}(S,x)$.
Let $S \subseteq X$ be a set. The *positive dual cone of* S is defined as

$$S^+ := \{x^* \in X^* \mid x^*(x) \geq 0\ \forall x \in S\}.$$

The *negative dual cone of* S is defined as

$$S^- := \{x^* \in X^* \mid x^*(x) \leq 0\ \forall x \in S\}.$$

S is called *affine set*, if

$$(1-\lambda)x + \lambda y \in S \qquad \forall x,y \in S, \lambda \in \mathbb{R}.$$

The set

$$\begin{aligned}\operatorname{aff}(S) &:= \bigcap\{M \mid M \text{ is an affine set and } S \subseteq M\} \\ &= \left\{\sum_{i=1}^{m} \lambda_i x_i \mid m \in \mathbb{N},\ \sum_{i=1}^{m} \lambda_i = 1,\ x_i \in S,\ i=1,\ldots,m\right\}.\end{aligned}$$

is called *affine hull* of S. The set

$$\mathrm{relint}(S) \quad := \quad \{x \mid \exists \varepsilon > 0 : U_\varepsilon(x) \cap \mathrm{aff}(S) \subseteq S\}$$

is called *relative interior of* S. The set $\mathrm{cl}(S) \setminus \mathrm{relint}(S)$ is called *relative boundary of* S.
The following results for $X = \mathbb{R}^n$ can be found in Section 6 of Rockafellar [Roc70]. Let $S \subseteq \mathbb{R}^n$ be convex, $x \in \mathrm{relint}(S)$, and $y \in \mathrm{cl}(S)$. Then, $(1-\lambda)x + \lambda y \in \mathrm{relint}(S)$ for $0 \leq \lambda < 1$. If $S \subseteq \mathbb{R}^n$ is convex and $S \neq \emptyset$ then $\mathrm{relint}(S) \neq \emptyset$. Furthermore, the three convex sets $\mathrm{relint}(S), S$, and $\mathrm{cl}(S)$ have the same affine hull, the same relative interior and the same closure. Furthermore, $\mathrm{cl}(\mathrm{relint}(S)) = \mathrm{cl}(S)$ and $\mathrm{relint}(\mathrm{cl}(S)) = \mathrm{relint}(S)$. Let $A : \mathbb{R}^n \to \mathbb{R}^m$ be a linear transformation and $S \subseteq \mathbb{R}^n$ convex. Then $\mathrm{relint}(A(S)) = A(\mathrm{relint}(S))$ and $A(\mathrm{cl}(S)) \subseteq \mathrm{cl}(A(S))$. For any convex sets $S, T \subseteq \mathbb{R}^n$ it holds

$$\mathrm{relint}(S+T) = \mathrm{relint}(S) + \mathrm{relint}(T), \qquad \mathrm{cl}(S) + \mathrm{cl}(T) \subseteq \mathrm{cl}(S+T).$$

2.6 Separation Theorems

The proof of the necessary Fritz-John conditions is based on the separability of two convex sets by a hyperplane.
Let $f : X \to \mathbb{R}$ be a non-zero, linear functional and $\gamma \in \mathbb{R}$. Then the set

$$H := \{x \in X \mid f(x) = \gamma\}$$

is called a *hyperplane.* If f is continuous, the hyperplane is closed. The set

$$H_+ := \{x \in X \mid f(x) \geq \gamma\}$$

is called *positive half space of* H. Similarly,

$$H_- := \{x \in X \mid f(x) \leq \gamma\}$$

is the *negative half space of* H.
A hyperplane *separates* the sets A and B, if

$$\begin{aligned} f(x) &\leq \gamma, & \forall x \in A \\ f(x) &\geq \gamma, & \forall x \in B. \end{aligned}$$

A hyperplane *strictly separates* the sets A and B, if

$$\begin{aligned} f(x) &< \gamma, & \forall x \in A \\ f(x) &> \gamma, & \forall x \in B. \end{aligned}$$

A hyperplane H *properly separates* the sets A and B, if not both sets are contained in H.

Theorem 2.6.1 *Let $T : X \to Y$ be a linear and continuous operator and $R := im(T) \subset Y$ the image of T. Furthermore, assume that $y_0 \in Y$ is a point with $y_0 \notin R$. Then there exists a linear functional f with $f(r) = 0$ for $r \in R$ and $f(y_0) \neq 0$.*

Remark 2.6.2 *Notice, that f is not necessarily continuous. The hyperplane $\{y \mid f(y) = 0\}$ separates R and y_0.*

Proof. Consider the linear subspace $L := \{ty_0 \mid t \in \mathbb{R}\} \subset Y$. It holds $L \neq \emptyset$ and $L \neq Y$, since $L \not\subset R$. Hence, Y can be written as $Y = L \oplus S$ with a subspace $S \subset Y$. Define the functional $f : Y \to \mathbb{R}$ by

$$f(s + ty_0) := t.$$

Then, $f(s) = 0$ for $s \in S$ and $f(y_0) = 1$. Furthermore, f is linear because

$$\begin{aligned} f(\lambda(s+ty_0)) &= f(\underbrace{\lambda s}_{\in S} + \underbrace{\lambda t}_{\in \mathbb{R}} y_0) = \lambda t = \lambda f(s+ty_0), \\ f((s_1+t_1y_0)+(s_2+t_2y_0)) &= f(\underbrace{s_1+s_2}_{\in S} + \underbrace{(t_1+t_2)}_{\in \mathbb{R}} y_0) = t_1+t_2 \\ &= f(s_1+t_1y_0)+f(s_2+t_2y_0). \end{aligned}$$

This completes the proof. ■

If the subspace is closed, every point not contained in this subspace can be separated by a continuous non-zero linear functional.

Theorem 2.6.3 *Let X be a Banach space, M a closed subspace of X, and $\hat{x} \in X$, $\hat{x} \notin M$. Then there exists $f \in X^*$ with $f(\hat{x}) \neq 0$ and $f(x) = 0$ for all $x \in M$.*

Proof. cf. Werner [Wer95], Cor. III.1.8, p.98. ■

The following results are concerned with separation in $\mathbb{R}^n$ and can be found in Section 11 of Rockafellar [Roc70].

Theorem 2.6.4 *Let $A, B \subseteq \mathbb{R}^n$ be non-empty convex sets. There exists a hyperplane separating A and B properly, if and only if relint$(A) \cap$ relint$(B) = \emptyset$.*

Theorem 2.6.5 *Let $A, B \subseteq \mathbb{R}^n$ be non-empty convex sets. There exists a hyperplane separating A and B strictly, if and only if $\Theta_n \notin$ cl$(A - B)$, i.e. if*

$$\inf\{\|a - b\| \mid a \in A,\ b \in B\} > 0.$$

We are in order to list separation theorems for vector spaces, which can be found in Lempio [Lem71b, Lem71a]. Notice, that relint and int are to be understood in the purely algebraic sense and that the functional defining the hyperplane is not necessarily continuous.

Theorem 2.6.6 *Let A and B be non-empty convex subsets of a vector space X. If relint$(A) \neq \emptyset$, A has finite defect, and relint$(A) \cap B = \emptyset$, then there exists a non-zero linear functional f separating A and B.*

Theorem 2.6.7 *Let A and B be non-empty convex subsets of a vector space X with int$(A) \neq \emptyset$. Then there exists a non-zero linear functional f separating A and B if and only if int$(A) \cap B = \emptyset$.*

Theorem 2.6.8 *Let A and B be non-empty convex subsets of a vector space X with relint$(A) \neq \emptyset$ and relint$(B) \neq \emptyset$. Then there exists a non-zero linear functional f separating A and B if and only if either $A \cup B$ are contained in one hyperplane or relint$(A) \cap$ relint$(B) = \emptyset$.*

The subsequent separation theorems in Banach spaces can be found in Bonnans and Shapiro [BS00]. Here, the functional defining the hyperplane is continuous and hence an element of the dual space.

Theorem 2.6.9 (cf. Bonnans and Shapiro [BS00], Th. 2.13)
Let A and B be convex subsets of a Banach space X, where A has non-empty interior. Then there exists a non-zero functional $f \in X^*$ separating A and B if and only if $int(A) \cap B = \emptyset$.

Theorem 2.6.10 (cf. Bonnans and Shapiro [BS00], Th. 2.14)
Let A and B be closed convex subsets of a Banach space X, where A is compact and $A \cap B = \emptyset$. Then there exists a non-zero functional $f \in X^*$ separating A and B strictly.

Theorem 2.6.11 (cf. Bonnans and Shapiro [BS00], Th. 2.17)
Let A be a convex subset of a Banach space X with non-empty relative interior and $x_0 \in X$ a point with $x_0 \notin relint(A)$. Then there exists a non-zero functional $f \in X^*$ separating A and $\{x_0\}$.

The proofs exploit the Hahn-Banach theorem (cf. Werner [Wer95], Th. III.1.2, Luenberger [Lue69], p. 111).

Theorem 2.6.12 (Hahn-Banach)
Let X be a vector space, M a linear subspace of X, $f : M \to \mathbb{R}$ a linear functional on M, and $g : X \to \mathbb{R}$ a subadditive and positively homogeneous function, i.e. a function satisfying

$$\begin{aligned} g(x+y) &\leq g(x) + g(y), \\ g(\alpha x) &= \alpha g(x), \end{aligned}$$

for all $\alpha \geq 0$ and $x, y \in X$. Furthermore, let g majorize f on M, i.e.

$$f(x) \leq g(x) \qquad \forall x \in M.$$

Then there exists a linear functional $\hat{f} : X \to \mathbb{R}$ with

$$\begin{aligned} \hat{f}(x) &= f(x), & \forall x \in M, \\ \hat{f}(x) &\leq g(x), & \forall x \in X. \end{aligned}$$

As a consequence of the Hahn-Banach theorem it follows that continuous linear functionals defined on a linear subspace of X can be extended to the whole space X. This is because continuous linear functionals are bounded, i.e. it holds $|f(x)| \leq C\|x\|_X$ on M. The subadditive and positively homogeneous function $C\|\cdot\|_X$ plays the role of the function g in the Hahn-Banach theorem.

2.7 Derivatives

Definition 2.7.1

(i) *$F : X \to Y$ is called* directionally differentiable *at x in direction $h \in X$ if the limit*

$$F'(x;h) = \lim_{t \downarrow 0} \frac{1}{t} \left(F(x+th) - F(x) \right))$$

exists. $F'(x;h)$ is called directional derivative of F at x in direction h.

(ii) *$F : X \to Y$ is called* Gateaux-differentiable at x, *if there exists a continuous and linear operator $\delta F(x) : X \to Y$ with*

$$\lim_{t \downarrow 0} \frac{F(x+th) - F(x) - t\delta F(x)(h)}{t} = \Theta_Y$$

for all $h \in X$. The operator $\delta F(x)$ is called Gateaux differential of F at x.

(iii) $F : X \to Y$ *is called* Fréchet-differentiable at $x \in X$ *or* differentiable at $x \in X$, *if there exists a continuous linear operator* $F'(x) : X \to Y$ *with*

$$\lim_{\|h\|_X \to 0} \frac{F(x+h) - F(x) - F'(x)(h)}{\|h\|_X} = \Theta_Y$$

for all $h \in X$.

F *is called* regular at $x \in X$, *if* F *is Fréchet-differentiable and* $im(F'(x)) = Y$.

If $x \mapsto F'(x)$ *is continuous in the strong topology of* $\mathcal{L}(X, Y)$, *then* F *is called* continuously differentiable.

(iv) $F : X \times Y \to Z$ *is called* (partially) Fréchet-differentiable w.r.t. x at $(x^*, y^*) \in X \times Y$, *if* $F(\cdot, y^*)$ *is Fréchet-differentiable at* x^*. *We denote the partial derivative of* F *w.r.t.* x *at* (x^*, y^*) *by* $F'_x(x^*, y^*)$. *A similar definition holds for the component* y.

If $F : X \times Y \to Z$ is Fréchet-differentiable at (x^*, y^*), then F is also (partially) Fréchet-differentiable w.r.t. x and y and it holds

$$F'(x^*, y^*)(x, y) = F'_x(x^*, y^*)(x) + F'_y(x^*, y^*)(y)$$

for all $(x, y) \in X \times Y$.
The following statements can be found in Ioffe and Tihomirov [IT79].

- If $F : X \to Y$ is Gateaux-differentiable at x, then $F'(x; h) = \delta F(x)(h)$.
- If $F : X \to Y$ is Fréchet-differentiable at x, then F is continuous and Gateaux-differentiable at x with $F'(x)(h) = \delta F(x)(h)$.
- Let $F : X \to Y$ be continuous in a neighborhood of $x_0 \in X$ and let F be Gateaux-differentiable for every x in this neighborhood. Furthermore, let the mapping $x \mapsto \delta F(x)$ be continuous. Then F is Fréchet-differentiable in this neighborhood and $F'(x) = \delta F(x)$.
- Chain rule for $H = G \circ F$: It holds $H'(x) = G'(F(x)) \circ F'(x)$.
- Mean-value theorem (cf. Ioffe and Tihomirov [IT79], p. 27):

  Let X, Y be linear topological spaces, F Gateaux-differentiable for every point in $[x, x+h] \subseteq U$ and $z \mapsto F'(z)(h)$ continuous. Then

  $$F(x+h) - F(x) = \int_0^1 F'(x + th)(h)dt.$$

  If X, Y are Banach spaces, it holds

  $$F(x+h) - F(x) \leq \sup_{0 \leq t \leq 1} \|F'(x + th)\| \cdot \|h\|.$$

- Implicit function theorem (cf. Ioffe and Tihomirov [IT79], p. 29):

  Let X, Y, Z be Banach spaces and $F : U \subseteq X \times Y \to Z$ continuously differentiable. Let $F(x_0, y_0) = \Theta_Z$ and let the partial derivative $F'_y(x_0, y_0)$ be regular. Then there exists some neighborhood V of x_0 and a mapping $y : V \to Y$ with $y(x_0) = y_0$ and

  $$F(x, y(x)) = \Theta_Z \qquad \text{for all } x \in V.$$

  Moreover, $y(\cdot)$ is Fréchet-differentiable in V and

  $$y'(x) = -(F'_y(x, y(x)))^{-1}(F'_x(x, y(x))) \qquad \text{for all } x \in V.$$

2.8 Variational Equalities and Inequalities

We summarize some results on variational equalities and inequalities. These results are exploited during the proof of the local minimum principles for optimal control problems.

Lemma 2.8.1 *Let $f, g \in L^{\infty}([t_0, t_f], \mathbb{R})$, $s \in C([t_0, t_f], \mathbb{R})$, and $\mu \in BV([t_0, t_f], \mathbb{R})$. If*

$$\int_{t_0}^{t_f} f(t)h(t) + g(t)\dot{h}(t)dt + \int_{t_0}^{t_f} s(t)h(t)d\mu(t) = 0$$

holds for every $h \in W^{1,\infty}([t_0, t_f], \mathbb{R})$ with $h(t_0) = h(t_f) = 0$, then there exists a function $\hat{g} \in BV([t_0, t_f], \mathbb{R})$ such that $\hat{g}(t) = g(t)$ a.e. in $[t_0, t_f]$ and $\hat{g}(t) = C + \int_{t_0}^{t} f(\tau)d\tau + \int_{t_0}^{t} s(\tau)d\mu(\tau)$.

Proof. Define $F(t) := \int_{t_0}^{t} f(\tau)d\tau$. Then

$$\int_{t_0}^{t_f} f(t)h(t)dt = \int_{t_0}^{t_f} h(t)dF(t) = [F(t)h(t)]_{t_0}^{t_f} - \int_{t_0}^{t_f} F(t)dh(t) = -\int_{t_0}^{t_f} F(t)\dot{h}(t)dt,$$

since $h(t_0) = h(t_f) = 0$. Similarly, with $G(t) := \int_{t_0}^{t} s(\tau)d\mu(\tau)$ we find

$$\int_{t_0}^{t_f} s(t)h(t)d\mu(t) = \int_{t_0}^{t_f} h(t)dG(t) = [G(t)h(t)]_{t_0}^{t_f} - \int_{t_0}^{t_f} G(t)dh(t) = -\int_{t_0}^{t_f} G(t)\dot{h}(t)dt.$$

Furthermore, for any constant c and every function $h \in W^{1,\infty}([t_0, t_f], \mathbb{R})$ with $h(t_0) = h(t_f) = 0$, it holds

$$\int_{t_0}^{t_f} c\dot{h}(t)dt = 0.$$

Hence, we conclude that

$$\begin{aligned} 0 &= \int_{t_0}^{t_f} \left(f(t)h(t) + (g(t) + c)\dot{h}(t) \right) dt + \int_{t_0}^{t_f} s(t)h(t)d\mu(t) \\ &= \int_{t_0}^{t_f} \left(-F(t) - G(t) + g(t) + c \right) \dot{h}(t)dt \end{aligned}$$

holds for every $h \in W^{1,\infty}([t_0, t_f], \mathbb{R})$ with $h(t_0) = h(t_f) = 0$. Choose

$$\begin{aligned} c &= \frac{1}{t_f - t_0} \int_{t_0}^{t_f} \left(F(t) + G(t) - g(t) \right) dt, \\ h(t) &= \int_{t_0}^{t} \left(-F(\tau) - G(\tau) + g(\tau) + c \right) d\tau. \end{aligned}$$

Observe, that $h(t_0) = h(t_f) = 0$. Then,

$$\begin{aligned} 0 &= \int_{t_0}^{t_f} \left(f(t)h(t) + g(t)\dot{h}(t) \right) dt + \int_{t_0}^{t_f} s(t)h(t)d\mu(t) \\ &= \int_{t_0}^{t_f} \left(-F(t) - G(t) + g(t) + c \right) \dot{h}(t)dt \\ &= \int_{t_0}^{t_f} \left(-F(t) - G(t) + g(t) + c \right)^2 dt. \end{aligned}$$

This implies $-F(t) - G(t) + g(t) + c = 0$ a.e. in $[t_0, t_f]$ and thus $g(t) = F(t) + G(t) - c = \int_{t_0}^{t} f(\tau)d\tau + \int_{t_0}^{t} s(\tau)d\mu(\tau) - c$ a.e. in $[t_0, t_f]$. Setting $\hat{g}(t) = F(t) + G(t) - c$ yields the assertion. ■

Remark 2.8.2 *Replacing F, G, h, and c in the proof by $\int_t^{t_f} f(\tau)d\tau$, $\int_t^{t_f} s(\tau)d\mu(\tau)$, $\int_t^{t_f}(F(\tau) + G(\tau) + g(\tau) + c)d\tau$, and $\frac{1}{t_f - t_0}\int_{t_0}^{t_f}(-F(t) - G(t) - g(t))dt$, respectively, yields $g(t) = -F(t) - G(t) - c = -\int_t^{t_f} f(\tau)d\tau - \int_t^{t_f} s(\tau)d\mu(\tau) - c$.*

Setting $g(t) \equiv 0$ and $\mu \equiv 0$ in the previous result and exploiting $W^{1,\infty}([t_0, t_f], \mathbb{R}) \subset L^\infty([t_0, t_f], \mathbb{R})$, we immediately obtain the following special result.

Lemma 2.8.3 *Let $f \in L^\infty([t_0, t_f], \mathbb{R})$. If*

$$\int_{t_0}^{t_f} f(t)h(t)dt = 0$$

holds for every $h \in L^\infty([t_0, t_f], \mathbb{R})$, then $f(t) = 0$ a.e. in $[t_0, t_f]$.

Lemma 2.8.4 *Let $f \in L^1([t_0, t_f], \mathbb{R})$. If*

$$\int_{t_0}^{t_f} f(t)h(t)dt \geq 0$$

holds for every $h \in L^\infty([t_0, t_f], \mathbb{R})$ with $h(t) \geq 0$ a.e. in $[t_0, t_f]$, then $f(t) \geq 0$ a.e. in $[t_0, t_f]$.

Proof. Assume the contrary, i.e. there is a set $M \subseteq [t_0, t_f]$ with measure greater than 0 and $f(t) < 0$ for $t \in M$. Choose h to be one on M and zero otherwise. Then

$$\int_{t_0}^{t_f} f(t)h(t)dt = \int_M f(t)dt < 0$$

in contradiction to the assumption. ■

Lemma 2.8.5 *Let $\mu \in BV([t_0, t_f], \mathbb{R})$. If*

$$\int_{t_0}^{t_f} f(t)d\mu(t) \geq 0$$

holds for every $f \in C([t_0, t_f], \mathbb{R})$, then μ is non-decreasing in $[t_0, t_f]$.

Proof. According to the assumption of the Lemma, for every continuous function f it holds $S = \int_{t_0}^{t_f} f(t)d\mu(t) \geq 0$. Using the definition of the Riemann-Stieltjes integral, this is equivalent with the following. For every continuous function and every $\varepsilon > 0$ there exists $\delta > 0$ such that for every partition $\mathbb{G}_N = \{t_0 < t_1 < \ldots < t_N = t_f\}$ with $\max_{i=1,\ldots,N}\{t_i - t_{i-1}\} < \delta$ it holds

$$-\varepsilon \leq -\varepsilon + S < \sum_{i=1}^{N} f(\xi_i)\left(\mu(t_i) - \mu(t_{i-1})\right) < \varepsilon + S, \tag{2.8.1}$$

where ξ_i is an arbitrary point in $[t_{i-1}, t_i]$.
Now, assume that μ is not non-decreasing. Then there are points $\underline{t} < \overline{t}$ with $\mu(\underline{t}) = \mu(\overline{t}) + \gamma$, $\gamma > 0$. Let $\varepsilon := \gamma/2$ and $\delta > 0$ be arbitrary and $\mathbb{G}_N = \{t_0 < t_1 < \ldots < t_p := \underline{t} < \ldots < t_q := \overline{t} < \ldots < t_N = t_f\}$ with $\max_{i=1,\ldots,N}\{t_i - t_{i-1}\} < \delta$. Then there exists a continuous non-negative function f with $f(t) \equiv 1$ in $[\underline{t}, \overline{t}]$, $f(t_i) = 0$ for $i \notin \{p, \ldots, q\}$, and f linear in $[t_{p-1}, t_p]$ and $[t_q, t_{q+1}]$. Then,

$$\sum_{i=1}^{N} f(t_{i-1})\left(\mu(t_i) - \mu(t_{i-1})\right) = \mu(\overline{t}) - \mu(\underline{t}) = -\gamma < -\varepsilon.$$

This contradicts (2.8.1). ■

Lemma 2.8.6 *Let $\mu \in BV([t_0, t_f], \mathbb{R})$ be non-decreasing and $f \in C([t_0, t_f], \mathbb{R})$ non-positive. If*

$$\int_{t_0}^{t_f} f(t) d\mu(t) = 0$$

holds, then μ is constant on every interval $[a, b] \subseteq [t_0, t_f]$ with $a < b$ and $f(t) < 0$ in $[a, b]$.

Proof. Since $S = 0$ by assumption we find that for every $\varepsilon > 0$ there exists $\delta > 0$ such that for every partition $\mathbb{G}_N = \{t_0 < t_1 < \ldots < t_N = t_f\}$ with $\max_{i=1,\ldots,N}\{t_i - t_{i-1}\} < \delta$ it holds

$$-\varepsilon < \sum_{i=1}^{N} f(\xi_i) \left(\mu(t_i) - \mu(t_{i-1})\right) < \varepsilon, \tag{2.8.2}$$

where ξ_i is an arbitrary point in $[t_{i-1}, t_i]$.
Assume, that μ is not constant in intervals $[a, b]$ of the above type. Then, there exist points $\underline{t} < \overline{t}$ with $f(t) \leq -\alpha < 0$ for all $t \in [\underline{t}, \overline{t}]$ and $\mu(\underline{t}) = \mu(\overline{t}) - \gamma$, $\gamma > 0$ (because μ is non-decreasing). Choose $\varepsilon := \alpha\gamma/2$. Let $\delta > 0$ be arbitrary and $\mathbb{G}_N = \{t_0 < t_1 < \ldots < t_p = \underline{t} < \ldots < t_q = \overline{t} < \ldots < t_N = t_f\}$ with $\max_{i=1,\ldots,N}\{t_i - t_{i-1}\} < \delta$. Then,

$$\begin{aligned}
\sum_{i=1}^{N} f(\xi_{i-1}) \left(\mu(t_i) - \mu(t_{i-1})\right) &= \sum_{i=1}^{p} \underbrace{f(\xi_{i-1})}_{\leq 0} \underbrace{\left(\mu(t_i) - \mu(t_{i-1})\right)}_{\geq 0} \\
&\quad + \sum_{i=p+1}^{q} f(\xi_{i-1}) \left(\mu(t_i) - \mu(t_{i-1})\right) \\
&\quad + \sum_{i=q+1}^{N} \underbrace{f(\xi_{i-1})}_{\leq 0} \underbrace{\left(\mu(t_i) - \mu(t_{i-1})\right)}_{\geq 0} \\
&\leq \sum_{i=p+1}^{q} f(\xi_{i-1}) \left(\mu(t_i) - \mu(t_{i-1})\right) \\
&\leq -\alpha \sum_{i=p+1}^{q} \left(\mu(t_i) - \mu(t_{i-1})\right) \\
&= -\alpha \left(\mu(\overline{t}) - \mu(\underline{t})\right) \\
&= -\alpha\gamma < -\varepsilon.
\end{aligned}$$

This contradicts (2.8.2). ■

Chapter 3

Infinite and Finite Dimensional Optimization Problems

Throughout this chapter $(X, \|\cdot\|_X), (Y, \|\cdot\|_Y)$, and $(Z, \|\cdot\|_Z)$ denote Banach spaces over $\mathbb{K} = \mathbb{R}$. Without loss of generality we will exclusively consider minimization problems, since maximization problems always can be transformed into equivalent minimization problems. Necessary conditions for certain classes of optimization problems are summarized.

3.1 Problem Classes

We investigate general nonlinear optimization problems of the form given in Problem 3.1.1 below.

Problem 3.1.1 (General Optimization Problem)
Let $f : X \to \mathbb{R}$ be a functional, and $\emptyset \neq \Sigma \subseteq X$ a set. Find $\hat{x} \in \Sigma$ such that $f(x)$ is minimized subject to the constraint $x \in \Sigma$.

The following terminology is used.

- f is called *objective function.*
- A vector x is called *admissible* or *feasible* for Problem 3.1.1, if $x \in \Sigma$. Σ is called *admissible set* or *feasible set* for Problem 3.1.1.
- $\hat{x} \in \Sigma$ is called *global minimum* of Problem 3.1.1, if
$$f(\hat{x}) \leq f(x) \quad \forall x \in \Sigma. \tag{3.1.1}$$
$\hat{x} \in X$ is called *strict global minimum* of Problem 3.1.1, if '$<$' holds in (3.1.1) for all $x \in \Sigma$, $x \neq \hat{x}$.
- $\hat{x} \in \Sigma$ is called *local minimum* of Problem 3.1.1, if there exists a $\varepsilon > 0$ such that
$$f(\hat{x}) \leq f(x) \quad \forall x \in \Sigma \cap U_\varepsilon(\hat{x}). \tag{3.1.2}$$
$\hat{x} \in \Sigma$ is called *strict local minimum* of Problem 3.1.1, if '$<$' holds in (3.1.2) for all $x \in \Sigma \cap U_\varepsilon(\hat{x})$, $x \neq \hat{x}$.
- Problem 3.1.1 is called *unconstrained*, if $\Sigma = X$ holds.
- Problem 3.1.1 is called *convex*, if f and Σ are convex.

In convex optimization problems every local minimum is also a global one. For, assume that $\hat{x}$ is a local minimum but not a global minimum. Then there exists $x \in \Sigma$ with $f(x) < f(\hat{x})$. The convexity of f yields

$$f(\alpha x + (1-\alpha)\hat{x}) \leq \alpha f(x) + (1-\alpha) f(\hat{x}) < f(\hat{x}) \qquad \forall 0 \leq \alpha \leq 1,$$

which contradicts the local minimality of $\hat{x}$.
The subsequent special case of Problem 3.1.1 is of importance in practical applications, in deriving numerical algorithms, or in the statement of necessary (and sufficient) conditions. The set Σ is described by inequalities $g(x) \in K$ and equalities $h(x) = \Theta_Z$.

Problem 3.1.2 (Standard Nonlinear Optimization Problem)
Let $f : X \to \mathbb{R}$ be a functional, $g : X \to Y$, $h : X \to Z$ operators, $S \subseteq X$ a closed, convex set, and $K \subseteq Y$ a closed convex cone with vertex at Θ_Y. Find $\hat{x}$ such that $f(x)$ is minimized subject to the constraints $x \in S$, $g(x) \in K$, and $h(x) = \Theta_Z$.

Notice, that the admissible set in Problem 3.1.2 is given by

$$\Sigma = S \cap g^{-1}(K) \cap h^{-1}(\Theta_Z),$$

where $g^{-1}(K) := \{x \in X \mid g(x) \in K\}$ denotes the preimage of K under g and $h^{-1}(\Theta_Z) := \{x \in X \mid h(x) = \Theta_Z\}$ is the preimage of Θ_Z under h.
In order to be able to derive necessary conditions for a local minimum of Problem 3.1.2 the set $S \subseteq X$ usually cannot be an arbitrary set but has to fulfill additional conditions, e.g. S has to be closed and convex with nonempty interior. Nevertheless, there are also practically important problems, e.g. optimal control problems or mixed integer optimization problems, where S is a discrete set. Fortunately, it is possible to derive necessary conditions for such problems as well, but among other things additional convexity or differentiability conditions have to be imposed on the remaining constraints and the objective function. Application of these necessary conditions to optimal control problems will result in the famous maximum principle for optimal control problems.

3.2 Existence of a Solution

The existence of a solution for the optimization problem 3.1.1 where $\Sigma \subseteq X$ is a compact set and f is lower semicontinuous, is ensured by the famous Weierstrass Theorem:

Theorem 3.2.1 (Weierstrass)
Let Σ be a compact subset of a normed vector space X and let $f : \Sigma \to \mathbb{R}$ be lower semicontinuous. Then, f achieves its minimum on Σ.

Proof. (cf. Luenberger [Lue69], p. 40)
Assume, that f is not bounded from below on Σ. Then, there is a sequence $\{x_i\}$ in Σ with $f(x_i) \leq -i$. Since Σ is compact, there exists a convergent subsequence $\lim_{k\to\infty} x_{i_k} = \hat{x}$ with $f(x_{i_k}) \leq -i_k$ for all $k \in \mathbb{N}$. Since f is lower semicontinuous it follows $f(\hat{x}) \leq \liminf_{k\to\infty} f(x_{i_k})$. Hence, $f(x_{i_k})$ is bounded from below by $f(\hat{x}) \in \mathbb{R}$. This contradicts $f(x_{i_k}) \leq -i_k \to -\infty$. This shows that f is bounded from below on Σ.
This in turn implies that $\hat{f} = \inf_{x\in\Sigma} f(x)$ is a real number and for any $i \in \mathbb{N}$ there exists a $x_i \in \Sigma$ with $f(x_i) \leq \hat{f} + \frac{1}{i}$. Since Σ is compact, there exists a convergent subsequence $x_{i_k} \to \hat{x}$ with $f(x_{i_k}) \leq \hat{f} + \frac{1}{i_k}$ for all $k \in \mathbb{N}$. Since f is lower semicontinuous it follows $\hat{f} \leq f(\hat{x}) \leq \liminf_{k\to\infty} f(x_{i_k}) \leq \hat{f}$. Hence, f assumes its minimum on Σ. ■

A generalization is given by, cf. Alt [Alt02],

Theorem 3.2.2
Let $\Sigma \subseteq X$ and let $f : \Sigma \to \mathbb{R}$ be a lower semicontinuous function on Σ. Let the set

$$lev(f, f(w)) \cap \Sigma = \{x \in \Sigma \mid f(x) \leq f(w)\}$$

be nonempty and compact for some $w \in \Sigma$. Then, f achieves its minimum on Σ.

Proof. According to the Theorem of Weierstrass there exists $\hat{x} \in lev(f, f(w)) \cap \Sigma$ with $f(\hat{x}) \leq f(x)$ for all $x \in lev(f, f(w)) \cap \Sigma$. For $x \in \Sigma \backslash (lev(f, f(w)) \cap \Sigma) = \Sigma \backslash lev(f, f(w))$ it holds $f(x) > f(w) \geq f(\hat{x})$. Hence, $\hat{x}$ is a minimum of f on Σ. ■

3.3 Conical Approximation of Sets

Conical approximations to sets play an important role in the formulation of necessary conditions for constrained optimization problems. We will summarize some important cones. Throughout this section, Σ denotes a non-empty set in X.
The *(sequential) tangent cone to* Σ *at* $\hat{x} \in \Sigma$ is given by

$$T(\Sigma, x) = \left\{ d \in X \;\middle|\; \begin{array}{l} \text{there exist sequences } \{\alpha_k\}_{k\in\mathbb{N}}, \alpha_k \downarrow 0 \text{ and} \\ \{x_k\}_{k\in\mathbb{N}}, x_k \in \Sigma \text{ with } \lim_{k\to\infty} x_k = x, \text{ such that} \\ \lim_{k\to\infty}(x_k - x)/\alpha_k = d \text{ holds.} \end{array} \right\}. \tag{3.3.1}$$

Figure 3.1 illustrates the tangent cone in several situations.

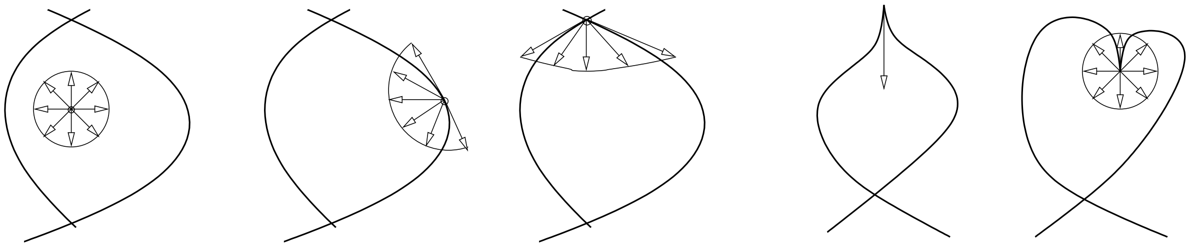


Figure 3.1: Tangent cones to different sets.

The tangent cone is a closed cone with vertex at zero. If x happens to be an interior point of Σ, then $T(\Sigma, x) = X$. If Σ is a convex set, then the tangent cone is convex and can be written as

$$\begin{aligned} T(\Sigma, x) &= \{d \in X \mid \exists \alpha_k \downarrow 0,\ d_k \to d \ :\ x + \alpha_k d_k \in \Sigma\} \\ &= \overline{\{d \in X \mid \exists \alpha > 0 \ :\ x + \alpha d \in \Sigma\}}. \end{aligned}$$

It is easy to show, cf. Kirsch et al. [KWW78], p. 31, that the tangent cone can be written as

$$T(\Sigma, x) = \left\{ d \in X \;\middle|\; \begin{array}{l} \text{there exist } \sigma > 0 \text{ and a mapping } r : (0, \sigma] \to X \text{ with} \\ \lim_{\varepsilon\downarrow 0} \frac{r(\varepsilon)}{\varepsilon} = \Theta_X \text{ and a sequence } \{\alpha_k\}_{k\in\mathbb{N}}, \alpha_k \downarrow 0 \text{ with} \\ x + \alpha_k d + r(\alpha_k) \in \Sigma \text{ for all } k \in \mathbb{N}. \end{array} \right\}.$$

Theorem 3.3.1 (Ljusternik)
Let X, Z be Banach spaces, $h : X \to Z$ a mapping, and $\hat{x} \in M := \{x \in X \mid h(x) = \Theta_Z\}$. Let h be continuous in a neighborhood of $\hat{x}$ and continuously Gateaux-differentiable at $\hat{x}$. Let the Gateaux-derivative $\delta h(\hat{x})$ be surjective. Let $\hat{d} \in X$ be given with $\delta h(\hat{x})(\hat{d}) = \Theta_Z$. Then, there exist $\varepsilon_0 > 0$ and a mapping

$$r : (0, \varepsilon_0] \to X, \qquad \lim_{\varepsilon\downarrow 0} \frac{r(\varepsilon)}{\varepsilon} = \Theta_X$$

such that

$$h(\hat{x} + \varepsilon\hat{d} + r(\varepsilon)) = \Theta_Z$$

holds for every $\varepsilon \in (0, \varepsilon_0]$. In particular, it holds

$$\{d \in X \mid \delta h(\hat{x})(d) = \Theta_Z\} = T(M, \hat{x}).$$

Proof. cf. Kirsch et al. [KWW78], p. 40 ■

Recall, that given a cone $C \subseteq X$ with vertex at zero, the positive and negative dual cones are given by

$$\begin{aligned} C^+ &:= \{x^* \in X^* \mid x^*(x) \geq 0 \text{ for all } x \in C\}, \\ C^- &:= \{x^* \in X^* \mid x^*(x) \leq 0 \text{ for all } x \in C\}, \end{aligned}$$

respectively. Sometimes these cones are also called positive and negative polar cones or normal cones or conjugate cones, respectively. Functionals from C^+ are called positive on C, and functionals from C^- are called negative on C. C^+ and C^- are non-empty closed convex cones.

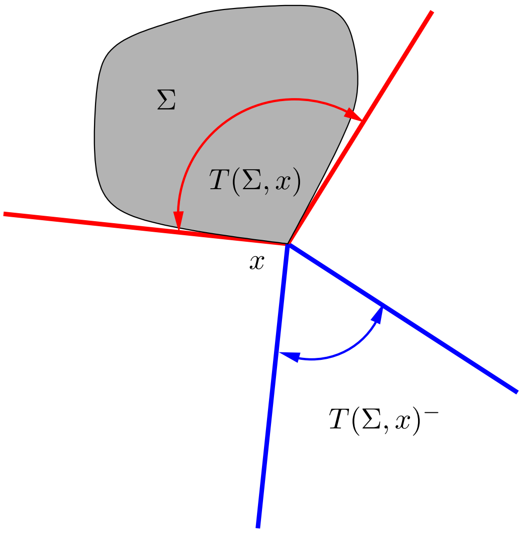


Figure 3.2: Tangent cone to Σ and its negative dual cone in $X = \mathbb{R}^n$.

The upcoming necessary conditions involve a conic approximation of the set S and a linearization of the constraint $g(x) \in K$ in Problem 3.1.2.

Let g be Fréchet-differentiable. The *linearizing cone of K and S at x* is given by

$$T_{lin}(K, S, x) := \{d \in \text{cone}(S, x) \mid g'(x)(d) \in \text{cone}(K, g(x)),\ h'(x)(d) = \Theta_Z\}.$$

A relation between tangent cone and linearizing cone is given by

Corollary 3.3.2 *Let g and h be Fréchet-differentiable. Let cone(S, x) and cone$(K, g(x))$ be closed. Then it holds $T(\Sigma, x) \subseteq T_{lin}(K, S, x)$.*

Proof. Let $d \in T(\Sigma, x)$. Then, there are sequences $\varepsilon_k \downarrow 0$ and $x_k \to x$ with $x_k \in S, g(x_k) \in K, h(x_k) = \Theta_Z$ and $(x_k - x)/\varepsilon_k \to d$. Since $(x_k - x)/\varepsilon_k \in \text{cone}(S, x)$ and cone(S, x) is closed it holds $d \in \text{cone}(S, x)$. The continuity of $h'(x)(\cdot)$ implies

$$\Theta_Z = \lim_{k \to \infty} h(x)((x_k - x)/\varepsilon_k) = h(x)(d).$$

Furthermore, since g is Fréchet-differentiable it holds

$$g(x_k) = g(x) + g'(x)(x_k - x) + \alpha_k \|x_k - x\|_X$$

with some sequence $\{\alpha_k\}_{k\in\mathbb{N}} \subseteq Y$, $\alpha_k \to \Theta_Y$. Hence,

$$\underbrace{\frac{g(x_k)-g(x)}{\varepsilon_k}}_{\in \text{cone}(K,g(x))} = g'(x)\left(\frac{x_k - x}{\varepsilon_k}\right) + \alpha_k \left\|\frac{x_k - x}{\varepsilon_k}\right\|_X.$$

Since $\text{cone}(K, g(x))$ is closed, the term on the left converges to an element in $\text{cone}(K, g(x))$, while the term on the right converges to $g'(x)(d)$. Thus, $g'(x)(d) \in \text{cone}(K, g(x))$. Together with $d \in \text{cone}(S, x)$ and $h'(x)(d) = \Theta_Z$ this shows $d \in T_{lin}(K, S, x)$. ■

3.4 First Order Necessary Conditions of Fritz-John Type

First, we will discuss a geometrically motivated first order necessary condition for Problem 3.1.1, which involves the tangent cone. It holds

Theorem 3.4.1 *Let $f : X \to \mathbb{R}$ be Fréchet-differentiable at $\hat{x}$ and let $\hat{x}$ be a local minimum of Problem 3.1.1. Then*

$$f'(\hat{x})(d) \geq 0 \qquad \forall d \in T(\Sigma, \hat{x}).$$

Proof. Let $d \in T(\Sigma, \hat{x})$. Then there are sequences $\alpha_k \downarrow 0$, $x_k \to \hat{x}$, $x_k \in \Sigma$ with $d = \lim_{k\to\infty}(x_k - \hat{x})/\alpha_k$. Since $\hat{x}$ is a local minimum and f is Fréchet-differentiable at $\hat{x}$ it follows

$$0 \leq f(x_k) - f(\hat{x}) = f'(\hat{x})(x_k - \hat{x}) + o(\|x_k - \hat{x}\|_X).$$

Division by $\alpha_k > 0$ yields

$$0 \leq f'(\hat{x})\underbrace{\left(\frac{x_k - \hat{x}}{\alpha_k}\right)}_{\to d} + \underbrace{\left\|\frac{x_k - \hat{x}}{\alpha_k}\right\|_X}_{\to \|d\|_X} \cdot \underbrace{\frac{o(\|x_k - \hat{x}\|_X)}{\|x_k - \hat{x}\|_X}}_{\to 0}.$$

■

The subsequent theorem stating necessary conditions for Problem 3.1.2 provides the basis for the minimum principle for optimal control problems and can be found in Lempio [Lem72].

Theorem 3.4.2 (First Order Necessary Conditions)
Let $f : X \to \mathbb{R}$ and $g : X \to Y$ be Fréchet-differentiable and $h : X \to Z$ continuously Fréchet-differentiable. Let $\hat{x}$ be a local minimum of Problem 3.1.2, int$(S) \neq \emptyset$, and int$(K) \neq \emptyset$. Assume that im$(h'(\hat{x}))$ is not a proper dense subset of Z. Then there exist nontrivial multipliers $(l_0, \lambda^*, \mu^*) \in \mathbb{R} \times Y^* \times Z^*$, $(l_0, \lambda^*, \mu^*) \neq (0, \Theta_{Y^*}, \Theta_{Z^*})$ such that

$$\begin{aligned} l_0 &\geq 0, && (3.4.1)\\ \lambda^* &\in K^+, && (3.4.2)\\ \lambda^*(g(\hat{x})) &= 0, && (3.4.3)\\ l_0 f'(\hat{x})(d) - \lambda^*(g'(\hat{x})(d)) - \mu^*(h'(\hat{x})(d)) &\geq 0, \quad \forall d \in S - \{\hat{x}\}. && (3.4.4) \end{aligned}$$

Proof. Consider the linearized problem

$$\begin{array}{ll} \text{Minimize} & f(\hat{x}) + f'(\hat{x})(x - \hat{x}) \\ \text{w.r.t.} & x \in S \\ \text{subject to} & g(\hat{x}) + g'(\hat{x})(x - \hat{x}) \in K, \\ & h(\hat{x}) + h'(\hat{x})(x - \hat{x}) = \Theta_Z. \end{array}$$

For the first part of the proof we assume that the mapping $h'(\hat{x})(\cdot) : X \to Z$ is surjective and that there exists a feasible $x \in \text{int}(S)$ for the linearized problem satisfying

$$\begin{aligned} g(\hat{x}) + g'(\hat{x})(x - \hat{x}) &\in \text{int}(K), &(3.4.5)\\ h(\hat{x}) + h'(\hat{x})(x - \hat{x}) &= \Theta_Z, &(3.4.6)\\ f(\hat{x}) + f'(\hat{x})(x - \hat{x}) &< f(\hat{x}). &(3.4.7) \end{aligned}$$

(i) Since $h'(\hat{x})$ is surjective and $h'(\hat{x})(x - \hat{x}) = \Theta_Z$ holds, we may apply Ljusternik's theorem. Hence, there exist $t_0 > 0$ and a mapping

$$r : [0, t_0] \to X, \qquad \lim_{t \downarrow 0} \frac{r(t)}{t} = \Theta_X, \; r(0) := \Theta_X,$$

such that

$$h(\hat{x} + t(x - \hat{x}) + r(t)) = \Theta_Z$$

holds for every $t \in [0, t_0]$. This means, that the curve

$$x(t) = \hat{x} + t(x - \hat{x}) + r(t)$$

remains feasible for the nonlinear equality constraints for every $t \in [0, t_0]$. Furthermore, it holds

$$x(0) = \hat{x}, \quad x'(0) = x - \hat{x}.$$

The latter holds because

$$x'(0) = \lim_{t \downarrow 0} \frac{x(t) - x(0)}{t} = \lim_{t \downarrow 0} \frac{x(t) - \hat{x}}{t} = x - \hat{x} + \lim_{t \downarrow 0} \frac{r(t)}{t} = x - \hat{x}.$$

(ii) Now, we consider the inequality constraints at $\hat{x}$. Since g is Fréchet-differentiable at $\hat{x}$ it holds

$$\begin{aligned} g(x(t)) &= g(\hat{x}) + tg'(\hat{x})(x - \hat{x}) + o(t)\\ &= t(\underbrace{g(\hat{x}) + g'(\hat{x})(x - \hat{x})}_{\in \text{int}(K)}) + (1 - t)\underbrace{g(\hat{x})}_{\in K} + o(t). \end{aligned}$$

Since $g(\hat{x}) + g'(\hat{x})(x - \hat{x}) \in \text{int}(K)$ there exists $\delta_1 > 0$ such that

$$U_{\delta_1}(g(\hat{x}) + g'(\hat{x})(x - \hat{x})) \subseteq K.$$

Furthermore, the convexity of K yields

$$(1 - t)g(\hat{x}) + ty \in K$$

for all $y \in U_{\delta_1}(g(\hat{x}) + g'(\hat{x})(x - \hat{x}))$ and all $0 \le t \le 1$. Since for $t \downarrow 0$ it holds $o(t)/t \to 0$ there exists $\delta_2 > 0$ with $\|o(t)/t\|_Y < \delta_1$ for all $0 < t < \delta_2$. Hence,

$$g(x(t)) = t(\underbrace{g(\hat{x}) + g'(\hat{x})(x - \hat{x}) + o(t)/t}_{\in U_{\delta_1}(g(\hat{x}) + g'(\hat{x})(x - \hat{x}))}) + (1 - t)g(\hat{x}) \in K.$$

Hence, for sufficiently small $t > 0$ the curve $x(t)$ stays feasible for the nonlinear inequality constraints.

(iii) Now, the objective function is investigated. It holds

$$\frac{d}{dt} f(x(t))\Big|_{t=0} = f'(\hat{x}) \cdot \frac{dx}{dt}(0) = f'(\hat{x})(x - \hat{x}) \overset{(3.4.7)}{<} 0.$$

Hence, $x - \hat{x}$ is a direction of descent of f, i.e. it holds $f(x(t)) < f(\hat{x})$ for $t > 0$ sufficiently small. (this will contradict the local minimality of $\hat{x}$ for Problem 3.1.2).

(iv) The point x in the linear problem fulfilling (3.4.5)-(3.4.7) is assumed to be an interior point of S, cf. Figure 3.3.

Then, since S is assumed to be convex, there exists a neighborhood $U_\delta(x)$ such that $\hat{x} + t(z - \hat{x}) \in S$ for all $0 \leq t \leq 1$ and all $z \in U_\delta(x)$. Since

$$\lim_{t \downarrow 0} \frac{r(t)}{t} = \Theta_X$$

there exists $\varepsilon > 0$ with

$$\left\| \frac{r(t)}{t} \right\|_X < \delta$$

for all $0 < t < \varepsilon$. Hence,

$$x(t) = \hat{x} + t(x - \hat{x}) + r(t) = \hat{x} + t\Big(\underbrace{x + \frac{r(t)}{t}}_{\in U_\delta(x)} - \hat{x} \Big) \in S$$

for $0 < t < \varepsilon$.

Items (i)-(iv) showed the following: If $\hat{x}$ is a local minimum of Problem 3.1.2 and $h'(\hat{x})$ is surjective, then

$$f'(\hat{x})(d) \geq 0 \quad \forall d \in T(\hat{x}) := \left\{ d \in \operatorname{int}(\operatorname{cone}(S, \hat{x})) \;\middle|\; \begin{array}{l} g'(\hat{x})(d) \in \operatorname{int}(\operatorname{cone}(K, g(\hat{x}))), \\ h'(\hat{x})(d) = \Theta_Z \end{array} \right\}. \tag{3.4.8}$$

Consider the non-empty convex set

$$A := \left\{ \begin{pmatrix} f'(\hat{x})(d) + r \\ g'(\hat{x})(d) - k \\ h'(\hat{x})(d) \end{pmatrix} \;\middle|\; d \in \operatorname{int}(\operatorname{cone}(S, \hat{x})),\ k \in \operatorname{int}(\operatorname{cone}(K, g(\hat{x}))),\ r > 0 \right\}.$$

If $h'(\hat{x})$ is not surjective and $\operatorname{im}(h'(\hat{x}))$ is not a proper dense subset of Z, the set

$$M := cl \left(\left\{ \begin{pmatrix} r \\ y \\ h'(\hat{x})(x) \end{pmatrix} \;\middle|\; r \in \mathbb{R},\ y \in Y,\ x \in X \right\} \right)$$

is a proper closed subspace of $\mathbb{R} \times Y \times Z$. According to Theorem 2.6.3 there is a non-zero functional $\lambda \in (\mathbb{R} \times Y \times Z)^*$ with $\lambda(r, y, z) = 0$ for all $(r, y, z) \in M$. Hence, the hyperplane

$$\{(r, y, z) \in \mathbb{R} \times Y \times Z \mid \lambda(r, y, z) = 0\}$$

trivially separates the sets A and the point $(0, \Theta_Y, \Theta_Z)$, since both are contained in M.

A can be decomposed into $A = A_1 + A_2$ with

$$\begin{aligned} A_1 &:= \left\{ \begin{pmatrix} f'(\hat{x})(d) \\ g'(\hat{x})(d) \\ h'(\hat{x})(d) \end{pmatrix} \;\middle|\; d \in \text{int}(\text{cone}(S, \hat{x})) \right\}, \\ A_2 &:= \left\{ \begin{pmatrix} r \\ -k \\ \Theta_Z \end{pmatrix} \;\middle|\; k \in \text{int}(\text{cone}(K, g(\hat{x}))),\ r > 0 \right\}. \end{aligned}$$

If $h'(\hat{x})$ is surjective, the projection of A_1 onto Z contains interior points in Z according to the open mapping theorem. Hence, $A_1 + A_2$ contains interior points in $\mathbb{R} \times Y \times Z$. The considerations in (i)-(iv) showed that $(0, \Theta_Y, \Theta_Z) \notin \text{int}(A)$. For, let us assume $(0, \Theta_Y, \Theta_Z) \in \text{int}(A)$. Then there exists $d \in \text{int}(\text{cone}(S, \hat{x}))$ with $f'(\hat{x})(d) < 0$, $g'(\hat{x})(d) \in \text{int}(\text{cone}(K, g(\hat{x})))$, and $h'(\hat{x})(d) = \Theta_Z$ contradicting (3.4.8).

According to the separation theorem 2.6.11 there exists a hyperplane separating A and $(0, \Theta_Y, \Theta_Z)$, i.e. there exist multipliers

$$(0, \Theta_{Y^*}, \Theta_{Z^*}) \neq (l_0, \lambda^*, \mu^*) \in \mathbb{R} \times Y^* \times Z^* = (\mathbb{R} \times Y \times Z)^*$$

such that

$$0 \;\leq\; l_0 \left(f'(\hat{x})(d) + r \right) - \lambda^* \left(g'(\hat{x})(d) - y \right) - \mu^* \left(h'(\hat{x})(d) \right)$$

for all $d \in \text{int}(\text{cone}(S, \hat{x}))$, $y \in \text{int}(\text{cone}(K, g(\hat{x})))$, $r > 0$. Owing to the continuity of the functionals $l_0(\cdot)$, $\lambda^*(\cdot)$, $\mu^*(\cdot)$ and the linear operators $f'(\hat{x})(\cdot)$, $g'(\hat{x})(\cdot)$, and $h'(\hat{x})(\cdot)$ this inequality also holds for all $d \in \text{cone}(S, \hat{x})$, $y \in \text{cone}(K, g(\hat{x}))$, $r \geq 0$.

Choosing $d = \Theta_X \in \text{cone}(S, \hat{x})$ and $y = \Theta_Y \in \text{cone}(K, g(\hat{x}))$ yields $l_0 r \geq 0$ for all $r \geq 0$ and thus $l_0 \geq 0$. The choices $d = \Theta_X$ and $r = 0$ imply $\lambda^*(y) \geq 0$ for all $y \in \text{cone}(K, g(\hat{x})) = \{k - \alpha g(\hat{x}) \mid k \in K,\ \alpha \geq 0\}$, i.e. $\lambda^*(k - \alpha g(\hat{x})) \geq 0$ for all $k \in K$, $\alpha \geq 0$. This in turn implies $\lambda^* \in K^+$ (choose $\alpha = 0$) and $\lambda^*(g(\hat{x})) = 0$ (choose $k = \Theta_Y$, $\alpha = 1$ and observe that $g(\hat{x}) \in K$). ■

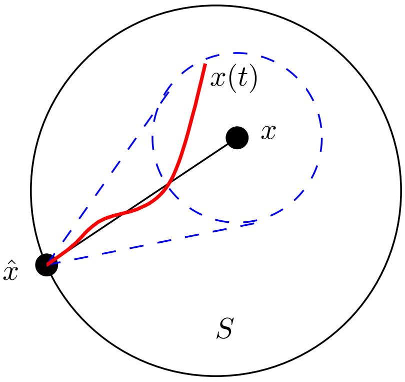


Figure 3.3: Fritz-John conditions under set constraints $x \in S$. The assumption $\text{int}(S) \neq \emptyset$ is essential. For sufficiently small $t > 0$ the curve $x(t)$ stays feasible.

Every point $(x, l_0, \lambda^*, \mu^*) \in X \times \mathbb{R} \times Y^* \times Z^*$, $(l_0, \lambda^*, \mu^*) \neq \Theta$ satisfying the Fritz-John conditions (3.4.1)-(3.4.4) is called *Fritz-John point* of Problem 3.1.2. The multipliers l_0, λ^*, and μ^* are

called *Lagrange multipliers* or simply *multipliers*. The main statement of the theorem is that there exists a *nontrivial* vector $(l_0, \lambda^*, \mu^*) \neq \Theta$. Notice, that $(l_0, \lambda^*, \mu^*) = \Theta$ trivially fulfills the Fritz-John conditions. Unfortunately, the case $l_0 = 0$ may occur. In this case, the objective function f does not enter into in the Fritz-John conditions. In case of $l_0 \neq 0$ we call the Fritz-John point $(x, l_0, \lambda^*, \mu^*)$ *Karush-Kuhn-Tucker (KKT) point.*
First order necessary conditions of Fritz-John type for general cone constrained problems

$$f(x) \to \min \quad \text{s.t.} \quad x \in S,\ g(x) \in K$$

can be found in Kurcyusz [Kur76]. Essentially, the existence of non-trivial multipliers (defining a separating hyperplane) can be guaranteed, if

$$\text{im}(g'(\hat{x})) + \text{cone}(K, g(\hat{x}))$$

is not a proper dense subset of Y.

3.5 Constraint Qualifications

Conditions which ensure that the multiplier l_0 in Theorem 3.4.2 is not zero are called *regularity conditions* or *constraint qualifications.* In this case, without loss of generality l_0 can be normalized to one due to the linearity in the multipliers.
The following regularity condition was postulated by Robinson [Rob76] in the context of stability analysis for generalized inequalities.

Definition 3.5.1 (Regularity condition of Robinson [Rob76])
The regularity condition of Robinson *is satisfied at $\hat{x}$ if*

$$\begin{pmatrix} \Theta_Y \\ \Theta_Z \end{pmatrix} \in int \left\{ \begin{pmatrix} g(\hat{x}) + g'(\hat{x})(x - \hat{x}) - k \\ h'(\hat{x})(x - \hat{x}) \end{pmatrix} \middle|\ x \in S,\ k \in K \right\}. \tag{3.5.1}$$

The validity of the regularity condition ensures that the multiplier l_0 is not zero.

Theorem 3.5.2 (KKT-conditions)
Let the assumptions of Theorem 3.4.2 be satisfied. In addition let the regularity condition of Robinson be satisfied at $\hat{x}$. Then the assertions of Theorem 3.4.2 hold with $l_0 = 1$.

Proof. Let us assume that the necessary conditions are valid with $l_0 = 0$, i.e.

$$\begin{aligned} \lambda^*(k) &\geq 0, && \forall k \in K, \\ \lambda^*(g(\hat{x})) &= 0, && \\ \lambda^*(g'(\hat{x})(d)) + \mu^*(h'(\hat{x})(d)) &\leq 0, && \forall d \in S - \{\hat{x}\}. \end{aligned}$$

These conditions imply

$$\lambda^*(g(\hat{x}) + g'(\hat{x})(d) - k) + \mu^*(h'(\hat{x})(d)) \leq 0, \qquad \forall d \in S - \{\hat{x}\},\ k \in K.$$

Hence, the functional $\eta^*(y, z) := \lambda^*(y) + \mu^*(z)$ separates (Θ_Y, Θ_Z) from the set

$$\left\{ \begin{pmatrix} g(\hat{x}) + g'(\hat{x})(x - \hat{x}) - k \\ h'(\hat{x})(x - \hat{x}) \end{pmatrix} \middle|\ x \in S,\ k \in K \right\}$$

since $\eta^*(\Theta_Y, \Theta_Z) = 0$. But, (Θ_Y, Θ_Z) is assumed to be an interior point of this set. Hence, a separation is impossible and the assumption $l_0 = 0$ was wrong. ■

Zowe and Kurcyusz [ZK79] show that the regularity condition of Robinson is stable under perturbations of the constraints and ensures the boundedness of the multipliers.
A sufficient condition for the regularity condition of Robinson is the surjectivity constraint qualification.

Corollary 3.5.3 *Let $\hat{x} \in int(S)$ and let the operator*

$$T : X \to Y \times Z, \qquad T := (g'(\hat{x}), h'(\hat{x}))$$

be surjective. Then the regularity condition of Robinson (3.5.1) holds.

Proof. Since $\hat{x} \in \text{int}(S)$ there exists an open ball $U_\varepsilon(\Theta_X) \subseteq S - \{\hat{x}\}$. The operator T is linear, continuous, surjective, and $T(\Theta_X) = (\Theta_Y, \Theta_Z)$. According to the open mapping theorem 2.2.3 the set $T(U_\varepsilon(\Theta_X))$ is open in $Y \times Z$. Hence,

$$\begin{pmatrix} \Theta_Y \\ \Theta_Z \end{pmatrix} \in \text{int} \left\{ \begin{pmatrix} g'(\hat{x})(x - \hat{x}) \\ h'(\hat{x})(x - \hat{x}) \end{pmatrix} \middle| \, x \in S \right\}.$$

Since $g(\hat{x}) \in K$ and thus $\Theta_Y \in K - g(\hat{x})$ it follows the regularity condition of Robinson (3.5.1). ■

The subsequent Mangasarian-Fromowitz like condition is sufficient for the regularity condition of Robinson.

Corollary 3.5.4 *Let $g : X \to Y$ and $h : X \to Z$ be Fréchet-differentiable at $\hat{x}$, $K \subseteq Y$ a closed convex cone with vertex at zero and $\text{int}(K) \neq \emptyset$, $g(\hat{x}) \in K$, $h(\hat{x}) = \Theta_Z$. Furthermore, let the following conditions be fulfilled:*

(i) Let $h'(\hat{x})$ be surjective.

(ii) Let there exist some $\hat{d} \in int(S - \{\hat{x}\})$ with

$$h'(\hat{x})(\hat{d}) = \Theta_Z, \tag{3.5.2}$$
$$g'(\hat{x})(\hat{d}) \in int(K - \{g(\hat{x})\}). \tag{3.5.3}$$

Then the regularity condition of Robinson (3.5.1) holds.

Proof. Since $g(\hat{x}) + g'(\hat{x})(\hat{d}) \in \text{int}(K)$ there exists a ball $U_{\varepsilon_1}(g(\hat{x}) + g'(\hat{x})(\hat{d}))$ with radius $\varepsilon_1 > 0$ and center $g(\hat{x}) + g'(\hat{x})(\hat{d})$ which lies in $\text{int}(K)$. Since subtraction viewed as a mapping from $Y \times Y$ into Y is continuous and observing, that

$$g(\hat{x}) + g'(\hat{x})(\hat{d}) = g(\hat{x}) + g'(\hat{x})(\hat{d}) - \Theta_Y,$$

there exist balls $U_{\varepsilon_2}(g(\hat{x}) + g'(\hat{x})(\hat{d}))$ and $U_{\varepsilon_3}(\Theta_Y)$ with

$$U_{\varepsilon_2}(g(\hat{x}) + g'(\hat{x})(\hat{d})) - U_{\varepsilon_3}(\Theta_Y) \subseteq U_{\varepsilon_1}(g(\hat{x}) + g'(\hat{x})(\hat{d})).$$

Since $g'(\hat{x})$ is a continuous linear mapping, there exists a ball $U_{\delta_1}(\hat{d})$ in X with

$$g(\hat{x}) + g'(\hat{x})(U_{\delta_1}(\hat{d})) \subseteq U_{\varepsilon_2}(g(\hat{x}) + g'(\hat{x})(\hat{d})).$$

Since $\hat{d} \in \mathrm{int}(S - \{\hat{x}\})$, eventually after diminishing δ_1 to $\delta_2 > 0$, we find

$$g(\hat{x}) + g'(\hat{x})(U_{\delta_2}(\hat{d})) \subseteq U_{\varepsilon_2}(g(\hat{x}) + g'(\hat{x})(\hat{d}))$$

and

$$U_{\delta_2}(\hat{d}) \subseteq \mathrm{int}(S - \{\hat{x}\}).$$

Since $h'(\hat{x})$ is continuous and surjective by the open mapping theorem, there exists a ball $U_{\varepsilon_4}(\Theta_Z)$ in Z with

$$U_{\varepsilon_4}(\Theta_Z) \subseteq h'(\hat{x})(U_{\delta_2}(\hat{d})).$$

Summarizing, we found the following: For every $\tilde{y} \in U_{\varepsilon_3}(\Theta_Y)$, $\tilde{z} \in U_{\varepsilon_4}(\Theta_Z)$ there exists $d \in U_{\delta_2}(\hat{d})$, in particular $d \in \mathrm{int}(S - \{\hat{x}\})$, with

$$h'(\hat{x})(d) = \tilde{z}, \qquad g(\hat{x}) + g'(\hat{x})(d) \in U_{\varepsilon_2}(g(\hat{x}) + g'(\hat{x})(\hat{d})),$$

hence

$$g(\hat{x}) + g'(\hat{x})(d) - \tilde{y} \in U_{\varepsilon_1}(g(\hat{x}) + g'(\hat{x})(\hat{d})) \subseteq \mathrm{int}(K),$$

i.e. there exists $k \in \mathrm{int}(K)$ with

$$g(\hat{x}) + g'(\hat{x})(d) - \tilde{y} = k.$$

Thus, we proved

$$\tilde{y} = g(\hat{x}) + g'(\hat{x})(d) - k, \qquad \tilde{z} = h'(\hat{x})(d)$$

with some $k \in \mathrm{int}(K)$, $d \in \mathrm{int}(S - \{\hat{x}\})$. ■

Remark 3.5.5

- *Condition (ii) in Corollary 3.5.4 can be replaced by the following assumption: Let there exist some $\hat{d} \in int(cone(S, \hat{x}))$ with*

$$\begin{aligned} h'(\hat{x})(\hat{d}) &= \Theta_Z, \\ g'(\hat{x})(\hat{d}) &\in int(cone(K, g(\hat{x}))). \end{aligned}$$

- *It is possible to show that the above Mangasarian-Fromowitz condition is equivalent to Robinson's condition for problems of type 3.1.2.*

Second order necessary and sufficient conditions for infinite optimization problem are discussed in Maurer and Zowe [MZ79] and Maurer [Mau81]. In Maurer [Mau81] also the so-called two norm discrepancy is addressed. This problem occurs if the sufficient conditions do not hold for the norm $\|\cdot\|_X$. On the other hand, these conditions may be satisfied for an alternative norm $\|\cdot\|_p$ on the space X. Unfortunately, the appearing functions usually are not differentiable anymore in this alternative norm and the proof techniques are more intricate.

3.6 Necessary and Sufficient Conditions in Finte Dimensions

In this section we address the finite dimensional case with $X = \mathbb{R}^{n_x}$, $Y = \mathbb{R}^{n_g}$, $Z = \mathbb{R}^{n_h}$, $S \subseteq \mathbb{R}^{n_x}$, and continuously differentiable functions

$$\begin{aligned} f &: \mathbb{R}^{n_x} \to \mathbb{R}, \\ g = (g_1, \ldots, g_{n_g})^\top &: \mathbb{R}^{n_x} \to \mathbb{R}^{n_g}, \\ h = (h_1, \ldots, h_{n_h})^\top &: \mathbb{R}^{n_x} \to \mathbb{R}^{n_h}. \end{aligned}$$

We will discuss necessary and sufficient conditions as well as the sequential quadratic programming (SQP) method for nonlinear optimization problems of the form

Problem 3.6.1 (Nonlinear Optimization Problem)
Find $x \in \mathbb{R}^{n_x}$ such that $f(x)$ is minimized subject to the constraints

$$\begin{array}{rcll} g_i(x) & \leq & 0, & i = 1,\ldots,n_g, \\ h_j(x) & = & 0, & j = 1,\ldots,n_h, \\ x & \in & S. & \end{array}$$

The feasible set of Problem 3.6.1 is

$$\Sigma := \{x \in S \mid g_i(x) \leq 0,\ i = 1,\ldots,n_g,\ h_j(x) = 0,\ j = 1,\ldots,n_h\}.$$

The set

$$A(x) := \{i \mid g_i(x) = 0,\ 1 \leq i \leq n_g\}$$

is called *index set of active inequality constraints* at $x \in \Sigma$.
The subsequent results are collected from the monographs [BS79], [GMW81], [FM90], [Spe93], [Man94], [GK99], [Alt02], [GK02].
By reformulating the infinite Fritz-John conditions in Theorem 3.4.2 for the finite dimensional problem 3.6.1 we find the following finite dimensional version of the Fritz-John conditions. Notice, that the assumption that $\mathrm{im}(h'(\hat{x}))$ be not a proper dense subset of $Z = \mathbb{R}^{n_h}$ is trivially fulfilled in finite dimensions, since then the image of a linear mapping is closed. Furthermore, recall that $\mathbb{R}^n$ is a Hilbert space and hence can be identified with its dual space. We will use the *Lagrange function*

$$L(x, l_0, \lambda, \mu) := l_0 f(x) + \sum_{i=1}^{n_g} \lambda_i g_i(x) + \sum_{j=1}^{n_h} \mu_j h_j(x).$$

Theorem 3.6.2 (First Order Necessary Conditions, finite case)
Let $\hat{x}$ be a local minimum of Problem 3.6.1 and S closed and convex with $int(S) \neq \emptyset$. Then there exist multipliers $l_0 \geq 0$, $\lambda = (\lambda_1,\ldots,\lambda_{n_g})^\top \in \mathbb{R}^{n_g}$ and $\mu = (\mu_1,\ldots,\mu_{n_h})^\top \in \mathbb{R}^{n_h}$ not all zero such that

$$L'_x(\hat{x}, l_0, \lambda, \mu)(x - \hat{x}) \geq 0 \quad \text{for all } x \in S, \tag{3.6.1}$$
$$g_i(\hat{x}) \leq 0,\ i = 1,\ldots,n_g, \tag{3.6.2}$$
$$h_j(\hat{x}) = 0,\ j = 1,\ldots,n_h, \tag{3.6.3}$$
$$\lambda_i g_i(\hat{x}) = 0,\ i = 1,\ldots,n_g, \tag{3.6.4}$$
$$\lambda_i \geq 0,\ i = 1,\ldots,n_g. \tag{3.6.5}$$

The constraint qualification of Mangasarian-Fromowitz is defined as in Corollary 3.5.4.

Definition 3.6.3 (Constraint Qualification of Mangasarian-Fromowitz)
The constraint qualification of Mangasarian-Fromowitz is satisfied at $\hat{x}$, if the following conditions are fulfilled:

(a) The derivatives $h'_j(\hat{x})$, $j = 1,\ldots,n_h$ are linearly independent;

(b) There exists a vector $\hat{d} \in int(S - \{\hat{x}\})$ with

$$g'_i(\hat{x})(\hat{d}) < 0 \text{ for } i \in A(\hat{x}) \text{ and } h'_j(\hat{x})(\hat{d}) = 0 \text{ for } j = 1,\ldots,n_h.$$

Although the following theorem is a consequence of Theorem 3.5.2 and Corollary 3.5.4, we will give an alternative proof for finite dimensions.

Theorem 3.6.4 (Karush-Kuhn-Tucker (KKT) Conditions I)
Let the assumptions of Theorem 3.6.2 be satisfied and let the constraint qualification of Mangasarian-Fromowitz be fulfilled at $\hat{x}$. Then, the assertions of Theorem 3.6.2 hold with $l_0 = 1$.

Proof. Assume, that the constraint qualification of Mangasarian-Fromowitz holds and that the Fritz-John conditions hold at $\hat{x}$ with $l_0 = 0$, that is, there exist multipliers $\lambda_i \geq 0$, $i = 1, \ldots, n_g$, μ_j, $j = 1, \ldots, n_h$ not all zero with $\lambda_i g_i(\hat{x}) = 0$ for all $i = 1, \ldots, n_g$ and

$$\left(\sum_{i=1}^{n_g} \lambda_i g_i'(\hat{x}) + \sum_{j=1}^{n_h} \mu_j h_j'(\hat{x}) \right) (d) \geq 0 \quad \text{for all } d \in S - \{\hat{x}\}. \tag{3.6.6}$$

Let $\hat{d}$ denote the vector in the constraint qualification of Mangasarian-Fromowitz, then we find

$$\sum_{i=1}^{n_g} \lambda_i g_i'(\hat{x})(\hat{d}) = \sum_{i \in A(\hat{x})} \lambda_i g_i'(\hat{x})(\hat{d}) \quad \geq \quad 0.$$

Since $\lambda_i \geq 0$ and $g_i'(\hat{x})(\hat{d}) < 0$ for $i \in A(\hat{x})$ this inequality can only be valid if $\lambda_i = 0$ holds for every $i \in A(\hat{x})$. Since $\lambda_i = 0$ for $i \notin A(\hat{x})$ we found $\lambda_i = 0$ for all $i = 1, \ldots, n_g$.
Thus, inequality (3.6.6) reduces to

$$0 \leq \sum_{j=1}^{n_h} \mu_j h_j'(\hat{x})(d) = \sum_{j=1}^{n_h} \mu_j (h_j'(\hat{x})(d) - h_j'(\hat{x})(\hat{d})) = \sum_{j=1}^{n_h} \mu_j h_j'(\hat{x})(d - \hat{d})$$

for all $d \in S - \{\hat{x}\}$. Since $\hat{d} \in \text{int}(S - \{\hat{x}\})$ and $h_j'(\hat{x})$, $j = 1, \ldots, n_h$ are linearly independent this inequality can only hold for $\mu_j = 0$, $j = 1, \ldots, n_h$.
Hence, we derived $(l_0, \lambda, \mu) = 0_{1+n_g+n_h}$, which is a contradiction to the statement that not all multipliers are zero. ■

Likewise, the linear independence constraint qualification is defined according to Corollary 3.5.3. Actually, this condition implies that of Mangasarian-Fromowitz.

Definition 3.6.5 (Linear Independence Constraint Qualification (LICQ))
The linear independence constraint qualification is satisfied at $\hat{x}$, if the following conditions are fulfilled:

(a) $\hat{x} \in int(S)$;

(b) The derivatives $g_i'(\hat{x})$, $i \in A(\hat{x})$, and $h_j'(\hat{x})$, $j = 1, \ldots, n_h$ are linearly independent.

The first part of the following theorem is a consequence of Theorem 3.5.2 and Corollary 3.5.3. Nevertheless, we will give an alternative proof.

Theorem 3.6.6 (Karush-Kuhn-Tucker (KKT) Conditions II)
Let the assumptions of Theorem 3.6.2 be satisfied and let the linear independence constraint qualification be fulfilled at $\hat{x}$. Then, the assertions of Theorem 3.6.2 hold with $l_0 = 1$ and in particular

$$\nabla_x L(\hat{x}, l_0, \lambda, \mu) = 0_{n_x}.$$

Furthermore, the multipliers λ and μ are unique.

Proof. Again, we assume that the Fritz-John conditions hold with $l_0 = 0$. Since $\hat{x} \in \text{int}(S)$, inequality (3.6.1) has to hold for all $x \in \mathbb{R}^{n_x}$ and thus

$$\sum_{i=1}^{n_g} \lambda_i g_i'(\hat{x}) + \sum_{j=1}^{n_h} \mu_j h_j'(\hat{x}) = \sum_{i \in A(\hat{x})} \lambda_i g_i'(\hat{x}) + \sum_{j=1}^{n_h} \mu_j h_j'(\hat{x}) = 0_{n_x}.$$

The linear independence of the derivatives implies $\lambda_i = \mu_j = 0$ for all $i = 1, \ldots, n_g,\ j = 1, \ldots, n_h$. Again, this is a contradiction to $(l_0, \lambda, \mu) \neq 0_{1+n_g+n_h}$.
The uniqueness of the Lagrange multipliers follows from the following considerations. Assume, that there are Lagrange multipliers λ_i, $i = 1, \ldots, n_g$, μ_j, $j = 1, \ldots, n_h$ and $\tilde{\lambda}_i$, $i = 1, \ldots, n_g$, $\tilde{\mu}_j$, $j = 1, \ldots, n_h$ satisfying the KKT conditions. Again, $\hat{x} \in \text{int}(S)$ particularly implies

$$\begin{aligned} 0_{n_x} &= f'(\hat{x}) + \sum_{i=1}^{n_g} \lambda_i g_i'(\hat{x}) + \sum_{j=1}^{n_h} \mu_j h_j'(\hat{x}), \\ 0_{n_x} &= f'(\hat{x}) + \sum_{i=1}^{n_g} \tilde{\lambda}_i g_i'(\hat{x}) + \sum_{j=1}^{n_h} \tilde{\mu}_j h_j'(\hat{x}). \end{aligned}$$

Subtracting these equations leads to

$$0_{n_x} = \sum_{i=1}^{n_g} (\lambda_i - \tilde{\lambda}_i) g_i'(\hat{x}) + \sum_{j=1}^{n_h} (\mu_j - \tilde{\mu}_j) h_j'(\hat{x}).$$

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 $0 = \mu_j - \tilde{\mu}_j$, $j = 1, \ldots, n_h$. Hence, the Lagrange multipliers are unique. ■

Remark 3.6.7 *There exist several other constraint qualifications. One of the weakest conditions is the* constraint qualification of Abadie *postulating*

$$T_{lin}(K, S, \hat{x}) \subseteq T(\Sigma, \hat{x}),$$

where $K = \{y \in \mathbb{R}^{n_g} \mid y \leq 0_{n_g}\}$ *and*

$$\begin{aligned} T_{lin}(K, S, x) \;=\; \{d \in cone(S, x) \mid\ & g_i'(x)(d) \leq 0,\ i \in A(x), \\ & h_j'(x)(d) = 0,\ j = 1, \ldots, n_h\} \end{aligned}$$

denotes the linearizing cone at $\hat{x}$.
The condition of Abadie is weaker as the previously discussed constraint qualifications since those imply the condition of Abadie but not vice versa.
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* $T_{lin}(K, S, x)$ *depends on the functions* g_i *and* h_j *describing* Σ.

In the sequel, we restrict the discussion to the case $S = \mathbb{R}^{n_x}$. This special case is particularly convenient for the design of numerical methods, e.g. SQP methods. Setting $S = \mathbb{R}^{n_x}$, inequality (3.6.1) has to hold for all $x \in \mathbb{R}^{n_x}$ and thus it is equivalent with the equality

$$\nabla_x L(\hat{x}, l_0, \lambda, \mu) = 0_{n_x}.$$

The previously discussed regularity conditions additionally ensure $l_0 = 1$.

To decide, whether a given point that fulfills the necessary conditions is optimal we need *sufficient conditions.* We need the so-called critical cone

$$T_C(\hat{x}) := \left\{ d \in \mathbb{R}^{n_x} \;\middle|\; \begin{array}{rcl} g_i'(\hat{x})(d) & \leq & 0,\ i \in A(\hat{x}), \lambda_i = 0, \\ g_i'(\hat{x})(d) & = & 0,\ i \in A(\hat{x}), \lambda_i > 0, \\ h_j'(\hat{x})(d) & = & 0,\ j = 1, \ldots, n_h \end{array} \right\}.$$

The term 'critical cone' is due to the following reasoning. Directions d with $g_i'(\hat{x})(d) > 0$ for some $i \in A(\hat{x})$ or $h_j'(\hat{x})(d) \neq 0$ for some $j \in \{1, \ldots, n_h\}$ are infeasible directions. So, consider only feasible directions d with $g_i'(\hat{x})(d) \leq 0$ for $i \in A(\hat{x})$ and $h_j'(\hat{x})(d) = 0$ for $j = 1, \ldots, n_h$. For such directions the KKT conditions yield

$$\nabla f(\hat{x})^\top d + \sum_{i \in A(\hat{x})} \underbrace{\lambda_i \nabla g_i(\hat{x})^\top d}_{\leq 0} + \sum_{j=1}^{n_h} \mu_j \underbrace{\nabla h_j(\hat{x})^\top d}_{=0} = 0_{n_x},$$

and thus $f'(\hat{x})(d) \geq 0$. If even $f'(\hat{x})(d) > 0$ holds, then d is a direction of ascent and the direction d is not interesting for the investigation of sufficient conditions. So, let $f'(\hat{x})(d) = 0$. This is the critical case. In this critical case it holds

$$\sum_{i \in A(\hat{x})} \lambda_i g_i'(\hat{x})(d) = \sum_{i \in A(\hat{x}), \lambda_i > 0} \lambda_i g_i'(\hat{x})(d) = 0,$$

and thus $g_i'(\hat{x})(d) = 0$ for all $i \in A(\hat{x})$ with $\lambda_i > 0$. Hence, $d \in T_C(\hat{x})$ and for such directions we need additional assumptions about the curvature (2nd derivative!).
A second order sufficient condition, cf., e.g., Geiger and Kanzow [GK02], Th. 2.55, p. 67, Alt [Alt02], Th. 7.3.1, p. 281, and Bazaraa et al. [BSS93], Th. 4.4.2, p. 169, is given by

Theorem 3.6.8 (Second Order Sufficient Condition)
Let f, g_i, $i = 1, \ldots, n_g$, and h_j, $j = 1, \ldots, n_h$ be twice continuously differentiable. Let $S = \mathbb{R}^{n_x}$ and let $(\hat{x}, \hat{\lambda}, \hat{\mu})$ be a KKT point of Problem 3.6.1 with

$$L''_{xx}(\hat{x}, \hat{\lambda}, \hat{\mu})(d, d) > 0 \qquad \forall d \in T_C(\hat{x}),\ d \neq 0_{n_x}. \tag{3.6.7}$$

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.

(a) Let $d \in T(\Sigma, \hat{x})$, $d \neq 0_{\mathbb{R}^{n_x}}$. 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

$$0 \geq \frac{g_i(x_k) - g_i(\hat{x})}{\alpha_k} = g_i'(\xi_k) \left(\frac{x_k - \hat{x}}{\alpha_k} \right) \to g_i'(\hat{x})(d)$$

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

$$f'(\hat{x})(d) = -\sum_{i=1}^{n_g} \hat{\lambda}_i g_i'(\hat{x})(d) - \sum_{j=1}^{n_h} \hat{\mu}_j h_j'(\hat{x})(d) \geq 0.$$

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

(b) Assume, that the statement of the theorem is wrong. Then for any ball around $\hat{x}$ with radius $1/i$ there exists a point $x_i \in \Sigma$ with $x_i \neq \hat{x}$ and

$$f(x_i) - f(\hat{x}) < \frac{1}{i}\|x_i - \hat{x}\|^2 , \quad \|x_i - \hat{x}\| \leq \frac{1}{i} \qquad \forall i \in \mathbb{N}. \tag{3.6.8}$$

Since the unit ball w.r.t. $\|\cdot\|$ is compact in $\mathbb{R}^{n_x}$, there exists a convergent subsequence $\{x_{i_k}\}$ with

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

Hence, $d \in T(\Sigma, \hat{x}) \setminus \{0_{n_x}\}$. Taking the limit in (3.6.8) yields

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

Together with (a) we have

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

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

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

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

According to (3.6.8) it holds

$$\lim_{k\to\infty} \frac{f(x_{i_k}) - f(\hat{x})}{\|x_{i_k} - \hat{x}\|^2} \leq \lim_{k\to\infty} \frac{1}{i_k} = 0 \tag{3.6.9}$$

for the direction d. Furthermore, it is ($l_0 = 1$)

$$\begin{aligned}
L(x_{i_k}, l_0, \hat{\lambda}, \hat{\mu}) &= f(x_{i_k}) + \sum_{i=1}^{n_g} \hat{\lambda}_i g_i(x_{i_k}) + \sum_{j=1}^{n_h} \hat{\mu}_j h_j(x_{i_k}) \leq f(x_{i_k}) , \\
L(\hat{x}, l_0, \hat{\lambda}, \hat{\mu}) &= f(\hat{x}) + \sum_{i=1}^{n_g} \hat{\lambda}_i g_i(\hat{x}) + \sum_{j=1}^{n_h} \hat{\mu}_j h_j(\hat{x}) = f(\hat{x}) , \\
L_x'(\hat{x}, l_0, \hat{\lambda}, \hat{\mu}) &= f'(\hat{x}) + \sum_{i=1}^{n_g} \hat{\lambda}_i g_i'(\hat{x}) + \sum_{j=1}^{n_h} \hat{\mu}_j h_j'(\hat{x}) = 0_{n_x}^\top .
\end{aligned}$$

Taylor expansion of L with $l_0 = 1$ w.r.t. to x at $\hat{x}$ yields

$$\begin{aligned}
f(x_{i_k}) \geq L(x_{i_k}, l_0, \hat{\lambda}, \hat{\mu}) &= L(\hat{x}, l_0, \hat{\lambda}, \hat{\mu}) + L_x'(\hat{x}, l_0, \hat{\lambda}, \hat{\mu})(x_{i_k} - \hat{x}) \\
&\quad + \frac{1}{2} L_{xx}''(\xi_k, l_0, \hat{\lambda}, \hat{\mu})(x_{i_k} - \hat{x}, x_{i_k} - \hat{x}) \\
&= f(\hat{x}) + \frac{1}{2} L_{xx}''(\xi_k, l_0, \hat{\lambda}, \hat{\mu})(x_{i_k} - \hat{x}, x_{i_k} - \hat{x}),
\end{aligned}$$

where ξ_k is some point between $\hat{x}$ and x_{i_k}. Division by $\|x_{i_k} - \hat{x}\|^2$ and taking the limit, yields together with (3.6.9)

$$0 \geq \frac{1}{2} L_{xx}''(\hat{x}, l_0, \hat{\lambda}, \hat{\mu})(d, d).$$

This contradicts the assumption $L_{xx}''(\hat{x}, l_0, \hat{\lambda}, \hat{\mu})(d, d) > 0$ for all $d \in T_C(\hat{x})$, $d \neq 0_{n_x}$.

■

A second order necessary condition involving the critical cone can be found in, e.g., Geiger and Kanzow [GK02], Th. 2.54, p. 65.

3.7 Perturbed Nonlinear Optimization Problems

In view of the convergence analysis of the SQP method and for real-time approximations we need results about the sensitivity of solutions under perturbations. We restrict the discussion to the finite dimensional case. Results for infinite optimization problems can be found in Lempio and Maurer [LM80] and Bonnans and Shapiro [BS00] and the literature cited therein. We investigate parametric optimization problems:

Problem 3.7.1 (Parametric Optimization Problem $NLP(p)$)
Let $p \in \mathbb{R}^{n_p}$ be a given parameter. Find $x \in \mathbb{R}^{n_x}$ such that $f(x,p)$ is minimized subject to the constraints

$$\begin{aligned} g_i(x,p) &\leq 0, \quad i = 1,\dots,n_g, \\ h_j(x,p) &= 0, \quad j = 1,\dots,n_h. \end{aligned}$$

Herein, $f, g_1, \dots, g_{n_g}, h_1, \dots, h_{n_h} : \mathbb{R}^{n_x} \times \mathbb{R}^{n_p} \to \mathbb{R}$ are sufficiently smooth functions. Let $\hat{p}$ denote a fixed *nominal parameter.* We are interested in the behavior of the optimal solutions $\hat{x}(p)$ as functions of p in a neighborhood of the nominal parameter $\hat{p}$.
The admissible set of $NLP(p)$ is defined by

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

The index set of active inequality constraints is given by

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

Definition 3.7.2 (Strongly Regular Local Solution)
A local minimum $\hat{x}$ of $NLP(p)$ is called strongly regular *if the following properties hold:*

- *$\hat{x}$ is admissible, i.e. $\hat{x} \in \Sigma(p)$.*
- *$\hat{x}$ fulfills the linear independence constraint qualification, i.e. the gradients $\nabla_x g_i(\hat{x},p)$, $i \in A(\hat{x},p)$, $\nabla_x h_j(\hat{x},p)$, $j = 1,\dots,n_h$, are linearly independent.*
- *The KKT conditions hold at $(\hat{x}, \hat{\lambda}, \hat{\mu})$.*
- *The strict complementarity condition holds, i.e. $\hat{\lambda}_i - g_i(\hat{x},p) > 0$ for all $i = 1,\dots,n_g$.*
- *The second order sufficient condition (3.6.7) holds.*

The following result is based on Fiacco [Fia83, FM90] and Spellucci [Spe93]. Further results can be found in Bank et al. [BGK+83] and Klatte [Kla90].

Theorem 3.7.3 (Sensitivity Theorem)
Let $f, g_1, \dots, g_{n_g}, h_1, \dots, h_{n_h} : \mathbb{R}^{n_x} \times \mathbb{R}^{n_p} \to \mathbb{R}$ be twice continuously differentiable and $\hat{p}$ a nominal parameter. Let $\hat{x}$ be a strongly regular local minimum of $NLP(\hat{p})$, $\hat{\lambda}$, $\hat{\mu}$ denote the corresponding Lagrange multipliers. Then there exist neighborhoods $V_\epsilon(\hat{p})$ and $U_\delta(\hat{x}, \hat{\lambda}, \hat{\mu})$, such that $NLP(p)$ has a unique strongly regular local minimum

$$(x(p), \lambda(p), \mu(p)) \in U_\delta(\hat{x}, \hat{\lambda}, \hat{\mu})$$

for each $p \in V_\epsilon(\hat{p})$. Furthermore, it holds $A(\hat{x},\hat{p}) = A(x(p),p)$. In addition, $(x(p),\lambda(p),\mu(p))$ is continuously differentiable w.r.t. p with

$$\begin{pmatrix} \frac{dx}{dp}(\hat{p}) \\ \frac{d\lambda}{dp}(\hat{p}) \\ \frac{d\mu}{dp}(\hat{p}) \end{pmatrix} = - \begin{pmatrix} L''_{xx} & (g'_x)^\top & (h'_x)^\top \\ \hat{\Lambda}\cdot g'_x & \hat{\Gamma} & \Theta \\ h'_x & \Theta & \Theta \end{pmatrix}^{-1} \cdot \begin{pmatrix} L''_{xp} \\ \hat{\Lambda}\cdot g'_p \\ h'_p \end{pmatrix} \tag{3.7.1}$$

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

Proof. Consider the nonlinear equation

$$F(x,\lambda,\mu,p) := \begin{pmatrix} L'_x(x,\lambda,\mu,p)^\top \\ \Lambda\cdot g(x,p) \\ h(x,p) \end{pmatrix} = 0_{n_x+n_g+n_h}, \tag{3.7.2}$$

where $\Lambda := \operatorname{diag}(\lambda_1,\ldots,\lambda_{n_g})$. F is continuously differentiable and it holds

$$F(\hat{x},\hat{\lambda},\hat{\mu},\hat{p}) = 0_{n_x+n_g+n_h}.$$

We intend to apply the implicit function theorem. Hence, we have to show the non-singularity of

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

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

$$\hat{\Lambda}_2 := \operatorname{diag}(\hat{\lambda}_{l+1},\ldots,\hat{\lambda}_{n_g}) \qquad \text{and} \qquad \hat{\Gamma}_1 := \operatorname{diag}(g_1(\hat{x},\hat{p}),\ldots,g_l(\hat{x},\hat{p})).$$

Consider the linear equation

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

for $v_1 \in \mathbb{R}^{n_x}$, $v_2 = (v_{21},v_{22})^\top \in \mathbb{R}^{l+(n_g-l)}$, and $v_3 \in \mathbb{R}^{n_h}$. 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 $v_{21} = 0_l$. With this, it remains to investigate the reduced system

$$\begin{pmatrix} A & B^\top & C^\top \\ B & \Theta & \Theta \\ C & \Theta & \Theta \end{pmatrix} \begin{pmatrix} v_1 \\ v_{22} \\ v_3 \end{pmatrix} = \begin{pmatrix} 0_{n_x} \\ 0_{n_g-l} \\ 0_{n_h} \end{pmatrix}$$

with $A := L''_{xx}(\hat{x}, \hat{\lambda}, \hat{\mu}, \hat{p})$, $B := (g'_{x,i}(\hat{x}, \hat{p}))_{i=l+1,\ldots,n_g}$, and $C := h'_x(\hat{x}, \hat{p})$. Notice, that the second block equation has been multiplied with $\hat{\Lambda}_2^{-1}$. The last two block equations yield $Bv_1 = 0_{n_g-l}$ and $Cv_1 = 0_{n_h}$. 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 $T_C(\hat{x}) \setminus \{0_{n_x}\}$, i.e. it holds $d^\top A d > 0$ for all $d \neq 0_{n_x}$ with $Bd = 0_{n_g-l}$ and $Cd = 0_{n_h}$, it follows $v_1 = 0_{n_x}$. Taking this property into account, the first block equation reduces to $B^\top v_{22} + C^\top v_3 = 0_{n_x}$. By the linear independence of the gradients $\nabla g_i(\hat{x}, \hat{p})$, $i \in A(\hat{x}, \hat{p})$ and $\nabla h_j(\hat{x}, \hat{p})$, $j = 1, \ldots, n_h$ we obtain $v_{22} = 0_{n_g-l}$, $v_3 = 0_{n_h}$. Putting all together, the above linear equation has the unique solution $v_1 = 0_{n_x}$, $v_2 = 0_{n_g}$, $v_3 = 0_{n_h}$, which implies that the matrix F'_x is non-singular and the implicit function theorem is applicable.
By the implicit function theorem there exist neighborhoods $V_\epsilon(\hat{p})$ and $U_\delta(\hat{x}, \hat{\lambda}, \hat{\mu})$, and uniquely defined functions

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

satisfying

$$F(x(p), \lambda(p), \mu(p), p) = 0_{n_x+n_g+n_h} \tag{3.7.3}$$

for all $p \in V_\epsilon(\hat{p})$. Furthermore, these functions are continuously differentiable and (3.7.1) arises by differentiation of the identity (3.7.3) w.r.t. p.
It remains to verify, that $x(p)$ actually is a strongly regular local minimum of $NLP(p)$. The continuity of the functions $x(p), \lambda(p)$ and g together with $\lambda_i(\hat{p}) = \hat{\lambda}_i > 0$, $i = l+1, \ldots, n_g$ and $g_i(x(\hat{p}), \hat{p}) = g_i(\hat{x}, \hat{p}) < 0$, $i = 1, \ldots, l$ guarantees $\lambda_i(p) > 0$, $i = l+1, \ldots, n_g$ and $g_i(x(p), p) < 0$, $i = 1, \ldots, l$ for p sufficiently close to $\hat{p}$. From (3.7.3) it follows $g_i(x(p), p) = 0$, $i = l+1, \ldots, n_g$, and $h_j(x(p), p) = 0$, $j = 1, \ldots, n_h$. Thus, $x(p) \in \Sigma(p)$ and the KKT conditions are satisfied. Furthermore, the index set $A(x(p), p) = A(\hat{x}, \hat{p})$ remains unchanged in a neighborhood of $\hat{p}$.
Finally, we have to show that $L''_{xx}(x(p), \lambda(p), \mu(p), p)$ remains positive definite on $T_C(x(p))$ for p sufficiently close to $\hat{p}$. Notice, that the critical cone $T_C(x(p))$ varies with p. By now, we only know that $L''_{xx}(\hat{p}) := L''_{xx}(x(\hat{p}), \lambda(\hat{p}), \mu(\hat{p}), \hat{p})$ is positive definite on $T_C(x(\hat{p}))$ which is equivalent to the existence of some $\alpha > 0$ with $d^\top L''_{xx}(\hat{p}) d \geq \alpha \|d\|^2$ for all $d \in T_C(x(\hat{p}))$. Owing to the strict complementarity in a neighborhood of $\hat{p}$ it holds

$$T_C(x(p)) = \left\{ d \in \mathbb{R}^{n_x} \;\middle|\; \begin{array}{rcl} g'_i(x(p), p)(d) & = & 0, \; i \in A(\hat{x}, \hat{p}), \\ h'_j(x(p), p)(d) & = & 0, \; j = 1, \ldots, n_h \end{array} \right\}$$

around $\hat{p}$. Assume, that for every $i \in \mathbb{N}$ there exists some $p^i \in \mathbb{R}^{n_p}$ with $\|p^i - \hat{p}\| \leq \frac{1}{i}$ such that for all $j \in \mathbb{N}$ there exists some $d^{i_j} \in T_C(x(p^i))$, $d^{i_j} \neq 0_{n_x}$, with

$$(d^{i_j})^\top L''_{xx}(p^{i_j}) d^{i_j} < \frac{1}{j} \|d^{i_j}\|^2.$$

Since the unit ball w.r.t. $\|\cdot\|$ is compact in $\mathbb{R}^{n_x}$, there exists a convergent subsequence $\{p^{i_{j_k}}\}$ with $\lim_{k\to\infty} p^{i_{j_k}} = \hat{p}$ and

$$\lim_{k\to\infty} \frac{d^{i_{j_k}}}{\|d^{i_{j_k}}\|} = \hat{d}, \quad \|\hat{d}\| = 1, \quad \hat{d} \in T_C(x(\hat{p}))$$

and

$$\hat{d}^\top L''_{xx}(\hat{p}) \hat{d} \leq 0.$$

This contradicts the positive definiteness of $L''_{xx}(\hat{p})$. ■

Büskens and Maurer [BM01b] exploit Theorem 3.7.3 and equation (3.7.1) to construct an approximation for the optimal solution of a perturbed optimization problem in real-time. This real-time approximation is based on a linearization of the nominal solution for a given nominal parameter $\hat{p}$. Under the assumptions of Theorem 3.7.3 the solution $x(p)$ of $NLP(p)$ is continuously differentiable in some neighborhood of $\hat{p}$. Hence, it holds

$$x(p) = x(\hat{p}) + \frac{dx}{dp}(\hat{p})(p - \hat{p}) + o(\|p - \hat{p}\|)$$

with dx/dp from (3.7.1).
An approximation of the optimal solution $x(p)$ is obtained by the linear approximation

$$x(p) \approx x(\hat{p}) + \frac{dx}{dp}(\hat{p})(p - \hat{p}). \tag{3.7.4}$$

The time consuming computation of the nominal solution $x(\hat{p})$ and the sensitivity dx/dp is done offline. The evaluation of the right handside in (3.7.4) is very cheap since only a matrix-vector product and two vector additions are necessary.
The linearization in (3.7.4) is only justified locally in some neighborhood of the nominal parameter $\hat{p}$. Unfortunately, Theorem 3.7.3 does not indicate how large this neighborhood is. In particular, if the index set of active inequality constraints changes then Theorem 3.7.3 is not applicable anymore.

3.8 Numerical Methods

We analyze the Lagrange-Newton-Method and the SQP-Method more closely. The SQP-Method is discussed in, e.g., [Han77], [Pow78], [GMW81], [Sto85], [Sch81], [Sch83], [Alt02], [GK02]. The SQP-Method exists in several implementations, e.g. [Sch85], [Kra88], [GMSW98], [GMS02]. Special adaptations of the SQP method to discretized optimal control problems are described in [GMS94], [Sch96], [Ste95], [BH99]. Again, we restrict the discussion to the finite dimensional case. A version of the SQP method working in general Banach spaces can be found in Alt [Alt91]. A SQP method for ODE optimal control problems is discussed in Machielsen [Mac88] and for PDE optimal control problems in Tröltzsch [Trö05].

3.8.1 Lagrange-Newton-Method

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

Problem 3.8.1 *Find $x \in \mathbb{R}^{n_x}$ such that $f(x)$ is minimized subject to the constraints*

$$h_j(x) = 0, \quad j = 1, \ldots, n_h.$$

The functions $f : \mathbb{R}^{n_x} \to \mathbb{R}$ and $h_j : \mathbb{R}^{n_x} \to \mathbb{R}$, $j = 1, \ldots, n_h$, are assumed to be twice continuously differentiable. Let $\hat{x}$ be a local minimum of Problem 3.8.1 and let the gradients $\nabla h_j(\hat{x})$, $j = 1, \ldots, n_h$ be linearly independent. Then the KKT conditions are valid: There exist multipliers $\hat{\mu} = (\hat{\mu}_1, \ldots, \hat{\mu}_{n_h})^\top \in \mathbb{R}^{n_h}$ such that

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

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

$$F(\hat{x},\hat{\mu}) = 0_{\mathbb{R}^{n_x+n_h}}, \tag{3.8.1}$$

where $F : \mathbb{R}^{n_x} \times \mathbb{R}^{n_h} \to \mathbb{R}^{n_x+n_h}$ and

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

The Lagrange-Newton is based on the application of Newton's method to solve the necessary conditions (3.8.1). This leads to the following algorithm:

Algorithm 3.8.2 (Lagrange-Newton Method)

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

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

(iii) Solve the linear equation

$$\begin{pmatrix} L''_{xx}(x^{(k)},\mu^{(k)}) & h'(x^{(k)})^\top \\ h'(x^{(k)}) & \Theta \end{pmatrix} \cdot \begin{pmatrix} d \\ v \end{pmatrix} = -\begin{pmatrix} \nabla_x L(x^{(k)},\mu^{(k)}) \\ h(x^{(k)}) \end{pmatrix} \tag{3.8.2}$$

and set

$$x^{(k+1)} = x^{(k)} + d, \qquad \mu^{(k+1)} = \mu^{(k)} + v. \tag{3.8.3}$$

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

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

Theorem 3.8.3 (Local Convergence)

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

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

(iii) Let the matrix

$$\begin{pmatrix} L''_{xx}(\hat{x},\hat{\mu}) & h'(\hat{x})^\top \\ h'(\hat{x}) & \Theta \end{pmatrix} \tag{3.8.4}$$

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.

Remark 3.8.4

- *The matrix in (3.8.4) is called Kuhn-Tucker-matrix (KT-matrix). The KT-matrix is non-singular, if the following conditions are satisfied:*

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

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

- *The convergence is at least super-linear, if the second derivatives of f and h_j, $j = 1, \ldots, n_h$, exist: 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 all sufficiently large k.

3.8.2 Sequential Quadratic Programming (SQP)

The linear equation (3.8.2) in item (iii) of the Lagrange-Newton method can be obtained in a different way.

Let us again consider the equality constrained problem 3.8.1. We assume, that the problem can be approximated locally at some point $(x^{(k)}, \mu^{(k)})$ by the quadratic optimization problem

$$\begin{array}{ll} \min\limits_{d\in\mathbb{R}^{n_x}} & \frac{1}{2} d^\top L''_{xx}(x^{(k)}, \mu^{(k)})d + \nabla f(x^{(k)})^\top d \\ \text{s.t.} & h(x^{(k)}) + h'(x^{(k)})d \;=\; 0_{n_h}. \end{array}$$

The Lagrange function for the quadratic problem is given by

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

The evaluation of the first order necessary conditions leads to

$$\begin{pmatrix} L''_{xx}(x^{(k)}, \mu^{(k)}) & h'(x^{(k)})^\top \\ h'(x^{(k)}) & \Theta \end{pmatrix} \cdot \begin{pmatrix} d \\ \eta \end{pmatrix} = - \begin{pmatrix} \nabla f(x^{(k)}) \\ h(x^{(k)}) \end{pmatrix}. \tag{3.8.5}$$

If we subtract $h'(x^{(k)})^\top \mu^{(k)}$ on both sides of the first equation in (3.8.5) we get the linear equation

$$\begin{pmatrix} L''_{xx}(x^{(k)}, \mu^{(k)}) & h'(x^{(k)})^\top \\ h'(x^{(k)}) & \Theta \end{pmatrix} \cdot \begin{pmatrix} d \\ \eta - \mu^{(k)} \end{pmatrix} = - \begin{pmatrix} \nabla_x L(x^{(k)}, \mu^{(k)}) \\ h(x^{(k)}) \end{pmatrix}. \tag{3.8.6}$$

A comparison of (3.8.6) with (3.8.2) reveals that these two linear equation are identical, if we set $v := \eta - \mu^{(k)}$. According to (3.8.3) it follows that the new iterates are given by

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

Hence, 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 Lagrange-Newton method for Problem 3.6.1 with $S = \mathbb{R}^{n_x}$.

Algorithm 3.8.5 (Local SQP Method)

(i) *Choose* $(x^{(0)}, \lambda^{(0)}, \mu^{(0)}) \in \mathbb{R}^{n_x} \times \mathbb{R}^{n_g} \times \mathbb{R}^{n_h}$ *and set* $k = 0$.

(ii) *If* $(x^{(k)}, \lambda^{(k)}, \mu^{(k)})$ *is a KKT point of Problem 3.6.1 with* $S = \mathbb{R}^{n_x}$*, STOP.*

(iii) *Compute a KKT point* $(d^{(k)}, \lambda^{(k+1)}, \mu^{(k+1)}) \in \mathbb{R}^{n_x} \times \mathbb{R}^{n_g} \times \mathbb{R}^{n_h}$ *of the the quadratic programming problem* $QP(x^{(k)}, \lambda^{(k)}, \mu^{(k)})$ *given by*

$$\begin{array}{lrcll} \min\limits_{d \in \mathbb{R}^{n_x}} & \multicolumn{4}{l}{\frac{1}{2} d^\top L''_{xx}(x^{(k)}, \lambda^{(k)}, \mu^{(k)}) d + f'(x^{(k)}) d} \\ s.t. & g_i(x^{(k)}) + g_i'(x^{(k)}) d & \leq & 0, & i = 1, \ldots, n_g, \\ & h_j(x^{(k)}) + h_j'(x^{(k)}) d & = & 0, & j = 1, \ldots, n_h. \end{array}$$

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

Remark 3.8.6

- *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. [GM78], [GI83], [GMSW91], [Spe93].*

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

Theorem 3.8.7 (Local Convergence of SQP Method)

(i) *Let* $\hat{x}$ *be a local minimum of Problem 3.6.1 with* $S = \mathbb{R}^{n_x}$.

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

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

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

(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 $0_{n_x} \neq d \in \mathbb{R}^{n_x}$ *with*

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

Then there exists $\varepsilon > 0$ *such that for arbitrary initial 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 3.7.3 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.

(a) We consider $QP(\hat{x}, \hat{\lambda}, \hat{\mu})$ as the unperturbed quadratic problem with nominal parameter $\hat{p} = (\hat{x}, \hat{\lambda}, \hat{\mu})$. Furthermore, we notice, that the KKT conditions for $QP(\hat{p})$ and Problem 3.6.1 with $S = \mathbb{R}^{n_x}$ coincide for $\hat{d} = 0_{n_x}$. Hence, $(0_{n_x}, \hat{\lambda}, \hat{\mu})$ is a KKT point of $QP(\hat{p})$. In addition, assumptions (iii)-(v) guarantee that $\hat{d}$ is a strongly regular local minimum of $QP(\hat{p})$.

Hence, we may apply Theorem 3.7.3: There exists a neighborhood $V_\varepsilon(\hat{p})$ such that $QP(p)$ has a unique strongly regular local minimum $(d(p), \lambda(p), \mu(p))$ for each $p \in V_\varepsilon(\hat{p})$. Herein, $(d(p), \lambda(p), \mu(p))$ is continuously differentiable w.r.t. p.

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

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

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

We can apply the Lagrange-Newton method and under the assumptions (i)-(v) Theorem 3.8.3 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

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

(c) Let δ denote the radius of convergence of the Lagrange-Newton method. Let $r := \min\{\varepsilon, \delta\}$. For $p^{(0)} \in U_r(\hat{p})$ all subsequent iterates $p^{(k)}$ 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 $QP(p^{(k)})$, cf. (3.8.5) and (3.8.6). According to (a), the solution $(d(p^{(k)}), \lambda(p^{(k)}), \mu(p^{(k)}))$ of $QP(p^{(k)})$ is unique. Hence, the SQP iteration coincides with the Lagrange-Newton iteration.

■

Remark 3.8.8 (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 [Pow78] suggested to use the modified BFGS-update formula

$$B_{k+1} = B_k + \frac{q^{(k)}(q^{(k)})^\top}{(q^{(k)})^\top s^{(k)}} - \frac{B_k s^{(k)}(s^{(k)})^\top B_k}{(s^{(k)})^\top B_k s^{(k)}}, \tag{3.8.7}$$

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.

3.8.3 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 Problem 3.6.1 with $S = \mathbb{R}^{n_x}$. 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* or *merit 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.

Often, one of the following merit functions is used for globalization:

- The non-differentiable ℓ_1-penalty function

$$\ell_1(x;\alpha) := f(x) + \alpha \sum_{i=1}^{n_g} \max\{0, g_i(x)\} + \alpha \sum_{j=1}^{n_h} |h_j(x)|$$

  was used by, e.g., Powell [Pow78].

- A commonly used differentiable penalty function is the *augmented Lagrange function*

$$
\begin{aligned}
L_a(x,\lambda,\mu;\alpha) &= f(x) + \mu^\top h(x) + \frac{\alpha}{2}\|h(x)\|^2 \\
&\quad + \frac{1}{2\alpha}\sum_{i=1}^{n_g}\left((\max\{0,\lambda_i+\alpha g_i(x)\})^2 - \lambda_i^2\right) \\
&= f(x) + \sum_{j=1}^{n_h}\left(\mu_j h_j(x) + \frac{\alpha}{2}h_j(x)^2\right) \qquad (3.8.8)\\
&\quad + \sum_{i=1}^{n_g}\begin{cases} \lambda_i g_i(x) + \frac{\alpha}{2}g_i(x)^2, & \text{if } \lambda_i + \alpha g_i(x) \geq 0, \\ -\frac{\lambda_i^2}{2\alpha}, & \text{otherwise.}\end{cases}
\end{aligned}
$$

  A SQP method employing the augmented Lagrange function is discussed in [Sch81], [Sch83].

Both functions are exact under suitable assumptions, i.e. there exists a finite parameter $\hat{\alpha} > 0$, such that every local minimum $\hat{x}$ of Problem 3.6.1 with $S = \mathbb{R}^{n_x}$ is also a local minimum of the penalty function for all $\alpha \geq \hat{\alpha}$.
Without specifying all details, a globalized version of the SQP method employing the ℓ_1 penalty function and an Armijo rule for step length determination reads as follows.

Algorithm 3.8.9 (Globalized SQP Method)

(i) *Choose* $(x^{(0)},\lambda^{(0)},\mu^{(0)}) \in \mathbb{R}^{n_x}\times\mathbb{R}^{n_g}\times\mathbb{R}^{n_h}$, $B_0 \in \mathbb{R}^{n_x\times n_x}$ *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 Problem 3.6.1 with* $S = \mathbb{R}^{n_x}$, *STOP.*

(iii) *Compute a KKT point* $(d^{(k)},\lambda^{(k+1)},\mu^{(k+1)}) \in \mathbb{R}^{n_x}\times\mathbb{R}^{n_g}\times\mathbb{R}^{n_h}$ *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*

$$\ell_1(x^{(k)} + t_k d^{(k)};\alpha) \leq \ell_1(x^{(k)};\alpha) + \sigma t_k \ell_1'(x^{(k)};d^{(k)};\alpha).$$

(v) *Compute* B_{k+1} *according to (3.8.7) and set* $x^{(k+1)} := x^{(k)} + t_k d^{(k)}$.

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

Remark 3.8.10

- *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 [Sch83] and [Pow78].*

- *So far, we always assumed that the QP problem has a solution. This assumption is not always justified, even if the original problem is feasible. To overcome this problem, Powell [Pow78] suggested to relax the constraints of the QP problem in such a way, that the relaxed QP problem possesses admissible points. The infeasible QP problem is then replaced by the feasible relaxed problem. A convergence analysis for a SQP method using the augmented Lagrangian can be found in Schittkowski [Sch81, Sch83].*

3.8.4 Nonsmooth Newton's Method for Quadratic Programs

One approach to solve the quadratic program is to employ primal or dual active set methods as in Gill et al. [GM78, GMSW91] or Goldfarb and Idnani [GI83]. The alternative approach of Lemke [Lem62] uses pivot strategies to solve the linear complementarity problem resulting from the KKT conditions. We will discuss a third strategy.

For simplicity we consider quadratic programs (QP) of type

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}x^\top Qx + c^\top x \qquad \text{s.t.} \qquad g(x) := Ax - u \leq 0_m,$$

where $Q \in \mathbb{R}^{n\times n}$ is symmetric and positive definite, $A \in \mathbb{R}^{m\times n}$, $c \in \mathbb{R}^n$, and $u \in \mathbb{R}^m$. Additional equality constraints can be easily added and will not effect the subsequent analysis.

Let x^* be a local minimum of QP. Then, the first order necessary Karush-Kuhn-Tucker (KKT) conditions read as follows: There exists a multiplier $\lambda \in \mathbb{R}^m$ such that

$$Qx + c + A^\top\lambda = 0_n, \quad \lambda \geq 0, \quad \lambda^\top g(x) = 0, \quad g(x) \leq 0_m. \tag{3.8.9}$$

The idea is to restate the KKT conditions (3.8.9) as the equivalent nonlinear equation

$$F(z) = 0_{n+m} \tag{3.8.10}$$

with variables $z = (x, \lambda)^\top$ and

$$F(z) := \begin{pmatrix} Qx + c + A^\top\lambda \\ \varphi(-g_1(x), \lambda_1) \\ \vdots \\ \varphi(-g_m(x), \lambda_m) \end{pmatrix} = 0_{n+m}. \tag{3.8.11}$$

Herein, the function $\varphi : \mathbb{R}^2 \to \mathbb{R}$ is a so-called *NCP function* with the property

$$\varphi(a, b) = 0 \qquad \Leftrightarrow \qquad a \geq 0,\ b \geq 0,\ a \cdot b = 0.$$

Obviously, the KKT conditions (3.8.9) are satisfied if and only if (3.8.11) holds. It is easy to check that the convex *Fischer-Burmeister function*

$$\varphi(a, b) := \sqrt{a^2 + b^2} - a - b.$$

is a particular NCP function. Unfortunately, the Fischer-Burmeister function φ is not differentiable at the origin $(a, b) = (0, 0)$. Hence, the function F in (3.8.10) is not differentiable at points with $(g_i(x), \lambda_i) = (0, 0)$ and the common Newton's method is not applicable to the nonlinear equation (3.8.11). In the sequel we will discuss a nonsmooth version of Newton's method adapted to the particular equation in (3.8.11). Nonsmooth Newton's method for general equations are discussed in Qi [Qi93], Qi and Sun [QS93], Jiang [Jia99], Xu and Glover [XG97], Xu and Chang [XC97], Han et al. [HPR92], Ralph [Ral94], and Dingguo and Weiwen [DW02]. Similar ideas are also used to solve nonlinear complementarity problems, cf. Fischer [Fis97], Facchinei and Kanzow [FK97], Yamashita and Fukushima [YF97] and Billups and Ferris [BF97].

The components of F are convex functions (g_i is linear, φ is convex) and thus F is locally Lipschitz continuous. According to Rademacher's theorem, F is differentiable almost everywhere and we may define the *B(ouligand)-differential*

$$\partial_B F(z) := \left\{ V \,\middle|\, V = \lim_{\substack{z_i \in D_F \\ z_i \to z}} F'(z_i) \right\},$$

where D_F denotes the set of points where F is differentiable. Notice, that Clarke's [Cla83] generalized Jacobian ∂F is the convex hull of the B-differential:

$$\partial F(z) := \operatorname{conv}(\partial_B F(z)).$$

The generalized Jacobian $\partial F(z)$ is a non-empty, convex, and compact set and as a function of z it is upper semicontinuous at z, cf. Clarke [Cla83], Prop. 2.6.2, p. 70.

Definition 3.8.11 (Qi [Qi93])
Let $F : \mathbb{R}^n \to \mathbb{R}^m$.

- *∂F and $\partial_B F$, respectively, are called* non-singular *at z, if every $V \in \partial F(z)$ resp. every $V \in \partial_B F(z)$ is non-singular.*
- *F is called* B(ouligand)-differentiable at x *if it is directionally differentiable, i.e. the limit*

  $$F'(x;h) = \lim_{t\downarrow 0} \frac{F(x+th) - F(x)}{t}$$

  exists for every direction h, and it holds

  $$\lim_{h\to 0} \frac{F(x+h) - F(x) - F'(x;h)}{\|h\|} = 0$$

  for every direction h.
- *Let F be B-differentiable in some neighborhood of x. $F'(\cdot;\cdot)$ is* semicontinuous at x, *if for every $\varepsilon > 0$ there exists a neighborhood N of x such that*

  $$\|F'(x+h;h) - F'(x;h)\| \leq \varepsilon \|h\| \qquad \forall x+h \in N.$$

  $F'(\cdot;\cdot)$ is semicontinuous of degree 2 at x, *if there exist a constant L and a neighborhood N of x such that*

  $$\|F'(x+h;h) - F'(x;h)\| \leq L\|h\|^2 \qquad \forall x+h \in N.$$
- *F is called* semismooth at x, *if*

  $$\lim_{\substack{V\in\partial F(x+th')\\ h'\to h, t\downarrow 0}} Vh'$$

  exists for every $h \in \mathbb{R}^n$.

We summarize results from Qi [Qi93] and Qi and Sun [QS93]. In finite dimensional spaces Shapiro showed that a locally Lipschitz continuous function is B-differentiable if and only if it is directionally differentiable.

Lemma 3.8.12 (Qi [Qi93], Qi and Sun [QS93])
Let $F : \mathbb{R}^n \to \mathbb{R}^m$ be directionally differentiable in some neighborhood of x.

(a) The following statements are equivalent:

  - *F is semismooth at x.*
  - *$F'(\cdot;\cdot)$ is semicontinuous at x.*

- *For every* $V \in \partial F(x+h)$, $h \to 0$ *it holds*

$$Vh - F'(x;h) = o(\|h\|).$$

- $\displaystyle\lim_{\substack{x+h \in D_F \\ h \to 0}} \frac{F'(x+h;h) - F'(x;h)}{\|h\|} = 0.$

(b) The following statements are equivalent:

- $F'(\cdot;\cdot)$ *is semicontinuous of degree* 2 *at* x.
- *For every* $V \in \partial F(x+h)$, $h \to 0$ *it holds*

$$Vh - F'(x;h) = \mathcal{O}(\|h\|^2). \tag{3.8.12}$$

In either case F *is B-differentiable of degree* 2 *at* x.

In particular, convex functions and smooth functions are semismooth. Scalar products, sums, and compositions of semismooth functions are semismooth. The *min*-Operator is semismooth.

Local Convergence

First, we investigate a local nonsmooth Newton's method and its convergence.

Algorithm 3.8.13 (Local Nonsmooth Newton's Method)

(0) Choose z^0 *and set* $k = 0$.

(1) If some stopping criterion is satisfied, stop.

(2) Choose $V_k \in \partial_B F(z^k)$ *and compute the search direction* d^k *as the solution of the linear equation*

$$V_k d = -F(z^k).$$

(3) Set $z^{k+1} = z^k + d^k$, $k = k+1$, *and goto (1).*

The following convergence theorem can be found in Qi [Qi93].

Theorem 3.8.14 (Qi [Qi93], Th. 3.1)
Let z^* *satisfy (3.8.10). Let* F *be locally Lipschitz continuous and semismooth at* z^* *and let* $\partial_B F(z^*)$ *be nonsingular. Then, there exists some* $r > 0$ *such that for any* $z^0 \in U_r(z^*)$ *Algorithm 3.8.13 is well-defined and the sequence* $\{z^k\}$ *converges superlinearly to* z^*.
If $F(z^k) \neq 0$ *for all* k, *then*

$$\lim_{k \to \infty} \frac{\|F(z^{k+1})\|}{\|F(z^k)\|} = 0.$$

If in addition F *is directionally differentiable in a neighborhood of* z^* *and* $F'(\cdot;\cdot)$ *is semicontinuous of degree* 2 *at* z^*, *then the convergence is quadratic.*

Qi and Sun [QS93], Th. 3.2 proved a similar result, if V_k is chosen from $\partial F(z^k)$ in step (2) of Algorithm 3.8.13.

Global Convergence

One reason that makes the Fischer-Burmeister function appealing is the fact that its square

$$\psi(a,b) := \varphi(a,b)^2 = \left(\sqrt{a^2+b^2} - a - b\right)^2$$

is continuously differentiable. This allows to globalize the local Newton's method by introducing the merit function

$$\begin{aligned}\Psi(z) &:= \frac{1}{2}\|F(z)\|^2 = \frac{1}{2}F(z)^\top F(z) = \frac{1}{2}\sum_{i=1}^{n+m} F_i(z)^2 \\ &= \frac{1}{2}\left(\|Qx + c + A^\top\lambda\|^2 + \sum_{i=1}^{m}\psi(-g_i(x),\lambda_i)\right).\end{aligned}$$

Ψ is continuously differentiable and it holds

$$\nabla\Psi(z) = V^\top F(z),$$

where V is an arbitrary element of $\partial F(z)$.

Algorithm 3.8.15 (Globalized Nonsmooth Newton's Method)

(0) Choose z^0, $\beta \in (0,1)$, $\sigma \in (0,1/2)$ and set $k = 0$.

(1) If some stopping criterion is satisfied, stop.

(2) Compute the search direction d^k as the solution of the linear equation

$$V_k d = -F(z^k),$$

where $V_k \in \partial F(z^k)$.

(3) Find the smallest $i_k \in \mathbb{N}_0$ with

$$\Psi(z^k + \beta^{i_k} d^k) \leq \Psi(z^k) + \sigma\beta^{i_k}\nabla\Psi(z^k)^\top d^k$$

and set $\alpha_k = \beta^{i_k}$.

(4) Set $z^{k+1} = z^k + \alpha_k d^k$, $k = k+1$, and goto (1).

The following result is Theorem 4.1 in Jiang [JQ97].

Theorem 3.8.16 *Let the assumptions of Theorem 3.8.14 be satisfied. Let the linear equation in step (2) be solvable and let z^* be an accumulation point of the sequence $\{z^k\}$ generated by Algorithm 3.8.15.*
Then:

(i) z^* is a zero of F if $\{d^k\}$ is bounded.

(ii) z^* is a zero of F and $\{z^k\}$ converges to z^* superlinearly if $\sigma \in (0,1/2)$ and if $\partial F(z^*)$ is nonsingular.

There are two assumptions that ought to be discussed. The first one is the nonsingularity of $\partial F(z)$. The second one is the existence of an accumulation point. An accumulation point exists, if the sequence $\{z^k\}$ is bounded. This would be the case, if the level set

$$\{z \mid \Psi(z) \leq \Psi(z^0)\}$$

is bounded. Unfortunately, the question of whether the level set is bounded or not remains open. The problem is, that any point (x,λ) satisfying the complementarity conditions $-g_i(x) \geq 0$, $\lambda_i \geq 0$, $\lambda_i g_i(x) = 0$ yields $\varphi(-g_i(x), \lambda_i) = 0$. As a consequence, it may happen that a sequence $\{(x^k, \lambda^k)\}$ with $\|(x^k,\lambda^k)\| \to \infty$ and $\Psi(x^k, \lambda^k) < \infty$ exists.

Lemma 3.8.17 *The Fischer-Burmeister function φ is both, semismooth and semicontinuous of degree 2.*

Proof. For $(a,b) \neq (0,0)$, φ is smooth in some neighborhood and thus

$$\varphi'(a,b;h_1,h_2) = \varphi'(a,b)h.$$

For $(a,b) = (0,0)$, φ is directionally differentiable with

$$\varphi'(0,0;h_1,h_2) = \lim_{t \downarrow 0} \frac{\varphi(th_1,th_2) - \varphi(0,0)}{t} = \lim_{t\downarrow 0} \frac{\sqrt{t^2h_1^2 + t^2h_2^2} - th_1 - th_2}{t} = \varphi(h_1,h_2).$$

Semismoothness: We show the equivalent characterizations

$$Vh - \varphi'(a,b;h_1,h_2) = o(\|h\|) \qquad \forall V \in \partial\varphi(a+h_1, b+h_2) \text{ as } h \to 0$$

for semismoothness and

$$Vh - \varphi'(a,b;h_1,h_2) = \mathcal{O}(\|h\|^2) \qquad \forall V \in \partial\varphi(a+h_1, b+h_2) \text{ as } h \to 0.$$

for semicontinuity of degree 2.
Let $(a,b) \neq (0,0)$. Then since φ is twice continuously differentiable in a neighborhood of (a,b):

$$\begin{aligned} |Vh - \varphi'(a,b;h_1,h_2)| &= |\varphi'(a+h_1,b+h_2)h - \varphi'(a,b)h| \\ &\leq \left|\varphi'(a+h_1,b+h_2) - \varphi'(a,b)\right| \|h\| \\ &\leq L\|h\|^2. \end{aligned}$$

Let $(a,b) = (0,0)$ and $h \neq 0$. Then

$$Vh - \varphi'(0,0;h_1,h_2) = \varphi'(h_1,h_2)h - \varphi(h_1,h_2) = \frac{h_1^2}{\|h\|} - h_1 + \frac{h_2^2}{\|h\|} - h_2 - \|h\| + h_1 + h_2 = 0.$$

This proves the assertion. ■

The chain law for directionally differentiable functions reads as follows, cf. Geiger and Kanzow [GK02], Theorem 5.33, p. 251: Given three functions $h : \mathbb{R}^n \to \mathbb{R}^m$, $g : \mathbb{R}^m \to \mathbb{R}^p$, and $f = g \circ h$, where h and g are directionally differentiable at x resp. $h(x)$ and g is locally Lipschitz continuous at $h(x)$. Then f is directionally differentiable at x with

$$f'(x;d) = g'(h(x);h'(x;d)).$$

Hence, application of this result to our situation yields that F is also semismooth and F' is semicontinuous of degree 2.

Nonsingularity of Subdifferential

We first start with the investigation of the subdifferential of φ. If $(a,b) \neq (0,0)$, then φ is differentiable and

$$\varphi'(a,b) = \left(\frac{a}{\sqrt{a^2+b^2}} - 1, \frac{b}{\sqrt{a^2+b^2}} - 1 \right).$$

In order to obtain the B-subdifferential we have to investigate all limits of $\varphi'(a,b)$ for $(a,b) \to (0,0)$. It holds

$$\|\varphi'(a,b) + (1,1)\| \leq 1 \qquad \forall (a,b) \neq (0,0)$$

and thus

$$\varphi'(a,b) \in M := \{(s,t) \mid (s+1)^2 + (t+1)^2 \leq 1\} \qquad \forall (a,b) \neq (0,0).$$

Consequently,

$$\partial_B \varphi(0,0) \subseteq \partial \varphi(0,0) \subseteq M.$$

On the other hand, using the sequences $(a_i, b_i) = \frac{1}{i}(\cos\alpha, \sin\alpha)$ with arbitrary $\alpha \in [0, 2\pi)$ it follows

$$\lim_{i\to\infty} \varphi'(a_i, b_i) = (\cos\alpha - 1, \sin\alpha - 1).$$

Hence,

$$\partial \varphi(0,0) = M.$$

Any element from $\partial_B \varphi(0,0)$ resp. $\partial\varphi(0,0)$ can be represented as (s,t) with $(s+1)^2+(t+1)^2 \leq 1$. On the other hand, we did not show that every such (s,t) is actually an element of $\partial_B\varphi(0,0)$. Summarizing, we obtained

$$\partial\varphi(a,b) = \begin{cases} \left\{ \left(\frac{a}{\sqrt{a^2+b^2}} - 1, \frac{b}{\sqrt{a^2+b^2}} - 1 \right) \right\}, & \text{if } (a,b) \neq (0,0), \\ \left\{ (s,t) \mid (s+1)^2 + (t+1)^2 \leq 1 \right\}, & \text{if } (a,b) = (0,0). \end{cases}$$

For the forthcoming computations we used the particular subgradient with $s = 0, t = 1$, which is an element of $\partial_B\varphi(0,0)$.

The first n components of F in (3.8.11) are continuously differentiable. The last m components of F in (3.8.11) are composite functions of the locally Lipschitz continuous function φ and the affine function $(x,\lambda) \mapsto \begin{pmatrix} -g_i(x) \\ \lambda_i \end{pmatrix}$. Hence we need a chain rule for differentiation. Application of Theorem 2.6.6 in Clarke [Cla83] yields

$$\partial F_{n+i}(x,\lambda) \subseteq \operatorname{co}\left\{ \partial\varphi(-g_i(x), \lambda_i) \cdot \begin{pmatrix} -a_i & 0 \\ 0 & e_i \end{pmatrix} \right\}$$

for $i = 1, \ldots, m$, where a_i denotes the i-th row of A and e_i the i-th unity vector of dimension $\mathbb{R}^m$. The convex hull can be omitted since from convex analysis it is well-known that the product of a convex set by a matrix is again a convex set, i.e. if C is convex and A is a matrix of appropriate dimension, then $C \cdot A$ is convex. Hence, for $i = 1, \ldots, m$ it holds

$$\partial F_{n+i}(x,\lambda) \subseteq \partial\varphi(-g_i(x), \lambda_i) \cdot \begin{pmatrix} -a_i & 0 \\ 0 & e_i \end{pmatrix}.$$

Using Proposition 2.6.2 of Clarke [Cla83] we finally find

$$\partial F(x,\lambda) \subseteq \begin{pmatrix} Q & A^\top \\ -SA & T \end{pmatrix},$$

where $S := \text{diag}(s_1,\ldots,s_m)$, $T := \text{diag}(t_1,\ldots,t_m)$, and $(s_i,t_i) \in \partial\varphi(-g_i(x),\lambda_i)$. In particular,

$$(s_i+1)^2 + (t_i+1)^2 \leq 1,$$

cf., eg., Jiang [Jia99]. Now, we will show that any such matrix is nonsingular, if Q is positive definite and if the constraints do not contain redundant information. Define

$$\alpha(x) := \{i \in \{1,\ldots,m\} \mid g_i(x) = 0\}.$$

Theorem 3.8.18 *Let Q be positive definite and let the constraints $Ax \leq u$ be regular in the sense that for any $x \in \mathbb{R}^n$ the rows $\{a_i \mid i \in \alpha(x)\}$ are linearly independent. Then, $\partial F(x,\lambda)$ is nonsingular for any (x,λ).*

Proof. Any element from $\partial F(x,\lambda)$ can be represented by

$$V := \begin{pmatrix} Q & A^\top \\ -SA & T \end{pmatrix},$$

where $S := \text{diag}(s_1,\ldots,s_m)$, $T := \text{diag}(t_1,\ldots,t_m)$, and

$$(s_i+1)^2 + (t_i+1)^2 \leq 1,$$

Define

$$\begin{aligned} I &:= \{i \in \{1,\ldots,m\} \mid s_i = 0\}, \\ J &:= \{i \in \{1,\ldots,m\} \mid t_i = 0\}, \\ K &:= \{i \in \{1,\ldots,m\} \mid i \notin I \cup J\}. \end{aligned}$$

Notice, that $I \cap J = \emptyset$ because of $(s_i+1)^2 + (t_i+1)^2 \leq 1$. In particular it holds $t_i \neq 0$ for $i \in I$, $s_i \neq 0$ for $i \in J$, and $s_i, t_i \in [-2,0)$ for $i \in K$.

In order to show the nonsingularity of V we show that the linear equation $V \begin{pmatrix} u \\ v \end{pmatrix} = 0_{n+m}$ only admits the trivial solution. So, let

$$\begin{aligned} Qu + A^\top v &= 0_n, \\ -SAu + Tv &= 0_m \end{aligned}$$

resp.

$$\begin{aligned} Qu + \sum_{i\in I} a_i^\top v_i + \sum_{j \in J} a_j^\top v_j + \sum_{k\in K} a_k^\top v_k &= 0_{n+m}, \\ -s_i a_i u + t_i v_i &= 0, \qquad i \in I, \\ -s_j a_j u + t_j v_j &= 0, \qquad j \in J, \\ -s_k a_k u + t_k v_k &= 0, \qquad k \in K. \end{aligned}$$

Since $s_i = 0, t_i \neq 0$ for $i \in I$ the second equality immediately yields $v_i = 0$ for $i \in I$. Similarly, the third equation yields $a_j u = 0$ for $j \in J$ because $s_j \neq 0, t_j = 0$ for $j \in J$. In particular, $v_j a_j u = 0$, $j \in J$.

Multiplication of the last equation by v_k yields

$$\frac{t_k}{s_k} v_k^2 = v_k a_k u,$$

where $t_k/s_k > 0$. Recall that $s_k, t_k \in [-2, 0)$.
Multiplication of the first equation and exploitation of the above relations yields

$$0 = u^\top Q u + \sum_{i \in I} u^\top a_i^\top v_i + \sum_{j \in J} u^\top a_j^\top v_j + \sum_{k \in K} u^\top a_k^\top v_k = u^\top Q u + \sum_{k \in K} \frac{t_k}{s_k} v_k^2.$$

Due to the positive definiteness of Q this equality only holds for $u = 0$ and $v_k = 0$, $k \in K$.
By now, we obtained $u = 0$, $v_i = 0$, $i \in I$, and $v_k = 0$, $k \in K$. With this, again the first equation yields

$$\sum_{j \in J} a_j^\top v_j = 0.$$

Recall that $j \in J$ means $t_j = 0$ and $(s_j, t_j) \in \partial\varphi(-g_j(x), \lambda_j)$ and $(s_j + 1)^2 \leq 1$. The value $t_j = 0$ is only possible, if $g_j(x) = a_j x - u_j = 0$ holds, i.e. $j \in \alpha(x)$. According to our assumption the set $\{a_j \mid j \in \alpha(x)\}$ is linearly independent. Thus, $v_j = 0$, $j \in J$. This completes the proof. ■

Example 3.8.19 We will test the performance of the nonsmooth Newton's method on two randomly generated test sets with strictly convex objective function. In addition, we will compare the results in view of iterations needed with those obtained by the active set method QUADPROG from MATLAB. It has to be mentioned, that the respective methods approximately need the same amount of time per iteration, provided that the structure of the occurring linear equations is exploited.

- The first test set consists of 941 small scale problems. The problem sizes range from 2 to 20 variables and from 1 to 20 constraints.
- The second test set consists of 261 medium to large scale problems. The problem sizes range from 2 to 452 variables and from 1 to 801 constraints.

Table 3.1 summarizes the results of the active set method and the nonsmooth Newton's method for the first test set with few constraints. The active set method always needed less iterations.

Table 3.1: Primal active set method vs. nonsmooth Newton's method: Iterations for problems with few constraints.

method	average	min	max
active set	5	1	25
newton	8	2	37

Table 3.2 summarizes the results of the active set method and the nonsmooth Newton's method for the second test set with many constraints. The nonsmooth Newton's method performs significantly better for many constraints than the active set method. This suggests that the nonsmooth Newton's method might be superior for problems with many constraints.

Table 3.2: Primal active set method vs. nonsmooth Newton's method: Iterations for problems with many constraints.

method	average	min	max
active set	197	1	987
newton	12	2	56

Example 3.8.20 (Discretized Elliptic PDE Optimal Control Problem)
We consider the subsequent elliptic PDE Optimal Control Problem with an objective function of tracking type:

$$\min_{y,u\in L^2(\Omega)} \frac{1}{2}\|y - y_d\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u\|^2_{L^2(\Omega)}$$

subject to the elliptic PDE

$$\begin{array}{rcll} -\Delta y &=& \beta \cdot u & \text{in } \Omega := (0,1)\times(0,1),\\ y &=& 0 & \text{on } \Gamma = \partial\Omega \end{array}$$

and the box constraints

$$b_\ell(x) \leq u(x) \leq b_u(x).$$

The problem is discretized using the 5-point star approximation

$$-\Delta y(x_{ij}) \approx \frac{4y_{ij} - y_{i-1,j} - y_{i,j-1} - y_{i+1,j} - y_{i,j+1}}{h^2}$$

on the equidistant grid $x_{ij} = (ih, jh)$, $i,j = 0,\ldots,N$, $h = 1/N$ with approximations $y_{ij} \approx y(x_{ij})$, $u_{ij} \approx u(x_{ij})$. The integral is approximated by

$$\int_\Omega y(x)dx \approx h^2 \sum_{ij} y_{ij}.$$

The discretized problem reads as

$$\min \frac{1}{2}\sum_{i,j=1}^{N-1}\Big((y_{ij} - y_d(x_{ij}))^2 + \alpha u_{ij}^2\Big)$$

subject to

$$A_h y = B_h u, \qquad b_\ell(x_{ij}) \leq u_{ij} \leq b_u(x_{ij}),$$

where A_h is a banded matrix of bandwidth N with approximately $5/(N-1)\%$ of nonzero elements. B_h is a diagonal matrix with entries $\beta(x_{ij})$.
The discretized problem is equivalent to the quadratic program

$$\min_z \frac{1}{2}z^\top Q z + c^\top z \quad \text{s.t.} \quad A_{eq}z = 0,\ A_{in}z \leq b$$

where

$$Q = \begin{pmatrix} I & \Theta \\ \Theta & \alpha I \end{pmatrix}, \qquad A_{eq} = \begin{pmatrix} A_h & -B_h \end{pmatrix}, \qquad A_{in} = \begin{pmatrix} \Theta & I \\ \Theta & -I \end{pmatrix}$$

and $z = (y_{ij}, u_{ij})^\top \in \mathbb{R}^{2(N-1)^2}$, $c = (-y_d(x_{ij}), 0, \ldots, 0)^\top$, $b = (b_u(x_{ij}), -b_l(x_{ij})))\top$. The matrices Q, A_{eq}, and A_{in} are large but sparse. Hence, instead of treating the Newton equation

$$V_k d = -F(x^k)$$

by dense lu-decomposition a conjugate gradient method is applied. Actually, we solved

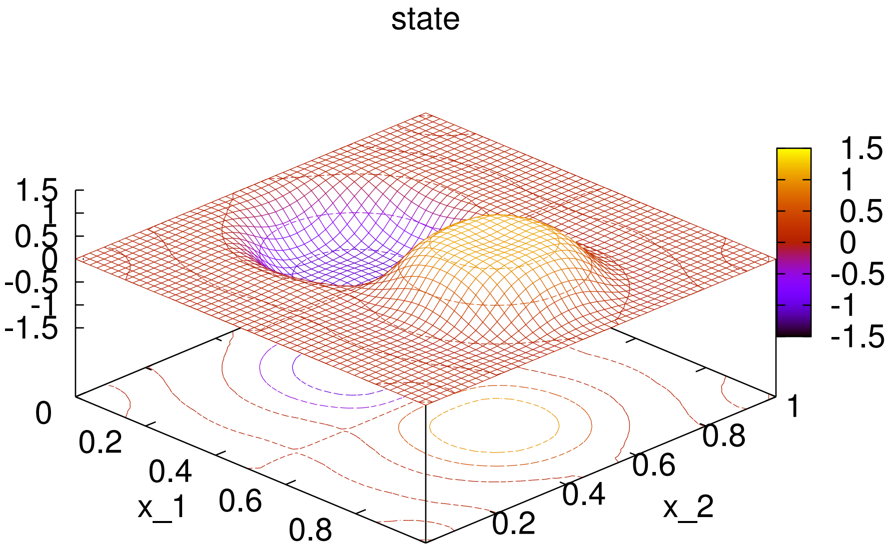
$$V_k^\top V_k d = -V_k^\top F(x^k)$$

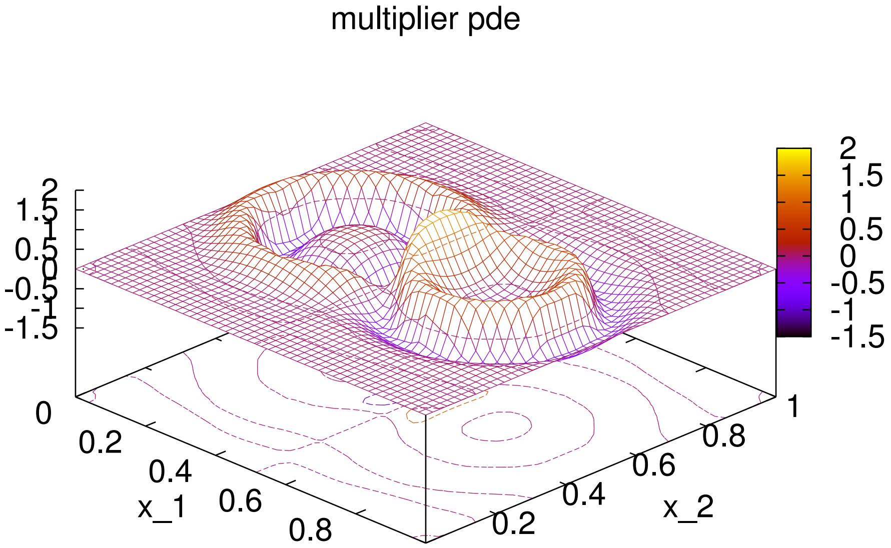
using the cg-method. Since V_k is nonsingular, both equations are equivalent but $V_k^\top V_k$ is positive definite.

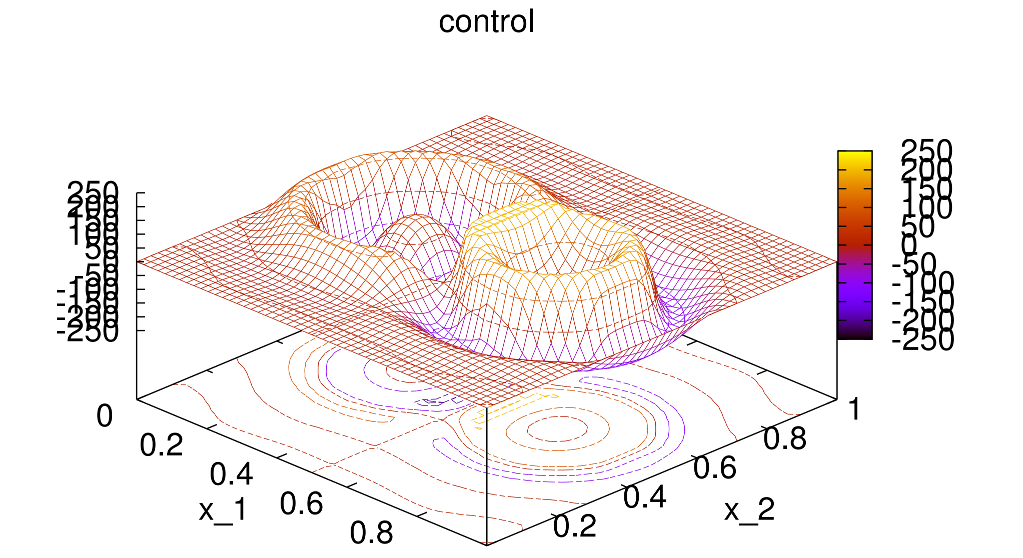
The following numerical results are obtained for the data $\beta \equiv 1$, $b_\ell \equiv -200$, $b_u \equiv 200$, $\alpha = 10^{-6}$, $N = 60$, and

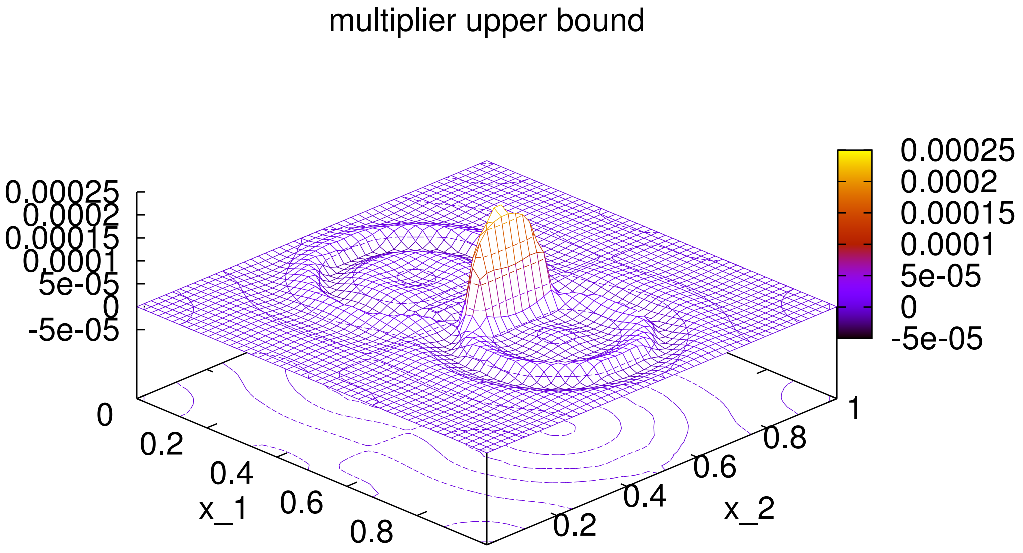
$$y_d(x_1, x_2) = \begin{cases} 1, & \text{if } (x_1 - 0.7)^2 + (x_2 - 0.5)^2 \leq 0.2^2, \\ -1, & \text{if } (x_1 - 0.3)^2 + (x_2 - 0.5)^2 \leq 0.2^2, \\ 0, & \text{otherwise} \end{cases}$$

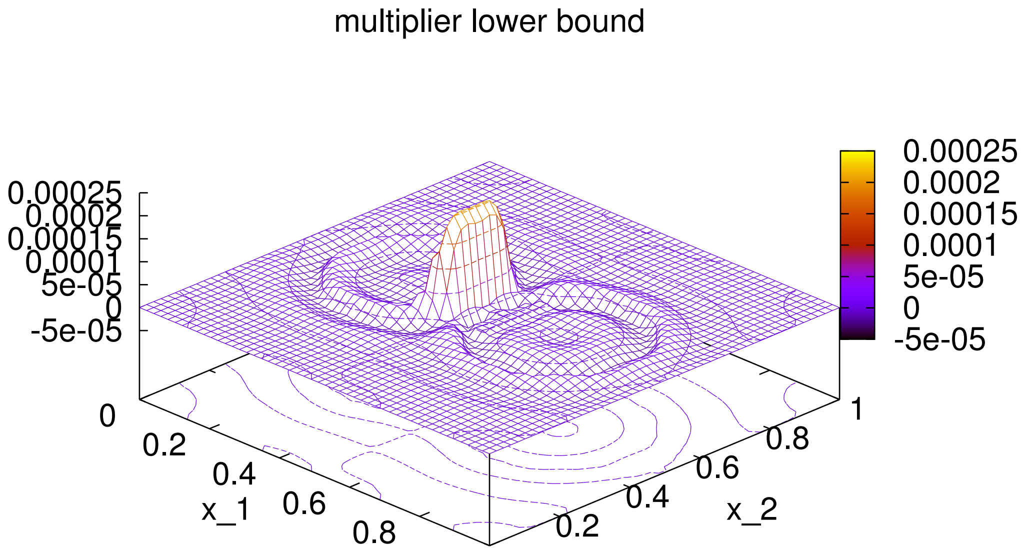
The resulting QP problem has 6962 variables, 3481 equality constraints, and 6962 inequality constraints. The nonsmooth Newton's method has 17405 equations, 209045 nonzero elements in the Jacobian, i.e. ≈ 0.07 % nonzero elements.






control
250
200
150
100
50
0
-50
-100
-150
-200
-250
0
0.2
0.4
x_1
0.6
0.8
0.2
0.4
0.6
x_2
0.8
1
250
200
150
100
50
0
-50
-100
-150
-200
-250


multiplier upper bound
0.00025
0.0002
0.00015
0.0001
5e-05
0
-5e-05
0
0.2
0.4
x_1
0.6
0.8
0.2
0.4
0.6
x_2
0.8
1
0.00025
0.0002
0.00015
0.0001
5e-05
0
-5e-05



3.9 Duality

We consider Problem 3.6.1 and refer to it as the *primal optimization problem.* In contrast to the previous discussions, we do not assume that the functions f, g, and h are continuous or even differentiable. Similarly, the set $S \subseteq \mathbb{R}^{n_x}$ may be an arbitrary non-empty set.
In order to derive the dual problem we consider the

Problem 3.9.1 (Perturbed Optimization Problem)
Find $x \in \mathbb{R}^{n_x}$ such that $f(x)$ is minimized subject to the constraints

$$\begin{array}{rcll} g_i(x) & \leq & y_i, & i = 1, \ldots, n_g, \\ h_j(x) & = & z_j, & j = 1, \ldots, n_h, \\ x & \in & S. & \end{array}$$

Herein, $y = (y_1, \ldots, y_{n_g})^\top \in \mathbb{R}^{n_g}$ and $z = (z_1, \ldots, z_{n_h})^\top \in \mathbb{R}^{n_h}$ are arbitrary vectors. Notice, that the original (unperturbed) problem arises for $y = 0_{n_g}$ and $z = 0_{n_h}$.

Definition 3.9.2 (Minimum Value Function)
The function

$$\Phi : \mathbb{R}^{n_g+n_h} \to \bar{\mathbb{R}} := \mathbb{R} \cup \{\infty, -\infty\}$$

with

$$\Phi(y, z) := \inf\{f(x) \mid g_i(x) \leq y_i,\ i = 1, \ldots, n_g,\ h_j(x) = z_j,\ j = 1, \ldots, n_h,\ x \in S\}$$

is called minimum value function *for the perturbed problem.*

The dual problem is motivated by the task of approximating the graph of the minimum value function Φ from below by a hyperplane

$$\lambda^\top y + \mu^\top z + r = \gamma \tag{3.9.1}$$

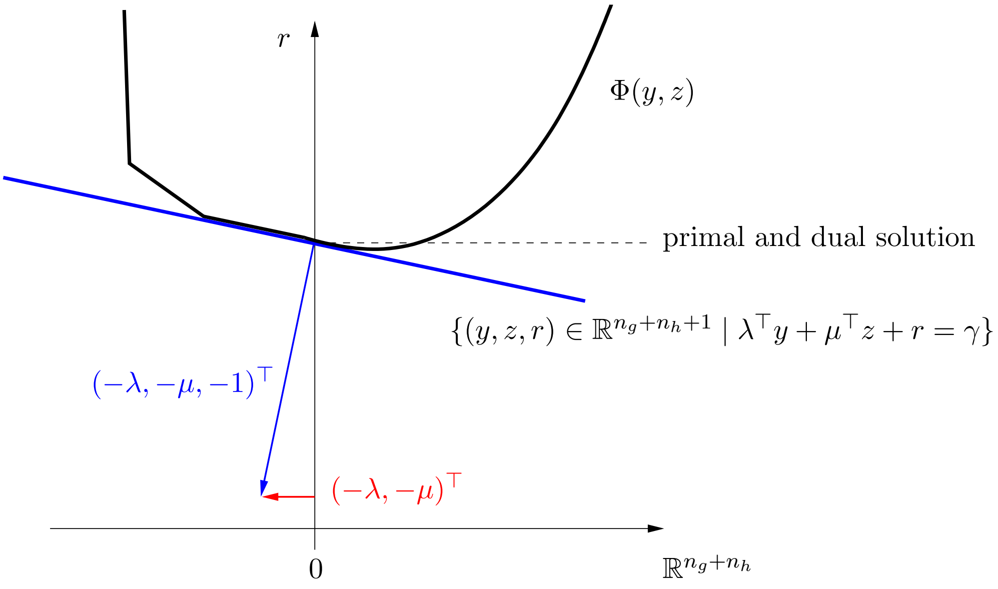

r
$\Phi(y,z)$
primal and dual solution
$\{(y,z,r) \in \mathbb{R}^{n_g+n_h+1} \mid \lambda^\top y + \mu^\top z + r = \gamma\}$
$(-\lambda,-\mu,-1)^\top$
$(-\lambda,-\mu)^\top$
0
$\mathbb{R}^{n_g+n_h}$

Figure 3.4: Graphical interpretation of the dual problem: Approximation of the graph of the minimum value function from below by a hyperplane with normal vector $(\lambda,\mu,1)^\top$ respectively $(-\lambda,-\mu,-1)^\top$.

with normal vector $(\lambda,\mu,1)^\top \in \mathbb{R}^{n_g+n_h+1}$ and variables $(y,z,r)^\top \in \mathbb{R}^{n_g+n_h+1}$, compare Figure 3.4. The intersection of this hyperplane with the r-axis at $y = 0_{n_g}, z = 0_{n_h}$ is $(0_{n_g}, 0_{n_h}, \gamma)^\top$. The dual problem is given by

$$\begin{array}{lll} \text{Maximize} & \gamma & \\ \text{w.r.t.} & \lambda \in \mathbb{R}^{n_g}, \mu \in \mathbb{R}^{n_h} & \\ \text{s.t.} & r \le \Phi(y,z) & \forall y \in \mathbb{R}^{n_g}, z \in \mathbb{R}^{n_h}. \end{array}$$

where r from (3.9.1) is given by $r = \gamma - \lambda^\top y - \mu^\top z$. Hence, the dual problem can be reformulated as

$$\begin{array}{lll} \text{Maximize} & \gamma & \\ \text{w.r.t.} & \lambda \in \mathbb{R}^{n_g}, \mu \in \mathbb{R}^{n_h} & \\ \text{s.t.} & \gamma - \lambda^\top y - \mu^\top z \le \Phi(y,z) & \forall y \in \mathbb{R}^{n_g}, z \in \mathbb{R}^{n_h}. \end{array}$$

By the definition of Φ we get the equivalent problem

$$\begin{array}{lll} \text{Maximize} & \gamma & \\ \text{w.r.t.} & \lambda \in \mathbb{R}^{n_g}, \mu \in \mathbb{R}^{n_h} & \\ \text{s.t.} & \gamma \le f(x) + \lambda^\top y + \mu^\top z & \forall y \in \mathbb{R}^{n_g}, z \in \mathbb{R}^{n_h},\ x \in S,\ g(x) \le y,\ h(x) = z. \end{array}$$

Since $h(x) = z$ this problem is equivalent to

$$\begin{array}{lll} \text{Maximize} & \gamma & \\ \text{w.r.t.} & \lambda \in \mathbb{R}^{n_g}, \mu \in \mathbb{R}^{n_h} & \\ \text{s.t.} & \gamma \le f(x) + \lambda^\top y + \mu^\top h(x) & \forall y \in \mathbb{R}^{n_g},\ x \in S,\ g(x) \le y. \end{array}$$

We distinguish two cases. First, we investigate the constraint

$$\gamma \le f(x) + \lambda^\top y + \mu^\top h(x) \qquad \forall y \in \mathbb{R}^{n_g},\ x \in S,\ g(x) \le y \tag{3.9.2}$$

and assume, that there exists a component $\lambda_i < 0$. Then, $\lambda_i y_i \to -\infty$ for $y_i \to \infty$ and for $y_i \to \infty$ the constraint $g_i(x) \leq y_i$ is fulfilled for $x \in S$. Thus, it follows that (3.9.2) is only fulfilled for $\gamma = -\infty$. Since γ is to be maximized this case is of no interest for the dual problem. Consequently, we can replace the condition $\lambda \in \mathbb{R}^{n_g}$ by $\lambda \geq 0_{n_g}$. For $\lambda \geq 0_{n_g}$ condition (3.9.2) is equivalent with the condition

$$\gamma \leq f(x) + \lambda^\top g(x) + \mu^\top h(x) \qquad \forall x \in S. \tag{3.9.3}$$

Since (3.9.2) holds for the particular choices $y = g(x)$, $x \in S$ we immediately obtain (3.9.3). On the other hand, if (3.9.3) holds for some $\lambda \geq 0_{n_g}$, then for any y with $g(x) \leq y$ it follows $\lambda^\top g(x) \leq \lambda^\top y$ and (3.9.3) implies (3.9.2).
Therefore, the dual problem is equivalent with

$$\begin{array}{ll} \text{Maximize} & \gamma \\ \text{w.r.t.} & \lambda \geq 0_{n_g}, \mu \in \mathbb{R}^{n_h} \\ \text{s.t.} & \gamma \leq f(x) + \lambda^\top g(x) + \mu^\top h(x) \qquad \forall x \in S. \end{array}$$

With the Lagrange function

$$L(x, \lambda, \mu) := f(x) + \lambda^\top g(x) + \mu^\top h(x)$$

and the *dual objective function*

$$\psi(\lambda, \mu) := \inf_{x \in S} L(x, \lambda, \mu)$$

we finally obtain the

Problem 3.9.3 (Dual Problem)
Find $\lambda \geq 0_{n_g}$ and $\mu \in \mathbb{R}^{n_h}$ such that $\psi(\lambda, \mu)$ is maximized, i.e. solve

$$\max_{\lambda \geq 0_{n_g}, \mu \in \mathbb{R}^{n_h}} \psi(\lambda, \mu) = \max_{\lambda \geq 0_{n_g}, \mu \in \mathbb{R}^{n_h}} \inf_{x \in S} L(x, \lambda, \mu).$$

Notice, that ψ as an infimum of affine functions is concave. Therefore, Problem 3.9.3 is a concave maximization problem.
We consider some examples.

Example 3.9.4

- Consider the primal linear optimization problem

  $$\min c^\top x \qquad \text{s.t.} \qquad Ax = b,\ x \in S = \{x \in \mathbb{R}^{n_x} \mid x \geq 0_{n_x}\}.$$

  The dual objective function is given by

  $$\begin{aligned} \psi(\lambda) &= \inf_{x \geq 0_{n_x}} \left(c^\top x + \lambda^\top (b - Ax) \right) \\ &= \inf_{x \geq 0_{n_x}} \left(c^\top - \lambda^\top A \right) x + \lambda^\top b \\ &= \begin{cases} \lambda^\top b, & \text{if } c^\top - \lambda^\top A \geq 0_{n_x}, \\ -\infty, & \text{otherwise.} \end{cases} \end{aligned}$$

  The dual problem is given by

  $$\max b^\top \lambda \qquad \text{s.t.} \qquad A^\top \lambda \leq c.$$

- Consider the primal quadratic optimization problem

  $$\min \frac{1}{2}x^\top Qx + c^\top x \qquad \text{s.t.} \qquad Ax \leq b$$

  with symmetric and positive definite matrix Q. The dual objective function is given by

  $$\psi(\lambda) = \inf_{x \in \mathbb{R}^{n_x}} \left(\frac{1}{2}x^\top Qx + c^\top x + \lambda^\top (Ax - b) \right).$$

  The infimum can be calculated: A necessary and sufficient condition is

  $$Qx + A^\top \lambda + c = 0, \quad \text{resp.} \quad x = -Q^{-1}(c + A^\top \lambda).$$

  With the abbreviations

  $$W = -AQ^{-1}A^\top, \qquad v = -b - AQ^{-1}c$$

  we obtain the dual problem

  $$\max_{\lambda \geq 0} \left(\frac{1}{2}\lambda^\top W\lambda + \lambda^\top v - \frac{1}{2}c^\top Q^{-1}c \right).$$

  The advantage of the dual problem is the simpler structure of the constraints (only sign conditions).

The following theorem can be exploited for Branch & Bound methods to construct lower bounds for the primal problem.

Theorem 3.9.5 (Weak duality theorem)
Let x be admissible for the primal problem and (λ, μ) be admissible for the dual problem. Then the minimal value $w(P)$ of the primal problem and the maximal value $w(D)$ of the dual problem satisfy

$$\psi(\lambda, \mu) \leq w(D) \leq w(P) \leq f(x).$$

Proof. It holds

$$\psi(\lambda, \mu) = \inf_{z \in S} L(z, \lambda, \mu) \leq f(x) + \underbrace{\lambda^\top}_{\geq 0_{n_g}} \underbrace{g(x)}_{\leq 0_{n_g}} + \mu^\top \underbrace{h(x)}_{=0_{n_h}} \leq f(x).$$

The left side is independent of x, while the right side is independent of λ, μ. Taking the supremum w.r.t. $\lambda \geq 0_{n_g}, \mu$ on the left and the infimum w.r.t. $x \in S, g(x) \leq 0_{n_g}, h(x) = 0_{n_h}$ on the right proves the statement. ■

If the primal objective function value for some primal feasible solution equals the dual objective function value for some dual feasible solution, then both problems are solved optimally already, because we have

Theorem 3.9.6 (Sufficient optimality criterion)
Let $\psi(\hat{\lambda}, \hat{\mu}) = f(\hat{x})$, where $\hat{x}$ is admissible for the primal problem and $(\hat{\lambda}, \hat{\mu})$ is admissible for the dual problem. Then $\hat{x}$ is optimal for the primal problem and $(\hat{\lambda}, \hat{\mu})$ is optimal for the dual problem. In addition, the complementary slackness condition *holds:*

$$\hat{\lambda}_i = 0, \ \text{if } g_i(\hat{x}) < 0 \qquad (i = 1, \ldots, n_g).$$

Proof. The first statement is a direct consequence of the weak duality theorem. Complementary slackness condition: Assume that $g_i(\hat{x}) < 0$ and $\hat{\lambda}_i > 0$ for at least one $1 \leq i \leq n_g$. Then

$$\psi(\hat{\lambda}, \hat{\mu}) = \inf_{x \in S} L(x, \hat{\lambda}, \hat{\mu}) \leq f(\hat{x}) + \underbrace{\hat{\lambda}^\top g(\hat{x})}_{<0} + \underbrace{\hat{\mu}^\top h(\hat{x})}_{=0} < f(\hat{x})$$

in contradiction to the assumption $\psi(\hat{\lambda}, \hat{\mu}) = f(\hat{x})$. ■

The following example shows, that the case $w(D) < w(P)$ in fact occurs. In this case a *duality gap* occurs, compare Figure 3.5.

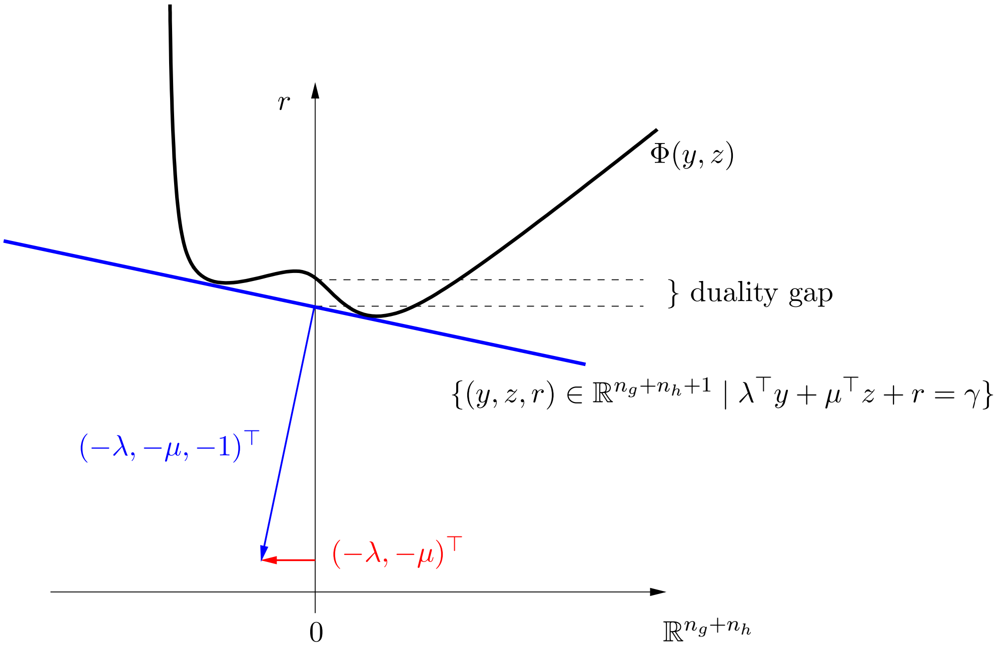

r
$\Phi(y, z)$
} duality gap
$\{(y, z, r) \in \mathbb{R}^{n_g+n_h+1} \mid \lambda^\top y + \mu^\top z + r = \gamma\}$
$(-\lambda, -\mu, -1)^\top$
$(-\lambda, -\mu)^\top$
0
$\mathbb{R}^{n_g+n_h}$

Figure 3.5: Graphical interpretation of the dual problem: Approximation of the graph of the minimum value function from below by a hyperplane with normal vector $(\lambda, \mu, 1)^\top$ respectively $(-\lambda, -\mu, -1)^\top$ in the presence of a duality gap.

Example 3.9.7 *Duality gap (cf. [BS79], p. 181)*

$$\begin{array}{ll} \textit{Minimize} & f(x_1, x_2) = -2x_1 + x_2 \\ \textit{s.t.} & (x_1, x_2) \in S = \{(0,0), (0,4), (4,4), (4,0), (1,2), (2,1)\}, \\ & 0 = h(x_1, x_2) = x_1 + x_2 - 3. \end{array}$$

The solution of the primal problem is $(2, 1)$ *with* $f(2, 1) = -3$*. The dual objective function is given by*

$$\psi(\mu) = \min\{-2x_1 + 3 + \mu(x_1 + x_2 - 3) \mid (x_1, x_2) \in S\} = \begin{cases} -4 + 5\mu, & \textit{if } \mu \leq -1, \\ -8 + \mu, & \textit{if } -1 \leq \mu \leq 2, \\ -3\mu, & \textit{if } \mu \geq 2. \end{cases}$$

The optimal solution of the dual problem $\max_{\mu \in \mathbb{R}} \psi(\mu)$ *is given by* $\mu = 2$ *with value* -6*. Since* $-6 < -3$ *there exists a duality gap.*

There remains the question under which conditions a duality gap does not occur. The following theorem states a sufficient condition.

Theorem 3.9.8 (Strong duality theorem)
Let $S \subseteq \mathbb{R}^{n_x}$ be nonempty and convex. Let the functions f and g_i, $i = 1, \dots, n_g$ be convex. Let the functions $h_j(x) = a_j^\top x - \gamma_j$, $j = 1, \dots, n_h$ be affine linear. Let $w(P)$ and $w(D)$ denote the optimal primal and dual objective function values, respectively. Let $w(P)$ be finite and let there exist $y \in relint(S)$ with

$$\begin{array}{rcll} g_i(y) & < & 0, & i = 1, \dots, n_g, \\ h_j(y) & = & 0, & j = 1, \dots, n_h \end{array}$$

*(*Slater condition*). Then the dual problem is solvable and it holds $w(P) = w(D)$.*

Proof. cf. Geiger and Kanzow [GK02], p. 323. ■

Remark 3.9.9 *The assumptions of the strong duality theorem can be weakened further. Only those inequalities, which are not affine, have to be satisfied strictly by some feasible $y \in relint(S)$, cf., e.g., Blum and Oettli [BO75].*

The advantage of the dual problem is, that the dual objective function $\psi(\lambda, \mu)$ is *concave* on its *domain*

$$dom(\psi) := \{(\lambda, \mu)^\top \in \mathbb{R}^{n_g + n_h} \mid \lambda \geq 0_{n_g},\ \psi(\lambda, \mu) > -\infty\}.$$

Hence, the problem $\min_{\lambda \geq 0_{n_g}, \mu \in \mathbb{R}^{n_h}} -\psi(\lambda, \mu)$ is a convex problem. In special cases this problem can be solved more easily than the primal problem. Furthermore, each solution of the dual problem is already a global maximum. If, in addition, the occurrence of a duality gap can be excluded, then in view of Theorem 3.9.6 the primal problem can be solved indirectly by solving the dual problem. If a duality gap cannot be excluded, then the dual problem at least provides lower bounds for the optimal primal objective function value, cf. Theorem 3.9.5.

3.10 Mixed-Integer Nonlinear Programs and Branch&Bound

So far, the variables x in the nonlinear optimization problems under consideration were real-valued and could assume any real number as value which belongs to a convex subset of $\mathbb{R}^n$ with non-empty interior. In addition to these continuous optimization variables many optimization problems of practical interest include optimization variables which are restricted to a discrete set consisting of at most countably many elements, e.g. $\{0, 1\}^n$ or $\mathbb{Z}^n$. These variables are called discrete variables. In reality, discrete variables often are used to model on/off switches, decisions, gear shifts, or discrete resource assignments.
Let sets

$$S := \{x \in \mathbb{R}^{n_x} \mid x_l \leq x \leq x_u\}, \qquad x_l, x_u \in \mathbb{R}^{n_x},\ x_l \leq x_u,$$

and

$$\mathcal{Y} := \{y \in \mathbb{Z}^{n_y} \mid y_l \leq y \leq y_u\}, \qquad y_l, y_u \in \mathbb{Z}^{n_y},\ y_l \leq y_u$$

be given. Furthermore, let the functions

$$\begin{array}{rcl} f & : & \mathbb{R}^{n_x} \times \mathcal{Y} \to \mathbb{R}, \\ g = (g_1, \dots, g_{n_g})^\top & : & \mathbb{R}^{n_x} \times \mathcal{Y} \to \mathbb{R}^{n_g}, \\ h = (h_1, \dots, h_{n_h})^\top & : & \mathbb{R}^{n_x} \times \mathcal{Y} \to \mathbb{R}^{n_h} \end{array}$$

be continuously differentiable w.r.t. the first n_x components. We consider

Problem 3.10.1 (Mixed Integer Nonlinear Program (MINLP))
Find $x \in \mathbb{R}^{n_x}$ and $y \in \mathbb{Z}^{n_y}$ such that $f(x,y)$ is minimized subject to the constraints

$$\begin{array}{rcll} g_i(x,y) & \leq & 0, & i = 1,\ldots,n_g, \\ h_j(x,y) & = & 0, & j = 1,\ldots,n_h, \\ x & \in & S, & \\ y & \in & \mathcal{Y}. & \end{array}$$

Since $\mathcal{Y}$ is a finite set, the simplest idea to solve Problem 3.10.1 would be to solve Problem 3.10.1 for all fixed choices $y \in \mathcal{Y}$. For fixed $y \in \mathcal{Y}$ Problem 3.10.1 is a common differentiable nonlinear optimization problem with optimization variables x and can be solved numerically by, e.g., SQP methods, if a solution exists at all for the particular choice y.

Just to give an idea how long this complete enumeration technique may take, consider the special case $\mathcal{Y} = \{1,2,3,4,5\}^{n_y}$ with $n_y \in \mathbb{N}$. Depending on the dimension n_y of y, the number of admissible points in $\mathcal{Y}$ is given by 5^{n_y}. Assume that each numerical solution of Problem 3.10.1 with fixed vector $y \in \mathcal{Y}$ requires one second. Then, for $n_y = 10$ this would lead to an overall solution time of $5^{10}/(3600 \cdot 24) \approx 113$ days. For $n_y = 20$ it takes approximately $3 \cdot 10^6$ years, for $n_y = 40$ even $3 \cdot 10^{20}$ years. Hence, this idea of complete enumeration only works for small combinatorial problems.

3.10.1 Branch&Bound

The Branch&Bound algorithm basically is a tree-search algorithm combined with a rule for pruning subtrees. The root of the tree corresponds to the original problem 3.10.1. All other nodes of the tree correspond to the original problem subject to additional constraints for the discrete variables. More precisely, the successor nodes are constructed in such a way that the union of all feasible sets of the successors yields the feasible set of the father's node. Hence, the optimal objective function values of the nodes are non-decreasing when traversing the tree downwards starting at the root.

For each node a lower bound of the objective function of the corresponding problem has to be determined. Usually this is done by solving a relaxed problem or by evaluating the objective function of the dual problem resp. the dual problem of the relaxed problem. If, coincidently, one finds a feasible point for Problem 3.10.1, then the respective objective function value provides an upper bound for the optimal objective function value of Problem 3.10.1 and the node is explored. Usually, this will not be the case and the value serves as a lower bound for the subtree emanating from the current node. If the lower bound is less than the upper bound provided by the best solution found so far, then all successors of the current node are generated by adding additional constraints (branching). Afterward, all successors have to be investigated in the same way. If at some node the lower bound is greater or equal than the upper bound provided by the best solution found so far, then this node needs not to be explored any further and the subtree can be pruned, since the optimal objective function values of all nodes in the subtree are greater or equal than the lower bound.

For an implementable Branch&Bound algorithm the following components have to be specified in detail:

- algorithm for determining a lower and an upper bound of the optimal objective function value of Problem 3.10.1;
- branching rule for creating new nodes;
- traversing rule for determining the order in which new nodes are generated;

- fathoming rules for pruning subtrees.

3.10.2 Bounding

Every feasible point (x, y) of Problem 3.10.1 provides an upper bound $U := f(x, y)$ for the optimal objective value of Problem 3.10.1.

A common way to determine a lower bound is to solve a relaxed problem and to use its objective function value as a lower bound. Often there is one technical difficulty involved, namely, the functions f, g, and h are only defined on $\mathbb{R}^{n_x} \times \mathcal{Y}$ and not on $\mathbb{R}^{n_x} \times \mathbb{R}^{n_y}$ as it would be necessary for the relaxation. In the sequel, we assume that there is a way to extend the domain of f, g, and h to $\mathbb{R}^{n_x} \times \mathbb{R}^{n_y}$, e.g. by some interpolation procedure or by building convex combinations, and that the functions are continuously differentiable on this space.

We concretize the term relaxation. For given vectors $y_l \leq r_l \leq r_u \leq y_u$ with $r_l, r_u \in \mathbb{Z}^{n_y}$ define

$$\mathcal{Y}(r_l, r_u) := \{y \in \mathbb{Z}^{n_y} \mid r_l \leq y \leq r_u\}$$

and

$$\hat{\mathcal{Y}}(r_l, r_u) := \{y \in \mathbb{R}^{n_y} \mid r_l \leq y \leq r_u\}$$

and consider

Problem 3.10.2 ($MINLP(r_l, r_u)$)
Find $x \in \mathbb{R}^{n_x}$ and $y \in \mathbb{Z}^{n_y}$ such that $f(x, y)$ is minimized subject to the constraints

$$\begin{aligned} g_i(x, y) &\leq 0, \quad i = 1, \ldots, n_g, \\ h_j(x, y) &= 0, \quad j = 1, \ldots, n_h, \\ x &\in S, \\ y &\in \mathcal{Y}(r_l, r_u) \end{aligned}$$

and the (continuous) nonlinear optimization problem

Problem 3.10.3 ($NLP(r_l, r_u)$)
Find $x \in \mathbb{R}^{n_x}$ and $y \in \mathbb{R}^{n_y}$ such that $f(x, y)$ is minimized subject to the constraints

$$\begin{aligned} g_i(x, y) &\leq 0, \quad i = 1, \ldots, n_g, \\ h_j(x, y) &= 0, \quad j = 1, \ldots, n_h, \\ x &\in S, \\ y &\in \hat{\mathcal{Y}}(r_l, r_u). \end{aligned}$$

Evidently,

$$\begin{aligned} \mathcal{Y} = \mathcal{Y}(y_l, y_u) &\subseteq \hat{\mathcal{Y}}(y_l, y_u), \quad (3.10.1) \\ \mathcal{Y}(r_l, r_u) &\subseteq \hat{\mathcal{Y}}(r_l, r_u). \quad (3.10.2) \end{aligned}$$

The set $\hat{\mathcal{Y}}(r_l, r_u)$ is called a *relaxation of* $\mathcal{Y}(r_l, r_u)$, because the constraint $y \in \mathbb{Z}^{n_y}$ is relaxed respectively dropped. Correspondingly, $NLP(r_l, r_u)$ is called a *relaxation of* $MINLP(r_l, r_u)$. Let f_{r_l,r_u} denote the optimal objective function value of $MINLP(r_l, r_u)$ and $\hat{f}_{r_l,r_u}$ the optimal objective function value of $NLP(r_l, r_u)$. As a consequence of (3.10.2) it holds

$$\hat{f}_{r_l,r_u} \leq f_{r_l,r_u} \quad (3.10.3)$$

and hence, $\hat{f}_{r_l,r_u}$ may serve as a lower bound for f_{r_l,r_u}.

A second way to find a lower bound for f_{r_l,r_u} is to find a feasible point of the dual problem of Problem 3.10.2. The objective function of the dual problem is given by

$$\psi(\lambda,\mu) = \inf_{x\in S,\ y\in\mathcal{Y}(r_l,r_u)} L(x,y,\lambda,\mu)$$

where

$$L(x,y,\lambda,\mu) = f(x,y) + \lambda^\top g(x,y) + \mu^\top h(x,y).$$

cf. Problem 3.9.3. According to the weak duality theorem 3.9.5 it holds

$$\psi(\lambda,\mu) \leq f_{r_l,r_u}$$

for all dual feasible points $\lambda \geq 0_{n_g}$ and $\mu \in \mathbb{R}^{n_h}$. Notice, that the evaluation of the dual objective function ψ requires to solve a mixed integer nonlinear program as well. Since this problem is usually as hard as the primal problem, instead we consider the dual problem of the relaxed problem 3.10.3. This problem possesses the objective function

$$\hat{\psi}(\lambda,\mu) = \inf_{x\in S,\ y\in\hat{\mathcal{Y}}(r_l,r_u)} L(x,y,\lambda,\mu).$$

According to the weak duality theorem 3.9.5 and (3.10.3) it holds

$$\hat{\psi}(\lambda,\mu) \leq \hat{f}_{r_l,r_u} \leq f_{r_l,r_u}$$

for all dual feasible points $\lambda \geq 0_{n_g}$ and $\mu \in \mathbb{R}^{n_h}$. Hence, $\hat{\psi}(\lambda,\mu)$ actually is a lower bound whenever λ and μ are dual feasible.

Evaluating $\hat{\psi}$ corresponds to solving a continuous optimization problem with 'simple' constraints $x \in S$, $y \in \hat{\mathcal{Y}}(r_l,r_u)$ only, which usually can be easier solved than the relaxed problem itself (often S imposes only box constraints).

In the sequel we concentrate on the relaxation technique. Nevertheless, all methods are applicable for the dual approach as well and can be adapted in a straightforward way.

3.10.3 Branching Rule

We focus on how to create new nodes. Each node corresponds to a mixed-integer nonlinear program $MINLP(r_l,r_u)$ with

$$r_l = (r_{l,1},\ldots,r_{l,n_y})^\top \in \mathbb{Z}^{n_y},\ r_u = (r_{u,1},\ldots,r_{u,n_y})^\top \in \mathbb{Z}^{n_y},\ r_l \neq r_u,\ y_l \leq r_l \leq r_u \leq y_u.$$

Let $(\hat{x},\hat{y})$ with $\hat{y} = (\hat{y}_1,\ldots,\hat{y}_{n_y})^\top$ be an optimal solution of the relaxation $NLP(r_l,r_u)$ of $MINLP(r_l,r_u)$ such that $\hat{y}$ is not integral, i.e. $\hat{y} \notin \mathcal{Y}(r_l,r_u)$. Then, branching consists of the following steps.

(i) Determine some index k with $r_{l,k} < \hat{y}_k < r_{u,k}$ and maximal distance of $\hat{y}_k$ to the closest integer, i.e. k minimizes the expression $|\lfloor\hat{y}_k\rfloor + 1/2 - \hat{y}_k|$. Herein, $\lfloor x\rfloor$ denotes the largest integer not greater than x.

(ii) Create two new successor nodes $MINLP(r_l^{(j)},r_u^{(j)})$, $j=1,2$ with

$$r_{l,i}^{(1)} := r_{l,i}, \qquad r_{u,i}^{(1)} := \begin{cases} r_{l,i}, & \text{if } i \neq k, \\ \lfloor\hat{y}_k\rfloor, & \text{if } i = k, \end{cases} \qquad i = 1,\ldots,n_y,$$

and

$$r_{l,i}^{(2)} := \begin{cases} r_{l,i}, & \text{if } i \neq k, \\ \lfloor\hat{y}_k\rfloor + 1, & \text{if } i = k, \end{cases}, \qquad r_{u,i}^{(2)} := r_{u,i}, \qquad i = 1,\ldots,n_y.$$

A recursive application of the branching rule starting with $MINLP(y_l, y_u)$ as the root generates a finite tree structure. Each edge corresponds to imposing one additional constraint to the set $\mathcal{Y}(r_l, r_u)$. Notice, that the union of the sets $\mathcal{Y}(r_l^{(j)}, r_u^{(j)})$, $j = 1, 2$ is just the set $\mathcal{Y}(r_l, r_u)$.
The branching rule is applicable in the same way, if the dual problem (of the relaxed problem) is used to determine lower bounds. In this case, the evaluation of the dual problem yields a point

$$(\hat{x}, \hat{y}) = \arg \min_{x \in S,\, y \in \hat{\mathcal{Y}}(r_l, r_u)} L(x, y, \lambda, \mu),$$

provided that such a point exists at all.

3.10.4 Fathoming Rules

A node $MINLP(r_l, r_u)$ is explored or fathomed, i.e. no further branching has to be done, if one of the following events occur, cf. Leyffer [Ley01]:

- $NLP(r_l, r_u)$ is infeasible. Hence, $MINLP(r_l, r_u)$ is infeasible as well. This implies that all nodes of the subtree are infeasible as well, since the feasible sets in the subtree are restricted further by additional constraints.

- $NLP(r_l, r_u)$ has an integer solution. Hence, this solution is optimal for $MINLP(r_l, r_u)$ and feasible for $MINLP$, since the feasible set of $MINLP(r_l, r_u)$ is a subset of the feasible set of $MINLP$. The corresponding optimal objective function value U serves as an upper bound for MINLP, that is $U \geq f_{y_l, y_u}$.

- The optimal objective function value $\hat{f}_{r_l, r_u}$ of $NLP(r_l, r_u)$ is greater or equal than the best upper bound U found so far. Due to the branching rule, the optimal objective function values are non-decreasing for the nodes of the subtree and consequently, the subtree can be pruned, since no better solution exists in the subtree.

3.10.5 Traversing Rule

We use a depth-first search as the *traversing rule* in order to find admissible solutions for MINLP and hence upper bounds as fast as possible. In particular, that subproblem in step (ii) of the branching rule with the smallest objective function value is investigated first. Alternative strategies are breadth-first search or some problem-specific search strategy.

3.10.6 Branch&Bound Algorithm

The overall Branch&Bound algorithm is as follows.

Algorithm 3.10.4 (Branch&Bound Algorithm)

```
Init:
    – Let $U$ be an upper bound for the optimal objective function value of MINLP
      (if none is known, let $U = \infty$).
    – Let the root of the tree $MINLP(y_l, y_u)$ be active.
while (there are active nodes) do
    Select an active node $MINLP(r_l, r_u)$, say, according to the traversing rule.
    Solve $NLP(r_l, r_u)$ (if possible) and let $(\hat{x}, \hat{y})$ be an optimal solution and
    $\hat{f} := f(\hat{x}, \hat{y})$ the optimal objective function value of $NLP(r_l, r_u)$.
    if (infeasible) then Mark node as explored. endif
    if ($\hat{y} \in \mathcal{Y}$) then
        if ($\hat{f} < U$) then Save ($\hat{x}, \hat{y}$) as best solution and set $U = \hat{f}$. endif
        Mark node as explored.
    endif
    if ($\hat{f} \geq U$) then Mark node as explored. endif
    if ($\hat{f} < U$) then
        Apply branching rule, mark all successors as active, and mark current
        node as inactive.
    endif
end
```

Since the set $\mathcal{Y}$ in our case is finite, the Branch&Bound algorithm will stop after finitely many steps. Unfortunately, the number of steps may be very large depending on the problem. Hence, it may be necessary to stop the algorithm after a maximal number of nodes have been generated. An advantage of the Branch&Bound algorithm in this case is, that it provides an upper bound and a lower bound of the optimal objective function value of MINLP. The upper bound is given by U in the algorithm. A lower bound arises, if the minimum L over the lower bounds of all active nodes is taken. Moreover, the knowledge of an upper bound U and a lower bound L for the optimal objective function value f_{y_l,y_u} for the current tree allows to use a stopping criterion of type $U - L < tol$ for a given tolerance $tol > 0$. This guarantees $-tol < L - U \leq f_{y_l,y_u} - U \leq 0$ and hence, the best solution found so far, which provides the upper bound U, has an objective function value that lies within the tolerance tol of f_{y_l,y_u}.

Leyffer [Ley01] describes an algorithm, where the tree search and the NLP solution is interlaced. He finds that his strategy leads to improved lower bounds and is up to three times faster than common Branch&Bound. A different approach to solve convex MINLP without equality constraints is the outer approximation method depicted in the survey article of Grossmann and Kravanja [GK97], cf. also Fletcher and Leyffer [FL94] and Duran and Grossmann [DG86b]. According to Grossmann and Kravanja [GK97] the outer approximation method may lead to satisfactory results even for non-convex problems if some heuristics are added. The outer approximation method works iteratively. In each iteration a mixed integer linear optimization problem (master problem), which is a linearization of MINLP at the current iterate, has to be solved in order to obtain a new iterate for the integer variables. Then, a continuous nonlinear programming problem with fixed integer variables has to be solved. The master problems yield a non-decreasing sequence of lower bounds, while the continuous programming problems yield upper bounds. The iteration is stopped, if the upper and lower bounds are within a given tolerance.

The determination of a global minimum of a nonconvex optimization problem is a difficult task. For the Branch&Bound-Algorithm it is not necessary to find the global optimum exactly but instead it is sufficient to find a lower bound. This can be achieved by constructing convex underestimators, cf. Meyer and Floudas [MF05], Akrotirianakis and Floudas [AF04a, AF04b], Adjiman and Floudas [AF96, AF01], Ryoo and Sahinidis [RS96], Androulakis et al. [AMF95], Sahinidis [Sah96], and Ghildyal and Sahinidis [GS01].

Chapter 4

Local Minimum Principles

Optimal control problems subject to ordinary differential equations have a wide range of applications in different disciplines like engineering sciences, chemical engineering, and economics. Necessary conditions known as 'Maximum principles' or 'Minimum principles' have been investigated intensively since the 1950's. Early proofs of the maximum principle are given by Pontryagin [PBGM64] and Hestenes [Hes66]. Necessary conditions with pure state constraints are discussed in, e.g., Jacobsen et al. [JLS71], Girsanov [Gir72], Knobloch [Kno75], Maurer [Mau77, Mau79], Ioffe and Tihomirov [IT79], and Kreindler [Kre82]. Neustadt [Neu76] and Zeidan [Zei94] discuss optimal control problems with mixed control-state constraints. Hartl et al. [HSV95] provide a survey on maximum principles for optimal control problems with state constraints including an extensive list of references. Necessary conditions for variational problems, i.e. smooth optimal control problems, are developed in Bryson and Ho [BH75]. Second order necessary conditions and sufficient conditions are stated in Zeidan [Zei94]. Sufficient conditions are also presented in Maurer [Mau81], Malanowski [Mal97], Maurer and Pickenhain [MP95b], and Malanowski et al. [MMP04]. Necessary conditions for optimal control problems subject to index-1 DAE systems without state constraints and without mixed control-state constraints can be found in de Pinho and Vinter [dPV97]. Implicit control systems are discussed in Devdariani and Ledyaev [DL99]. Recently, Backes [Bac06] derived necessary conditions for optimal control problems subject to nonlinear quasi-linear DAEs without control and state constraints. Necessary and sufficient conditions for linear-quadratic DAE optimal control problems can be found in Mehrmann [Meh91], Kunkel and Mehrmann [KM97], Kurina and März [KM04], and Backes [Bac06]. Roubicek and Valásek [RV02] discuss optimal control problems subject to Hessenberg DAEs up to index three and obtain results that are closely related to our results. Their technique for proving the results was based on an index reduction. Semi-explicit Index-1 systems often occur in process system engineering, cf. Hinsberger [Hin97], but also in vehicle simulation, cf. Gerdts [Ger03b], and many other fields of applications. A very important subclass of index two DAEs is the stabilized descriptor form describing the motion of mechanical multi-body systems.

Necessary conditions are not only interesting from a theoretical point of view, but also provide the basis of the so-called indirect approach for solving optimal control problems numerically. In this approach the minimum principle is exploited and usually leads to a multi-point boundary value problem, which is solved numerically by, e.g., the multiple shooting method. Nevertheless, even for the direct approach, which is based on a suitable discretization of the optimal control problem, the minimum principle is very important for the post-optimal approximation of adjoints. In this context, the multipliers resulting from the formulation of the necessary Fritz-John conditions for the finite dimensional discretized optimal control problem have to be related to the multipliers of the original infinite dimensional optimal control problem in an appropriate way. It is evident that this requires the knowledge of necessary conditions for the optimal control problem.

In the sequel necessary conditions in terms of local minimum principles are derived for optimal control problems subject to index-1 and index-2 DAEs, pure state constraints and mixed control-state constraints. The local minimum principles are based on necessary optimality conditions

for infinite optimization problems. The special structure of the optimal control problems under consideration is exploited and allows to obtain more regular representations for the multipliers involved. An additional Mangasarian-Fromowitz like constraint qualification for the optimal control problem ensures the regularity of a local minimum.
Although, index-1 DAEs are easier to handle, we will start with the index-2 case and present a complete proof. The proof for the index-1 case works similarly and will be omitted here. It can be found in Gerdts [Ger05a].

4.1 Local Minimum Principles for Index-2 Problems

In this section we investigate the special case of Problem 1.1, where the algebraic constraint g does not depend on y and u. Again, let $[t_0, t_f] \subset \mathbb{R}$ be a non-empty and bounded interval with fixed time points $t_0 < t_f$ and $\mathcal{U} \subseteq \mathbb{R}^{n_u}$ a closed and convex set with non-empty interior. Let

$$\begin{aligned}
\varphi &: \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \to \mathbb{R},\\
f_0 &: [t_0, t_f] \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_u} \to \mathbb{R},\\
f &: [t_0, t_f] \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_u} \to \mathbb{R}^{n_x},\\
g &: [t_0, t_f] \times \mathbb{R}^{n_x} \to \mathbb{R}^{n_y},\\
\psi &: \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \to \mathbb{R}^{n_\psi},\\
c &: [t_0, t_f] \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_u} \to \mathbb{R}^{n_c},\\
s &: [t_0, t_f] \times \mathbb{R}^{n_x} \to \mathbb{R}^{n_s}
\end{aligned}$$

be mappings. We consider

Problem 4.1.1 (Higher Index DAE Optimal Control Problem)
Find a state variable $x \in W^{1,\infty}([t_0, t_f], \mathbb{R}^{n_x})$*, an* algebraic variable $y \in L^\infty([t_0, t_f], \mathbb{R}^{n_y})$*, and a* control variable $u \in L^\infty([t_0, t_f], \mathbb{R}^{n_u})$ *such that the* objective function

$$F(x, y, u) := \varphi(x(t_0), x(t_f)) + \int_{t_0}^{t_f} f_0(t, x(t), y(t), u(t))dt \tag{4.1.1}$$

is minimized subject to the higher index semi-explicit differential algebraic equation (DAE)

$$\begin{aligned}
\dot{x}(t) &= f(t, x(t), y(t), u(t)) \quad \textit{a.e. in } [t_0, t_f], && (4.1.2)\\
0_{n_y} &= g(t, x(t)) \quad \textit{in } [t_0, t_f], && (4.1.3)
\end{aligned}$$

the boundary conditions

$$\psi(x(t_0), x(t_f)) = 0_{n_\psi}, \tag{4.1.4}$$

the mixed control-state constraints

$$c(t, x(t), y(t), u(t)) \leq 0_{n_c} \qquad \textit{a.e. in } [t_0, t_f], \tag{4.1.5}$$

the pure state constraints

$$s(t, x(t)) \leq 0_{n_s} \qquad \textit{in } [t_0, t_f], \tag{4.1.6}$$

and the set constraints

$$u(t) \in \mathcal{U} \qquad \textit{a.e. in } [t_0, t_f]. \tag{4.1.7}$$

Some definitions and terminologies are in order. $(x, y, u) \in W^{1,\infty}([t_0, t_f], \mathbb{R}^{n_x}) \times L^\infty([t_0, t_f], \mathbb{R}^{n_y}) \times L^\infty([t_0, t_f], \mathbb{R}^{n_u})$ is called *admissible* or *feasible* for the optimal control problem 4.1.1, if the constraints (4.1.2)-(4.1.7) are fulfilled. An admissible pair

$$(\hat{x}, \hat{y}, \hat{u}) \in W^{1,\infty}([t_0, t_f], \mathbb{R}^{n_x}) \times L^\infty([t_0, t_f], \mathbb{R}^{n_y}) \times L^\infty([t_0, t_f], \mathbb{R}^{n_u}) \tag{4.1.8}$$

is called a *weak local minimum of Problem 4.1.1*, if there exists $\varepsilon > 0$ such that

$$F(\hat{x}, \hat{y}, \hat{u}) \leq F(x, y, u) \tag{4.1.9}$$

holds for all admissible (x, y, u) with $\|x - \hat{x}\|_{1,\infty} < \varepsilon$, $\|y - \hat{y}\|_\infty < \varepsilon$, and $\|u - \hat{u}\|_\infty < \varepsilon$. An admissible pair (4.1.8) is called *strong local minimum of Problem 4.1.1*, if there exists $\varepsilon > 0$ such that (4.1.9) holds for all admissible (x, y, u) with $\|x - \hat{x}\|_\infty < \varepsilon$.

Notice, that strong local minima are also weak local minima. The converse is not true. Strong local minima are minimal w.r.t. a larger class of algebraic variables and controls. Weak local minima are only optimal w.r.t. all algebraic variables and controls in a L^∞-neighborhood.

For notational convenience throughout this chapter we will use the abbreviations

$$\varphi'_{x_0} := \varphi'_{x_0}(\hat{x}(t_0), \hat{x}(t_f)), \quad f'_x[t] := f'_x(t, \hat{x}(t), \hat{y}(t), \hat{u}(t)),$$

and in a similar way $\varphi'_{x_f}, f'_{0,x}[t], f'_{0,y}[t], f'_{0,u}[t], c'_x[t], c'_y[t], c'_u[t], s'_x[t], f'_y[t], f'_u[t], g'_x[t], \psi'_{x_0}, \psi'_{x_f}$ for the respective derivatives.

Remark 4.1.2

- *Since the algebraic component y is missing in (4.1.3), the DAE (4.1.2)-(4.1.3) is not an index-1 DAE. In contrast to the index-1 case in (4.2.3) the algebraic constraint (4.1.3) cannot be solved for y and hence it has to be differentiated at least once w.r.t. time in order to obtain additional information about y. Consequently, the index of (4.1.2)-(4.1.3) is at least two and the DAE is called 'higher index DAE'. From a formal point of view, y can be considered as an additional control.*

- *Using the technique in Remark 1.8 we may assume that the control u does not appear in the algebraic constraint (4.1.3). This assumption is essential for the subsequent analysis. Without this simplification additional smoothness assumptions for the control have to be imposed.*

So far, DAEs (4.1.2)-(4.1.3) without any additional properties of the functions f and g are not well investigated and are extremely difficult to handle in view of existence and representation of solutions. In the sequel we will exclusively consider index-2 DAEs which are characterized by the following assumption, cf. Definition 1.5 and the remarks thereafter and Hestenes [Hes66], p. 352.

Assumption 4.1.3 (Index-2 Assumption)
Let the matrix

$$M(t) := g'_x[t] \cdot f'_y[t]$$

be non-singular *a.e. in $[t_0, t_f]$ and let M^{-1} be essentially bounded in $[t_0, t_f]$.*

The optimal control problem 4.1.1 is to be reformulated as an infinite optimization problem in appropriate Banach spaces and necessary conditions for a weak local minimum are derived.

The variables (x, y, u) are taken as elements from the Banach space

$$X := W^{1,\infty}([t_0, t_f], \mathbb{R}^{n_x}) \times L^{\infty}([t_0, t_f], \mathbb{R}^{n_y}) \times L^{\infty}([t_0, t_f], \mathbb{R}^{n_u})$$

endowed with the norm $\|(x, y, u)\|_X := \max\{\|x\|_{1,\infty}, \|y\|_{\infty}, \|u\|_{\infty}\}$. The objective function $F : X \to \mathbb{R}$ is given by (4.1.1). If φ and f_0 are continuous w.r.t. all arguments and continuously differentiable w.r.t. to x, y, and u, then F is Fréchet-differentiable at $(\hat{x}, \hat{y}, \hat{u})$ with

$$\begin{aligned} F'(\hat{x}, \hat{y}, \hat{u})(x, y, u) &= \varphi'_{x_0} x(t_0) + \varphi'_{x_f} x(t_f) \\ &\quad + \int_{t_0}^{t_f} f'_{0,x}[t]x(t) + f'_{0,y}[t]y(t) + f'_{0,u}[t]u(t)dt, \end{aligned}$$

cf. Kirsch et al. [KWW78], p. 94. Now we collect the equality constraints. The space

$$Z := L^{\infty}([t_0, t_f], \mathbb{R}^{n_x}) \times W^{1,\infty}([t_0, t_f], \mathbb{R}^{n_y}) \times \mathbb{R}^{n_\psi}$$

endowed with the norm $\|(z_1, z_2, z_3)\|_Z := \max\{\|z_1\|_{\infty}, \|z_2\|_{1,\infty}, \|z_3\|_2\}$ is a Banach space. The equality constraints of the optimal control problem are given by

$$H(x, y, u) = \Theta_Z,$$

where $H = (H_1, H_2, H_3) : X \to Z$ with

$$\begin{aligned} H_1(x, y, u) &= f(\cdot, x(\cdot), y(\cdot), u(\cdot)) - \dot{x}(\cdot), \\ H_2(x, y, u) &= g(\cdot, x(\cdot)), \\ H_3(x, y, u) &= -\psi(x(t_0), x(t_f)). \end{aligned}$$

Remark 4.1.4 *Notice, that we consider the function $g[\cdot]$ in the space of absolutely continuous functions and not in the space of continuous functions. The reason is that we will need the surjectivity of the operator (H_1', H_2'). But in the space of continuous functions this operator is not surjective, cf. Lemma 4.1.6 below.*

If f, g, and ψ are continuous w.r.t. all arguments and continuously differentiable w.r.t. to x, y, and u, then H is continuously Fréchet-differentiable at $(\hat{x}, \hat{y}, \hat{u})$ with $H' = (H_1', H_2', H_3')$ and

$$\begin{aligned} H_1'(\hat{x}, \hat{y}, \hat{u})(x, y, u) &= f_x'[\cdot]x(\cdot) + f_y'[\cdot]y(\cdot) + f_u'[\cdot]u(\cdot) - \dot{x}(\cdot), \\ H_2'(\hat{x}, \hat{y}, \hat{u})(x, y, u) &= g_x'[\cdot]x(\cdot), \\ H_3'(\hat{x}, \hat{y}, \hat{u})(x, y, u) &= -\psi'_{x_0} x(t_0) - \psi'_{x_f} x(t_f), \end{aligned}$$

cf. Kirsch et al. [KWW78], p. 95. Now we collect the inequality constraints. The space

$$Y := L^{\infty}([t_0, t_f], \mathbb{R}^{n_c}) \times C([t_0, t_f], \mathbb{R}^{n_s})$$

endowed with the norm $\|(y_1, y_2)\|_Y := \max\{\|y_1\|_{\infty}, \|y_2\|_{\infty}\}$ is a Banach space. The inequality constraints of the optimal control problem are given by

$$G(x, y, u) \in K,$$

where $G = (G_1, G_2) : X \to Y$ and

$$G_1(x, y, u) = -c(\cdot, x(\cdot), y(\cdot), u(\cdot)), \quad G_2(x, y, u) = -s(\cdot, x(\cdot)).$$

The cone $K := K_1 \times K_2 \subseteq Y$ is defined by

$$\begin{aligned} K_1 &:= \{z \in L^\infty([t_0,t_f],\mathbb{R}^{n_c}) \mid z(t) \geq 0_{n_c} \text{ a.e. in } [t_0,t_f]\}, \\ K_2 &:= \{z \in C([t_0,t_f],\mathbb{R}^{n_s}) \mid z(t) \geq 0_{n_s} \text{ in } [t_0,t_f]\}. \end{aligned}$$

If c and s are continuous w.r.t. all arguments and continuously differentiable w.r.t. to x, y, u, then G is continuously Fréchet-differentiable at $(\hat{x},\hat{y},\hat{u})$ with $G' = (G_1', G_2')$ and

$$\begin{aligned} G_1'(\hat{x},\hat{y},\hat{u})(x,y,u) &= -c_x'[\cdot]x(\cdot) - c_y'[\cdot]y(\cdot) - c_u'[\cdot]u(\cdot), \\ G_2'(\hat{x},\hat{y},\hat{u})(x,y,u) &= -s_x'[\cdot]x(\cdot). \end{aligned}$$

Finally, the set $S \subseteq X$ is given by

$$S := W^{1,\infty}([t_0,t_f],\mathbb{R}^{n_x}) \times L^\infty([t_0,t_f],\mathbb{R}^{n_y}) \times U_{ad},$$

where

$$U_{ad} := \{u \in L^\infty([t_0,t_f],\mathbb{R}^{n_u}) \mid u(t) \in \mathcal{U} \text{ a.e. in } [t_0,t_f]\}.$$

Summarizing, the optimal control problem 4.1.1 is equivalent with

Problem 4.1.5 *Find $(x,y,u) \in X$ such that $F(x,y,u)$ is minimized subject to the constraints*

$$G(x,y,u) \in K, \quad H(x,y,u) = \Theta_Z, \quad (x,y,u) \in S.$$

We intend to apply the necessary conditions in Theorem 3.4.2 to Problem 4.1.5. In order to show the non-density of the linearized equality constraints, we need the following results about linear differential algebraic equations and boundary value problems.

Lemma 4.1.6 Consider the linear DAE

$$\dot{x}(t) = A_1(t)x(t) + B_1(t)y(t) + h_1(t), \tag{4.1.10}$$

$$0_{n_y} = A_2(t)x(t) + h_2(t), \tag{4.1.11}$$

with time dependent matrix functions $A_1(\cdot) \in L^\infty([t_0,t_f],\mathbb{R}^{n_x \times n_x})$, $A_2(\cdot) \in W^{1,\infty}([t_0,t_f],\mathbb{R}^{n_y \times n_x})$, $B_1(\cdot) \in L^\infty([t_0,t_f],\mathbb{R}^{n_x \times n_y})$ and time dependent vector functions $h_1(\cdot) \in L^\infty([t_0,t_f],\mathbb{R}^{n_x})$, $h_2(\cdot) \in W^{1,\infty}([t_0,t_f],\mathbb{R}^{n_y})$. Let $C(t) := A_2(t) \cdot B_1(t)$ be non-singular almost everywhere in $[t_0,t_f]$ and let C^{-1} be essentially bounded. Define

$$\begin{aligned} A(t) &:= A_1(t) - B_1(t)C(t)^{-1}Q(t), \\ h(t) &:= h_1(t) - B_1(t)C(t)^{-1}q(t), \\ Q(t) &:= \dot{A}_2(t) + A_2(t)A_1(t), \\ q(t) &:= \dot{h}_2(t) + A_2(t)h_1(t). \end{aligned}$$

(a) There exist consistent initial values $x(t_0) = x_0$ satisfying (4.1.11) and every consistent x_0 possesses the representation

$$x_0 = \Pi h_2(t_0) + \Gamma w,$$

where $\Pi \in \mathbb{R}^{n_x \times n_y}$ satisfies $(I + A_2(t_0)\Pi)h_2(t_0) = 0_{n_y}$ and the columns of $\Gamma \in \mathbb{R}^{n_x \times (n_x - n_y)}$ define an orthonormal basis of $\ker(A_2(t_0))$, i.e. $A_2(t_0)\Gamma = \Theta$. Vice versa, every such x_0 is consistent for arbitrary $w \in \mathbb{R}^{n_x - n_y}$.

(b) The initial value problem given by (4.1.10)-(4.1.11) together with the consistent initial value $x(t_0) = x_0 = \Pi h_2(t_0) + \Gamma w$ has a unique solution $x(\cdot) \in W^{1,\infty}([t_0, t_f], \mathbb{R}^{n_x})$, for every $w \in \mathbb{R}^{n_x - n_y}$, every $h_1(\cdot) \in L^{\infty}([t_0, t_f], \mathbb{R}^{n_x})$ and every $h_2(\cdot) \in W^{1,\infty}([t_0, t_f], \mathbb{R}^{n_y})$. The solution is given by

$$\begin{aligned} x(t) &= \Phi(t)\left(\Pi h_2(t_0) + \Gamma w + \int_{t_0}^{t} \Phi^{-1}(\tau) h(\tau) d\tau\right) \qquad \text{in } [t_0, t_f], & (4.1.12) \\ y(t) &= -C(t)^{-1}\left(q(t) + Q(t)x(t)\right) \quad \text{a.e. in } [t_0, t_f], & (4.1.13) \end{aligned}$$

where the fundamental system $\Phi(t) \in \mathbb{R}^{n_x \times n_x}$ is the unique solution of

$$\dot{\Phi}(t) = A(t)\Phi(t), \qquad \Phi(t_0) = I_{n_x}. \tag{4.1.14}$$

(c) Let a vector $b \in \mathbb{R}^r$ and matrices $C_0, C_f \in \mathbb{R}^{r \times n_x}$ be given, such that

$$\operatorname{rank}\left(\left(C_0\Phi(t_0) + C_f\Phi(t_f)\right)\Gamma\right) = r$$

holds for the fundamental solution Φ from (b). Then, the boundary value problem given by (4.1.10)-(4.1.11) together with the boundary condition

$$C_0 x(t_0) + C_f x(t_f) = b \tag{4.1.15}$$

has a solution for every $b \in \mathbb{R}^r$.

Proof.

(a) Let $x(t_0) = x_0$ be consistent, i.e. x_0 satisfies $0_{n_y} = A_2(t_0)x_0 + h_2(t_0)$. Since $C(t_0) = A_2(t_0) \cdot B_1(t_0)$ is supposed to be non-singular, $A_2(t_0)$ has full row rank. Hence, there exists a QR decomposition

$$A_2(t_0)^\top = P\begin{pmatrix} R \\ 0 \end{pmatrix}, \qquad P = (\Pi_1, \Gamma) \in \mathbb{R}^{n_x \times n_x},$$

where $R \in \mathbb{R}^{n_y \times n_y}$ is non-singular, P is orthogonal, $\Pi_1 \in \mathbb{R}^{n_x \times n_y}$ is a orthonormal basis of $\operatorname{im}(A_2(t_0)^\top)$, $\Gamma \in \mathbb{R}^{n_x \times (n_x - n_y)}$ is a orthonormal basis of $\operatorname{im}(A_2(t_0)^\top)^\perp = \ker(A_2(t_0))$. Every $x_0 \in \mathbb{R}^{n_x}$ can be expressed uniquely as $x_0 = \Pi_1 v + \Gamma w$ with $v \in \mathbb{R}^{n_y}$ and $w \in \mathbb{R}^{n_x - n_y}$. Introducing this expression into the algebraic equation (4.1.11) yields

$$0_{n_y} = A_2(t_0)(\Pi_1 v + \Gamma w) + h_2(t_0) = R^\top v + h_2(t_0) \quad \Rightarrow \quad v = -R^{-\top} h_2(t_0).$$

Hence, the consistent values are characterized by

$$x_0 = \Pi h_2(t_0) + \Gamma w, \qquad \Pi := -\Pi_1 R^{-\top}.$$

(b) We differentiate the algebraic equation (4.1.11) and obtain

$$\begin{aligned} 0_{n_y} &= \dot{A}_2(t)x(t) + A_2(t)\dot{x}(t) + \dot{h}_2(t) \\ &= \left(\dot{A}_2(t) + A_2(t)A_1(t)\right)x(t) + A_2(t)B_1(t)y(t) + \dot{h}_2(t) + A_2(t)h_1(t) \\ &= Q(t)x(t) + C(t)y(t) + q(t). \end{aligned}$$

We exploit the non-singularity of $C(t) = A_2(t) \cdot B_1(t)$ in order to solve the equation w.r.t. y and obtain

$$y(t) = -C(t)^{-1}\left(q(t) + Q(t)x(t)\right).$$

Introducing this expression into (4.1.10) yields

$$\dot{x}(t) = \left(A_1(t) - B_1(t)C(t)^{-1}Q(t)\right)x(t) + h_1(t) - B_1(t)C(t)^{-1}q(t) = A(t)x(t) + h(t).$$

Considering the representation of consistent initial values x_0 in (a) we are in the situation as in Hermes and Lasalle [HL69], p. 36, and the assertions follow likewise.

(c) Part (c) exploits the solution formulas in (a) and (b). The boundary conditions (4.1.15) are satisfied, if

$$\begin{aligned} b &= C_0x(t_0) + C_fx(t_f) \\ &= C_0\left(\Pi h_2(t_0) + \Gamma w\right) + C_f\Phi(t_f)\left(\Pi h_2(t_0) + \Gamma w + \int_{t_0}^{t_f} \Phi^{-1}(\tau)h(\tau)d\tau\right) \\ &= \left(C_0 + C_f\Phi(t_f)\right)\Gamma w + \left(C_0 + C_f\Phi(t_f)\right)\Pi h_2(t_0) \\ &\quad + C_f\Phi(t_f)\int_{t_0}^{t_f} \Phi^{-1}(\tau)h(\tau)d\tau. \end{aligned}$$

Rearranging terms and exploiting $\Phi(t_0) = I_{n_x}$ yields

$$\begin{aligned} \left(C_0\Phi(t_0) + C_f\Phi(t_f)\right)\Gamma w &= b - \left(C_0 + C_f\Phi(t_f)\right)\Pi h_2(t_0) \\ &\quad - C_f\Phi(t_f)\int_{t_0}^{t_f} \Phi^{-1}(\tau)h(\tau)d\tau. \end{aligned}$$

This equation is solvable for every $b \in \mathbb{R}^r$, if the matrix $\left(C_0\Phi(t_0) + C_f\Phi(t_f)\right)\Gamma$ is of rank r. Then, for every $b \in \mathbb{R}^r$ there exists a w such that (4.1.15) is satisfied. Application of part (b) completes the proof.

■

Part (b) of Lemma 4.1.6 enables us to formulate necessary conditions for the optimal control problem 4.1.1 if Assumption 4.1.3 is valid.

Theorem 4.1.7 (Necessary Conditions)
Let the following assumptions be fulfilled for the optimal control problem 4.1.1.

(i) Let the functions $\varphi, f_0, f, \psi, c, s$ be continuous w.r.t. all arguments and continuously differentiable w.r.t. x, y, and u. Let g be continuously differentiable w.r.t. all arguments.

(ii) Let $\mathcal{U} \subseteq \mathbb{R}^{n_u}$ be a closed and convex set with non-empty interior.

(iii) Let $(\hat{x}, \hat{y}, \hat{u}) \in X$ be a weak local minimum of the optimal control problem.

(iv) Let Assumption 4.1.3 be valid.

Then there exist non-trivial multipliers $l_0 \in \mathbb{R}$, $\eta^* \in Y^*$, $\lambda^* \in Z^*$ with

$$l_0 \geq 0, \tag{4.1.16}$$

$$\eta^* \in K^+, \tag{4.1.17}$$

$$\eta^*(G(\hat{x}, \hat{y}, \hat{u})) = 0, \tag{4.1.18}$$

$$\begin{aligned} &l_0F'(\hat{x}, \hat{y}, \hat{u})(x - \hat{x}, y - \hat{y}, u - \hat{u}) \\ &-\eta^*(G'(\hat{x}, \hat{y}, \hat{u})(x - \hat{x}, y - \hat{y}, u - \hat{u})) \\ &-\lambda^*(H'(\hat{x}, \hat{y}, \hat{u})(x - \hat{x}, y - \hat{y}, u - \hat{u})) \geq 0 \quad \forall (x, y, u) \in S. \end{aligned} \tag{4.1.19}$$

Proof. We show, that all assumptions of Theorem 3.4.2 are satisfied. Observe, that K is a closed convex cone with vertex at zero, $\mathrm{int}(K)$ is non-empty, the functions F and G are Fréchet-differentiable, and H is continuously Fréchet-differentiable due to the smoothness assumptions. Since $\mathcal{U}$ is supposed to be closed and convex with non-empty interior, the set S is closed and convex with non-empty interior. It remains to show that $\mathrm{im}(H'(\hat{x},\hat{y},\hat{u}))$ is not a proper dense subset of Z. According to part (b) of Lemma 4.1.6 the operator $(H_1'(\hat{x},\hat{y},\hat{u}),H_2'(\hat{x},\hat{y},\hat{u}))$ given by

$$\left(H_1'(\hat{x},\hat{y},\hat{u}),H_2'(\hat{x},\hat{y},\hat{u})\right)(x,y,u)=\left(f_x'[\cdot]x(\cdot)+f_y'[\cdot]y(\cdot)+f_u'[\cdot]u(\cdot)-\dot{x}(\cdot),g_x'[\cdot]x(\cdot)\right)$$

is continuous, linear and surjective. Thus, we can apply Theorem 2.2.8, which yields that the image of $H'(\hat{x},\hat{y},\hat{u})$ is closed in Z and hence $\mathrm{im}(H'(\hat{x},\hat{y},\hat{u}))$ is not a proper dense subset in Z. Hence, all assumptions of Theorem 3.4.2 are satisfied and Theorem 3.4.2 yields (4.1.16)-(4.1.19). ∎

Notice, that the multipliers are elements of the following dual spaces:

$$\begin{aligned}\eta^*:=(\eta_1^*,\eta_2^*)&\in Y^*=(L^\infty([t_0,t_f],\mathbb{R}^{n_c}))^*\times(C([t_0,t_f],\mathbb{R}^{n_s}))^*,\\ \lambda^*:=(\lambda_f^*,\lambda_g^*,\sigma)&\in Z^*=(L^\infty([t_0,t_f],\mathbb{R}^{n_x}))^*\times\left(W^{1,\infty}([t_0,t_f],\mathbb{R}^{n_y})\right)^*\times\mathbb{R}^{n_\psi}.\end{aligned}$$

Hence, by Riesz' representation theorem 2.4.2 the functional η_2^* admits the following representation

$$\eta_2^*(h)=\sum_{i=1}^{n_s}\int_{t_0}^{t_f}h_i(t)d\mu_i(t) \tag{4.1.20}$$

for every continuous function $h\in C([t_0,t_f],\mathbb{R}^{n_s})$. Herein, μ_i, $i=1,\ldots,n_s$, are functions of bounded variation. To make the representations unique, we choose μ_i, $i=1,\ldots,n_s$, from the space $NBV([t_0,t_f],\mathbb{R})$, i.e. the space of normalized functions of bounded variation which are continuous from the right in (t_0,t_f) and satisfy $\mu_i(t_0)=0$, $i=1,\ldots,n_s$.

In the sequel, the special structure of the functions F, G, and H in (4.1.19) is exploited. Furthermore, the fact, that the variational inequality (4.1.19) holds for all $(x,y,u)\in S$, i.e. for all $x\in W^{1,\infty}([t_0,t_f],\mathbb{R}^{n_x})$, all $y\in L^\infty([t_0,t_f],\mathbb{R}^{n_y})$, and all $u\in U_{ad}$, is used to derive three separate variational equalities and inequalities, respectively.

Evaluation of the Fréchet-derivatives in (4.1.19), using (4.1.20) and setting $y=\hat{y},u=\hat{u}$ yields the variational equality

$$\begin{aligned}0\;=\;&\left(l_0\varphi_{x_0}'+\sigma^\top\psi_{x_0}'\right)x(t_0)+\left(l_0\varphi_{x_f}'+\sigma^\top\psi_{x_f}'\right)x(t_f)\\&+\int_{t_0}^{t_f}l_0f_{0,x}'[t]x(t)dt+\sum_{i=1}^{n_s}\int_{t_0}^{t_f}s_{i,x}'[t]x(t)d\mu_i(t)\\&+\eta_1^*(c_x'[\cdot]x(\cdot))+\lambda_f^*(\dot{x}(\cdot)-f_x'[\cdot]x(\cdot))-\lambda_g^*(g_x'[\cdot]x(\cdot))\end{aligned} \tag{4.1.21}$$

which holds for all $x\in W^{1,\infty}([t_0,t_f],\mathbb{R}^{n_x})$. Similarly, we find

$$0=\int_{t_0}^{t_f}l_0f_{0,y}'[t]y(t)dt+\eta_1^*(c_y'[\cdot]y(\cdot))-\lambda_f^*(f_y'[\cdot]y(\cdot)) \tag{4.1.22}$$

for all $y\in L^\infty([t_0,t_f],\mathbb{R}^{n_y})$ and

$$\int_{t_0}^{t_f}l_0f_{0,u}'[t]u(t)dt+\eta_1^*(c_u'[\cdot]u(\cdot))-\lambda_f^*(f_u'[\cdot]u(\cdot))\geq 0 \tag{4.1.23}$$

for all $u\in U_{ad}-\{\hat{u}\}$.

4.1.1 Representation of Multipliers

At a first glance, the necessary conditions seem to be of little practical use since the multipliers η_1^*, λ_f^*, and λ_g^* as elements of the dual spaces of L^∞ resp. $W^{1,\infty}$ do not possess a useful representation. However, in this section we will derive nice representations for them. Herein, properties of Stieltjes integrals will be exploited extensively. These results about Stieltjes integrals can be found in the books of Natanson [Nat75] and Widder [Wid46].

According to Lemma 4.1.6 the initial value problem

$$\begin{aligned} \dot{x}(t) &= f_x'[t]x(t) + f_y'[t]y(t) + h_1(t), \qquad x(t_0) = \Pi h_2(t_0), & (4.1.24)\\ 0_{n_y} &= g_x'[t]x(t) + h_2(t) & (4.1.25) \end{aligned}$$

has a solution for every $h_1 \in L^\infty([t_0,t_f],\mathbb{R}^{n_x})$, $h_2 \in W^{1,\infty}([t_0,t_f],\mathbb{R}^{n_y})$. Notice, that $\Pi h_2(t_0)$ is a special consistent initial value with $w = 0_{n_x-n_y}$. According to (4.1.12) and (4.1.13) in Lemma 4.1.6 the solution is given by

$$\begin{aligned} x(t) &= \Phi(t)\left(\Pi h_2(t_0) + \int_{t_0}^t \Phi^{-1}(\tau)h(\tau)d\tau\right), & (4.1.26)\\ y(t) &= -M(t)^{-1}\left(q(t) + Q(t)x(t)\right) \\ &= -M(t)^{-1}\left(q(t) + Q(t)\Phi(t)\left(\Pi h_2(t_0) + \int_{t_0}^t \Phi^{-1}(\tau)h(\tau)d\tau\right)\right), & (4.1.27) \end{aligned}$$

where

$$h(t) = h_1(t) - f_y'[t]M(t)^{-1}q(t),\ Q(t) = \frac{d}{dt}g_x'[t] + g_x'[t]f_x'[t],\ q(t) = \dot{h}_2(t) + g_x'[t]h_1(t),$$

and Φ is the solution of

$$\dot{\Phi}(t) = A(t)\Phi(t), \quad \Phi(t_0) = I_{n_x}, \quad A(t) = f_x'[t] - f_y'[t]M(t)^{-1}Q(t).$$

Now, let $h_1 \in L^\infty([t_0,t_f],\mathbb{R}^{n_x})$, $h_2 \in W^{1,\infty}([t_0,t_f],\mathbb{R}^{n_y})$ be arbitrary. Adding equations (4.1.22) and (4.1.21) and exploiting the linearity of the functionals leads to

$$\begin{aligned} &\left(l_0\varphi_{x_0}' + \sigma^\top\psi_{x_0}'\right)x(t_0) + \left(l_0\varphi_{x_f}' + \sigma^\top\psi_{x_f}'\right)x(t_f) \\ &+ \int_{t_0}^{t_f} l_0 f_{0,x}'[t]x(t) + l_0 f_{0,y}'[t]y(t)dt + \sum_{i=1}^{n_s}\int_{t_0}^{t_f} s_{i,x}'[t]x(t)d\mu_i(t) \\ &\qquad + \eta_1^*(c_x'[\cdot]x(\cdot) + c_y'[\cdot]y(\cdot)) + \lambda_f^*(h_1(\cdot)) + \lambda_g^*(h_2(\cdot)) = 0. \qquad (4.1.28) \end{aligned}$$

Introducing the solution formulas (4.1.26)-(4.1.27) into equation (4.1.28) and combining terms leads to the expression

$$\begin{aligned}
&\left(l_0\varphi'_{x_0} + \sigma^\top \psi'_{x_0}\right)\Pi h_2(t_0) + \left(l_0\varphi'_{x_f} + \sigma^\top \psi'_{x_f}\right)\Phi(t_f)\Pi h_2(t_0) \\
&+ \int_{t_0}^{t_f} l_0 \hat{f}_0[t]\Phi(t)\Pi h_2(t_0)dt + \sum_{i=1}^{n_x}\int_{t_0}^{t_f} s'_{i,x}[t]\Phi(t)\Pi h_2(t_0)d\mu_i(t) \\
&+ \left(l_0\varphi'_{x_f} + \sigma^\top \psi'_{x_f}\right)\Phi(t_f)\int_{t_0}^{t_f}\Phi^{-1}(t)h_1(t)dt \\
&- \left(l_0\varphi'_{x_f} + \sigma^\top \psi'_{x_f}\right)\Phi(t_f)\int_{t_0}^{t_f}\Phi^{-1}(t)f'_y[t]M(t)^{-1}q(t)dt \\
&+ \int_{t_0}^{t_f} l_0\hat{f}_0[t]\Phi(t)\left(\int_{t_0}^{t}\Phi^{-1}(\tau)h_1(\tau)d\tau\right)dt \\
&- \int_{t_0}^{t_f} l_0\hat{f}_0[t]\Phi(t)\left(\int_{t_0}^{t}\Phi^{-1}(\tau)f'_y[\tau]M(\tau)^{-1}q(\tau)d\tau\right)dt \\
&- \int_{t_0}^{t_f} l_0 f'_{0,y}[t]M(t)^{-1}q(t)dt \\
&+ \sum_{i=1}^{n_s}\int_{t_0}^{t_f} s'_{i,x}[t]\Phi(t)\left(\int_{t_0}^{t}\Phi^{-1}(\tau)h_1(\tau)d\tau\right)d\mu_i(t) \\
&- \sum_{i=1}^{n_s}\int_{t_0}^{t_f} s'_{i,x}[t]\Phi(t)\left(\int_{t_0}^{t}\Phi^{-1}(\tau)f'_y[\tau]M(\tau)^{-1}q(\tau)d\tau\right)d\mu_i(t) \\
&+ \eta_1^*(c'_x[\cdot]x(\cdot) + c'_y[\cdot]y(\cdot)) + \lambda_f^*(h_1(\cdot)) + \lambda_g^*(h_2(\cdot)) = 0,
\end{aligned} \tag{4.1.29}$$

where

$$\hat{f}_0[t] := f'_{0,x}[t] - f'_{0,y}[t]M(t)^{-1}Q(t).$$

Integration by parts yields

$$\int_{t_0}^{t_f}\hat{f}_0[t]\Phi(t)\left(\int_{t_0}^{t}\Phi^{-1}(\tau)h_1(\tau)d\tau\right)dt = \int_{t_0}^{t_f}\left(\int_{t}^{t_f}\hat{f}_0[\tau]\Phi(\tau)d\tau\right)\Phi^{-1}(t)h_1(t)dt$$

and

$$\begin{aligned}
&\int_{t_0}^{t_f}\hat{f}_0[t]\Phi(t)\left(\int_{t_0}^{t}\Phi^{-1}(\tau)f'_y[\tau]M(\tau)^{-1}q(\tau)d\tau\right)dt \\
&= \int_{t_0}^{t_f}\left(\int_{t}^{t_f}\hat{f}_0[\tau]\Phi(\tau)d\tau\right)\Phi^{-1}(t)f'_y[t]M(t)^{-1}q(t)dt.
\end{aligned}$$

The Riemann-Stieltjes integrals are to be transformed using integration by parts. Therefore, let $a : [t_0, t_f] \to \mathbb{R}^n$ be continuous, $b : [t_0, t_f] \to \mathbb{R}^n$ absolutely continuous and $\mu : [t_0, t_f] \to \mathbb{R}$ of bounded variation. Then, using integration by parts for Riemann-Stieltjes integrals leads to

$$\begin{aligned}
\int_{t_0}^{t_f} a(t)^\top b(t) d\mu(t) &= \sum_{i=1}^{n} \int_{t_0}^{t_f} a_i(t) b_i(t) d\mu(t) \\
&= \sum_{i=1}^{n} \int_{t_0}^{t_f} b_i(t) d\left(\int_{t_0}^{t} a_i(\tau) d\mu(\tau)\right) \\
&= \sum_{i=1}^{n} \left(\left[\left(\int_{t_0}^{t} a_i(\tau) d\mu(\tau) \right) \cdot b_i(t) \right]_{t_0}^{t_f} \right. \\
&\qquad \left. - \int_{t_0}^{t_f} \left(\int_{t_0}^{t} a_i(\tau) d\mu(\tau) \right) d b_i(t) \right) \\
&= \left[\left(\int_{t_0}^{t} a(\tau)^\top d\mu(\tau) \right) \cdot b(t) \right]_{t_0}^{t_f} - \int_{t_0}^{t_f} \left(\int_{t_0}^{t} a(\tau)^\top d\mu(\tau) \right) b'(t) dt \\
&= \left(\int_{t_0}^{t_f} a(\tau)^\top d\mu(\tau) \right) \cdot b(t_f) - \int_{t_0}^{t_f} \left(\int_{t_0}^{t} a(\tau)^\top d\mu(\tau) \right) b'(t) dt.
\end{aligned}$$

Application of this formula to (4.1.29) where $a(t)^\top = s'_{i,x}[t]\Phi(t)$ and

$$b(t) = \int_{t_0}^{t} \Phi^{-1}(\tau) h_1(\tau) d\tau \quad \text{resp.} \quad b(t) = \int_{t_0}^{t} \Phi^{-1}(\tau) f'_y[\tau] M(\tau)^{-1} q(\tau) d\tau$$

yields

$$\int_{t_0}^{t_f} s'_{i,x}[t]\Phi(t) \left(\int_{t_0}^{t} \Phi^{-1}(\tau) h_1(\tau) d\tau \right) d\mu_i(t) = \int_{t_0}^{t_f} \left(\int_{t}^{t_f} s'_{i,x}[\tau]\Phi(\tau) d\mu_i(\tau) \right) \Phi^{-1}(t) h_1(t) dt$$

and

$$\begin{aligned}
&\int_{t_0}^{t_f} s'_{i,x}[t]\Phi(t) \left(\int_{t_0}^{t} \Phi^{-1}(\tau) f'_y[\tau] M(\tau)^{-1} q(\tau) d\tau \right) d\mu_i(t) \\
&\qquad = \int_{t_0}^{t_f} \left(\int_{t}^{t_f} s'_{i,x}[\tau]\Phi(\tau) d\mu_i(\tau) \right) \Phi^{-1}(t) f'_y[t] M(t)^{-1} q(t) dt.
\end{aligned}$$

Substitution into (4.1.29) yields

$$\begin{aligned}
&\left(l_0 \varphi'_{x_0} + \sigma^\top \psi'_{x_0} \right) \Pi h_2(t_0) + \left(l_0 \varphi'_{x_f} + \sigma^\top \psi'_{x_f} \right) \Phi(t_f) \Pi h_2(t_0) \\
&\quad + \int_{t_0}^{t_f} l_0 \hat{f}_0[t] \Phi(t) \Pi h_2(t_0) dt + \sum_{i=1}^{n_x} \int_{t_0}^{t_f} s'_{i,x}[t] \Phi(t) \Pi h_2(t_0) d\mu_i(t) \\
&+ \int_{t_0}^{t_f} \left(\left(l_0 \varphi'_{x_f} + \sigma^\top \psi'_{x_f} \right) \Phi(t_f) + \int_{t}^{t_f} l_0 \hat{f}_0[\tau] \Phi(\tau) d\tau \right. \\
&\qquad \left. + \sum_{i=1}^{n_s} \left(\int_{t}^{t_f} s'_{i,x}[\tau] \Phi(\tau) d\mu_i(\tau) \right) \right) \Phi^{-1}(t) h_1(t) dt \\
&- \int_{t_0}^{t_f} \left(\left(l_0 \varphi'_{x_f} + \sigma^\top \psi'_{x_f} \right) \Phi(t_f) + \int_{t}^{t_f} l_0 \hat{f}_0[\tau] \Phi(\tau) d\tau \right. \\
&\qquad \left. + \sum_{i=1}^{n_s} \left(\int_{t}^{t_f} s'_{i,x}[\tau] \Phi(\tau) d\mu_i(\tau) \right) \right) \Phi^{-1}(t) f'_y[t] M(t)^{-1} q(t) dt \\
&\qquad - \int_{t_0}^{t_f} l_0 f'_{0,y}[t] M(t)^{-1} q(t) dt \\
&\qquad + \eta_1^*(c'_x[\cdot]x(\cdot) + c'_y[\cdot]y(\cdot)) + \lambda_f^*(h_1(\cdot)) + \lambda_g^*(h_2(\cdot)) = 0.
\end{aligned} \tag{4.1.30}$$

Equation (4.1.30) is equivalent with

$$\begin{aligned}\zeta^\top h_2(t_0) + \int_{t_0}^{t_f} p_f(t)^\top h_1(t)dt + \int_{t_0}^{t_f} p_g(t)^\top q(t)dt & \\ +\eta_1^*(c_x'[\cdot]x(\cdot) + c_y'[\cdot]y(\cdot)) + \lambda_f^*(h_1(\cdot)) + \lambda_g^*(h_2(\cdot)) &= 0, \qquad (4.1.31)\end{aligned}$$

where

$$\begin{aligned}\zeta^\top &:= \Bigg(\left(l_0\varphi_{x_0}' + \sigma^\top\psi_{x_0}'\right) + \left(l_0\varphi_{x_f}' + \sigma^\top\psi_{x_f}'\right)\Phi(t_f) \\ &\qquad + \int_{t_0}^{t_f} l_0\hat{f}_0[t]\Phi(t)dt + \sum_{i=1}^{n_x}\int_{t_0}^{t_f} s_{i,x}'[t]\Phi(t)d\mu_i(t)\Bigg)\Pi \\ &= \left(\left(l_0\varphi_{x_0}' + \sigma^\top\psi_{x_0}'\right) + p_f(t_0)^\top\right)\Pi, \qquad (4.1.32) \\ p_f(t)^\top &:= \Bigg(\left(l_0\varphi_{x_f}' + \sigma^\top\psi_{x_f}'\right)\Phi(t_f) + \int_t^{t_f} l_0\hat{f}_0[\tau]\Phi(\tau)d\tau \\ &\qquad + \sum_{i=1}^{n_s}\left(\int_t^{t_f} s_{i,x}'[\tau]\Phi(\tau)d\mu_i(\tau)\right)\Bigg)\Phi^{-1}(t), \qquad (4.1.33) \\ p_g(t)^\top &:= -l_0 f_{0,y}'[t]M(t)^{-1} \\ &\quad -\Bigg(\left(l_0\varphi_{x_f}' + \sigma^\top\psi_{x_f}'\right)\Phi(t_f) + \int_t^{t_f} l_0\hat{f}_0[\tau]\Phi(\tau)d\tau \\ &\qquad + \sum_{i=1}^{n_s}\left(\int_t^{t_f} s_{i,x}'[\tau]\Phi(\tau)d\mu_i(\tau)\right)\Bigg)\Phi^{-1}(t)f_y'[t]M(t)^{-1}. \qquad (4.1.34)\end{aligned}$$

(4.1.31) and (4.1.23) will be exploited in order to derive explicit representations of the functionals λ_f^*, λ_g^*, and η_1^*. This is possible if either there are no mixed control-state constraints (4.1.5) present in the optimal control problem, or if there are no set constraints (4.1.7), i.e. $\mathcal{U} = \mathbb{R}^{n_u}$. In case of no mixed control-state constraints we find

Corollary 4.1.8 *Let the assumptions of Theorem 4.1.7 be valid and let there be no mixed control-state constraints (4.1.5) in the optimal control problem 4.1.1. Then there exist $\zeta \in \mathbb{R}^{n_y}$ and functions $p_f(\cdot) \in BV([t_0,t_f],\mathbb{R}^{n_x})$, $p_g(\cdot) \in L^\infty([t_0,t_f],\mathbb{R}^{n_y})$ with*

$$\lambda_f^*(h_1(\cdot)) = -\int_{t_0}^{t_f}\left(p_f(t)^\top + p_g(t)^\top g_x'[t]\right)h_1(t)dt, \qquad (4.1.35)$$

$$\lambda_g^*(h_2(\cdot)) = -\zeta^\top h_2(t_0) - \int_{t_0}^{t_f} p_g(t)^\top \dot{h}_2(t)dt, \qquad (4.1.36)$$

for every $h_1 \in L^\infty([t_0,t_f],\mathbb{R}^{n_x})$ and every $h_2 \in W^{1,\infty}([t_0,t_f],\mathbb{R}^{n_y})$.

Proof. In the absence of mixed control-state constraints equation (4.1.31) is equivalent with

$$\zeta^\top h_2(t_0) + \int_{t_0}^{t_f} p_f(t)^\top h_1(t)dt + \int_{t_0}^{t_f} p_g(t)^\top q(t)dt + \lambda_f^*(h_1(\cdot)) + \lambda_g^*(h_2(\cdot)) = 0, \qquad (4.1.37)$$

where $q(t) = \dot{h}_2(t) + g_x'[t]h_1(t)$. Setting $h_2(\cdot) \equiv \Theta$ yields

$$\lambda_f^*(h_1(\cdot)) = -\int_{t_0}^{t_f} \left(p_f(t)^\top + p_g(t)^\top g_x'[t]\right) h_1(t)dt. \tag{4.1.38}$$

Likewise, setting $h_1(\cdot) \equiv \Theta$ yields

$$\lambda_g^*(h_2(\cdot)) = -\zeta^\top h_2(t_0) - \int_{t_0}^{t_f} p_g(t)^\top \dot{h}_2(t)dt. \tag{4.1.39}$$

The function $p_f(\cdot)$ is of bounded variation and $p_g(\cdot)$ is essentially bounded. ■

In the presence of mixed control-state constraints we find

Corollary 4.1.9 *Let the assumptions of Theorem 4.1.7 be valid and let $\mathcal{U} = \mathbb{R}^{n_u}$. Let*

$$rank(c_u'[t]) = n_c \tag{4.1.40}$$

hold almost everywhere in $[t_0, t_f]$. Furthermore, let the pseudo-inverse of $c_u'[t]$

$$(c_u'[t])^+ := c_u'[t]^\top \left(c_u'[t]c_u'[t]^\top\right)^{-1}$$

be essentially bounded and let the matrix

$$\hat{M}(t) := g_x'[t] \cdot \left(f_y'[t] - f_u'[t](c_u'[t])^+ c_y'[t]\right) \tag{4.1.41}$$

be non-singular with essentially bounded inverse $\hat{M}^{-1}$ almost everywhere in $[t_0, t_f]$. Then there exist $\zeta \in \mathbb{R}^{n_y}$ and functions

$$p_f(\cdot) \in BV([t_0, t_f], \mathbb{R}^{n_x}), \quad p_g(\cdot) \in L^\infty([t_0, t_f], \mathbb{R}^{n_y}), \quad \eta \in L^\infty([t_0, t_f], \mathbb{R}^{n_c})$$

with

$$\begin{aligned}
\lambda_f^*(h_1(\cdot)) &= -\int_{t_0}^{t_f} \left(p_f(t)^\top + p_g(t)^\top g_x'[t]\right) h_1(t)dt, \\
\lambda_g^*(h_2(\cdot)) &= -\zeta^\top h_2(t_0) - \int_{t_0}^{t_f} p_g(t)^\top \dot{h}_2(t)dt, \\
\eta_1^*(k(\cdot)) &= \int_{t_0}^{t_f} \eta(t)^\top k(t)dt
\end{aligned}$$

for every $h_1 \in L^\infty([t_0, t_f], \mathbb{R}^{n_x})$, every $h_2 \in W^{1,\infty}([t_0, t_f], \mathbb{R}^{n_y})$, and every $k \in L^\infty([t_0, t_f], \mathbb{R}^{n_c})$.

Proof. The assumption $\mathcal{U} = \mathbb{R}^{n_u}$ implies $U_{ad} = L^\infty([t_0, t_f], \mathbb{R}^{n_u})$ and hence, inequality (4.1.23) turns into the equality

$$\int_{t_0}^{t_f} l_0 f_{0,u}'[t]u(t)dt + \eta_1^*(c_u'[\cdot]u(\cdot)) - \lambda_f^*(f_u'[\cdot]u(\cdot)) = 0 \tag{4.1.42}$$

for all $u \in L^\infty([t_0, t_f], \mathbb{R}^{n_u})$. Equation (4.1.31) for the particular choices $h_1(\cdot) = f_u'[\cdot]u(\cdot)$ and $h_2(\cdot) \equiv \Theta$ implies $q(t) = g_x'[t]h_1(t)$ and

$$-\lambda_f^*(h_1(\cdot)) = \int_{t_0}^{t_f} \left(p_f(t)^\top + p_g(t)^\top g_x'[t]\right) h_1(t)dt + \eta_1^*(c_x'[\cdot]x(\cdot) + c_y'[\cdot]y(\cdot)), \tag{4.1.43}$$

where x and y are determined by

$$\begin{aligned} \dot{x}(t) &= f'_x[t]x(t) + f'_y[t]y(t) + f'_u[t]u(t), \qquad x(t_0) = 0_{n_x}, & (4.1.44)\\ 0_{n_y} &= g'_x[t]x(t). & (4.1.45) \end{aligned}$$

Notice, that $x(t_0) = 0_{n_x}$ is consistent with (4.1.45). Introducing (4.1.43) into (4.1.42) and exploiting the linearity of the functional η_1^* yields

$$\int_{t_0}^{t_f} \mathcal{H}'_u[t]u(t)dt + \eta_1^*(k(\cdot)) = 0 \tag{4.1.46}$$

with

$$\mathcal{H}'_u[t] := l_0 f'_{0,u}[t] + \Big(p_f(t)^\top + p_g(t)^\top g'_x[t]\Big) f'_u[t]$$

and

$$k(t) := c'_x[t]x(t) + c'_y[t]y(t) + c'_u[t]u(t). \tag{4.1.47}$$

Due to the rank assumption (4.1.40) equation (4.1.47) can be solved for u with

$$u(t) = (c'_u[t])^+ \left(k(t) - c'_x[t]x(t) - c'_y[t]y(t)\right), \tag{4.1.48}$$

where $(c'_u[t])^+$ denotes the pseudo-inverse of $c'_u[t]$. Using the relation (4.1.48), equation (4.1.46) becomes

$$\int_{t_0}^{t_f} \mathcal{H}'_u[t](c'_u[t])^+ \left(k(t) - c'_x[t]x(t) - c'_y[t]y(t)\right) dt + \eta_1^*(k(\cdot)) \quad = \quad 0. \tag{4.1.49}$$

Replacing u in (4.1.44) by (4.1.48) yields

$$\begin{aligned} \dot{x}(t) &= \hat{f}_x[t]x(t) + \hat{f}_y[t]y(t) + f'_u[t](c'_u[t])^+ k(t), \qquad x(t_0) = 0_{n_x}, & (4.1.50)\\ 0_{n_y} &= g'_x[t]x(t) & (4.1.51) \end{aligned}$$

where

$$\hat{f}_x[t] := f'_x[t] - f'_u[t](c'_u[t])^+ c'_x[t], \qquad \hat{f}_y[t] := f'_y[t] - f'_u[t](c'_u[t])^+ c'_y[t].$$

Notice, that $\hat{M}(t) = g'_x[t] \cdot \hat{f}_y[t]$ is assumed to be non-singular with essentially bounded inverse $\hat{M}^{-1}$ on $[t_0, t_f]$. Using the solution formula in Lemma 4.1.6, the solution of (4.1.50)-(4.1.51) can be written as

$$x(t) \quad = \quad \hat{\Phi}(t) \int_{t_0}^{t} \hat{\Phi}^{-1}(\tau)\hat{h}(\tau)d\tau, \qquad y(t) = -\hat{M}(t)^{-1}\Big(\hat{q}(t) + \hat{Q}(t)x(t)\Big),$$

where $\hat{\Phi}$ solves the initial value problem

$$\dot{\hat{\Phi}}(t) = \Big(\hat{f}_x[t] - \hat{f}_y[t]\hat{M}(t)^{-1}\hat{Q}(t)\Big)\,\hat{\Phi}(t), \quad \hat{\Phi}(t_0) = I_{n_x},$$

$\hat{h}$ and $\hat{Q}$ are given by

$$\hat{h}(t) = f'_u[t]\left(c'_u[t]\right)^+ k(t) - \hat{f}_y[t]\hat{M}(t)^{-1}\hat{q}(t), \quad \hat{Q}(t) = \frac{d}{dt}g'_x[t] + g'_x[t]\hat{f}_x[t],$$

and $\hat{q}(t) = g'_x[t]f'_u[t]\,(c'_u[t])^+ k(t)$. Introducing $\hat{q}$ and $\hat{h}$ into the solution formula leads to the representations

$$\begin{aligned} x(t) &= \hat{\Phi}(t) \int_{t_0}^{t} w(\tau)^\top k(\tau)d\tau, & (4.1.52)\\ y(t) &= -\nu_1(t)^\top x(t) - \nu_2(t)^\top k(t) & (4.1.53) \end{aligned}$$

where

$$\begin{aligned}
w(t)^\top &:= \hat{\Phi}^{-1}(t)\left(f_u'[t] - \hat{f}_y[t]\hat{M}(t)^{-1}g_x'[t]f_u'[t]\right)\left(c_u'[t]\right)^+,\\
\nu_1(t)^\top &:= \hat{M}(t)^{-1}\hat{Q}(t),\\
\nu_2(t)^\top &:= \hat{M}(t)^{-1}g_x'[t]f_u'[t]\left(c_u'[t]\right)^+.
\end{aligned}$$

Equations (4.1.49), (4.1.52), and (4.1.53) and integration by parts yields

$$\begin{aligned}
-\eta_1^*(k(\cdot)) &= \int_{t_0}^{t_f} \mathcal{H}_u'[t](c_u'[t])^+\left(I + c_y'[t]\nu_2(t)^\top\right)k(t)dt\\
&\quad - \int_{t_0}^{t_f} \mathcal{H}_u'[t](c_u'[t])^+\hat{c}[t]\hat{\Phi}(t)\left(\int_{t_0}^{t} w(\tau)^\top k(\tau)d\tau\right)dt\\
&= \int_{t_0}^{t_f} \mathcal{H}_u'[t](c_u'[t])^+\left(I + c_y'[t]\nu_2(t)^\top\right)k(t)dt\\
&\quad - \int_{t_0}^{t_f}\left(\int_{t}^{t_f} \mathcal{H}_u'[\tau](c_u'[\tau])^+\hat{c}[\tau]\hat{\Phi}(\tau)d\tau\right)w(t)^\top k(t)dt
\end{aligned}$$

where

$$\hat{c}[t] := c_x'[t] - c_y'[t]\nu_1(t)^\top = c_x'[t] - c_y'[t]\hat{M}(t)^{-1}\hat{Q}(t).$$

With the definition

$$\eta(t)^\top := \left(\int_t^{t_f} \mathcal{H}_u'[\tau](c_u'[\tau])^+\hat{c}[\tau]\hat{\Phi}(\tau)d\tau\right)w(t)^\top - \mathcal{H}_u'[t](c_u'[t])^+\left(I + c_y'[t]\nu_2(t)^\top\right)$$

we thus obtained the representation

$$\eta_1^*(k(\cdot)) = \int_{t_0}^{t_f} \eta(t)^\top k(t)dt,$$

where η is an element of $L^\infty([t_0,t_f],\mathbb{R}^{n_c})$.
Introducing this representation into (4.1.31), using integration by parts and the solution formulas (4.1.26), (4.1.27) leads to

$$\begin{aligned}
&-\lambda_f^*(h_1(\cdot)) - \lambda_g^*(h_2(\cdot))\\
&\quad = \zeta^\top h_2(t_0) + \int_{t_0}^{t_f} p_f(t)^\top h_1(t)dt + \int_{t_0}^{t_f} \hat{p}_g(t)^\top q(t)dt\\
&\qquad + \int_{t_0}^{t_f} \eta(t)^\top \nu_3(t)\Phi(t)\left(\Pi h_2(t_0) + \int_{t_0}^{t}\Phi^{-1}(\tau)h_1(\tau)d\tau\right)dt\\
&\qquad - \int_{t_0}^{t_f} \eta(t)^\top \nu_3(t)\Phi(t)\left(\int_{t_0}^{t}\Phi^{-1}(\tau)f_y'[\tau]M(\tau)^{-1}q(\tau)d\tau\right)dt\\
&\quad = \left(\zeta^\top + \int_{t_0}^{t_f}\eta(t)^\top\nu_3(t)\Phi(t)\Pi dt\right)h_2(t_0) + \int_{t_0}^{t_f} p_f(t)^\top h_1(t)dt + \int_{t_0}^{t_f}\hat{p}_g(t)^\top q(t)dt\\
&\qquad + \int_{t_0}^{t_f}\left(\int_t^{t_f}\eta(\tau)^\top\nu_3(\tau)\Phi(\tau)d\tau\right)\Phi^{-1}(t)h_1(t)dt\\
&\qquad - \int_{t_0}^{t_f}\left(\int_t^{t_f}\eta(\tau)^\top\nu_3(\tau)\Phi(\tau)d\tau\right)\Phi^{-1}(t)f_y'[t]M(t)^{-1}q(t)dt,
\end{aligned}$$

where

$$\hat{p}_g(t)^\top := p_g(t)^\top - \eta(t)^\top c'_y[t]M(t)^{-1}, \qquad \nu_3(t) := c'_x[t] - c'_y[t]M(t)^{-1}Q(t).$$

With the abbreviations

$$\begin{aligned}
\hat{\zeta}^\top &:= \zeta^\top + \int_{t_0}^{t_f} \eta(t)^\top \nu_3(t)\Phi(t)\Pi dt,\\
\hat{p}(t)^\top &:= p_f(t)^\top + \left(\int_t^{t_f} \eta(\tau)^\top \nu_3(\tau)\Phi(\tau)d\tau\right)\Phi^{-1}(t),\\
\tilde{p}(t)^\top &:= \hat{p}_g(t)^\top - \left(\int_t^{t_f} \eta(\tau)^\top \nu_3(\tau)\Phi(\tau)d\tau\right)\Phi^{-1}(t)f'_y[t]M(t)^{-1},
\end{aligned}$$

we obtain

$$-\lambda_f^*(h_1(\cdot)) - \lambda_g^*(h_2(\cdot)) = \hat{\zeta}^\top h_2(t_0) + \int_{t_0}^{t_f} \hat{p}(t)^\top h_1(t)dt + \int_{t_0}^{t_f} \tilde{p}(t)^\top q(t)dt. \tag{4.1.54}$$

Equation (4.1.54) has the same structure as (4.1.37). Thus, the assertion follows by repeating the proof of Corollary 4.1.8 for (4.1.54). ■

The remarkable result in the preceeding corollaries is, that they provide useful representations of the functionals λ_f^*, λ_g^*, and η_1^*. Originally, these functionals are elements of the dual spaces of L^∞ and $W^{1,\infty}$ which have a very complicated structure. The exploitation of the fact, that λ_f^*, λ_g^*, and η_1^* are multipliers in an optimization problem, showed that these functionals actually are more regular.

4.1.2 Local Minimum Principles

Theorem 4.1.7 and Corollary 4.1.8 yield the following necessary conditions for the optimal control problem 4.1.1.

The *Hamilton function* is defined by

$$\mathcal{H}(t,x,y,u,\lambda_f,\lambda_g,l_0) := l_0 f_0(t,x,y,u) + \lambda_f^\top f(t,x,y,u) + \lambda_g^\top \left(g'_t(t,x) + g'_x(t,x)f(t,x,y,u)\right). \tag{4.1.55}$$

Notice, that this Hamilton function does not use the algebraic constraint (4.1.3) but its time derivative.

Theorem 4.1.10 (Local Minimum Principle for Optimal Control Problems without Mixed Control-State Constraints)

Let the following assumptions be fulfilled for the optimal control problem 4.1.1.

(i) Let the functions φ, f_0, f, s, ψ be continuous w.r.t. all arguments and continuously differentiable w.r.t. x, y, and u. Let g be twice continuously differentiable w.r.t. all arguments.

(ii) Let $\mathcal{U} \subseteq \mathbb{R}^{n_u}$ be a closed and convex set with non-empty interior.

(iii) Let $(\hat{x}, \hat{y}, \hat{u}) \in X$ be a weak local minimum of the optimal control problem.

(iv) Let Assumption 4.1.3 be valid.

(v) Let there be no mixed control-state constraints (4.1.5) in the optimal control problem 4.1.1.

Then there exist multipliers

$l_0 \in \mathbb{R},\ \lambda_f \in BV([t_0,t_f],\mathbb{R}^{n_x}),\ \lambda_g \in L^\infty([t_0,t_f],\mathbb{R}^{n_y}),\ \mu \in NBV([t_0,t_f],\mathbb{R}^{n_s}),\ \zeta \in \mathbb{R}^{n_y},\ \sigma \in \mathbb{R}^{n_\psi}$

such that the following conditions are satisfied:

(i) $l_0 \geq 0$, $(l_0,\zeta,\sigma,\lambda_f,\lambda_g,\mu) \neq \Theta$,

(ii) Adjoint equations*:*

$$\begin{aligned}\lambda_f(t) &= \lambda_f(t_f) + \int_t^{t_f} \mathcal{H}'_x(\tau,\hat{x}(\tau),\hat{y}(\tau),\hat{u}(\tau),\lambda_f(\tau),\lambda_g(\tau),l_0)^\top d\tau \\ &\quad + \sum_{i=1}^{n_s}\int_t^{t_f} s'_{i,x}(\tau,\hat{x}(\tau))^\top d\mu_i(\tau) \quad \textit{in } [t_0,t_f], \qquad (4.1.56)\\ 0_{n_y} &= \mathcal{H}'_y(t,\hat{x}(t),\hat{y}(t),\hat{u}(t),\lambda_f(t),\lambda_g(t),l_0)^\top \quad \textit{a.e. in } [t_0,t_f]. \qquad (4.1.57)\end{aligned}$$

(iii) Transversality conditions*:*

$$\lambda_f(t_0)^\top = -\left(l_0\varphi'_{x_0}(\hat{x}(t_0),\hat{x}(t_f)) + \sigma^\top\psi'_{x_0}(\hat{x}(t_0),\hat{x}(t_f)) + \zeta^\top g'_x(t_0,\hat{x}(t_0))\right), \quad (4.1.58)$$

$$\lambda_f(t_f)^\top = \quad l_0\varphi'_{x_f}(\hat{x}(t_0),\hat{x}(t_f)) + \sigma^\top\psi'_{x_f}(\hat{x}(t_0),\hat{x}(t_f)). \quad (4.1.59)$$

(iv) Optimality condition*: Almost everywhere in* $[t_0,t_f]$ *for all* $u \in \mathcal{U}$ *it holds*

$$\mathcal{H}'_u(t,\hat{x}(t),\hat{y}(t),\hat{u}(t),\lambda_f(t),\lambda_g(t),l_0)(u-\hat{u}(t)) \geq 0. \quad (4.1.60)$$

(v) Complementarity condition*:*
μ_i *is monotonically increasing on* $[t_0,t_f]$ *and constant on every interval* (t_1,t_2) *with* $t_1 < t_2$ *and* $s_i(t,\hat{x}(t)) < 0$ *for all* $t \in (t_1,t_2)$*.*

Proof. Under the above assumptions, Corollary 4.1.8 (with $\lambda_f = p_f$, $\lambda_g = p_g$) guarantees the existence of $\zeta \in \mathbb{R}^{n_y}$, $\lambda_f \in BV([t_0,t_f],\mathbb{R}^{n_x})$, and $\lambda_g \in L^\infty([t_0,t_f],\mathbb{R}^{n_y})$ such that for every $h_1 \in L^\infty([t_0,t_f],\mathbb{R}^{n_x})$, $h_2 \in W^{1,\infty}([t_0,t_f],\mathbb{R}^{n_y})$ it holds

$$\begin{aligned}\lambda_f^*(h_1(\cdot)) &= -\int_{t_0}^{t_f}\left(\lambda_f(t)^\top + \lambda_g(t)^\top g'_x[t]\right)h_1(t)dt, \qquad (4.1.61)\\ \lambda_g^*(h_2(\cdot)) &= -\zeta^\top h_2(t_0) - \int_{t_0}^{t_f}\lambda_g(t)^\top \dot{h}_2(t)dt. \qquad (4.1.62)\end{aligned}$$

(a) Equation (4.1.17) is equivalent with

$$\begin{aligned}&\left(l_0\varphi'_{x_0} + \sigma^\top\psi'_{x_0} + \zeta^\top g'_x[t_0]\right)x(t_0) + \left(l_0\varphi'_{x_f} + \sigma^\top\psi'_{x_f}\right)x(t_f)\\ &+ \int_{t_0}^{t_f} l_0 f'_{0,x}[t]x(t)dt + \int_{t_0}^{t_f}\left(\lambda_f(t)^\top + \lambda_g(t)^\top g'_x[t]\right)\left(f'_x[t]x(t) - \dot{x}(t)\right)dt\\ &+ \int_{t_0}^{t_f}\lambda_g(t)^\top\frac{d}{dt}\left(g'_x[t]x(t)\right)dt + \sum_{i=1}^{n_s}\int_{t_0}^{t_f} s'_{i,x}[t]x(t)d\mu_i(t) = 0\end{aligned}$$

for all $x \in W^{1,\infty}([t_0,t_f],\mathbb{R}^{n_x})$. Exploitation of

$$\frac{d}{dt}\left(g'_x[t]x(t)\right) = \left(\frac{d}{dt}g'_x[t]\right)x(t) + g'_x[t]\dot{x}(t)$$

yields

$$\begin{aligned}\left(l_0\varphi'_{x_0}+\sigma^\top\psi'_{x_0}+\zeta^\top g'_x[t_0]\right)x(t_0)+\left(l_0\varphi'_{x_f}+\sigma^\top\psi'_{x_f}\right)x(t_f)&\\+\int_{t_0}^{t_f}l_0f'_{0,x}[t]x(t)dt+\int_{t_0}^{t_f}\lambda_f(t)^\top\left(f'_x[t]x(t)-\dot{x}(t)\right)dt&\\+\int_{t_0}^{t_f}\lambda_g(t)^\top\left(g'_x[t]f'_x[t]+\frac{d}{dt}g'_x[t]\right)x(t)dt+\sum_{i=1}^{n_s}\int_{t_0}^{t_f}s'_{i,x}[t]x(t)d\mu_i(t)&=0,\end{aligned}$$

which can be written equivalently as

$$\begin{aligned}\left(l_0\varphi'_{x_0}+\sigma^\top\psi'_{x_0}+\zeta^\top g'_x[t_0]\right)x(t_0)+\left(l_0\varphi'_{x_f}+\sigma^\top\psi'_{x_f}\right)x(t_f)&\\+\int_{t_0}^{t_f}\mathcal{H}'_x[t]x(t)dt+\sum_{i=1}^{n_s}\int_{t_0}^{t_f}s'_{i,x}[t]x(t)d\mu_i(t)-\int_{t_0}^{t_f}\lambda_f(t)^\top\dot{x}(t)dt&=0\end{aligned}$$

for all $x\in W^{1,\infty}([t_0,t_f],\mathbb{R}^{n_x})$. Application of the computation rules for Stieltjes integrals yields

$$\begin{aligned}\left(l_0\varphi'_{x_0}+\sigma^\top\psi'_{x_0}+\zeta^\top g'_x[t_0]\right)x(t_0)+\left(l_0\varphi'_{x_f}+\sigma^\top\psi'_{x_f}\right)x(t_f)&\\+\int_{t_0}^{t_f}\mathcal{H}'_x[t]x(t)dt+\sum_{i=1}^{n_s}\int_{t_0}^{t_f}s'_{i,x}[t]x(t)d\mu_i(t)-\int_{t_0}^{t_f}\lambda_f(t)^\top dx(t)&=0.\end{aligned}$$

Integration by parts of the last term leads to

$$\begin{aligned}\left(l_0\varphi'_{x_0}+\sigma^\top\psi'_{x_0}+\zeta^\top g'_x[t_0]+\lambda_f(t_0)^\top\right)x(t_0)&\\+\left(l_0\varphi'_{x_f}+\sigma^\top\psi'_{x_f}-\lambda_f(t_f)^\top\right)x(t_f)&\\+\int_{t_0}^{t_f}\mathcal{H}'_x[t]x(t)dt+\sum_{i=1}^{n_s}\int_{t_0}^{t_f}s'_{i,x}[t]x(t)d\mu_i(t)+\int_{t_0}^{t_f}x(t)^\top d\lambda_f(t)&=0\end{aligned}$$

for all $x\in W^{1,\infty}([t_0,t_f],\mathbb{R}^{n_x})$. This is equivalent with

$$\begin{aligned}\left(l_0\varphi'_{x_0}+\sigma^\top\psi'_{x_0}+\zeta^\top g'_x[t_0]+\lambda_f(t_0)^\top\right)x(t_0)&\\+\left(l_0\varphi'_{x_f}+\sigma^\top\psi'_{x_f}-\lambda_f(t_f)^\top\right)x(t_f)&\\+\int_{t_0}^{t_f}x(t)^\top d\left(\lambda_f(t)-\int_t^{t_f}\mathcal{H}'_x[\tau]^\top d\tau-\sum_{i=1}^{n_s}\int_t^{t_f}s'_{i,x}[\tau]^\top d\mu_i(\tau)\right)&=0\end{aligned}$$

for all $x\in W^{1,\infty}([t_0,t_f],\mathbb{R}^{n_x})$. This implies (4.1.58)-(4.1.59) and

$$C=\lambda_f(t)-\int_t^{t_f}\mathcal{H}'_x[\tau]^\top d\tau-\sum_{i=1}^{n_s}\int_t^{t_f}s'_{i,x}[\tau]^\top d\mu_i(\tau)$$

for some constant vector C, cf. Lemma 2.8.1 and Remark 2.8.2. Evaluation of the last equation at $t=t_f$ yields $C=\lambda_f(t_f)$ and thus (4.1.56).

(b) Equation (4.1.18) is equivalent with

$$\int_{t_0}^{t_f}\mathcal{H}'_y[t]y(t)dt\;=\;0$$

for all $y\in L^\infty([t_0,t_f],\mathbb{R}^{n_y})$. This implies (4.1.57), cf. Lemma 2.8.3.

(c) Introducing (4.1.35)-(4.1.36) into (4.1.19) leads to the variational inequality

$$\int_{t_0}^{t_f} \mathcal{H}'_u[t](u(t) - \hat{u}(t))dt \geq 0$$

for all $u \in U_{ad}$. This implies the optimality condition, cf. Kirsch et al. [KWW78], p. 102.

(d) According to Theorem 4.1.7, (4.1.16) it holds $\eta_2^* \in K_2^+$, i.e.

$$\eta_2^*(z) = \sum_{i=1}^{n_s} \int_{t_0}^{t_f} z_i(t) d\mu_i(t) \geq 0$$

for all $z \in K_2 = \{z \in C([t_0, t_f], \mathbb{R}^{n_s}) \mid z(t) \geq 0_{n_s} \text{ in } [t_0, t_f]\}$. This implies, that μ_i is monotonically increasing, cf. Lemma 2.8.5. Finally, the condition $\eta_2^*(s(\cdot, \hat{x}(\cdot))) = 0$, i.e.

$$\eta_2^*(s(\cdot, \hat{x}(\cdot))) = \sum_{i=1}^{n_s} \int_{t_0}^{t_f} s_i(t, \hat{x}(t)) d\mu_i(t) = 0,$$

together with the monotonicity of μ_i implies that μ_i is constant in intervals with $s_i(t, \hat{x}(t)) < 0$, cf. Lemma 2.8.6.

■

Likewise, Theorem 4.1.7 and Corollary 4.1.9 yield the following necessary conditions for the optimal control problem 4.1.1. The *augmented Hamilton function* is defined by

$$\hat{\mathcal{H}}(t, x, y, u, \lambda_f, \lambda_g, \eta, l_0) := \mathcal{H}(t, x, y, u, \lambda_f, \lambda_g, l_0) + \eta^\top c(t, x, y, u). \tag{4.1.63}$$

Theorem 4.1.11 (Local Minimum Principle for Optimal Control Problems without Set Constraints)

Let the following assumptions be fulfilled for the optimal control problem 4.1.1.

(i) Let the functions $\varphi, f_0, f, c, s, \psi$ be continuous w.r.t. all arguments and continuously differentiable w.r.t. x, y, and u. Let g be twice continuously differentiable w.r.t. all arguments.

(ii) Let $(\hat{x}, \hat{y}, \hat{u}) \in X$ be a weak local minimum of the optimal control problem.

(iii) Let

$$\text{rank}\left(c'_u(t, \hat{x}(t), \hat{y}(t), \hat{u}(t))\right) = n_c$$

almost everywhere in $[t_0, t_f]$.

(iv) Let the pseudo-inverse of $c'_u[t]$

$$(c'_u[t])^+ = c'_u[t]^\top \left(c'_u[t] c'_u[t]^\top\right)^{-1} \tag{4.1.64}$$

be essentially bounded and let the matrix

$$g'_x[t] \cdot \left(f'_y[t] - f'_u[t](c'_u[t])^+ c'_y[t]\right) \tag{4.1.65}$$

be non-singular almost everywhere with essentially bounded inverse in $[t_0, t_f]$.

(v) Let Assumption 4.1.3 be valid.

(vi) Let $\mathcal{U} = \mathbb{R}^{n_u}$.

Then there exist multipliers $l_0 \in \mathbb{R}$, $\zeta \in \mathbb{R}^{n_y}$, $\sigma \in \mathbb{R}^{n_\psi}$,

$$\lambda_f \in BV([t_0,t_f],\mathbb{R}^{n_x}),\ \lambda_g \in L^\infty([t_0,t_f],\mathbb{R}^{n_y}),\ \eta \in L^\infty([t_0,t_f],\mathbb{R}^{n_c}),\ \mu \in NBV([t_0,t_f],\mathbb{R}^{n_s})$$

such that the following conditions are satisfied:

(i) $l_0 \geq 0$, $(l_0,\zeta,\sigma,\lambda_f,\lambda_g,\eta,\mu) \neq \Theta$,

(ii) Adjoint equations*:*

$$\begin{aligned}\lambda_f(t) &= \lambda_f(t_f) + \int_t^{t_f} \hat{\mathcal{H}}'_x(\tau,\hat{x}(\tau),\hat{y}(\tau),\hat{u}(\tau),\lambda_f(\tau),\lambda_g(\tau),\eta(\tau),l_0)^\top d\tau \\ &\quad + \sum_{i=1}^{n_s} \int_t^{t_f} s'_{i,x}(\tau,\hat{x}(\tau))^\top d\mu_i(\tau) \quad \text{in } [t_0,t_f], \qquad (4.1.66)\\ 0_{n_y} &= \hat{\mathcal{H}}'_y(t,\hat{x}(t),\hat{y}(t),\hat{u}(t),\lambda_f(t),\lambda_g(t),\eta(t),l_0)^\top \text{ a.e. in } [t_0,t_f]. \qquad (4.1.67)\end{aligned}$$

(iii) Transversality conditions*:*

$$\begin{aligned}\lambda_f(t_0)^\top &= -\left(l_0\varphi'_{x_0}(\hat{x}(t_0),\hat{x}(t_f)) + \sigma^\top\psi'_{x_0}(\hat{x}(t_0),\hat{x}(t_f)) + \zeta^\top g'_x(t_0,\hat{x}(t_0))\right), \qquad (4.1.68)\\ \lambda_f(t_f)^\top &= \quad l_0\varphi'_{x_f}(\hat{x}(t_0),\hat{x}(t_f)) + \sigma^\top\psi'_{x_f}(\hat{x}(t_0),\hat{x}(t_f)). \qquad (4.1.69)\end{aligned}$$

(iv) Optimality condition*: A.e. in* $[t_0,t_f]$ *it holds*

$$\hat{\mathcal{H}}'_u(t,\hat{x}(t),\hat{y}(t),\hat{u}(t),\lambda_f(t),\lambda_g(t),\eta(t),l_0) = 0_{n_u}. \qquad (4.1.70)$$

(v) Complementarity conditions*:*
Almost everywhere in $[t_0,t_f]$ *it holds*

$$\eta(t)^\top c(t,\hat{x}(t),\hat{y}(t),\hat{u}(t)) \;=\; 0, \qquad \eta(t) \geq 0_{n_c}.$$

μ_i *is monotonically increasing on* $[t_0,t_f]$ *and constant on every interval* (t_1,t_2) *with* $t_1 < t_2$ *and* $s_i(t,\hat{x}(t)) < 0$ *for all* $t \in (t_1,t_2)$.

Proof. Corollary 4.1.9 yields the existence of the functions λ_f, λ_g, and η and provides representations of the functionals λ_f^*, λ_g^*, and η_1^*. Parts (a)-(c) and the second assertion of (d) follow as in the proof of Theorem 4.1.10.
From

$$\eta_1^* \in K_1^+ \quad \Leftrightarrow \quad \int_{t_0}^{t_f} \eta(t)^\top z(t)dt \geq 0 \quad \forall z \in K_1$$

and

$$\eta_1^*(c(\cdot,\hat{x}(\cdot),\hat{y}(\cdot),\hat{u}(\cdot))) = 0 \quad \Leftrightarrow \quad \int_{t_0}^{t_f} \eta(t)^\top c[t]dt = 0$$

we conclude that

$$\eta(t) \;\geq\; 0_{n_c}, \quad \eta(t)^\top c[t] = 0, \quad \text{a.e. in } [t_0,t_f],$$

cf. Lemma 2.8.5. ■

The following considerations apply to both, Theorem 4.1.10 and Theorem 4.1.11 and differ only in the Hamilton functions $\mathcal{H}$ and $\hat{\mathcal{H}}$, respectively. Hence, we restrict the discussion to the situation of Theorem 4.1.10.

The multiplier μ is of bounded variation. Hence, it has at most countably many jump points and μ can be expressed as $\mu = \mu_a + \mu_d + \mu_s$, where μ_a is absolutely continuous, μ_d is a jump function, and μ_s is singular (continuous, non-constant, $\dot{\mu}_s = 0$ a.e.). Hence, the adjoint equation (4.1.56) can be written as

$$\begin{aligned}\lambda_f(t) &= \lambda_f(t_f) + \int_t^{t_f} \mathcal{H}'_x(\tau, \hat{x}(\tau), \hat{y}(\tau), \hat{u}(\tau), \lambda_f(\tau), \lambda_g(\tau), l_0)^\top d\tau \\ &\quad + \sum_{i=1}^{n_s} \left(\int_t^{t_f} s'_{i,x}(\tau, \hat{x}(\tau))^\top d\mu_{i,a}(\tau) + \int_t^{t_f} s'_{i,x}(\tau, \hat{x}(\tau))^\top d\mu_{i,d}(\tau) \right. \\ &\qquad\qquad \left. + \int_t^{t_f} s'_{i,x}(\tau, \hat{x}(\tau))^\top d\mu_{i,s}(\tau) \right)\end{aligned}$$

for all $t \in [t_0, t_f]$. Notice, that λ_f is continuous from the right in (t_0, t_f) since μ is normalized. Let $\{t_j\}$, $j \in \mathcal{J}$, be the jump points of μ. Then, at every jump point t_j it holds

$$\begin{aligned}&\lim_{\varepsilon \downarrow 0} \left(\int_{t_j}^{t_f} s'_{i,x}(\tau, \hat{x}(\tau))^\top d\mu_{i,d}(\tau) - \int_{t_j-\varepsilon}^{t_f} s'_{i,x}(\tau, \hat{x}(\tau))^\top d\mu_{i,d}(\tau) \right) \\ &\qquad = -s'_{i,x}(t_j, \hat{x}(t_j))^\top \left(\mu_{i,d}(t_j) - \mu_{i,d}(t_j-) \right).\end{aligned}$$

Since μ_a is absolutely continuous and μ_s is continuous we obtain the *jump-condition*

$$\lambda_f(t_j) - \lambda_f(t_j-) = -\sum_{i=1}^{n_s} s'_{i,x}(t_j, \hat{x}(t_j))^\top \left(\mu_i(t_j) - \mu_i(t_j-) \right), \qquad j \in \mathcal{J}.$$

Furthermore, since every function of bounded variation is differentiable almost everywhere, μ and λ_f are differentiable almost everywhere. Exploitation of Lemma 2.4.1 proves

Corollary 4.1.12 *Let the assumptions of Theorem 4.1.10 be fulfilled. Then, λ_f is differentiable almost everywhere in $[t_0, t_f]$ with*

$$\dot{\lambda}_f(t) = -\mathcal{H}'_x(t, \hat{x}(t), \hat{y}(t), \hat{u}(t), \lambda_f(t), \lambda_g(t), l_0)^\top - \sum_{i=1}^{n_s} s'_{i,x}(t, \hat{x}(t))^\top \dot{\mu}_i(t). \tag{4.1.71}$$

Furthermore, the jump conditions

$$\lambda_f(t_j) - \lambda_f(t_j-) = -\sum_{i=1}^{n_s} s'_{i,x}(t_j, \hat{x}(t_j))^\top \left(\mu_i(t_j) - \mu_i(t_j-) \right) \tag{4.1.72}$$

hold at every point $t_j \in (t_0, t_f)$ of discontinuity of the multiplier μ.

Notice, that μ_i in (4.1.71) can be replaced by the absolutely continuous component $\mu_{i,a}$ since the derivatives of the jump component $\mu_{i,d}$ and the singular component $\mu_{i,s}$ are zero almost everywhere. However, $\lambda_f(t)$ cannot be reconstructed by simple integration of $\dot{\lambda}_f$.

A special case arises, if no state constraints are present. Then, the adjoint variable λ_f is even absolutely continuous, i.e. $\lambda_f \in W^{1,\infty}([t_0, t_f], \mathbb{R}^{n_x})$, and the adjoint equations (4.1.56)-(4.1.57) become

$$\begin{aligned}\dot{\lambda}_f(t) &= -\mathcal{H}'_x(t, \hat{x}(t), \hat{y}(t), \hat{u}(t), \lambda_f(t), \lambda_g(t), l_0)^\top \quad \text{a.e. in } [t_0, t_f], & (4.1.73) \\ 0_{n_y} &= \mathcal{H}'_y(t, \hat{x}(t), \hat{y}(t), \hat{u}(t), \lambda_f(t), \lambda_g(t), l_0)^\top \quad \text{a.e. in } [t_0, t_f]. & (4.1.74)\end{aligned}$$

The adjoint equations (4.1.73) and (4.1.74) form a DAE system of index one for λ_f and λ_g, where λ_f is the differential variable and λ_g denotes the algebraic variable. This follows from (4.1.74), which is given by

$$0_{n_y} = l_0 \left(f'_{0,y}[t]\right)^\top + \left(f'_y[t]\right)^\top \lambda_f(t) + \left(g'_x[t] \cdot f'_y[t]\right)^\top \lambda_g(t).$$

Since $g'_x[t] \cdot f'_y[t]$ is non-singular, we obtain

$$\lambda_g(t) = -\left(\left(g'_x[t] \cdot f'_y[t]\right)^{-1}\right)^\top \left(l_0 \left(f'_{0,y}[t]\right)^\top + \left(f'_y[t]\right)^\top \lambda_f(t)\right).$$

Remark 4.1.13 *Notice, that the adjoint system (4.1.73)-(4.1.74) is an index-1 DAE while the original DAE was index-2 according to Assumption 4.1.3.*

In this section we concentrated on local minimum principles only. The term 'local' is due to the fact, that the optimality conditions (4.1.60) and (4.1.70), respectively, can be interpreted as necessary conditions for a local minimum of the Hamilton function and the augmented Hamilton function, respectively. However, there are also *global minimum principles.* In a global minimum principle the optimality condition is given by

$$\hat{\mathcal{H}}(t, \hat{x}(t), \hat{y}(t), \hat{u}(t), \lambda_f(t), \lambda_g(t), \eta(t), l_0) \leq \hat{\mathcal{H}}(t, \hat{x}(t), \hat{y}(t), u, \lambda_f(t), \lambda_g(t), \eta(t), l_0)$$

for all $u \in \mathcal{U}$ with $c(t, \hat{x}(t), \hat{y}(t), u) \leq 0$ almost everywhere in $[t_0, t_f]$. If the Hamilton function and the function c is convex w.r.t. the control u, then both conditions are equivalent. Furthermore, the main important difference between local and global minimum principle is, that $\mathcal{U}$ can be an arbitrary subset of $\mathbb{R}^{n_u}$ in the global case, e.g. a discrete set. In our approach we had to assume that $\mathcal{U}$ is a convex set with non-empty interior. Proofs for global minimum resp. maximum principles in the context of ordinary differential equations subject to pure state constraints can be found in, e.g., Girsanov [Gir72] and Ioffe and Tihomirov [IT79], pp. 147-159, 241-253, compare also Hestenes [Hes66] for global maximum principles for some types of DAEs.

We favored the statement of necessary conditions in terms of local minimum principles because these conditions are closer to the finite dimensional case. A connection arises, if the infinite dimensional optimal control problem is discretized and transformed into a finite dimensional optimization problem. The formulation of the necessary Fritz-John conditions leads to the discrete minimum principle. These conditions are comparable to the infinite dimensional local conditions. This is different for the global minimum principle. In the finite dimensional case a comparable global minimum principle only holds approximately, cf. Mordukhovich [Mor88], or under additional convexity like assumptions, cf. Ioffe and Tihomirov [IT79], p. 278.

4.1.3 Regularity

We state conditions which ensure that the multiplier l_0 is not zero and, without loss of generality, can be normalized to one. Again, we consider the optimal control problem 4.1.1 and the optimization problem 4.1.5, respectively. The functions $\varphi, f_0, f, \psi, c, s$ are assumed to be continuous w.r.t. all arguments and continuously differentiable w.r.t. x, y, and u, g is assumed to be twice continuously differentiable w.r.t. all arguments.

As mentioned before, this implies that F is Fréchet-differentiable and G and H are continuously Fréchet-differentiable. According to Corollary 3.5.4 the following Mangasarian-Fromowitz conditions imply $l_0 = 1$:

(i) $H'(\hat{x}, \hat{y}, \hat{u})$ is surjective.

(ii) There exists some $(x, y, u) \in \operatorname{int}(S - \{(\hat{x}, \hat{y}, \hat{u})\})$ with

$$\begin{aligned} H'(\hat{x}, \hat{y}, \hat{u})(x, y, u) &= \Theta_Z, \\ G(\hat{x}, \hat{y}, \hat{u}) + G'(\hat{x}, \hat{y}, \hat{u})(x, y, u) &\in \operatorname{int}(K). \end{aligned}$$

Herein, the interior of $S = W^{1,\infty}([t_0, t_f], \mathbb{R}^{n_x}) \times L^\infty([t_0, t_f], \mathbb{R}^{n_y}) \times U_{ad}$ is given by

$$\operatorname{int}(S) = \{(x, y, u) \in X \mid \exists \varepsilon > 0 : B_\varepsilon(u(t)) \subseteq \mathcal{U} \text{ a.e. in } [t_0, t_f]\},$$

where $B_\varepsilon(u)$ denotes the open ball with radius ε around u. The interior of K_1 is given by

$$\begin{aligned} \operatorname{int}(K_1) &= \{z \in L^\infty([t_0, t_f], \mathbb{R}^{n_c}) \mid \exists \varepsilon > 0 : U_\varepsilon(z) \subseteq K_1\} \\ &= \{z \in L^\infty([t_0, t_f], \mathbb{R}^{n_c}) \mid \exists \varepsilon > 0 : z_i(t) \geq \varepsilon \text{ a.e. in } [t_0, t_f],\ i = 1, \ldots, n_c\}. \end{aligned}$$

The interior of K_2 is given by

$$\operatorname{int}(K_2) = \{z \in C([t_0, t_f], \mathbb{R}^{n_s}) \mid z(t) > 0_{n_s} \text{ in } [t_0, t_f]\}.$$

A sufficient condition for (i) to hold is given by the following lemma, which is an immediate consequence of part (c) of Lemma 4.1.6. It is an extension to DAEs of a lemma that has been used by several authors (e.g. K. Malanowski) for control problems with mixed control-state constraints or pure state constraints.

Lemma 4.1.14 *Let Assumption 4.1.3 be valid and let*

$$\mathit{rank}\left(\left(\psi'_{x_0}\Phi(t_0) + \psi'_{x_f}\Phi(t_f)\right)\Gamma\right) = n_\psi,$$

where Φ is the fundamental solution of the homogeneous linear differential equation

$$\dot{\Phi}(t) = A(t)\Phi(t), \quad \Phi(t_0) = I_{n_x}, \qquad t \in [t_0, t_f]$$

and the columns of Γ constitute an orthonormal basis of $\mathit{ker}(g'_x[t_0])$ and

$$\begin{aligned} M(t) &:= g'_x[t] \cdot f'_y[t], \\ A(t) &:= f'_x[t] - f'_y[t]M(t)^{-1}Q(t), \\ h(t) &:= h_1(t) - f'_y[t]M(t)^{-1}q(t), \\ Q(t) &:= \frac{d}{dt}g'_x[t] + g'_x[t] \cdot f'_x[t], \\ q(t) &:= \dot{h}_2(t) + g'_x[t]h_1(t). \end{aligned}$$

Then $H'(\hat{x}, \hat{y}, \hat{u})$ in Problem 4.1.5 is surjective.

Proof. Let $h_1 \in L^\infty([t_0, t_f], \mathbb{R}^{n_x})$, $h_2 \in W^{1,\infty}([t_0, t_f], \mathbb{R}^{n_y})$, and $h_3 \in \mathbb{R}^{n_\psi}$ be given. Consider the boundary value problem

$$\begin{aligned} H'_1(\hat{x}, \hat{y}, \hat{u})(x, y, u)(t) &= h_1(t) \quad \text{a.e. in } [t_0, t_f], \\ H'_2(\hat{x}, \hat{y}, \hat{u})(x, y, u)(t) &= h_2(t) \quad \text{in } [t_0, t_f], \\ H'_3(\hat{x}, \hat{y}, \hat{u})(x, y, u) &= h_3. \end{aligned}$$

More explicitly, the boundary value problem becomes

$$\begin{aligned}
-\dot{x}(t) + f_x'[t]x(t) + f_y'[t]y(t) + f_u'[t]u(t) &= h_1(t) \quad \text{a.e. in } [t_0, t_f], \\
g_x'[t]x(t) &= h_2(t) \quad \text{a.e. in } [t_0, t_f], \\
\psi_{x_0}'x(t_0) + \psi_{x_f}'x(t_f) &= -h_3.
\end{aligned}$$

By the rank assumption and Assumption 4.1.3 all assumptions of Lemma 4.1.6 are satisfied and the boundary value problem is solvable. This shows the surjectivity of the mapping $H'(\hat{x}, \hat{y}, \hat{u})$.

■

Condition (ii) is satisfied if there exist $x \in W^{1,\infty}([t_0, t_f], \mathbb{R}^{n_x})$, $y \in L^{\infty}([t_0, t_f], \mathbb{R}^{n_y})$, $u \in \text{int}(U_{ad} - \{\hat{u}\})$, and $\varepsilon > 0$ satisfying

$$\begin{aligned}
c[t] + c_x'[t]x(t) + c_y'[t]y(t) + c_u'[t]u(t) &\leq -\varepsilon \cdot e \quad \text{a.e. in } [t_0, t_f], && (4.1.75) \\
s[t] + s_x'[t]x(t) &< 0_{n_s} \quad \text{in } [t_0, t_f], && (4.1.76) \\
f_x'[t]x(t) + f_y'[t]y(t) + f_u'[t]u(t) - \dot{x}(t) &= 0_{n_x} \quad \text{a.e. in } [t_0, t_f], && (4.1.77) \\
g_x'[t]x(t) &= 0_{n_y} \quad \text{in } [t_0, t_f], && (4.1.78) \\
\psi_{x_0}'x(t_0) + \psi_{x_f}'x(t_f) &= 0_{n_\psi}, && (4.1.79)
\end{aligned}$$

where $e = (1, \ldots, 1)^\top \in \mathbb{R}^{n_c}$.
Hence, we conclude

Theorem 4.1.15 *Let the assumptions of Theorems 4.1.7, 4.1.10 or 4.1.11, and Lemma 4.1.14 be fulfilled. Furthermore, let there exist $x \in W^{1,\infty}([t_0, t_f], \mathbb{R}^{n_x})$, $y \in L^{\infty}([t_0, t_f], \mathbb{R}^{n_y})$, and $u \in int(U_{ad} - \{\hat{u}\})$ satisfying (4.1.75)-(4.1.79). Then it holds $l_0 = 1$ in Theorems 4.1.7 and 4.1.10 or 4.1.11, respectively.*

4.1.4 Example

We will apply the local minimum principle in Theorem 4.1.10 to the subsequent index-2 DAE optimal control problem without state constraints. The task is to minimize

$$\int_0^3 u(t)^2 dt$$

subject to the equations of motion of the mathematical pendulum in Gear-Gupta-Leimkuhler (GGL) formulation given by

$$\begin{aligned}
\dot{x}_1(t) &= x_3(t) - 2x_1(t)y_2(t), && (4.1.80)\\
\dot{x}_2(t) &= x_4(t) - 2x_2(t)y_2(t), && (4.1.81)\\
\dot{x}_3(t) &= \quad\;\; - 2x_1(t)y_1(t) + u(t)x_2(t), && (4.1.82)\\
\dot{x}_4(t) &= -g - 2x_2(t)y_1(t) - u(t)x_1(t), && (4.1.83)\\
0 &= x_1(t)x_3(t) + x_2(t)x_4(t), && (4.1.84)\\
0 &= x_1(t)^2 + x_2(t)^2 - 1, && (4.1.85)
\end{aligned}$$

and the boundary conditions

$$\psi(x(0), x(3)) := (x_1(0) - 1, x_2(0), x_3(0), x_4(0), x_1(3), x_3(3))^\top = 0_6. \qquad (4.1.86)$$

Herein, $g = 9.81$ denotes acceleration due to gravity. The control u is not restricted, i.e. $\mathcal{U} = \mathbb{R}$. With $x = (x_1, x_2, x_3, x_4)^\top$, $y = (y_1, y_2)^\top$, $f_0(u) = u^2$, and

$$\begin{aligned}
f(x, y, u) &= (x_3 - 2x_1y_2, x_4 - 2x_2y_2, -2x_1y_1 + ux_2, -g - 2x_2y_1 - ux_1)^\top,\\
g(x) &= \left(x_1x_3 + x_2x_4, x_1^2 + x_2^2 - 1\right)^\top
\end{aligned}$$

the problem has the structure of Problem 4.1.1. The matrix

$$\begin{aligned}
g'_x(x) \cdot f'_y(x, y, u) &= \begin{pmatrix} x_3 & x_4 & x_1 & x_2 \\ 2x_1 & 2x_2 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & -2x_1 \\ 0 & -2x_2 \\ -2x_1 & 0 \\ -2x_2 & 0 \end{pmatrix}\\
&= \begin{pmatrix} -2(x_1^2 + x_2^2) & -2(x_1x_3 + x_2x_4) \\ 0 & -4(x_1^2 + x_2^2) \end{pmatrix}\\
&= \begin{pmatrix} -2 & 0 \\ 0 & -4 \end{pmatrix}
\end{aligned}$$

is non-singular in a local minimum, hence the DAE has index two and Assumption 4.1.3 is satisfied. The remaining assumptions of Theorem 4.1.10 are satisfied as well and necessarily there exist functions $\lambda_f = (\lambda_{f,1}, \lambda_{f,2}, \lambda_{f,3}, \lambda_{f,4})^\top \in W^{1,\infty}([0, 3], \mathbb{R}^4)$, $\lambda_g = (\lambda_{g,1}, \lambda_{g,2})^\top \in L^\infty([0, 3], \mathbb{R}^2)$, and vectors $\zeta = (\zeta_1, \zeta_2)^\top$ and $\sigma = (\sigma_1, \ldots, \sigma_6)^\top$ such that the adjoint equations (4.1.73)-(4.1.74), the transversality conditions (4.1.58)-(4.1.59), and the optimality condition (4.1.60) are satisfied. The Hamilton function (4.1.55) calculates to

$$\begin{aligned}
\mathcal{H}(x, y, u, \lambda_f, \lambda_g, l_0) &= l_0u^2 + \lambda_{f,1}(x_3 - 2x_1y_2) + \lambda_{f,2}(x_4 - 2x_2y_2)\\
&\quad + \lambda_{f,3}(-2x_1y_1 + ux_2) + \lambda_{f,4}(-g - 2x_2y_1 - ux_1)\\
&\quad + \lambda_{g,1}\left(-gx_2 - 2y_1(x_1^2 + x_2^2) + x_3^2 + x_4^2 - 2y_2(x_1x_3 + x_2x_4)\right)\\
&\quad + \lambda_{g,2}\left(2(x_1x_3 + x_2x_4) - 4y_2(x_1^2 + x_2^2)\right).
\end{aligned}$$

In the sequel we assume $l_0 = 1$ (actually, the Mangasarian-Fromowitz condition is satisfied). Then, the optimality condition (4.1.60) yields

$$0 = 2u + \lambda_{f,3}x_2 - \lambda_{f,4}x_1 \quad \Rightarrow \quad u = \frac{\lambda_{f,4}x_1 - \lambda_{f,3}x_2}{2}. \tag{4.1.87}$$

The transversality conditions (4.1.58)-(4.1.59) are given by

$$\lambda_f(0) = (-\sigma_1 - 2\zeta_2, -\sigma_2, -\sigma_3 - \zeta_1, -\sigma_4)^\top, \quad \lambda_f(3) = (\sigma_5, 0, \sigma_6, 0)^\top.$$

The adjoint equations (4.1.73)-(4.1.74) yield

$$\begin{aligned}
\dot{\lambda}_{f,1} &= 2\left(\lambda_{f,1}y_2 + \lambda_{f,3}y_1\right) + \lambda_{f,4}u \\
&\quad -\lambda_{g,1}\left(-4y_1x_1 - 2y_2x_3\right) - \lambda_{g,2}\left(2x_3 - 8y_2x_1\right), && (4.1.88)\\
\dot{\lambda}_{f,2} &= 2\left(\lambda_{f,2}y_2 + \lambda_{f,4}y_1\right) - \lambda_{f,3}u \\
&\quad -\lambda_{g,1}\left(-g - 4y_1x_2 - 2y_2x_4\right) - \lambda_{g,2}\left(2x_4 - 8y_2x_2\right), && (4.1.89)\\
\dot{\lambda}_{f,3} &= -\lambda_{f,1} - \lambda_{g,1}\left(2x_3 - 2x_1y_2\right) - 2\lambda_{g,2}x_1, && (4.1.90)\\
\dot{\lambda}_{f,4} &= -\lambda_{f,2} - \lambda_{g,1}\left(2x_4 - 2x_2y_2\right) - 2\lambda_{g,2}x_2, && (4.1.91)\\
0 &= -2\left(\lambda_{f,3}x_1 + \lambda_{f,4}x_2 + \lambda_{g,1}(x_1^2 + x_2^2)\right), && (4.1.92)\\
0 &= -2\left(\lambda_{f,1}x_1 + \lambda_{f,2}x_2 + \lambda_{g,1}(x_1x_3 + x_2x_4) + 2\lambda_{g,2}(x_1^2 + x_2^2)\right). && (4.1.93)
\end{aligned}$$

Notice, that consistent initial values for $\lambda_{g_1}(0)$ and $\lambda_{g,2}(0)$ could be calculated from (4.1.92)-(4.1.93) and (4.1.84)-(4.1.85) by

$$\lambda_{g,1} = -\lambda_{f,3}x_1 - \lambda_{f,4}x_2, \quad \lambda_{g,2} = \frac{-\lambda_{f,1}x_1 - \lambda_{f,2}x_2}{2}.$$

The differential equations (4.1.80)-(4.1.85) and (4.1.88)-(4.1.93) with u replaced by (4.1.87) together with the boundary conditions (4.1.86) and $\lambda_{f,2}(3) = 0$, $\lambda_{f,4}(3) = 0$ form a two point boundary value problem (BVP). Notice that the DAE system has index-1 constraints (4.1.92)-(4.1.93) for λ_g as well as index-2 constraints (4.1.84)-(4.1.85) for y.

Numerically, the BVP is solved by a single shooting method as follows. Let $z = (\sigma_1 + 2\zeta_2, \sigma_2, \sigma_3 + \zeta_1, \sigma_4)^\top$ denote the unknown initial values of $-\lambda_f$ and let $x(t;z)$, $y(t;z)$, $\lambda_f(t;z)$, $\lambda_g(t;z)$ denote the solution of the initial value problem given by (4.1.80)-(4.1.85), (4.1.88)-(4.1.93) and the initial conditions $x(0) = (1,0,0,0)^\top$ and $\lambda_f(0) = -z$. Then, the BVP is solvable, if the nonlinear equation

$$G(z) := (x_1(3;z), x_3(3;z), \lambda_{f,2}(3;z), \lambda_{f,4}(3;z)) = (0,0,0,0)$$

is solvable. Numerically, the nonlinear equation is solved by Newton's method. The required Jacobian $G'_z(z)$ is obtained by a sensitivity analysis of the initial value problem w.r.t. z. Herein, the sensitivity DAE associated with the differential equations is employed. Figures 1–3 show the numerical solution obtained from Newton's method. Notice, that the initial conditions in (4.1.86) and the algebraic equations (4.1.84) and (4.1.85) contain redundant information. Hence, the multipliers σ and ζ are not unique, e.g. one may set $\zeta = 0_2$. In order to obtain unique multipliers, one could dispense with the first and third initial condition in (4.1.86), since these are determined by (4.1.84) and (4.1.85).

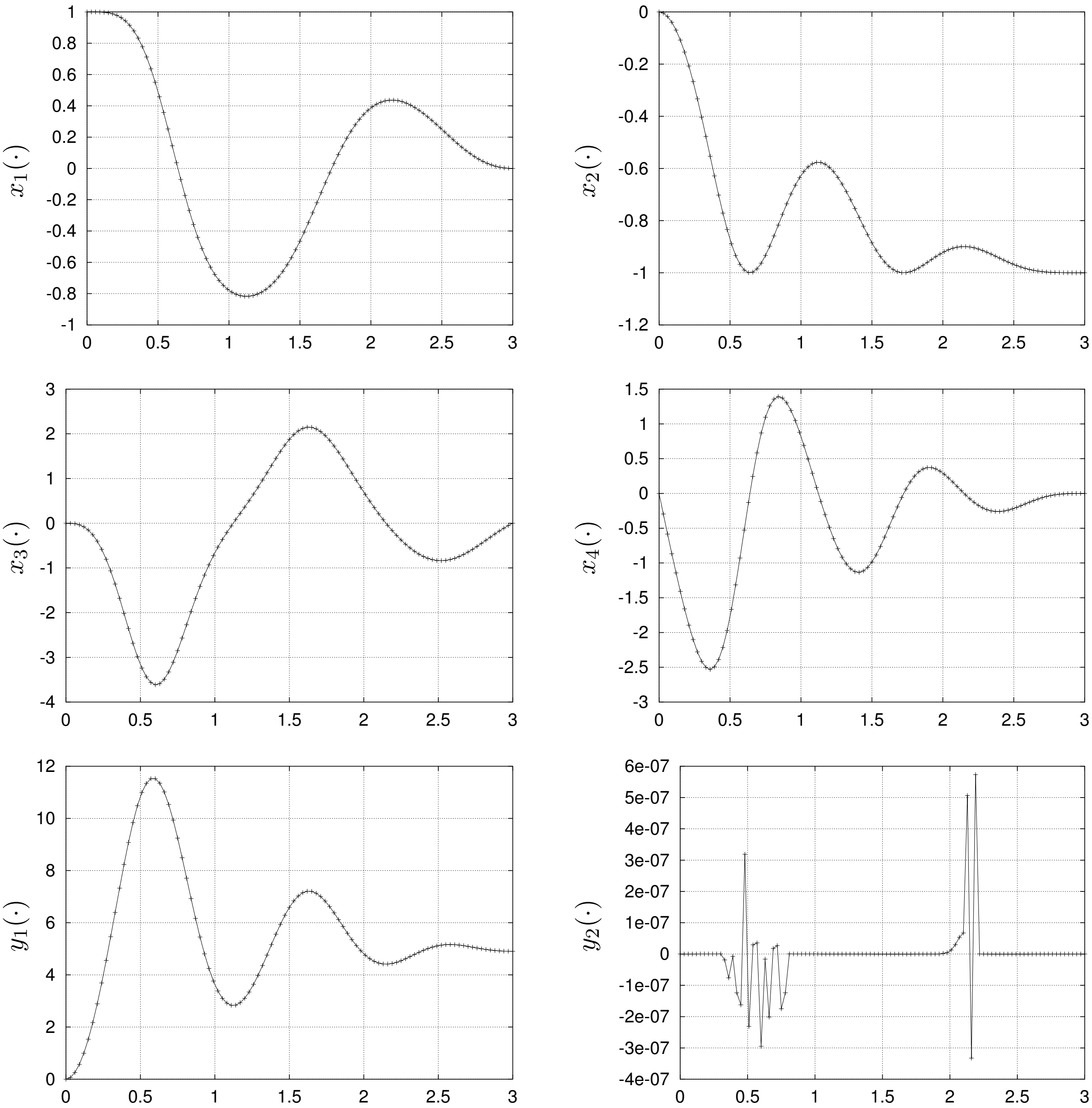


Figure 4.1: Numerical solution of BVP resulting from the minimum principle: Differential variable $x(t)$ and algebraic variable $y(t)$ for $t \in [0, 3]$.

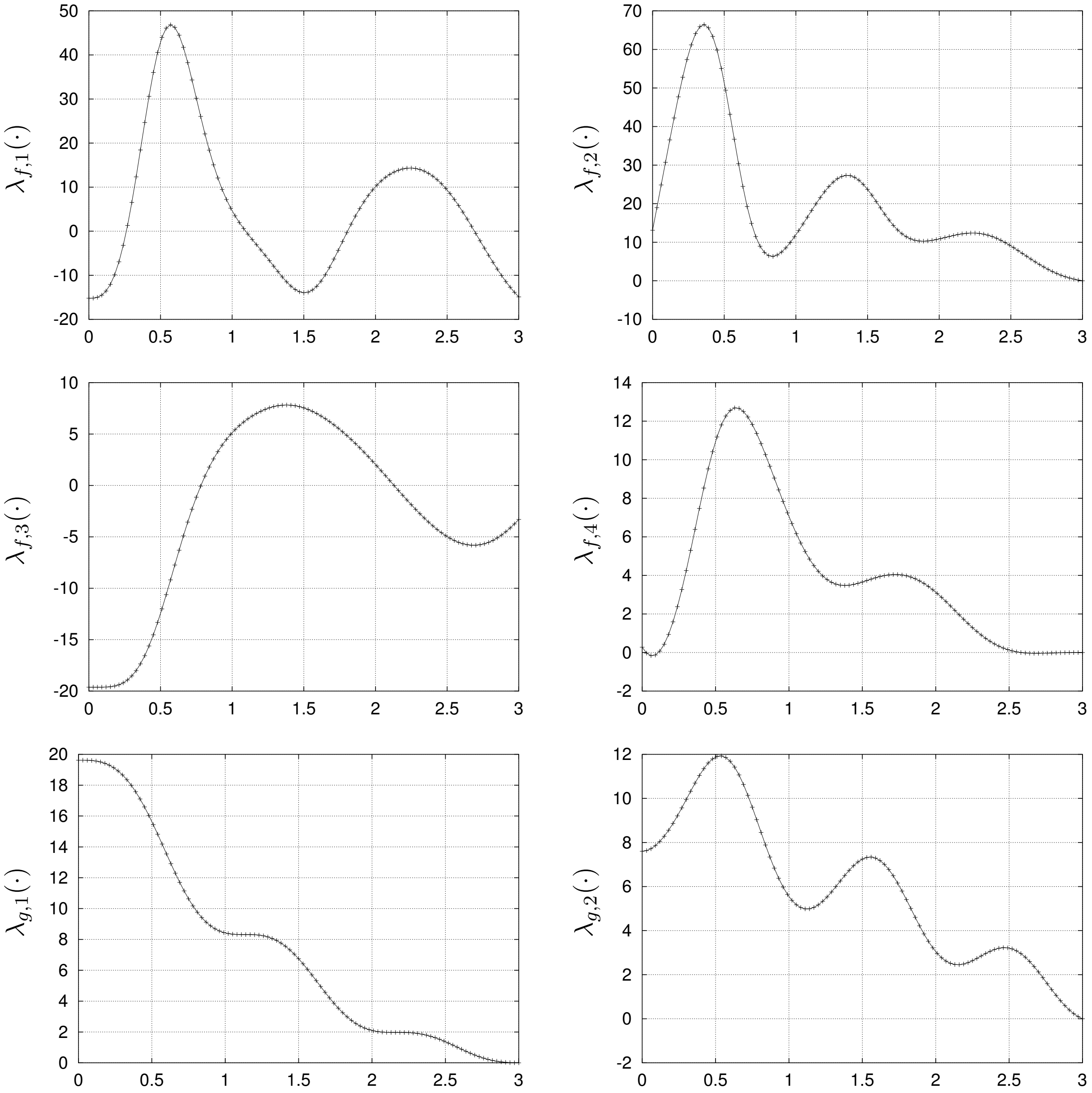


Figure 4.2: Numerical solution of BVP resulting from the minimum principle: Adjoint variables $\lambda_f(t)$ and $\lambda_g(t)$ for $t \in [0, 3]$.

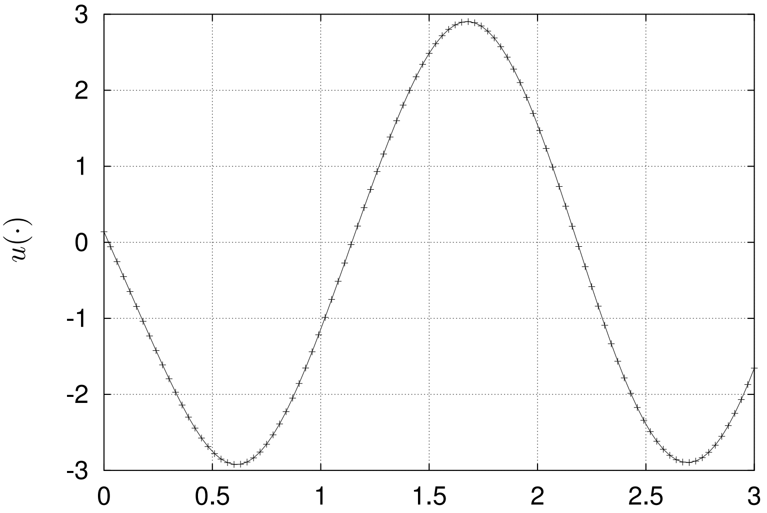


Figure 4.3: Numerical solution of BVP resulting from the minimum principle: Control $u(t)$ for $t \in [0, 3]$.

Finally, the output of Newton's method is depicted.

```
 ITER      L2 NORM OF RESIDUALS           FINAL L2 NORM OF THE RESIDUALS
 --------------------------------           0.3975362397848623E-10
   0        0.4328974468004262E+01        EXIT PARAMETER 1
   1        0.4240642504525387E+01        FINAL APPROXIMATE SOLUTION
   2        0.4193829901526831E+01          0.1520606630397864E+02
   3        0.4144234257066851E+01         -0.1310630040572912E+02
   4        0.4112295471233999E+01          0.1962000001208976E+02
   5        0.4044231539950495E+01         -0.2769958914561794E+00
   6        0.3923470866048804E+01
   7        0.3705978590921865E+01
   8        0.3690148344331338E+01
   9        0.3310555206607231E+01
  10        0.2881208390804098E+01
...
  22        0.4792724723868053E-01
  23        0.2221098477722281E-01
  24        0.6982985591502162E-04
  25        0.1271838284143635E-04
  26        0.5326278869236522E-06
  27        0.1478490489094870E-06
  28        0.1478490489094870E-06
  29        0.1478490489094870E-06
  30        0.4656214333457540E-08
  31        0.3079526214273727E-09
  32        0.1222130887100730E-09
  33        0.1222130887100730E-09
  34        0.4512679350279913E-10
  35        0.4512679350279913E-10
  36        0.3975362397848623E-10
  37        0.3975362397848623E-10
  38        0.3975362397848623E-10
```

Remark 4.1.16 *Usually, one would expect that the necessary conditions hold with the Hamilton function $H = l_0 f_0 + \lambda_f^\top f + \lambda_g^\top g$. The following example in Backes [Bac06] shows that this is not true:*
Minimize

$$\frac{1}{2}x_1^2(t_f) + \frac{1}{2}\int_0^{t_f} x_3(t)^2 + u(t)^2 dt$$

subject to

$$\begin{aligned} x_1'(t) &= u(t), \qquad x_1(0) = a, \\ x_2'(t) &= -x_3(t) + u(t), \qquad x_2(0) = 0, \\ 0 &= x_2(t), \end{aligned}$$

where $a \neq 0$. Solution:

$$x_1(t) = a\left(1 - \frac{t}{2+t_f}\right),\ x_2(t) = 0,\ x_3(t) = u(t) = -\frac{a}{2+t_f}.$$

4.2 Local Minimum Principles for Index-1 Problems

We summarize local minimum principles for the optimal control problem 1.1, i.e. for

Problem 4.2.1 (Index-1 DAE optimal control problem)
Find a state variable $x \in W^{1,\infty}([t_0,t_f],\mathbb{R}^{n_x})$, *an* algebraic variable $y \in L^\infty([t_0,t_f],\mathbb{R}^{n_y})$, *and a* control variable $u \in L^\infty([t_0,t_f],\mathbb{R}^{n_u})$ *such that the* objective function

$$F(x,y,u) := \varphi(x(t_0),x(t_f)) + \int_{t_0}^{t_f} f_0(t,x(t),y(t),u(t))dt \tag{4.2.1}$$

is minimized subject to the semi-explicit index-1 DAE

$$\begin{aligned} \dot{x}(t) &= f(t,x(t),y(t),u(t)) \qquad \text{a.e. in } [t_0,t_f], && (4.2.2) \\ 0_{n_y} &= g(t,x(t),y(t),u(t)) \qquad \text{a.e. in } [t_0,t_f], && (4.2.3) \end{aligned}$$

the boundary conditions

$$\psi(x(t_0),x(t_f)) = 0_{n_\psi}, \tag{4.2.4}$$

the mixed control-state constraints

$$c(t,x(t),y(t),u(t)) \leq 0_{n_c} \qquad \text{a.e. in } [t_0,t_f], \tag{4.2.5}$$

the pure state constraints

$$s(t,x(t)) \leq 0_{n_s} \qquad \text{in } [t_0,t_f], \tag{4.2.6}$$

and the set constraints

$$u(t) \in \mathcal{U} \qquad \text{a.e. in } [t_0,t_f]. \tag{4.2.7}$$

We assume:

Assumption 4.2.2 (Index-1 Assumption)
Let $g_y'[t]$ be non-singular a.e. in $[t_0,t_f]$ and let $g_y'[t]^{-1}$ be essentially bounded.

Assumption 4.2.2 implies that the index of the DAE (4.2.2)-(4.2.3) is one.

Remark 4.2.3 *Notice, that by Assumption 4.2.2 the algebraic equation (4.2.3) can be solved for $\hat{y}$ and $\hat{y}$ can be expressed as a function of $t, \hat{x}, \hat{u}$. By exploitation of the minimum principle known for ordinary differential equations this approach would lead to a second version of the proof of the minimum principle for DAE systems.*

Since the proof of the local minimum principles for index-1 DAE optimal control problems works in a similar way as in the index-2 case, we omit it here and we just summarize the results. The details of the proof can be found in a technical report of Gerdts [Ger05a].
The Hamilton function for Problem 4.2.1 reads as

$$\mathcal{H}(t,x,y,u,\lambda_f,\lambda_g,l_0) := l_0 f_0(t,x,y,u) + \lambda_f^\top f(t,x,y,u) + \lambda_g^\top g(t,x,y,u).$$

Theorem 4.2.4 (Local Minimum Principle for Optimal Control Problems without Mixed Control-State Constraints)

Let the following assumptions be fulfilled for the optimal control problem 4.2.1.

(i) Let the functions $\varphi, f_0, f, g, s, \psi$ be continuous w.r.t. all arguments and continuously differentiable w.r.t. x, y, and u.

(ii) Let $\mathcal{U} \subseteq \mathbb{R}^{n_u}$ be a closed and convex set with non-empty interior.

(iii) Let $(\hat{x}, \hat{y}, \hat{u}) \in X$ be a weak local minimum of Problem 4.2.1.

(iv) Let Assumption 4.2.2 be valid.

(v) Let there be no mixed control-state constraints (4.2.5) in the optimal control problem 4.2.1.

Then there exist multipliers $l_0 \in \mathbb{R}$, $\sigma \in \mathbb{R}^{n_\psi}$, $\lambda_f \in BV([t_0,t_f],\mathbb{R}^{n_x})$, $\lambda_g \in L^\infty([t_0,t_f],\mathbb{R}^{n_y})$, and $\mu \in NBV([t_0,t_f],\mathbb{R}^{n_s})$ such that the following conditions are satisfied:

(i) $l_0 \geq 0$, $(l_0,\sigma,\lambda_f,\lambda_g,\mu) \neq \Theta$,

(ii) Adjoint equations*:*

$$\begin{aligned} \lambda_f(t) &= \lambda_f(t_f) + \int_t^{t_f} \mathcal{H}'_x(\tau,\hat{x}(\tau),\hat{y}(\tau),\hat{u}(\tau),\lambda_f(\tau),\lambda_g(\tau),l_0)^\top d\tau \\ &\quad + \sum_{i=1}^{n_s} \int_t^{t_f} s'_{i,x}(\tau,\hat{x}(\tau))^\top d\mu_i(\tau) \quad \text{in } [t_0,t_f], \qquad (4.2.8) \\ 0_{n_y} &= \mathcal{H}'_y(t,\hat{x}(t),\hat{y}(t),\hat{u}(t),\lambda_f(t),\lambda_g(t),l_0)^\top \quad \text{a.e. in } [t_0,t_f]. \qquad (4.2.9) \end{aligned}$$

(iii) Transversality conditions*:*

$$\begin{aligned} \lambda_f(t_0)^\top &= -\left(l_0 \varphi'_{x_0}(\hat{x}(t_0),\hat{x}(t_f)) + \sigma^\top \psi'_{x_0}(\hat{x}(t_0),\hat{x}(t_f))\right), \qquad (4.2.10) \\ \lambda_f(t_f)^\top &= l_0 \varphi'_{x_f}(\hat{x}(t_0),\hat{x}(t_f)) + \sigma^\top \psi'_{x_f}(\hat{x}(t_0),\hat{x}(t_f)). \qquad (4.2.11) \end{aligned}$$

(iv) Optimality conditions*: Almost everywhere in $[t_0,t_f]$ for all $u \in \mathcal{U}$ it holds*

$$\mathcal{H}'_u(t,\hat{x}(t),\hat{y}(t),\hat{u}(t),\lambda_f(t),\lambda_g(t),l_0)(u - \hat{u}(t)) \geq 0. \qquad (4.2.12)$$

(v) Complementarity condition*:*
μ_i is monotonically increasing on $[t_0,t_f]$ and constant on every interval (t_1,t_2) with $t_1 < t_2$ and $s_i(t,\hat{x}(t)) < 0$ for all $t \in (t_1,t_2)$.

The augmented Hamilton function is given by

$$\hat{\mathcal{H}}(t,x,y,u,\lambda_f,\lambda_g,\eta,l_0) := \mathcal{H}(t,x,y,u,\lambda_f,\lambda_g,l_0) + \eta^\top c(t,x,y,u).$$

Theorem 4.2.5 (Local Minimum Principle for Optimal Control Problems without Set Constraints)

Let the following assumptions be fulfilled for the optimal control problem 4.2.1.

(i) Let the functions $\varphi, f_0, f, g, c, s, \psi$ be continuous w.r.t. all arguments and continuously differentiable w.r.t. x, y, and u.

(ii) Let $(\hat{x},\hat{y},\hat{u}) \in X$ be a weak local minimum of Problem 4.2.1.

(iii) Let

$$\mathit{rank}\left(c_u'(t,\hat{x}(t),\hat{y}(t),\hat{u}(t))\right) = n_c$$

almost everywhere in $[t_0,t_f]$.

(iv) Let the pseudo-inverse of $c_u'[t]$

$$(c_u'[t])^+ = c_u'[t]^\top \left(c_u'[t]c_u'[t]^\top\right)^{-1} \tag{4.2.13}$$

be essentially bounded and let the matrix

$$g_y'[t] - g_u'[t](c_u'[t])^+ c_y'[t] \tag{4.2.14}$$

be non-singular almost everywhere with essentially bounded inverse in $[t_0,t_f]$.

(v) Let Assumption 4.2.2 be valid.

(vi) Let $\mathcal{U} = \mathbb{R}^{n_u}$.

Then there exist multipliers $l_0 \in \mathbb{R}$, $\sigma \in \mathbb{R}^{n_\psi}$, $\lambda_f \in BV([t_0,t_f],\mathbb{R}^{n_x})$, $\lambda_g \in L^\infty([t_0,t_f],\mathbb{R}^{n_y})$, $\eta \in L^\infty([t_0,t_f],\mathbb{R}^{n_c})$, and $\mu \in NBV([t_0,t_f],\mathbb{R}^{n_s})$ such that the following conditions are satisfied:

(i) $l_0 \geq 0$, $(l_0,\sigma,\lambda_f,\lambda_g,\eta,\mu) \neq \Theta$,

(ii) Adjoint equations*:*

$$\begin{aligned} \lambda_f(t) &= \lambda_f(t_f) + \int_t^{t_f} \hat{\mathcal{H}}_x'(\tau,\hat{x}(\tau),\hat{y}(\tau),\hat{u}(\tau),\lambda_f(\tau),\lambda_g(\tau),\eta(\tau),l_0)^\top d\tau \\ &\quad + \sum_{i=1}^{n_s} \int_t^{t_f} s_{i,x}'(\tau,\hat{x}(\tau))^\top d\mu_i(\tau) \quad \textit{in } [t_0,t_f], & (4.2.15) \\ 0_{n_y} &= \hat{\mathcal{H}}_y'(t,\hat{x}(t),\hat{y}(t),\hat{u}(t),\lambda_f(t),\lambda_g(t),\eta(t),l_0)^\top \ \textit{a.e. in } [t_0,t_f]. & (4.2.16) \end{aligned}$$

(iii) Transversality conditions*:*

$$\begin{aligned} \lambda_f(t_0)^\top &= -\left(l_0\varphi_{x_0}'(\hat{x}(t_0),\hat{x}(t_f)) + \sigma^\top\psi_{x_0}'(\hat{x}(t_0),\hat{x}(t_f))\right), & (4.2.17) \\ \lambda_f(t_f)^\top &= \quad l_0\varphi_{x_f}'(\hat{x}(t_0),\hat{x}(t_f)) + \sigma^\top\psi_{x_f}'(\hat{x}(t_0),\hat{x}(t_f)). & (4.2.18) \end{aligned}$$

(iv) Optimality conditions*: It holds*

$$\hat{\mathcal{H}}'_u(t,\hat{x}(t),\hat{y}(t),\hat{u}(t),\lambda_f(t),\lambda_g(t),\eta(t),l_0) = 0_{n_u} \tag{4.2.19}$$

a.e. in $[t_0,t_f]$.

(v) Complementarity conditions*:*
It holds

$$\eta(t)^\top c(t,\hat{x}(t),\hat{y}(t),\hat{u}(t)) \;=\; 0, \qquad \eta(t) \geq 0_{n_c}$$

almost everywhere in $[t_0,t_f]$.

μ_i *is monotonically increasing on* $[t_0,t_f]$ *and constant on every interval* (t_1,t_2) *with* $t_1 < t_2$ *and* $s_i(t,\hat{x}(t)) < 0$ *for all* $t \in (t_1,t_2)$.

The following considerations apply to both, Theorem 4.2.4 and Theorem 4.2.5, and differ only in the Hamilton functions $\mathcal{H}$ and $\hat{\mathcal{H}}$, respectively. Hence, we restrict the discussion to the situation of Theorem 4.2.4.

Corollary 4.2.6 *Let the assumptions of Theorem 4.2.4 be fulfilled. Then,* λ_f *is differentiable almost everywhere in* $[t_0,t_f]$ *with*

$$\dot{\lambda}_f(t) = -\mathcal{H}'_x(t,\hat{x}(t),\hat{y}(t),\hat{u}(t),\lambda_f(t),\lambda_g(t),l_0)^\top - \sum_{i=1}^{n_s} s'_{i,x}(t,\hat{x}(t))^\top \dot{\mu}_i(t). \tag{4.2.20}$$

Furthermore, the jump conditions

$$\lambda_f(t_j) - \lambda_f(t_j-) = -\sum_{i=1}^{n_s} s'_{i,x}(t_j,\hat{x}(t_j))^\top (\mu_i(t_j) - \mu_i(t_j-)) \tag{4.2.21}$$

hold at every point $t_j \in (t_0,t_f)$ *of discontinuity of the multiplier* μ.

Notice, that μ_i in (4.2.20) can be replaced by the absolutely continuous component $\mu_{i,a}$ since the derivatives of the jump component $\mu_{i,d}$ and the singular component $\mu_{i,s}$ are zero almost everywhere. However, $\lambda_f(t)$ cannot be reconstructed by simple integration of $\dot{\lambda}_f$.
A special case arises, if no state constraints are present. Then, the adjoint variable λ_f is even absolutely continuous, i.e. $\lambda_f \in W^{1,\infty}([t_0,t_f],\mathbb{R}^{n_x})$, and the adjoint equations (4.2.8)-(4.2.9) become

$$\begin{aligned}
\dot{\lambda}_f(t) &= -\mathcal{H}'_x(t,\hat{x}(t),\hat{y}(t),\hat{u}(t),\lambda_f(t),\lambda_g(t),l_0)^\top \quad \text{a.e. in } [t_0,t_f], && (4.2.22)\\
0_{n_y} &= \mathcal{H}'_y(t,\hat{x}(t),\hat{y}(t),\hat{u}(t),\lambda_f(t),\lambda_g(t),l_0)^\top \quad \text{a.e. in } [t_0,t_f]. && (4.2.23)
\end{aligned}$$

The adjoint equations (4.2.22) and (4.2.23) form a DAE system of index one for λ_f and λ_g, where λ_f is the differential variable and λ_g denotes the algebraic variable. This follows from (4.2.23), which is given by

$$0_{n_y} = l_0 \left(f'_{0,y}[t]\right)^\top + \left(f'_y[t]\right)^\top \lambda_f(t) + \left(g'_y[t]\right)^\top \lambda_g(t).$$

Since $g'_y[t]$ is non-singular, we obtain

$$\lambda_g(t) = -\left(g'_y[t]^{-1}\right)^\top \left(l_0 \left(f'_{0,y}[t]\right)^\top + \left(f'_y[t]\right)^\top \lambda_f(t)\right).$$

Finally, we address the problem of regularity and state conditions which ensure that the multiplier l_0 is not zero and, without loss of generality, can be normalized to one.

Theorem 4.2.7 *Let the assumptions of Theorems 4.2.4 or 4.2.5 be fulfilled and let*

$$\mathit{rank}\left(\psi'_{x_0}\Phi(t_0)+\psi'_{x_f}\Phi(t_f)\right)=n_\psi,$$

where Φ is the fundamental solution of the homogeneous linear differential equation

$$\dot{\Phi}(t)=A(t)\Phi(t),\quad \Phi(t_0)=I_{n_x},\qquad t\in[t_0,t_f]$$

with

$$A(t):=f'_x[t]-f'_y[t]\left(g'_y[t]\right)^{-1}g'_x[t].$$

Furthermore, let there exist $\varepsilon>0$, $x\in W^{1,\infty}([t_0,t_f],\mathbb{R}^{n_x})$, $y\in L^\infty([t_0,t_f],\mathbb{R}^{n_y})$, and $u\in \mathit{int}(U_{ad}-\{\hat{u}\})$ satisfying

$$\begin{aligned}
c[t]+c'_x[t]x(t)+c'_y[t]y(t)+c'_u[t]u(t) &\le -\varepsilon\cdot e \quad \text{a.e. in } [t_0,t_f], && (4.2.24)\\
s[t]+s'_x[t]x(t) &< 0_{n_s} \quad \text{in } [t_0,t_f], && (4.2.25)\\
f'_x[t]x(t)+f'_y[t]y(t)+f'_u[t]u(t)-\dot{x}(t) &= 0_{n_x} \quad \text{a.e. in } [t_0,t_f], && (4.2.26)\\
g'_x[t]x(t)+g'_y[t]y(t)+g'_u[t]u(t) &= 0_{n_y} \quad \text{a.e. in } [t_0,t_f], && (4.2.27)\\
\psi'_{x_0}x(t_0)+\psi'_{x_f}x(t_f) &= 0_{n_\psi}, && (4.2.28)
\end{aligned}$$

where $e=(1,\ldots,1)^\top\in\mathbb{R}^{n_c}$. Then it holds $l_0=1$ in Theorems 4.2.4 or 4.2.5, respectively.

Chapter 5

Discretization Methods for ODEs and DAEs

Many discretization methods originally designed for ODEs, e.g. Runge-Kutta methods, Multistep methods, or extrapolation methods, can be adapted in a fairly straight forward way to solve DAEs as well. Naturally, the resulting methods are implicit methods due to the inherently implicit character of DAEs and it is necessary to solve nonlinear systems of equations in each integration step. The nonlinear systems usually are solved by Newton's method or well-known variants thereof, e.g. globalized or simplified Newton's method or quasi-Newton methods. To reduce the computational effort of Newton's method for nonlinear equations, methods based on a suitable linearization are considered, e.g. linearized implicit Runge-Kutta methods.

For notational convenience, we will write down the methods for general implicit DAE initial value problems of type

$$F(t, x(t), \dot{x}(t)) = 0_n, \qquad x(t_0) = x_0, \tag{5.1}$$

where $F : [t_0, t_f] \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \to \mathbb{R}^{n_x}$, $t_0 \leq t \leq t_f$, and x_0 is consistent. It is important to point out, that this does not imply automatically that the resulting discretizations actually converge to the exact solution. In contrast to the ODE case, where a quite general convergence theory exists, for DAEs the question of convergence highly depends on the structure of the system. In order to obtain convergence results it is necessary to assume a certain index of (5.1) or a special structure in (5.1), e.g. Hessenberg structure.

Generally, the discretization methods can be classified as one-step and multi-step methods. In the sequel, we will discuss Runge-Kutta methods (one-step) and BDF methods (multi-step), both being well investigated in the literature. In addition to these universal methods, there are also methods which are designed for special DAEs, e.g. for the simulation of mechanical multibody systems, cf. Simeon [Sim94], or the simulation of electric circuits, cf. Günther [Gün95].

5.1 General Discretization Theory

Let Banach spaces $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ and a mapping $G : X \to Y$ be given. The task is to find $\hat{x} \in X$ with

$$G(\hat{x}) = \Theta_Y.$$

This problem is referred to by the triple $\mathcal{P} = (X, Y, G)$. It is assumed that a unique solution $\hat{x}$ exists, that is, problem $\mathcal{P}$ is uniquely solvable.

Definition 5.1.1 (Discretization Method, Discretization)

(i) A discretization method $\mathcal{M}$ *for* $\mathcal{P}$ *is a sequence* $(X_N, Y_N, \Delta_N, \delta_N, \varphi_N)_{N\in\mathbb{N}}$*, where*

- $(X_N, \|\cdot\|_{X_N})$ *and* $(Y_N, \|\cdot\|_{Y_N})$ *are finite dimensional Banach spaces with* $dim(X_N) = dim(Y_N)$*;*
- $\Delta_N : X \to X_N$ *and* $\delta_N : Y \to Y_N$ *are linear mappings with*

$$\begin{aligned} \lim_{N\to\infty} \|\Delta_N(x)\|_{X_N} &= \|x\|_X, && (5.1.1)\\ \lim_{N\to\infty} \|\delta_N(y)\|_{Y_N} &= \|y\|_Y && (5.1.2)\end{aligned}$$

for each fixed $x \in X$ *and* $y \in Y$*;*

- $\varphi_N : (X \to Y) \to (X_N \to Y_N)$ *is a mapping with* G *in the domain of* φ_N *for all* $N \in \mathbb{N}$.

(ii) A discretization $\mathcal{D}$ *of the problem* $\mathcal{P}$ *generated by the discretization method* $\mathcal{M}$ *is a sequence* $(X_N, Y_N, G_N)_{N\in\mathbb{N}}$ *with* $G_N = \varphi_N(G) : X_N \to Y_N$.

A solution of the discretization $\mathcal{D}$ *is a sequence* $(x_N)_{N\in\mathbb{N}}$ *with*

$$G_N(x_N) = \Theta_{Y_N}, \qquad N \in \mathbb{N}.$$

Definition 5.1.2 (Consistency)
A discretization method $\mathcal{M} = (X_N, Y_N, \Delta_N, \delta_N, \varphi_N)_{N\in\mathbb{N}}$ *of the problem* $\mathcal{P}$ *is called* consistent at $x \in X$*, if* x *is in the domain of* G *and of* $\varphi_N(G)(\Delta_N(\cdot))$ *for* $N \in \mathbb{N}$ *and*

$$\lim_{N\to\infty} \|\varphi_N(G)(\Delta_N(x)) - \delta_N(G(x))\|_{Y_N} = 0.$$

It is called consistent of order p at x*, if*

$$\|\varphi_N(G)(\Delta_N(x)) - \delta_N(G(x))\|_{Y_N} = \mathcal{O}\left(\frac{1}{N^p}\right) \qquad \text{as } N \to \infty.$$

If $\mathcal{M}$ *is consistent (of order* p*) at* x*, the corresponding discretization* $\mathcal{D}$ *is called consistent (of order* p*) at* x.

Later on we will need consistency only in the exact solution $\hat{x}$ of the original problem $\mathcal{P}$. Notice, that the exact solution satisfies $\delta_N(G(\hat{x})) = \Theta_{Y_N}$. Thus, it suffices to investigate the local discretization error defined below.

Definition 5.1.3 (Local and Global Discretization Error, Convergence)
Let $\hat{x}$ *denote the solution of* $\mathcal{P}$.

(i) Let the discretization method $\mathcal{M}$ *be consistent. The* local discretization error of $\mathcal{M}$ and $\mathcal{D}$ *is defined as the sequence* $\{l_N\}_{N\in\mathbb{N}}$ *where*

$$l_N = \varphi_N(G)(\Delta_N(\hat{x})) \in Y_N, \qquad N \in \mathbb{N}.$$

(ii) Let the discretization $\mathcal{D}$ *of the problem* $\mathcal{P}$ *generated by the discretization method* $\mathcal{M}$ *possess unique solutions* $\hat{x}_N$*,* $N \in \mathbb{N}$*. The sequence* $\{e_N\}_{N\in\mathbb{N}}$ *with*

$$e_N = \hat{x}_N - \Delta_N(\hat{x}) \in X_N, \qquad N \in \mathbb{N}$$

is called global discretization error of $\mathcal{M}$ and $\mathcal{D}$.

(iii) $\mathcal{M}$ *and* $\mathcal{D}$ *are called* convergent*, if*

$$\lim_{N\to\infty} \|e_N\|_{X_N} = 0,$$

and convergent of order p*, if*

$$\|e_N\|_{X_N} = \mathcal{O}\left(\frac{1}{N^p}\right) \qquad \text{as } N \to \infty.$$

In view of a proof of convergence the discretization method has to be not only consistent but also stable.

Definition 5.1.4 (Stability)
The discretization $\mathcal{D}$ is called stable at $(x_N)_{N\in\mathbb{N}} \subset X_N$*, if there exist constants S and $r > 0$ independent of N such that uniformly for almost all $N \in \mathbb{N}$ it holds the following: Whenever $x_N^{(i)} \in X_N$, $i = 1, 2$, satisfy*

$$\|G_N(x_N^{(i)}) - G_N(x_N)\|_{Y_N} < r, \qquad i = 1,2$$

then it holds

$$\|x_N^{(1)} - x_N^{(2)}\|_{X_N} \leq S\|G_N(x_N^{(1)}) - G_N(x_N^{(2)})\|_{Y_N}.$$

S is called stability bound*, r is called* stability threshold*.*
The discretization method $\mathcal{M}$ is called stable, if the generated discretization $\mathcal{D}$ is stable at $(\Delta_N(\hat{x}))_{N\in\mathbb{N}}$, where $\hat{x}$ is the solution of the original problem $\mathcal{P}$.

So far, we only assumed the unique solvability of the original problem $\mathcal{P}$. It is not clear, that the discrete problems $G_N(x_N) = \Theta_{Y_N}$ are (uniquely) solvable. It will turn out that consistency, stability and a continuity argument will guarantee unique solvability. The proofs of the following statements can be found in Stetter [Ste73], pp. 11-15.

Lemma 5.1.5 (Stetter [Ste73], Lem. 1.2.1, p. 11)
Let $G_N : X_N \to Y_N$ be defined and continuous in

$$B_R(\hat{x}_N) = \{x_N \in X_N \mid \|x_N - \hat{x}_N\|_{X_N} < R\}$$

for $\hat{x}_N \in X_N$. Furthermore, for all $x_N^{(i)} \in B_R(\hat{x}_N)$, $i = 1, 2$, with

$$G_N(x_N^{(i)}) \in U_r(G_N(\hat{x}_N)) = \{y_N \in Y_N \mid \|y_N - G_N(\hat{x}_N)\|_{Y_N} < r\}$$

let

$$\|x_N^{(1)} - x_N^{(2)}\|_{X_N} \leq S\|G_N(x_N^{(1)}) - G_N(x_N^{(2)})\|_{Y_N}$$

for some constants $r, S > 0$. Then $G_N^{-1} : Y_N \to X_N$ exists and is Lipschitz continuous in $U_{r_0}(G_N(\hat{x}_N))$ with Lipschitz constant S, where $r_0 = \min\{r, R/S\}$.

Exploitation of this lemma allows to proof unique solvability. It holds

Theorem 5.1.6 (Stetter [Ste73], Th. 1.2.3, p.12)
Let $\hat{x}$ denote the unique solution of $\mathcal{P}$. Let the discretization method $\mathcal{M} = (X_N, Y_N, \Delta_N, \delta_N, \varphi_N)$ satisfy the following conditions.

(i) $G_N = \varphi_N(G)$ is defined and continuous in

$$B_R(\Delta_N(\hat{x})) = \{x_N \in X_N \mid \|x_N - \Delta_N(\hat{x})\|_{X_N} < R\}$$

with $R > 0$ independent of N.

(ii) $\mathcal{M}$ is consistent at $\hat{x}$.

(iii) $\mathcal{M}$ is stable.

Then the discretization $\mathcal{D} = (X_N, Y_N, G_N)$ possesses unique solutions $\hat{x}_N \in X_N$ for all sufficiently large $N \in \mathbb{N}$.

With Theorem 5.1.6 we can proof convergence.

Theorem 5.1.7 (Stetter [Ste73], Th. 1.2.4, p.13)
Let the assumptions of Theorem 5.1.6 be valid. Then, $\mathcal{M}$ is convergent. Furthermore, if $\mathcal{M}$ is consistent of order p, then it is convergent of order p.

Proof. Theorem 5.1.6 guarantees the existence of unique solutions $\hat{x}_N$ of $G_N(\hat{x}_N) = \Theta_{Y_N}$ for sufficiently large N. It holds

$$-l_N = -G_N(\Delta_N(\hat{x})) = \underbrace{G_N(\hat{x}_N)}_{=\Theta_{Y_N}} - G_N(\Delta_N(\hat{x})) = G_N(\Delta_N(\hat{x}) + e_N) - G_N(\Delta_N(\hat{x})).$$

Stability at $\Delta_N(\hat{x})$ implies

$$\|e_N\|_{X_N} = \|\hat{x}_N - \Delta_N(\hat{x})\|_{X_N} \leq S\|G_N(\hat{x}_N) - G_N(\Delta_N(\hat{x}))\|_{Y_N} = S\|l_N\|_{Y_N}$$

for $\|l_N\|_{Y_N} = \|G_N(\hat{x}_N) - G_N(\Delta_N(\hat{x}))\|_{Y_N} < r$. Since $\|l_N\|_{Y_N} \to 0$ respectively $\|l_N\|_{Y_N} = \mathcal{O}(N^{-p})$ holds, it follows $\|e_N\|_{X_N} \to 0$ respectively $\|e_N\|_{X_N} = \mathcal{O}(N^{-p})$. ■

Now, we reformulate the initial value problem (5.1) as an abstract root finding problem. The Banach spaces $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ given by

$$\begin{aligned} X &= C^1([t_0, t_f], \mathbb{R}^{n_x}), \\ \|x(\cdot)\|_X &= \max_{t_0 \leq t \leq t_f} \|x(t)\|, \\ Y &= \mathbb{R}^{n_x} \times C([t_0, t_f], \mathbb{R}^{n_x}), \\ \|(r, s(\cdot))\|_Y &= \|r\| + \max_{t_0 \leq t \leq t_f} \|s(t)\|. \end{aligned}$$

The mapping $G : X \to Y$ is defined by

$$G(x) = \begin{pmatrix} x(t_0) - x_0 \\ F(\cdot, x(\cdot), \dot{x}(\cdot)) \end{pmatrix}.$$

Then, (5.1) is equivalent with the problem

$$G(x) = \Theta_Y.$$

In order to discretize the abstract problem, we introduce a grid

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

depending on the discretization parameter $N \in \mathbb{N}$ with distinct grid points

$$t_0 < t_1 < \ldots < t_N =: t_f$$

and step lengths $h_j = t_{j+1} - t_j$, $j = 0, 1, \ldots, N-1$. It is assumed, that the step sizes linearly tend to zero as N approaches infinity, i.e. there exist constants $c_1, c_2 > 0$ such that

$$\frac{c_1}{N} \leq h_j \leq \frac{c_2}{N}, \qquad j = 0, 1, \ldots, N-1$$

for all $N \in \mathbb{N}$. Often, we will simply consider constant step lengths $h = (t_f - t_0)/N$. Finite dimensional Banach spaces $(X_N, \|\cdot\|_{X_N})$ and $(Y_N, \|\cdot\|_{Y_N})$ are given by

$$\begin{aligned} X_N &= \{x_N : \mathbb{G}_N \to \mathbb{R}^{n_x}\}, \\ \|x_N(\cdot)\|_{X_N} &= \max_{t \in \mathbb{G}_N} \|x_N(t)\|, \\ Y_N &= X_N, \\ \|\cdot\|_{Y_N} &= \|\cdot\|_{X_N}. \end{aligned}$$

The restriction operators $\Delta_N : X \to X_N$ and $\delta_N : Y \to Y_N$ are defined by

$$\begin{aligned} \Delta_N(x)(t_j) &:= x(t_j), \\ \delta_N(r, s(\cdot))(t_j) &:= \begin{cases} r, & \text{if } j = 0, \\ s(t_{j-1}), & \text{if } 1 \leq j \leq N. \end{cases} \end{aligned}$$

In the sequel, we focus on *one-step methods with increment function* Φ defined by

$$\begin{aligned} x_N(t_0) &= x_0, \\ \frac{x_N(t_j) - x_N(t_{j-1})}{h_{j-1}} &= \Phi(t_{j-1}, x_N(t_{j-1}), h_{j-1}), \qquad j = 1, 2, \ldots, N. \end{aligned} \tag{5.1.3}$$

For the one-step method (5.1.3) the mappings $\varphi_N : (X \to Y) \to (X_N \to Y_N)$ and $G_N = \varphi_N(G) : X_N \to Y_N$ are given by

$$\varphi_N(G)(x_N)(t_j) = \begin{cases} x_N(t_0) - x_0, & \text{if } j = 0, \\ \dfrac{x_N(t_j) - x_N(t_{j-1})}{h_{j-1}} - \Phi(t_{j-1}, x_N(t_{j-1}), h_{j-1}), & \text{if } j = 1, \ldots, N. \end{cases}$$

The following stability result holds for one-step methods.

Theorem 5.1.8 *Let Φ be a locally Lipschitz continuous function w.r.t. x uniformly with respect to t and h in the exact solution $\hat{x}$ for sufficiently small step sizes h. Then, the one-step method $\mathcal{M} = (X_N, Y_N, \Delta_N, \delta_N, \varphi_N)_{N \in \mathbb{N}}$ is stable.*

Proof. Since Φ is assumed to be locally Lipschitz continuous at $\hat{x}$, there exist constants $R > 0$, $h_R > 0$, and $L_R > 0$ such that

$$\|\Phi(t, x^{(1)}, h) - \Phi(t, x^{(2)}, h)\| \leq L_R \|x^{(1)} - x^{(2)}\| \tag{5.1.4}$$

holds for all $0 < h \leq h_R$ and all

$$(t, x^{(i)}) \in K, \quad i = 1, 2,$$

where

$$K := \{(t, x) \in \mathbb{R} \times \mathbb{R}^{n_x} \mid t_0 \leq t \leq t_f, \ \|x - \hat{x}(t)\| \leq R\}.$$

Let $\mathbb{G}_N$ be a grid with

$$\max_{j=0,\ldots,N-1} h_j \leq h_R. \tag{5.1.5}$$

Furthermore, let grid functions $x_N^{(i)} \in X_N$, $i = 1, 2$, be given with

$$\|x_N^{(i)} - \Delta_N(\hat{x})\|_{X_N} \leq R \qquad i = 1, 2. \tag{5.1.6}$$

Notice, that $(t_j, x_N^{(i)}(t_j)) \in K$ for all $t_j \in \mathbb{G}_N$, $i = 1, 2$. Define

$$\varepsilon_N^{(i)} := G_N(x_N^{(i)}), \qquad i = 1, 2.$$

Then, for $j = 0$ we have

$$x_N^{(1)}(t_0) - x_N^{(2)}(t_0) = x_0 + \varepsilon_N^{(1)}(t_0) - (x_0 + \varepsilon_N^{(2)}(t_0)) = \varepsilon_N^{(1)}(t_0) - \varepsilon_N^{(2)}(t_0).$$

For $j = 1, \ldots, N$ it holds

$$\begin{aligned}\|x_N^{(1)}(t_j) - x_N^{(2)}(t_j)\| &= \| \quad x_N^{(1)}(t_{j-1}) + h_{j-1}\Phi(t_{j-1}, x_N^{(1)}(t_{j-1}), h_{j-1}) + h_{j-1}\varepsilon_N^{(1)}(t_j) \\ &\qquad - x_N^{(2)}(t_{j-1}) - h_{j-1}\Phi(t_{j-1}, x_N^{(2)}(t_{j-1}), h_{j-1}) - h_{j-1}\varepsilon_N^{(2)}(t_j)\| \\ &\leq (1 + h_{j-1}L_R)\|x_N^{(1)}(t_{j-1}) - x_N^{(2)}(t_{j-1})\| + h_{j-1}\|\varepsilon_N^{(1)}(t_j) - \varepsilon_N^{(2)}(t_j)\|.\end{aligned}$$

Recursive evaluation leads to

$$\begin{aligned}\|x_N^{(1)}(t_j) - x_N^{(2)}(t_j)\| &\leq \left(\prod_{k=0}^{j-1}(1 + h_k L_R)\right)\|x_N^{(1)}(t_0) - x_N^{(2)}(t_0)\| \\ &\quad + \sum_{k=1}^{j} h_{k-1}\left(\prod_{l=k}^{j-1}(1 + h_l L_R)\right)\|\varepsilon_N^{(1)}(t_k) - \varepsilon_N^{(2)}(t_k)\| \\ &\leq \exp((t_j - t_0)L_R)\|\varepsilon_N^{(1)}(t_0) - \varepsilon_N^{(2)}(t_0)\| \\ &\quad + \max_{k=1,\ldots,j}\|\varepsilon_N^{(1)}(t_k) - \varepsilon_N^{(2)}(t_k)\|\exp((t_j - t_0)L_R)(t_j - t_0) \\ &\leq C\exp((t_j - t_0)L_R)\max_{k=0,\ldots,j}\|\varepsilon_N^{(1)}(t_k) - \varepsilon_N^{(2)}(t_k)\|, \qquad (5.1.7)\end{aligned}$$

where $C = \max\{t_f - t_0, 1\}$. Herein, we exploited

$$(1 + h_k L_R) \leq \exp(h_k L_R), \quad \sum_{k=0}^{j-1} h_k = t_j - t_0.$$

Finally, we found the estimate

$$\|x_N^{(1)} - x_N^{(2)}\|_{X_N} \leq S\|\varepsilon_N^{(1)} - \varepsilon_N^{(2)}\|_{Y_N} = S\|G_N(x_N^{(1)}) - G_N(x_N^{(2)})\|_{Y_N} \qquad (5.1.8)$$

with $S = C\exp((t_f - t_0)L_R)$, which holds for all grids with (5.1.5) and all $x_N^{(i)} \in X_N$, $i = 1, 2$, with (5.1.6).

To complete the proof it remains to show that (5.1.8) holds, whenever $x_N^{(i)} \in X_N$, $i = 1, 2$, satisfy

$$\|G_N(x_N^{(i)}) - G_N(\Delta_N(\hat{x}))\|_{Y_N} \leq r, \qquad i = 1, 2, \qquad (5.1.9)$$

for some $r > 0$. So, let (5.1.9) be valid with $r \leq \frac{R}{S}$. Then, the estimate (5.1.7) applied to $x_N^{(1)} = \Delta_N(\hat{x})$ and $x_N^{(2)} = \Delta_N(\hat{x})$, respectively, yields $\|x_N^{(i)}(t_j) - \hat{x}(t_j)\| \leq R$ for all $j = 0, 1, \ldots, N$ and $i = 1, 2$. Hence, (5.1.6) and thus (5.1.9) hold. This shows the stability of the one-step method with stability bound S and stability threshold R/S. ■

5.2 Backward Differentiation Formulae (BDF)

The Backward Differentiation Formulae (BDF) were introduced by Curtiss and Hirschfelder [CH52] and Gear [Gea71] and belong to the class of implicit linear multi-step methods.

We consider the integration step from t_{m+k-1} to t_{m+k}. Given approximations $x_N(t_m)$, $x_N(t_{m+1})$, …, $x_N(t_{m+k-1})$ for $x(t_m), x(t_{m+1}), \ldots, x(t_{m+k-1})$ and the approximation $x_N(t_{m+k})$ for $x(t_{m+k})$, which has to be determined, we compute the interpolating polynomial $Q(t)$ of degree k with $Q(t_{m+j}) = x_N(t_{m+j})$, $j = 0, \ldots, k$. The unknown value $x_N(t_{m+k})$ is determined by the postulation that the interpolating polynomial Q satisfies the DAE (5.1) at the time point t_{m+k}, i.e. $x_N(t_{m+k})$ satisfies

$$F(t_{m+k}, Q(t_{m+k}), \dot{Q}(t_{m+k})) = 0_{n_x}. \qquad (5.2.1)$$

The derivative $\dot{Q}(t_{m+k})$ can be expressed as

$$\dot{Q}(t_{m+k}) = \frac{1}{h_{m+k-1}} \sum_{j=0}^{k} \alpha_j x_N(t_{m+j}), \tag{5.2.2}$$

where the coefficients α_j depend on the step sizes $h_{m+j-1} = t_{m+j} - t_{m+j-1}$, $j = 1, \ldots, k$. Equation (5.2.1) together with (5.2.2) yields

$$F\left(t_{m+k}, x_N(t_{m+k}), \frac{1}{h_{m+k-1}} \sum_{j=0}^{k} \alpha_j x_N(t_{m+j})\right) = 0_{n_x}, \tag{5.2.3}$$

which is a nonlinear equation for $x = x_N(t_{m+k})$. Numerically, equation (5.2.3) is solved by the (globalized) Newton method given by

$$\begin{aligned} \left(F'_x + \frac{\alpha_k}{h_{m+k-1}} F'_{\dot{x}}\right) \Delta^{[l]} &= -F, \\ x^{[l+1]} &= x^{[l]} + \lambda_l \Delta^{[l]}, \quad l = 0, 1, \ldots, \end{aligned} \tag{5.2.4}$$

where the functions are evaluated at $(t_{m+k}, x^{[l]}, (\alpha_k x^{[l]} + \sum_{j=0}^{k-1} \alpha_j x_N(t_{m+j}))/h_{m+k-1})$.
BDF methods are appealing since they only require to solve a n_x-dimensional nonlinear system instead of a $n_x \cdot s$ dimensional system that arises for Runge-Kutta methods.
Brenan et al. [BCP96] develop in their code DASSL an efficient algorithm to set up the interpolating polynomial by use of a modified version of Newton's divided differences. In addition, a step size and order control based on an estimation of the local discretization error is implemented. A further improvement of the numerical performance is achieved by using some sort of simplified Newton's method, where the iteration matrix $F'_x + \alpha_k/h_{m+k-1} \cdot F'_{\dot{x}}$ is held constant as long as possible, even for several integration steps. For higher index systems the estimation procedure for the local discretization error has to be modified, cf. Petzold [Pet82b]. In practice, the higher index algebraic components are assigned very large error tolerances or are simply neglected in the error estimator. Führer [Füh88] and Führer and Leimkuhler [FL91] apply BDF to overdetermined DAEs. The implementation is called ODASSL and is an extension of DASSL. Gear and Petzold [GP84] and Petzold [Pet89] prove the convergence for BDF methods up to order 6, when applied to index-1 systems (5.1) with constant step size. Lötstedt and Petzold [LP86] show convergence up to order 6 for constant step size and certain DAEs arising in electric circuit simulation, fluid mechanics, and mechanics. Brenan and Engquist [BE88] prove convergence for DAEs of Hessenberg type (1.14) with index 2 and 3 at constant step size. Gear et al. [GLG85] extend the results in Gear and Petzold [GP84] to BDF methods with non-constant step size and non-constant order for index-2 DAEs (1.14) and index-1 DAEs (1.12)-(1.13). They also show, that BDF with non-constant step sizes applied to index-3 DAEs of Hessenberg type may yield wrong results in certain components of x. All these results assume consistent initial values.

5.3 Implicit Runge-Kutta Methods

An implicit Runge-Kutta (IRK) method with s stages is defined by the Butcher array

$$\begin{array}{c|ccc} c_1 & a_{11} & \cdots & a_{1s} \\ \vdots & \vdots & \ddots & \vdots \\ c_s & a_{s1} & \cdots & a_{ss} \\ \hline & b_1 & \cdots & b_s \end{array} \tag{5.3.1}$$

with real valued entries. An integration step of the IRK method from t_m to t_{m+1} with step size $h_m = t_{m+1} - t_m$ starting at an approximation $x_N(t_m)$ of $x(t_m)$ is defined by

$$x_N(t_{m+1}) = x_N(t_m) + h_m \sum_{j=1}^{s} b_j \, k_j(t_m, x_N(t_m), h_m). \tag{5.3.2}$$

It is a one-step method with increment function

$$\Phi(t_m, x_N(t_m), h_m) := \sum_{j=1}^{s} b_j \, k_j(t_m, x_N(t_m), h_m). \tag{5.3.3}$$

Herein, the stage derivatives $k_j = k_j(t_m, x_N(t_m), h_m)$, $j = 1, \dots, s$, are implicitly defined by the nonlinear equation

$$G(k, t_m, x_N(t_m), h_m) := \begin{pmatrix} F\left(t_m + c_1 h_m, x_{m+1}^{(1)}, k_1\right) \\ \vdots \\ F\left(t_m + c_s h_m, x_{m+1}^{(s)}, k_s\right) \end{pmatrix} = 0_{s \cdot n_x} \tag{5.3.4}$$

for $k = (k_1, \dots, k_s)^\top \in \mathbb{R}^{s \cdot n_x}$. The quantities

$$x_{m+1}^{(i)} = x_N(t_m) + h_m \sum_{j=1}^{s} a_{ij} \, k_j(t_m, x_N(t_m), h_m), \quad i = 1, \dots, s, \tag{5.3.5}$$

can be interpreted as approximations at the intermediate time points (stages) $t_m + c_i h_m$. Numerically, (5.3.4) is solved by the (globalized) Newton method defined by

$$\begin{aligned} G'_k(k^{[l]}, t_m, x_N(t_m), h_m) \, \Delta^{[l]} &= -G(k^{[l]}, t_m, x_N(t_m), h_m), \\ k^{[l+1]} &= k^{[l]} + \lambda_l \Delta^{[l]}, \quad l = 0, 1, \dots. \end{aligned} \tag{5.3.6}$$

The derivative G'_k evaluated at $(k^{[l]}, t_m, x_N(t_m), h_m)$ is given by

$$G'_k = \begin{pmatrix} F'_{\dot{x}}(\xi_1^{[l]}) & & \\ & \ddots & \\ & & F'_{\dot{x}}(\xi_s^{[l]}) \end{pmatrix} + h_m \begin{pmatrix} a_{11} F'_x(\xi_1^{[l]}) & \cdots & a_{1s} F'_x(\xi_1^{[l]}) \\ \vdots & \ddots & \vdots \\ a_{s1} F'_x(\xi_s^{[l]}) & \cdots & a_{ss} F'_x(\xi_s^{[l]}) \end{pmatrix}, \tag{5.3.7}$$

where

$$\xi_i^{[l]} = \left(t_m + c_i h_m, x_N(t_m) + h_m \sum_{j=1}^{s} a_{ij} k_j^{[l]}, k_i^{[l]} \right), \quad i = 1, \dots, s.$$

A Runge-Kutta method is called stiffly accurate, if it satisfies the conditions

$$c_s = 1 \quad \text{and} \quad a_{sj} = b_j, \quad j = 1, \dots, s,$$

cf. Strehmel and Weiner [SW95] and Hairer et al. [HLR89]. For instance, the RADAUIIA method is stiffly accurate. These methods are particularly well-suited for DAEs since the stage $x_{m+1}^{(s)}$ and the new approximation $x_N(t_{m+1})$ coincide. Since $x_{m+1}^{(s)}$ satisfies the DAE at $t_m + h$ so does $x_N(t_{m+1})$.

Given a Runge-Kutta method that is not stiffly accurate, the approximation (5.3.2) usually will not satisfy the DAE (5.1), although the stages $x_{m+1}^{(i)}$, $i = 1, \ldots, s$ satisfy it. To circumvent this problem, Ascher and Petzold [AP91] discuss projected Runge-Kutta methods for semi-explicit DAEs. Herein, $x_N(t_{m+1})$ is projected onto the algebraic equations and the projected value is taken as approximation at t_{m+1}.
The dimension of the equation (5.3.6) can be reduced for half-explicit Runge-Kutta methods, cf. Arnold [Arn98] and Arnold and Murua [AM98]. Pytlak [Pyt98] uses IRK methods in connection with implicit index-1 DAEs of type $F(\dot{x}, x, y) = 0$. Brenan et al. [BCP96] show convergence of IRK methods for index-1 systems (5.1). Brenan and Petzold [BP89] prove convergence of IRK methods for semi-explicit index-2 DAEs. Jay [Jay93, Jay95] proves convergence for semi-explicit Hessenberg index-3 DAEs. Consistent initial values and certain regularity and stability conditions of the Runge-Kutta methods under consideration are assumed, cf. also Hairer et al. [HLR89].

5.4 Linearized Implicit Runge-Kutta Methods

The nonlinear system (5.3.4) has to be solved in each integration step. In addition, the DAE (5.1) itself has to be solved many times during the iterative solution of discretized optimal control problems. Since large step sizes h may cause Newton's method to fail the implicit Runge-Kutta method usually is extended by an algorithm for automatic step size selection. The resulting method may be quite time consuming in practical computations. To reduce the computational cost for integration of (5.1) we introduce linearized implicit Runge-Kutta (LRK) methods, which only require to solve linear equations in each integration step and allow to use fixed step-sizes during integration. Numerical experiments show that these methods often lead to a speed-up in the numerical solution (with a reasonable accuracy) when compared to nonlinear implicit Runge-Kutta methods or BDF methods with step-size and order selection discussed in Section 5.2.
Consider the implicit Runge-Kutta method (5.3.2) with coefficients (5.3.1) and stage derivatives implicitly defined by (5.3.4). The linearized implicit Runge-Kutta method is based on the idea to perform only one iteration of Newton's method for (5.3.4). This leads to

$$G_k'(k^{[0]}, t_m, x_N(t_m), h_m) \cdot \left(k - k^{[0]}\right) = -G(k^{[0]}, t_m, x_N(t_m), h_m), \tag{5.4.1}$$

which is a linear equation for the stage derivatives $k = (k_1, \ldots, k_s)$. The derivative G_k' is given by (5.3.7) and the new approximation $x_N(t_{m+1})$ by (5.3.2).

Remark 5.4.1 *A similar idea is used for the construction of Rosenbrock-Wanner (ROW) methods, compare [HW96].*

The vector $k^{[0]} = (k_1^{[0]}, \ldots, k_s^{[0]})$ denotes an initial guess for k and will be specified in the sequel. It turns out that the choice of $k^{[0]}$ is important in view of the order of convergence.
To gain more insights into the convergence properties of the linearized Runge-Kutta method the discussion is restricted to explicit ordinary differential equations with

$$F(t, x(t), \dot{x}(t)) := \dot{x}(t) - f(t, x(t)) = 0_{n_x}, \qquad x(t_0) = x_0. \tag{5.4.2}$$

Later on, an extension to DAEs is suggested which yields promising numerical results although a convergence proof could not be obtained yet.
For the explicit ODE in (5.4.2) the derivative G_k' becomes

$$G_k'(k^{[0]}, t_m, x_N(t_m), h_m) = \begin{pmatrix} I & & \\ & \ddots & \\ & & I \end{pmatrix} - h_m \begin{pmatrix} a_{11} f_x'(\nu_1) & \cdots & a_{1s} f_x'(\nu_1) \\ \vdots & \ddots & \vdots \\ a_{s1} f_x'(\nu_s) & \cdots & a_{ss} f_x'(\nu_s) \end{pmatrix}, \tag{5.4.3}$$

where

$$\nu_i = \left(t_m + c_i h_m, x_N(t_m) + h_m \sum_{l=1}^{s} a_{il} k_l^{[0]} \right), \quad i = 1, \ldots, s.$$

Writing down the linear equation (5.4.1) in its components, we obtain

$$k_i = f(\nu_i) + h_m \sum_{j=1}^{s} a_{ij} f_x'(\nu_i) \left(k_j - k_j^{[0]} \right), \qquad i = 1, \ldots, s. \tag{5.4.4}$$

We are up to investigate the convergence properties of the LRK method for two different choices of the initial guess $k^{[0]}$. We intend to apply Theorem 5.1.7. Hence, to show convergence we have to prove consistency and stability both in the exact solution x of the initial value problem (5.4.2). The order of consistency is determined by Taylor expansion of the local discretization error l_N. The multivariate Taylor formula for a $p+1$ times continuously differentiable function $g : \mathbb{R}^n \to \mathbb{R}^m$ is given by

$$g(x+h) = \sum_{j=0}^{p} \frac{1}{j!} g^{(j)}(x) \cdot \underbrace{(h, \ldots, h)}_{j-fold} + \mathcal{O}(\|h\|^{p+1})$$

for any $x \in \mathbb{R}^n$ and sufficiently small $h \in \mathbb{R}^n$. Herein, the j^{th} derivative of g is a j-linear mapping with

$$g^{(j)}(x) \cdot (h_1, \ldots, h_j) = \sum_{i_1, \ldots, i_j = 1}^{n} \frac{\partial^j g(x)}{\partial x_{i_1} \cdots \partial x_{i_j}} h_{1,i_1} \cdots h_{j,i_j}.$$

For notational convenience we use the abbreviations $h = h_m$ and $x_m = x_N(t_m)$.

5.4.1 A first choice: Constant Prediction

The simplest idea one might have is to use $k^{[0]} = (k_1^{[0]}, \ldots, k_s^{[0]})$ with $k_i^{[0]} = 0_{n_x}$ for all $i = 1, \ldots, s$, which means, that we use x_m as predictor for $x_{m+1}^{(i)}$, $i = 1, \ldots, s$, in (5.3.5). Equation (5.4.4) reduces to

$$k_i = f(t_m + c_i h, x_m) + h \sum_{j=1}^{s} a_{ij} f_x'(t_m + c_i h, x_m)(k_j), \qquad i = 1, \ldots, s. \tag{5.4.5}$$

In matrix notation, Equation (5.4.5) is given by

$$\left(I - h B_1(t_m, x_m, h) \right) k = c_1(t_m, x_m, h), \tag{5.4.6}$$

where

$$\begin{aligned} B_1(t_m, x_m, h) &:= \begin{pmatrix} a^1 \otimes f_x'(t_m + c_1 h, x_m) \\ \vdots \\ a^s \otimes f_x'(t_m + c_s h, x_m) \end{pmatrix}, \\ c_1(t_m, x_m, h) &:= \begin{pmatrix} f(t_m + c_1 h, x_m) \\ \vdots \\ f(t_m + c_s h, x_m) \end{pmatrix}, \end{aligned} \tag{5.4.7}$$

and a^i denotes the i^{th} row of $A = (a_{ij})$ and $\otimes$ is the Kronecker product.

Taylor expansion of k_i at the exact solution x and recursive evaluation leads to

$$\begin{aligned}
k_i &= f + h\left(c_i f_t' + \sum_{j=1}^{s} a_{ij} f_x'(k_j)\right) + h^2\left(\frac{c_i^2}{2} f_{tt}'' + c_i \sum_{j=1}^{s} a_{ij} f_{xt}''(k_j)\right) + \mathcal{O}(h^3) \\
&= f + h\left(c_i f_t' + \sum_{j=1}^{s} a_{ij} f_x'\left(f + h\left(c_j f_t' + \sum_{l=1}^{s} a_{jl} f_x'(k_l)\right) + \mathcal{O}(h^2)\right)\right) \\
&\quad + h^2\left(\frac{c_i^2}{2} f_{tt}'' + c_i \sum_{j=1}^{s} a_{ij} f_{xt}''\left(f + \mathcal{O}(h)\right)\right) + \mathcal{O}(h^3) \\
&= f + h\left(c_i f_t' + \sum_{j=1}^{s} a_{ij} f_x'(f)\right) \\
&\quad + h^2\left(\sum_{j=1}^{s} a_{ij} c_j f_x'(f_t') + \sum_{j=1}^{s} a_{ij} \sum_{l=1}^{s} a_{jl} f_x'(f_x'(f)) + \frac{c_i^2}{2} f_{tt}'' + c_i \sum_{j=1}^{s} a_{ij} f_{xt}''(f)\right) \\
&\quad + \mathcal{O}(h^3),
\end{aligned}$$

where f is evaluated at $(t_m, x(t_m))$. Taylor expansion of the exact solution yields

$$\begin{aligned}
\frac{x(t_m+h) - x(t_m)}{h} &= \dot{x}(t_m) + \frac{h}{2!}\ddot{x}t_m + \frac{h^2}{3!}\dddot{x}(t_m) + \mathcal{O}(h^3) \\
&= f + \frac{h}{2!}\left(f_t' + f_x'(f)\right) \\
&\quad + \frac{h^2}{3!}\left(f_{tt}'' + 2f_{tx}'(f) + f_x'(f_t') + f_x'(f_x'(f)) + f_{xx}''(f,f)\right) + \mathcal{O}(h^3).
\end{aligned}$$

Hence, the Taylor expansion of the local discretization error is given by

$$\begin{aligned}
l_N &= \frac{x(t_m+h) - x(t_m)}{h} - \sum_{i=1}^{s} b_i k_i \\
&= \left(1 - \sum_{i=1}^{s} b_i\right) f + h\left(\left(\frac{1}{2} - \sum_{i=1}^{s} b_i c_i\right) f_t' + \left(\frac{1}{2} - \sum_{i=1}^{s} b_i \sum_{j=1}^{s} a_{ij}\right) f_x'(f)\right) + \mathcal{O}(h^2).
\end{aligned}$$

Notice, that the derivative f_{yy}'' does not occur in the Taylor expansion of the numerical solution. Hence, the maximum attainable order is 2 and it holds

Theorem 5.4.2 *Let f possess continuous and bounded partial derivatives up to order two. The LRK method (5.4.5) is consistent of order 1, if*

$$\sum_{i=1}^{s} b_i = 1$$

holds. It is consistent of order 2, if in addition

$$\sum_{i=1}^{s} b_i c_i = \frac{1}{2}, \quad \sum_{i=1}^{s} b_i \sum_{j=1}^{s} a_{ij} = \frac{1}{2}$$

hold. $p = 2$ is the maximum attainable order of consistency.

The LRK method (5.4.5) has a drawback. The method is not invariant under autonomization, i.e. it does not produce the same numerical results for the system (5.4.2) and the equivalent autonomous system

$$\begin{pmatrix} \dot{y}(t) \\ \dot{z}(t) \end{pmatrix} = \begin{pmatrix} f(z(t), y(t)) \\ 1 \end{pmatrix}, \quad \begin{pmatrix} y(0) \\ z(0) \end{pmatrix} = \begin{pmatrix} x_0 \\ t_0 \end{pmatrix}, \tag{5.4.8}$$

cf. Deuflhard and Bornemann [DB02], pp. 137-138. For, discretization of (5.4.2) and (5.4.8) with the LRK method (5.4.1) yields

$$k_i = f(t_m + c_i h, x_m) + h \sum_{j=1}^{s} a_{ij} f_x'(t_m + c_i h, x_m)(k_j), \quad i = 1, \ldots, s,$$

and

$$\begin{pmatrix} k_i^y \\ k_i^z \end{pmatrix} = \begin{pmatrix} f(z_m, y_m) + h \sum_{j=1}^{s} a_{ij} \left(f_x'(z_m, y_m)(k_j^y) + f_t'(z_m, y_m)(k_j^z) \right) \\ 1 \end{pmatrix}, \quad i = 1, \ldots, s,$$

respectively. Obviously, k_i and k_i^y are different if f explicitly depends on t.

5.4.2 A second choice: Linear Prediction

We use the initial guess $k^{[0]} = (k_1^{[0]}, \ldots, k_s^{[0]})$ with $k_i^{[0]} = f(t_m, x_m)$ for all $i = 1, \ldots, s$, which means, that we use $x_m + h \sum_{j=1}^{s} a_{ij} f(t_m, x_m)$ as predictor for $x_{m+1}^{(i)}$, $i = 1, \ldots, s$, in (5.3.5). Equation (5.4.4) reduces to

$$k_i = f(\nu_i) + h \sum_{j=1}^{s} a_{ij} f_x'(\nu_i)(k_j - f(t_m, x_m)), \qquad i = 1, \ldots, s, \tag{5.4.9}$$

where

$$\nu_i = \left(t_m + c_i h, x_m + h \sum_{l=1}^{s} a_{il} f(t_m, x_m) \right), \quad i = 1, \ldots, s.$$

It turns out that the method (5.4.9) is invariant under autonomization under appropriate assumptions, which are identical to those in Deuflhard and Bornemann [DB02], Lemma 4.16, p. 138.

Lemma 5.4.3 *Let the coefficients (5.3.1) fulfill the conditions*

$$\sum_{i=1}^{s} b_i = 1$$

and

$$c_i = \sum_{j=1}^{s} a_{ij}, \qquad i = 1, \ldots, s. \tag{5.4.10}$$

Then, the LRK method (5.4.9) is invariant under autonomization.

Proof. Application of the LRK method to (5.4.2) and (5.4.8) yields (5.4.9) and

$$\begin{pmatrix} k_i^y \\ k_i^z \end{pmatrix} = \begin{pmatrix} f(\tilde{\nu}_i) + h\sum_{j=1}^{s} a_{ij}\left(f_x'(\tilde{\nu}_i)(k_j^y - f(z_m,y_m)) + f_t'(\tilde{\nu}_i)(k_j^z - 1)\right) \\ 1 \end{pmatrix}, \quad i = 1,\ldots,s,$$

respectively, where

$$\tilde{\nu}_i = \left(z_m + h\sum_{l=1}^{s} a_{il}\cdot 1, y_m + h\sum_{l=1}^{s} a_{il} f(z_m,y_m)\right), \quad i = 1,\ldots,s.$$

k_i^y is equal to k_i in (5.4.9), if $\tilde{\nu}_i = \nu_i$ and $z_m = t_m$ hold for all i and all m. The latter is fulfilled if $\sum_{i=1}^s b_i = 1$ holds, since then $z_{m+1} = z_m + h\sum_{i=1}^s b_i k_i^z = z_m + h\sum_{i=1}^s b_i = z_m + h = t_{m+1}$. The first condition requires that $z_m + h\sum_{l=1}^s a_{il} = t_m + c_i h$, which is satisfied if (5.4.10) holds. ■

In matrix notation, equation (5.4.9) is given by

$$(I - hB_2(t_m,x_m,h))\,k = c_2(t_m,x_m,h), \tag{5.4.11}$$

where

$$\begin{aligned} B_2(t_m,x_m,h) &:= \begin{pmatrix} a^1 \otimes f_x'(t_m + c_1h, x_m + hc_1 f(t_m,x_m)) \\ \vdots \\ a^s \otimes f_x'(t_m + c_sh, x_m + hc_s f(t_m,x_m)) \end{pmatrix}, \\ c_2(t_m,x_m,h) &:= -hB_2(t_m,x_m,h)\cdot(e \otimes f(t_m,x_m)) \\ &\quad + \begin{pmatrix} f(t_m + c_1h, x_m + hc_1 f(t_m,x_m)) \\ \vdots \\ f(t_m + c_sh, x_m + hc_s f(t_m,x_m)) \end{pmatrix}, \end{aligned} \tag{5.4.12}$$

and a^i denotes the i^{th} row of $A = (a_{ij})$ and $e = (1,\ldots,1)^\top \in \mathbb{R}^s$. Notice, that we exploited (5.4.10).

Lemma 5.4.3 allows to restrict the discussion to autonomous initial value problems (5.4.2), where f does not depend explicitly on t. This simplifies the Taylor expansions considerably. From now on, it is assumed, that the assumptions of Lemma 5.4.3 are satisfied. We determine the order of consistency of the LRK method (5.4.9).

The Taylor expansion of the exact solution of the autonomous problem is given by

$$\begin{aligned} \frac{x(t_m+h) - x(t_m)}{h} &= f + \frac{h}{2!}f'f + \frac{h^2}{3!}\left(f''(f,f) + f'f'f\right) \\ &\quad + \frac{h^3}{4!}\left(f'''(f,f,f) + 3f''(f'f,f) + f'f''(f,f) + f'f'f'f\right) \\ &\quad + \frac{h^4}{5!}\Big(f^{(4)}(f,f,f,f) + 6f'''(f'f,f,f) + 4f''(f''(f,f),f) \\ &\qquad + 4f''(f'f'f,f) + 3f''(f'f,f'f) + f'f'''(f,f,f) \\ &\qquad + 3f'f''(f'f,f) + f'f'f''(f,f) + f'f'f'f'f\Big) + \mathcal{O}(h^5), \end{aligned}$$

where f is evaluated at $x(t_m)$. Notice, that we omitted brackets if no confusion was possible, e.g. we wrote $f'f'f$ instead of $f'(f'(f))$.

Now we derive the Taylor expansion of the approximate solution. The stages k_i of the linearized Runge-Kutta method (5.4.9) for the autonomous ODE (5.4.2) under consideration of (5.4.10) satisfy

$$\begin{aligned}
k_i &= h\sum_{j=1}^{s} a_{ij} f'(x_m + hc_i f)(k_j) + f(x_m + hc_i f) - h\sum_{j=1}^{s} a_{ij} f'(x_m + hc_i f)(f) \\
&= h\sum_{j=1}^{s} a_{ij} f'(x_m + hc_i f)(k_j) + f(x_m + hc_i f) - hc_i f'(x_m + hc_i f)(f) \\
&= h\sum_{k=0}^{p} \frac{h^k c_i^k}{k!} \sum_{j=1}^{s} a_{ij} f_n^{(k+1)}(k_j, \underbrace{f,\ldots,f}_{k-fold}) \\
&\quad + \sum_{k=0}^{p} \frac{h^k c_i^k}{k!} f_n^{(k)} \underbrace{(f,\ldots,f)}_{k-fold} - \sum_{k=0}^{p} \frac{h^{k+1} c_i^{k+1}}{k!} f_n^{(k+1)} \underbrace{(f,\ldots,f)}_{(k+1)-fold} + \mathcal{O}(h^{p+1}) \\
&= \sum_{k=1}^{p} \frac{h^k c_i^{k-1}}{(k-1)!} \sum_{j=1}^{s} a_{ij} f_n^{(k)}(k_j, \underbrace{f,\ldots,f}_{(k-1)-fold}) \\
&\quad + \sum_{k=0}^{p} \frac{h^k c_i^k}{k!} f_n^{(k)} \underbrace{(f,\ldots,f)}_{k-fold} - \sum_{k=1}^{p} \frac{h^k c_i^k}{(k-1)!} f_n^{(k)} \underbrace{(f,\ldots,f)}_{k-fold} + \mathcal{O}(h^{p+1}) \\
&= \sum_{k=1}^{p} \frac{h^k c_i^{k-1}}{(k-1)!} \sum_{j=1}^{s} a_{ij} f_n^{(k)}(k_j, \underbrace{f,\ldots,f}_{(k-1)-fold}) \\
&\quad + f + \sum_{k=1}^{p} h^k c_i^k \frac{1-k}{k!} f_n^{(k)} \underbrace{(f,\ldots,f)}_{k-fold} + \mathcal{O}(h^{p+1}).
\end{aligned}$$

In particular we get for $p = 4$:

$$\begin{aligned}
k_i &= f + h\sum_{j=1}^{s} a_{ij} f'(k_j) + h^2 \left(c_i \sum_{j=1}^{s} a_{ij} f''(k_j, f) - \frac{c_i^2}{2} f''(f,f) \right) \\
&\quad + h^3 \left(\frac{c_i^2}{2} \sum_{j=1}^{s} a_{ij} f'''(k_j, f, f) - \frac{c_i^3}{3} f'''(f,f,f) \right) \\
&\quad + h^4 \left(\frac{c_i^3}{6} \sum_{j=1}^{s} a_{ij} f^{(4)}(k_j, f, f, f) - \frac{c_i^4}{8} f^{(4)}(f,f,f,f) \right) + \mathcal{O}(h^5).
\end{aligned}$$

Recursive evaluation leads to

$$
\begin{aligned}
k_i \;=\;& f + h\sum_{j=1}^{s} a_{ij} f'\Bigg[f + h\sum_{l=1}^{s} a_{jl} f'(k_l) \\
&\qquad + h^2\left(c_j \sum_{l=1}^{s} a_{jl} f''(k_l,f) - \frac{c_j^2}{2} f''(f,f)\right) \\
&\qquad + h^3\left(\frac{c_j^2}{2} \sum_{l=1}^{s} a_{jl} f'''(k_l,f,f) - \frac{c_j^3}{3} f'''(f,f,f)\right) + \mathcal{O}(h^4)\Bigg] \\
& + h^2 \Bigg(c_i \sum_{j=1}^{s} a_{ij} f'' \Bigg[f + h\sum_{l=1}^{s} a_{jl} f'(k_l) \\
&\qquad + h^2\left(c_j \sum_{l=1}^{s} a_{jl} f''(k_l,f) - \frac{c_j^2}{2} f''(f,f)\right) + \mathcal{O}(h^3), f\Bigg] \\
&\quad - \frac{c_i^2}{2} f''(f,f)\Bigg) \\
& + h^3\left(\frac{c_i^2}{2} \sum_{j=1}^{s} a_{ij} f''' \left[f + h\sum_{l=1}^{s} a_{jl} f'(k_l) + \mathcal{O}(h^2), f, f\right] - \frac{c_i^3}{3} f'''(f,f,f)\right) \\
& + h^4\left(\frac{c_i^3}{6} \sum_{j=1}^{s} a_{ij} f^{(4)}(f,f,f,f) - \frac{c_i^4}{8} f^{(4)}(f,f,f,f)\right) + \mathcal{O}(h^5).
\end{aligned}
$$

Sorting for powers of h and exploiting $c_i = \sum_{j=1}^{s} a_{ij}$ yields

$$
\begin{aligned}
k_i \;=\;& f + hc_i f'(f) + h^2\left(\sum_{j=1}^{s} a_{ij} \sum_{l=1}^{s} a_{jl} f'(f'(k_l)) + \frac{c_i^2}{2} f''(f,f)\right) \\
& + h^3 \Bigg(\sum_{j=1}^{s} a_{ij} c_j \sum_{l=1}^{s} a_{jl} f'(f''(k_l,f)) - \sum_{j=1}^{s} a_{ij} \frac{c_j^2}{2} f'(f''(f,f)) \\
&\qquad + c_i \sum_{j=1}^{s} a_{ij} \sum_{l=1}^{s} a_{jl} f''(f'(k_l),f) + \frac{c_i^3}{6} f'''(f,f,f)\Bigg) \\
& + h^4 \Bigg(\sum_{j=1}^{s} a_{ij} \frac{c_j^2}{2} \sum_{l=1}^{s} a_{jl} f'(f'''(k_l,f,f)) - \sum_{j=1}^{s} a_{ij} \frac{c_j^3}{3} f'(f'''(f,f,f)) \\
&\qquad + c_i \sum_{j=1}^{s} a_{ij} c_j \sum_{l=1}^{s} a_{jl} f''(f''(k_l,f),f) - c_i \sum_{j=1}^{s} a_{ij} \frac{c_j^2}{2} f''(f''(f,f),f) \\
&\qquad + \frac{c_i^2}{2} \sum_{j=1}^{s} a_{ij} \sum_{l=1}^{s} a_{jl} f'''(f'(k_l),f,f) + \frac{c_i^4}{24} f^{(4)}(f,f,f,f)\Bigg) + \mathcal{O}(h^5).
\end{aligned}
$$

Again, a recursive evaluation leads to

$$\begin{aligned}
k_i &= f + hc_i f'(f) + h^2 \Bigg(\sum_{j=1}^{s} a_{ij} \sum_{l=1}^{s} a_{jl} f'\Bigg(f' \Bigg[f + hc_l f'(f) \\
&\qquad + h^2 \Bigg(\sum_{m=1}^{s} a_{lm} \sum_{r=1}^{s} a_{mr} f'(f'(k_r)) + \frac{c_l^2}{2} f''(f,f) + \mathcal{O}(h^3) \Bigg) \Bigg] + \frac{c_i^2}{2} f''(f,f) \Bigg) \\
&\quad + h^3 \Bigg(\sum_{j=1}^{s} a_{ij} c_j \sum_{l=1}^{s} a_{jl} f'(f'' \left[f + hc_l f'(f) + \mathcal{O}(h^2), f \right]) - \sum_{j=1}^{s} a_{ij} \frac{c_j^2}{2} f'(f''(f,f)) \\
&\qquad + c_i \sum_{j=1}^{s} a_{ij} \sum_{l=1}^{s} a_{jl} f''(f' \left[f + hc_l f'(f) + \mathcal{O}(h^2) \right], f) + \frac{c_i^3}{6} f'''(f,f,f) \Bigg) \\
&\quad + h^4 \Bigg(\sum_{j=1}^{s} a_{ij} \frac{c_j^3}{6} f'(f'''(f,f,f)) + c_i \sum_{j=1}^{s} a_{ij} \frac{c_j^2}{2} f''(f''(f,f),f) \\
&\qquad + \frac{c_i^2}{2} \sum_{j=1}^{s} a_{ij} c_j f'''(f'(f),f,f) + \frac{c_i^4}{24} f^{(4)}(f,f,f,f) \Bigg) + \mathcal{O}(h^5).
\end{aligned}$$

Sorting for powers of h and exploiting $c_i = \sum_{j=1}^{s} a_{ij}$ yields

$$\begin{aligned}
k_i &= f + hc_i f'(f) + h^2 \Bigg(\sum_{j=1}^{s} a_{ij} c_j f'(f'(f)) + \frac{c_i^2}{2} f''(f,f) \Bigg) \\
&\quad + h^3 \Bigg(\sum_{j=1}^{s} a_{ij} \sum_{l=1}^{s} a_{jl} c_l f'(f'(f'(f))) + \sum_{j=1}^{s} a_{ij} \frac{c_j^2}{2} f'(f''(f,f)) \\
&\qquad + c_i \sum_{j=1}^{s} a_{ij} c_j f''(f'(f),f) + \frac{c_i^3}{6} f'''(f,f,f) \Bigg) \\
&\quad + h^4 \Bigg(\sum_{j=1}^{s} a_{ij} \frac{c_j^3}{6} f'(f'''(f,f,f)) + c_i \sum_{j=1}^{s} a_{ij} \frac{c_j^2}{2} f''(f''(f,f),f) \\
&\qquad + \frac{c_i^2}{2} \sum_{j=1}^{s} a_{ij} c_j f'''(f'(f),f,f) + \frac{c_i^4}{24} f^{(4)}(f,f,f,f) \\
&\qquad + \sum_{j=1}^{s} a_{ij} \sum_{l=1}^{s} a_{jl} \sum_{m=1}^{s} a_{lm} c_m f'(f'(f'(f'(f)))) \\
&\qquad + \sum_{j=1}^{s} a_{ij} \sum_{l=1}^{s} a_{jl} \frac{c_l^2}{2} f'(f'(f''(f,f))) + \sum_{j=1}^{s} a_{ij} c_j \sum_{l=1}^{s} a_{jl} c_l f'(f''(f'(f),f)) \\
&\qquad + c_i \sum_{j=1}^{s} a_{ij} \sum_{l=1}^{s} a_{jl} c_l f''(f'(f'(f)),f) \Bigg) + \mathcal{O}(h^5).
\end{aligned}$$

Finally, we obtain the expansion

$$\begin{aligned}\sum_{i=1}^{s} b_i k_i &= \sum_{i=1}^{s} b_i f + h\sum_{i=1}^{s} b_i c_i f'f + h^2 \sum_{i=1}^{s} b_i \left(\sum_{j=1}^{s} a_{ij} c_j f'f'f + \frac{c_i^2}{2} f''(f,f) \right) \\ &\quad + h^3 \sum_{i=1}^{s} b_i \left(\sum_{j=1}^{s} a_{ij} \sum_{l=1}^{s} a_{jl} c_l f'f'f'f + \sum_{j=1}^{s} a_{ij} \frac{c_j^2}{2} f'f''(f,f) \right. \\ &\quad \left. + c_i \sum_{j=1}^{s} a_{ij} c_j f''(f'f,f) + \frac{c_i^3}{6} f'''(f,f,f) \right) \\ &\quad + h^4 \sum_{i=1}^{s} b_i \left(\sum_{j=1}^{s} a_{ij} \frac{c_j^3}{6} f'f'''(f,f,f) + c_i \sum_{j=1}^{s} a_{ij} \frac{c_j^2}{2} f''(f''(f,f),f) \right. \\ &\quad + \frac{c_i^2}{2} \sum_{j=1}^{s} a_{ij} c_j f'''(f'f,f,f) + \frac{c_i^4}{24} f^{(4)}(f,f,f,f) \\ &\quad + \sum_{j,l,m=1}^{s} a_{ij} a_{jl} a_{lm} c_m f'f'f'f'f + \sum_{j,l=1}^{s} a_{ij} a_{jl} \frac{c_l^2}{2} f'f'f''(f,f) \\ &\quad \left. + \sum_{j,l=1}^{s} a_{ij} c_j a_{jl} c_l f'f''(f'f,f) + c_i \sum_{j,l=1}^{s} a_{ij} a_{jl} c_l f''(f'f'f,f) \right) + \mathcal{O}(h^5).\end{aligned}$$

An investigation of the local discretization error

$$l_N = \frac{x(t_m+h) - x(t_m)}{h} - \sum_{i=1}^{s} b_i k_i$$

reveals that the terms involving the derivative $f''(f'f, f'f)$ are missing. Hence, the maximal attainable order of consistency is 4. We thus proved

Theorem 5.4.4 *Let f possess continuous and bounded partial derivatives up to order four and let (5.4.10) be valid. The LRK method (5.4.9) is consistent of order 1, if*

$$\sum_{i=1}^{s} b_i = 1$$

holds. It is consistent of order 2, if in addition

$$\sum_{i=1}^{s} b_i c_i = \frac{1}{2}$$

holds. It is consistent of order 3, if in addition

$$\sum_{i,j=1}^{s} b_i a_{ij} c_j = \frac{1}{6}, \qquad \sum_{i=1}^{s} b_i c_i^2 = \frac{1}{3}$$

hold. It is consistent of order 4, if in addition

$$\sum_{i,j,l=1}^{s} b_i a_{ij} a_{jl} c_l = \frac{1}{24}, \ \sum_{i,j=1}^{s} b_i a_{ij} c_j^2 = \frac{1}{12}, \ \sum_{i,j=1}^{s} b_i c_i a_{ij} c_j = \frac{1}{8}, \ \sum_{i=1}^{s} b_i c_i^3 = \frac{1}{4}.$$

hold. $p = 4$ is the maximal attainable order of consistency.

Remark 5.4.5 *The above conditions are the usual conditions known for general IRK methods. It has to be mentioned, that the LRK method is A-stable if the nonlinear IRK method (5.3.2) is A-stable, since A-stability is defined for linear differential equations. For linear differential equations, the LRK method coincides with the IRK method.*

In view of the completion of the convergence proof, we finally show the Lipschitz continuity of the increment function Φ for the two methods under suitable differentiability assumptions. Notice, that the LRK methods defined by (5.4.5) and (5.4.9) are one-step methods and we may apply Theorem 5.1.8. Notice, that implicit methods can be viewed as one-step methods as well, provided that the arising nonlinear resp. linear equations can be solved.

Theorem 5.4.6 *Let f be twice continuously differentiable. Then, Φ in (5.3.3) with k from (5.4.6) respectively (5.4.11) is locally Lipschitz continuous in the exact solution x.*

Proof. Since Φ is a linear combination of k_i, $i = 1, \ldots, s$, it suffices to show the local Lipschitz continuity of k. Since f is twice continuously differentiable, the functions B_i and c_i, $i = 1, 2$, in (5.4.7) and (5.4.12) are continuously differentiable w.r.t. t, x and h. In addition, for $i = 1, 2$ it holds

$$\begin{aligned}\lim_{h\to 0} B_i(t_m, x_m, h) &= A \otimes f_x'(t_m, x_m),\\ \lim_{h\to 0} c_i(t_m, x_m, h) &= e \otimes f(t_m, x_m).\end{aligned}$$

Hence, for sufficiently small h it holds $\|hB_i(\eta_n)\| < 1$, $i = 1, 2$. Neumann's lemma yields the non-singularity of $I - hB_i$, $i = 1, 2$ for all sufficiently small h. In either case, the implicit function theorem yields the existence of a continuously differentiable function $k = k(t_m, x_m, h)$ satisfying (5.4.6) respectively (5.4.11). Application of the mean-value theorem for vector functions yields

$$\|k(t_m, \hat{x}, h) - k(t_m, \tilde{x}, h)\| = \| \int_0^1 k_x'(t_m, \hat{x} + t(\tilde{x} - \hat{x}), h)(\tilde{x} - \hat{x}) dt\| \leq L_\delta \|\tilde{x} - \hat{x}\|,$$

where $L_\delta := \max_{\|\zeta - x(t)\| \leq \delta} \|k_x'(t, \zeta, h)\|$ exists for some $\delta > 0$ and all $t \in [t_0, t_f]$ and all sufficiently small h. ■

Exploiting Theorems 5.1.7, 5.1.8 we proved

Theorem 5.4.7 *Let the assumptions of Theorems 5.4.2, 5.4.6 and Theorems 5.4.4, 5.4.6, respectively, be valid. Then the linearized LRK methods (5.4.6) and (5.4.11), respectively, are convergent. The order of convergence is identical with the order of consistency of the respective method.*

Summarizing, the order of convergence of the LRK method depends on the choice of the initial guess for one step of Newton's method. An appropriate choice allows to create methods up to order 4. The methods are easy to implement and show a very good computational performance within the numerical solution of dynamic optimization problems. The following numerical examples show that the use of LRK methods reduces the computation time considerably compared to standard BDF integration schemes, while the accuracy remains the same.

5.4.3 Numerical Experiments for DAEs

Unfortunately, the theory for DAEs is not yet completed. Nevertheless, numerical tests show very promising results, cf. Gerdts [Ger04]. For implicit DAEs the problem of choosing the initial guess $k^{[0]} = (k_1^{[0]}, \ldots, k_s^{[0]})$ in (5.4.1) appropriately arises. Again, we consider the integration step from t_m to t_{m+1}. For ODEs the choices $k_i^{[0]} = f(t_m, x_N(t_m))$, $i = 1, \ldots, s$, produced the best results. Unfortunately, for implicit systems the right-hand side f defining the derivative is not available explicitly. Numerical experiments suggest that for stiffly accurate Runge-Kutta methods the choice

$$k_i^{[0]} := k_s(t_{m-1}, x_N(t_{m-1}), h_{m-1}), \quad i = 1, \ldots, s, \tag{5.4.13}$$

seems to be reasonable. Recall, that stiffly accurate methods, e.g. the RADAUIIA methods, satisfy

$$c_s = 1 \quad \text{and} \quad a_{sj} = b_j, \quad j = 1, \ldots, s,$$

cf. Strehmel and Weiner [SW95] and Hairer et al. [HLR89].
The quantity $k_s(t_{m-1}, x_N(t_{m-1}), h_{m-1})$ is calculated in the previous integration step from t_{m-1} to t_m and can be interpreted as a derivative at the time point $t_{m-1} + c_s h_{m-1} = t_m$ and hence plays at least approximately the role of f. For Runge-Kutta methods which are not stiffly accurate, an initial guess of type

$$k_i^{[0]} := \frac{x_N(t_m) - x_N(t_{m-1})}{h_{m-1}}, \quad i = 1, \ldots, s,$$

seems to be reasonable, since the right side is an approximation of the derivative at t_m.

Computational results for the well-known mathematical pendulum problem with $m = 1$ $[kg]$, $l = 1$ $[m]$ are to be presented, cf. Gerdts [Ger04]. The equations of motion of the pendulum are given by the index-3 Hessenberg DAE

$$\begin{aligned}
0 &= \dot{x}_1 - x_3 \\
0 &= \dot{x}_2 - x_4 \\
0 &= m \cdot \dot{x}_3 - (\qquad\qquad - 2 \cdot x_1 \cdot \lambda) \\
0 &= m \cdot \dot{x}_4 - (-m \cdot g - 2 \cdot x_2 \cdot \lambda) \\
0 &= x_1^2 + x_2^2 - l^2,
\end{aligned}$$

cf. Gerdts and Büskens [GB02]. The index reduced index-2 DAE arises, if the last equation is replaced by its time derivative

$$0 = 2(x_1 \cdot x_3 + x_2 \cdot x_4).$$

For the following numerical tests we used the initial value

$$(x_1(0), x_2(0), x_3(0), x_4(0), \lambda(0)) = (1, 0, 0, 0, 0).$$

Table 5.1 shows computational results for the linearized 2-stage RADAUIIA method with Butcher array

$$\begin{array}{c|cc}
1/3 & 5/12 & -1/12 \\
1 & 3/4 & 1/4 \\
\hline
 & 3/4 & 1/4
\end{array}$$

applied to the index reduced index-2 pendulum example with initial guess given by (5.4.13) and fixed step-sizes $h = 1/N$ on the time interval $[0, 1]$. The estimated order of convergence results are in agreement with those for the nonlinear RADAUIIA method (5.3.2) and (5.3.4) derived by Hairer et al. [HLR89] for Hessenberg systems, that is order 3 for the differential variables x_1, x_2 and x_3, x_4 and order 2 for the algebraic variable λ.

Table 5.1: Order of convergence for the linearized 2-stage RADAUIIA method applied to the index reduced index-2 pendulum test example.

N	max. ERR	max. ERR	Order	Order
	$x_1,\ldots,x_4$	λ	$x_1,\ldots,x_4$	λ
10	0.48083E-01	0.72142E+00	0.29990E+01	0.19993E+01
20	0.60145E-02	0.18044E+00	0.29999E+01	0.20000E+01
40	0.75184E-03	0.45111E-01	0.30000E+01	0.20000E+01
80	0.93981E-04	0.11278E-01	0.30000E+01	0.20000E+01
160	0.11748E-04	0.28194E-02	0.30000E+01	0.20000E+01
320	0.14684E-05	0.70485E-03	0.30000E+01	0.20000E+01
640	0.18356E-06	0.17621E-03	0.30000E+01	0.20000E+01
1280	0.22944E-07	0.44053E-04	0.30000E+01	0.20000E+01
2560	0.28681E-08	0.11013E-04	0.30000E+01	0.20000E+01

Table 5.2 shows computational results for the linearized 2-stage RADAUIIA method applied to the index-3 pendulum example with initial guess given by (5.4.13) and fixed step-sizes $h = 1/N$ on the time interval $[0, 1]$. The computational order results are in agreement with those for the nonlinear RADAUIIA method (5.3.2) and (5.3.4) derived in [HLR89] for Hessenberg systems, that is order 3 for the positions x_1, x_2, order 2 for the velocities x_3, x_4 and order 1 for the algebraic variable λ.

Table 5.2: Order of convergence for the linearized 2-stage RADAUIIA method applied to the index-3 pendulum test example.

N	max. ERR	max. ERR	max ERR	Order	Order	Order
	x_1, x_2	x_3, x_4	λ	x_1, x_2	x_3, x_4	λ
10	0.63312E-01	0.35222E+00	0.18016E+02	0.30701E+01	0.33231E+01	0.29223E+01
20	0.75389E-02	0.35195E-01	0.23766E+01	0.31923E+01	0.25480E+01	0.14415E+01
40	0.82474E-03	0.60181E-02	0.87504E+00	0.30845E+01	0.22292E+01	0.11858E+01
80	0.97231E-04	0.12835E-02	0.38465E+00	0.30349E+01	0.20446E+01	0.10938E+01
160	0.11863E-04	0.31110E-03	0.18022E+00	0.30153E+01	0.20205E+01	0.10480E+01
320	0.14672E-05	0.76676E-04	0.87158E-01	0.30071E+01	0.20097E+01	0.10243E+01
640	0.18251E-06	0.19041E-04	0.42850E-01	0.30033E+01	0.20046E+01	0.10123E+01
1280	0.22762E-07	0.47451E-05	0.21243E-01	0.29948E+01	0.20019E+01	0.10061E+01
2560	0.28556E-08	0.11847E-05	0.10577E-01	0.28912E+01	0.19995E+01	0.10028E+01

Similar computations for the linearized implicit Euler method always yield order one for all components.

Similar computations for the linearized 3-stage RADAUIIA method for the index reduced index-2 pendulum example yield only order 3 for the differential components x_1, x_2 and x_3, x_4 and order 2 for λ. According to [HLR89] the nonlinear method has orders 5 and 3, respectively.

In either case, the reference solution was obtained by RADAU5 with $atol = rtol = 10^{-12}$ and GGL-stabilization.

The use of the linearized implicit Runge-Kutta method for the discretization of optimal control problems allows to speed up the solution process significantly. An illustration of this statement can be found in Gerdts [Ger05c] where an optimal control problem resulting from automobile test-driving is solved numerically. Table 5.3 summarizes the CPU times for the numerical solution of the latter optimal control problem obtained by the third order linearized 2-stage RADAUIIA method with constant step size h and a standard BDF method (DASSL) with automatic step size and order selection. The relative error in Table 5.3 denotes the relative error in the respective optimal objective function values of the discretized optimal control problem. The number N denotes the number of discretization points used to discretize the optimal control problem. The speedup is the ratio of columns 4 and 2. Table 5.3 shows that the LRK method in average is 10 times faster than the BDF method. A comparison of the accuracies of the respective solutions reveals that the quality of the LRK solution is as good as the BDF solution.

Table 5.3: CPU times for the numerical solution of the discretized optimal control problem by the linearized 2-stage RADAUIIA method respectively the BDF method for different numbers N of control grid points.

N	CPU LRK	OBJ LRK	CPU BDF	OBJ BDF	RELERR	SPEEDUP
	(in [s])		(in [sec])		OBJ	FACTOR
26	2.50	7.718303	26.63	7.718305	0.00000026	10.7
51	8.15	7.787998	120.99	7.787981	0.00000218	14.8
101	18.02	7.806801	208.40	7.806798	0.00000038	11.6
201	21.24	7.819053	171.31	7.819052	0.00000013	8.1
251	198.34	7.817618	1691.15	7.817618	0.00000000	8.5
401	615.31	7.828956	4800.09	7.828956	0.00000000	7.8

Chapter 6

Discretization of Optimal Control Problems

The knowledge of necessary conditions in terms of a (local or global) minimum principle for optimal control problems subject to state and/or control constraints gives rise to the so-called 'indirect approach'. The indirect approach attempts to exploit the minimum principle and usually results in a multi-point boundary value problem for the state and adjoint variables. The resulting multi-point boundary value problem is efficiently solved by, e.g., the multiple shooting method. Multiple shooting methods and particularly the implementations MUMUS, cf. Hiltmann et al. [HCB93], and BOUNDSCO, cf. Oberle and Grimm [OG01], have shown their capability in several practical applications. Although the indirect approach usually leads to highly accurate numerical solutions, provided that the multiple shooting method converges, it suffers from two major drawbacks. The first drawback is that a very good initial guess for the approximate solution is needed in order to guarantee local convergence. The construction of a good initial guess is complicated, since this requires among other things a good estimate of the switching structure of the problem, i.e. the sequence of active and inactive state and/or control constraints. The second drawback lies in the fact, that for high dimensional systems it is often cumbersome and requires sophisticated knowledge about necessary or sufficient conditions to set up the optimality system, even if commercial algebra packages like MAPLE are used.

In contrast to indirect methods the 'direct methods' are based on a suitable discretization of the infinite dimensional optimal control problem. The resulting finite dimensional optimization problem can be solved by, e.g., SQP methods. This approach is particularly advantageous for large scale problems and for users which do not have deep insights in optimal control theory. Moreover, numerical experiences show that direct methods are robust and sufficiently accurate as well. Furthermore, by comparing the necessary conditions for the discretized problem with those of the original problem it is possible to derive approximations for adjoint variables based on the multipliers of the discretized problem. Consequently, these two approaches can often be combined by using the direct approach in order to find an accurate initial guess (including switching structure and adjoints) for the indirect approach. A numerical example of the indirect approach was already discussed in Example 4.1.4.

In the sequel, we will focus on the direct discretization approach for DAE optimal control problems. The aim of this section is to develop a general framework for the numerical solution of such problems. Therefore, the statement of the optimal control problem is kept as general as possible without assuming a special structure of the DAE in advance. Hence, as in Chapter 5 most of the time we will deal with general DAEs of type

$$F(t, x(t), \dot{x}(t), u(t)) = 0_{n_x} \qquad \text{a.e. in } [t_0, t_f]. \tag{6.1}$$

This has also the notational advantage that we don't have to distinguish between differential and algebraic variables as far as it is not essential. However, for numerical computations there are certain restrictions as pointed out in Remark 6.2 below.

We consider

Problem 6.1 (DAE Optimal Control Problem)

$$
\begin{array}{llll}
\textit{Minimize} & \varphi(x(t_0), x(t_f)) & & \\
\textit{s.t.} & \textit{DAE (6.1)}, & & \\
 & \psi(x(t_0), x(t_f)) = 0_{n_\psi}, & & \\
 & c(t, x(t), u(t)) \le 0_{n_c} & & \textit{a.e. in } [t_0, t_f], \\
 & s(t, x(t)) \le 0_{n_s} & & \textit{in } [t_0, t_f].
\end{array}
$$

Some remarks are in order.

Remark 6.2 *In several aspects the statement of the optimal control problem 6.1 and particularly the DAE (6.1) is too general. First of all, from Chapter 4 we may draw the conclusion that control and algebraic variables have the same properties (both are L^∞-functions). Hence, it makes no sense – at least theoretically – to allow that φ and ψ in Problem 6.1 depend on pointwise evaluations of the algebraic variables. Secondly, the numerical computation of consistent initial values, cf. Definition 1.7, only works efficiently for certain subclasses of the DAE (6.1), e.g. Hessenberg DAEs according to Definition 1.5. Thirdly, as it has been mentioned already in Chapter 5 numerical methods again only work well for certain subclasses of (6.1), e.g. Hessenberg DAEs up to index 3.*

6.1 Direct Discretization Methods

Direct discretization methods are based on a discretization of the infinite dimensional optimal control problem. The resulting discretized problem will be a finite dimensional optimization problem. All subsequently discussed methods work on the (not necessarily equidistant) grid

$$\mathbb{G}_N := \{t_0 < t_1 < \ldots < t_N = t_f\} \tag{6.1.1}$$

with step sizes $h_j = t_{j+1} - t_j$, $j = 0, \ldots, N-1$ and mesh-size $h := \max_{j=0,\ldots,N-1} h_j$. Often, $\mathbb{G}_N$ will be an equidistant partition of the interval $[t_0, t_f]$ with constant step size $h = (t_f - t_0)/N$ and grid points $t_i = t_0 + ih$, $i = 0, \ldots, N$.
A direct discretization method is essentially defined by the following operations:

- *Control discretization:*
  The control space $L^\infty([t_0, t_f], \mathbb{R}^{n_u})$ is replaced by some M-dimensional subspace

  $$L^\infty_M([t_0, t_f], \mathbb{R}^{n_u}) \subset L^\infty([t_0, t_f], \mathbb{R}^{n_u})$$

  where $M \in \mathbb{N}$ is finite. The dimension M usually depends on the number N of intervals in $\mathbb{G}_N$, i.e. $M = \mathcal{O}(N)$.

  Let $\mathcal{B} := \{B_1, \ldots, B_M\}$ be a basis of $L^\infty_M([t_0, t_f], \mathbb{R}^{n_u})$. Then, every $u_M \in L^\infty_M([t_0, t_f], \mathbb{R}^{n_u})$ is defined by

  $$u_M(\cdot) = \sum_{i=1}^{M} w_i B_i(\cdot) \tag{6.1.2}$$

  with coefficients $w := (w_1, \ldots, w_M)^\top \in \mathbb{R}^M$. The dependence on the vector w is indicated by the notation

  $$u_M(t) = u_M(t; w) = u_M(t; w_1, \ldots, w_M). \tag{6.1.3}$$

  Furthermore, we may identify u_M and w.

- *State discretization:*
  The DAE is discretized by a suitable discretization scheme, e.g. a one-step method (5.1.3) or a multi-step method (5.2.3).

- *Constraint discretization:*
  The control/state constraints are only evaluated on the grid $\mathbb{G}_N$.

- *Optimizer:*
  The resulting discretized optimal control problem has to be solved numerically, e.g. by a SQP method, cf. Section 3.8.2.

- *Calculation of gradients:*
  Most optimizers need the gradient of the objective function and the Jacobian of the constraints. There are basically three different methods for calculation: Approximation by finite differences, sensitivity equation approach, and adjoint equation approach.

In the sequel we illustrate the direct discretization method for a generic one-step method (5.1.3) which is supposed to be appropriate for the DAE (6.1). For a given (consistent) initial value x_0 and a given control approximation u_M as in (6.1.2) and (6.1.3) the one-step method generates values

$$\begin{aligned} x_N(t_0) &= x_0, \\ x_N(t_{j+1}) &= x_N(t_j) + h_j\Phi(t_j, x_N(t_j), w, h_j), \qquad j = 0, 1, \ldots, N-1. \end{aligned}$$

Notice, that Φ in general depends in a nonlinear way on the control parametrization w. We distinguish two approaches for the discretization of Problem 6.1: the *full discretization approach* and the *reduced discretization approach.* But before discussing them in detail, we have to address the problem of calculating a consistent initial value x_0.

6.1.1 Projection Method for Consistent Initial Values

Since we are dealing with DAEs only consistent initial values x_0 are admissible, cf. Definition 1.7. There are basically two approaches to tackle the computation of consistent initial values, cf. Gerdts [Ger01a, Ger03a] and Gerdts and Büskens [GB02]. The first approach was suggested by Schulz et al. [SBS98]. This approach relaxes the algebraic constraints of the DAE in such a way that the previously inconsistent value becomes consistent for the relaxed problem. For this approach it is necessary to introduce additional equality constraints to the optimization problem.

The second approach to be discussed in more detail projects inconsistent values onto consistent ones. In order to define a suitable projection we assume that the DAE (6.1) is a Hessenberg DAE of order k as in (1.14) in Definition 1.5, i.e. F is given by

$$\begin{array}{lcllllllll} \dot{x}_1(t) & = & f_1(t, & x_0(t), & x_1(t), & x_2(t), & \ldots, & x_{k-2}(t), & x_{k-1}(t), & u(t)), \\ \dot{x}_2(t) & = & f_2(t, & & x_1(t), & x_2(t), & \ldots, & x_{k-2}(t), & x_{k-1}(t), & u(t)), \\ & \vdots & & & & \ddots & & & & \\ \dot{x}_{k-1}(t) & = & f_{k-1}(t, & & & & & x_{k-2}(t), & x_{k-1}(t), & u(t)), \\ 0_{n_y} & = & g(t, & & & & & & x_{k-1}(t), & u(t)). \end{array}$$

Recall, that the component $y = x_0$ is the algebraic variable and $x^d = (x_1, \ldots, x_{k-1})^\top \in \mathbb{R}^{n_d}$ is the differential variable. According to Definition 1.7 a consistent initial value $x =$

$(x_0, x_1, \ldots, x_{k-1})^\top \in \mathbb{R}^{n_x}$ for the Hessenberg DAE has to satisfy not only the algebraic constraint but also the hidden constraints, i.e. x for $j = 0, 1, \ldots, k-2$ has to satisfy

$$G_j(x^d, w) := g^{(j)}\left(t_0, x_{k-1-j}, \ldots, x_{k-1}, u_M(t_0; w), \dot{u}_M(t_0; w), \ldots, u_M^{(j)}(t_0; w)\right) = 0_{n_y} \qquad (6.1.4)$$

and

$$G_{k-1}(y, x^d, w) := g^{(j)}\left(t_0, y, x^d, u_M(t_0; w), \dot{u}_M(t_0; w), \ldots, u_M^{(j)}(t_0; w)\right) = 0_{n_y} \qquad (6.1.5)$$

Notice, that only G_{k-1} depends on the algebraic variable y while G_j, $j = 0, \ldots, k-2$ do not depend on y. This suggests to split the computation of a consistent initial value into two parts. In the first part only the differential component x^d is computed consistently with the constraints G_j, $j = 0, \ldots, k-2$. Afterwards, the corresponding algebraic variable consistent with the equation G_{k-1} is computed.

For a given control parametrization w and a given estimate x_0^d of a consistent differential component (later on this will be an iterate of a SQP method) the closest consistent x^d is given by the nonlinear least-squares projection

$$X_0^d(x_0^d, w) := \arg\min_{z \in \mathbb{R}^{n_d}} \{\|z - x_0^d\|^2 \mid G_j(z, w) = 0_{n_y},\ j = 0, \ldots, k-2\}. \qquad (6.1.6)$$

This is a parametric nonlinear optimization problem with parameters x_0^d and w as in Section 3.7. Under the assumptions of Theorem 3.7.3 the derivatives $(X_0^d)'_x(x_0^d, w)$ and $(X_0^d)'_w(x_0^d, w)$ result from a sensitivity analysis of (6.1.6) w.r.t. x_0^d and w and are given by the linear equation (3.7.1). This completes the first part of the computation of a consistent initial value.

Now we focus on the equation (6.1.5). Thus, having computed $X_0^d(x_0^d, w)$ the corresponding consistent component y is given implicitly by

$$G_{k-1}(X_0^d(x_0^d, w), y, w) = 0_{n_y}.$$

Recall, that the matrix

$$\frac{\partial}{\partial y} G_{k-1}(y, x^d, w) = g'_{x_{k-1}}(\cdot) \cdot f'_{k-1, x_{k-2}}(\cdot) \cdots f'_{2, x_1}(\cdot) \cdot f'_{1, x_0}(\cdot),$$

cf. (1.15), is supposed to be non-singular since the DAE is Hessenberg of order k. By the implicit function theorem there exists a function $Y : \mathbb{R}^{n_d} \times \mathbb{R}^M \to \mathbb{R}^{n_y}$ with

$$G_{k-1}(Y(x^d, w), x^d, w) = 0_{n_y}$$

for all (x^d, w) in some neighborhood of $(X_0^d(x_0^d, w), w)$. Differentiation w.r.t. x^d and w and exploitation of the nonsingularity of $G'_{k-1,y}$ yields

$$\begin{aligned} Y'_x(x^d, w) &= -\left(G'_{k-1,y}(y, x^d, w)\right)^{-1} G'_{k-1,x}(y, x^d, w), \\ Y'_w(x^d, w) &= -\left(G'_{k-1,y}(y, x^d, w)\right)^{-1} G'_{k-1,w}(y, x^d, w) \end{aligned}$$

at $y = Y(x^d, w)$. Differentiation of $Y_0(x_0^d, w) := Y(X_0^d(x_0^d, w), w)$ w.r.t. x_0^d and w yields

$$\begin{aligned} Y'_{0,x}(x_0^d, w) &= Y'_x(X_0^d(x_0^d, w), w) \cdot (X_0^d)'_x(x_0^d, w), \\ Y'_{0,w}(x_0^d, w) &= Y'_x(X_0^d(x_0^d, w), w) \cdot (X_0^d)'_w(x_0^d, w) + Y'_w(X_0^d(x_0^d, w), w). \end{aligned}$$

Let us summarize the results. For Hessenberg DAEs a method for the computation of the consistent initial value

$$x(t_0) = X_0(x_0^d, w) := \begin{pmatrix} Y_0(x_0^d, w) \\ X_0^d(x_0^d, w) \end{pmatrix}$$

was developed. The derivatives $X'_{0,x}$, $X'_{0,w}$, $Y'_{0,x}$, $Y'_{0,w}$ result from a sensitivity analysis. Concerning general DAEs (6.1) the projection method will work whenever it is possible to identify the algebraic constraints and the hidden constraints. Attempts towards more general DAEs can be found in Leimkuhler et al. [LPG91] and Büskens and Gerdts [BG05]. From now on it is assumed that there exists a method to compute a consistent initial value $X_0(x_0, w)$ for the general DAE (6.1) for given initial guess x_0 and control parametrization w. The function X_0 is assumed to be at least continuously differentiable.

6.1.2 Full Discretization Approach

We obtain a discretization of the optimal control problem by replacing the DAE by the one-step method and discretizing the constraints on the grid $\mathbb{G}_N$:

Problem 6.1.1 (Full Discretization)
Find a grid function $x_N : \mathbb{G}_N \to \mathbb{R}^{n_x}$, $t_i \mapsto x_N(t_i) =: x_i$ and $w \in \mathbb{R}^M$ such that the objective function

$$\varphi(x_0, x_N)$$

is minimized subject to

$$\begin{array}{rcll} X_0(x_0, w) - x_0 & = & 0_{n_x}, & \\ x_j + h_j \Phi(t_j, x_j, w, h_j) - x_{j+1} & = & 0_{n_x}, & j = 0, 1, \ldots, N-1, \\ \psi(x_0, x_N) & = & 0_{n_\psi}, & \\ c(t_j, x_j, u_M(t_j; w)) & \leq & 0_{n_c}, & j = 0, 1, \ldots, N, \\ s(t_j, x_j) & \leq & 0_{n_s}, & j = 0, 1, \ldots, N. \end{array}$$

Problem 6.1.1 is a finite dimensional optimization problem of type

$$\min_z \quad F(z) \quad \text{s.t.} \quad G(z) \leq \Theta, \quad H(z) = \Theta \tag{6.1.7}$$

with the $n_x(N+1) + M$ dimensional optimization variable

$$z := (x_0, x_1, \ldots, x_N, w)^\top,$$

the objective function

$$F(z) := \varphi(x_0, x_N),$$

the $(n_c + n_s) \cdot (N+1)$ inequality constraints

$$G(z) := \begin{pmatrix} c(t_0, x_0, u_M(t_0; w)) \\ \vdots \\ c(t_N, x_N, u_M(t_N; w)) \\ s(t_0, x_0) \\ \vdots \\ s(t_N, x_N) \end{pmatrix},$$

and the $n_x \cdot (N+1) + n_\psi$ equality constraints

$$H(z) := \begin{pmatrix} X_0(x_0, w) - x_0 \\ x_0 + h_0 \Phi(t_0, x_0, w, h_0) - x_1 \\ \vdots \\ x_{N-1} + h_{N-1}\Phi(t_{N-1}, x_{N-1}, w, h_{N-1}) - x_N \\ \psi(x_0, x_N) \end{pmatrix}.$$

Structure of the Optimization Problem:
The size of Problem (6.1.7) depends on n_x, N, and M and can become very large. In practice dimensions up to a million of optimization variables or even more are not unrealistic. In order to handle these dimensions it is essential to exploit the sparse structure of (6.1.7), cf. Betts and Huffman [BH92, BH99]. On the other hand, an advantage of the full discretization approach is that it is easy to compute derivatives w.r.t. the optimization variables. It holds:

$$F'(z) = \left(\varphi'_{x_0} \,\middle|\, \Theta \,\middle|\, \cdots \,\middle|\, \Theta \,\middle|\, \varphi'_{x_f} \,\middle|\, \Theta \,\middle|\, \cdots \,\middle|\, \Theta \right), \tag{6.1.8}$$

$$G'(z) = \left(\begin{array}{ccc|ccc} c'_x[t_0] & & & c'_u[t_0] \cdot u'_{M,w}(t_0; w) & & \\ & \ddots & & & \ddots & \\ & & c'_x[t_N] & & & c'_u[t_N] \cdot u'_{M,w}(t_N; w) \\ \hline s'_x[t_0] & & & & & \\ & \ddots & & & \Theta & \\ & & s'_x[t_N] & & & \end{array} \right), \tag{6.1.9}$$

$$H'(z) = \left(\begin{array}{cccc|c} X'_{0,x_0} - I_{n_x} & & & & X'_{0,w} \\ M_0 & -I_{n_x} & & & h_0 \Phi'_w[t_0] \\ & \ddots & & & \vdots \\ & & M_{N-1} & -I_{n_x} & h_{N-1}\Phi'_w[t_{N-1}] \\ \hline \psi'_{x_0} & & & \psi'_{x_f} & \Theta \end{array} \right) \tag{6.1.10}$$

where $M_j := I_{n_x} + h_j \Phi'_x(t_j, x_j, w, h_j)$, $j = 0, \ldots, N-1$.

6.1.3 Reduced Discretization Approach

The reduced discretization approach is also based on the full discretization 6.1.1. But in the reduced approach the equations resulting from the one-step method are not explicitly taken as equality constraints in the discretized optimization problem. It is clear that the variable x_{i+1} is completely defined by (t_i, x_i, w). Solving the equations recursively we find:

$$\begin{array}{rcl} \tilde{x}_0 & = & X_0(x_0,w), \\ x_1 & = & \tilde{x}_0 + h_0\Phi(t_0,\tilde{x}_0,w,h_0) \\ & = & X_0(x_0,w) + h_0\Phi(t_0,X_0(x_0,w),w,h_0) \\ & =: & X_1(x_0,w), \\ x_2 & = & x_1 + h_1\Phi(t_1,x_1,w,h_1) \\ & = & X_1(x_0,w) + h_1\Phi(t_1,X_1(x_0,w),w,h_1) \\ & =: & X_2(x_0,w), \\ & \vdots & \\ x_N & = & x_{N-1} + h_{N-1}\Phi(t_{N-1},x_{N-1},w,h_{N-1}) \\ & = & X_{N-1}(x_0,w) + h_{N-1}\Phi(t_{N-1},X_{N-1}(x_0,w),w,h_{N-1}) \\ & =: & X_{N-1}(x_0,w). \end{array} \tag{6.1.11}$$

Of course, this is just the formal procedure of solving the DAE for given consistent initial value and control parametrization w. We obtain:

Problem 6.1.2 (Reduced Discretization)
Find $x_0 \in \mathbb{R}^{n_x}$ and $w \in \mathbb{R}^M$ such that the objective function

$$\varphi(X_0(x_0,w), X_N(x_0,w))$$

is minimized subject to

$$\begin{array}{rcll} \psi(X_0(x_0,w),X_N(x_0,w)) & = & 0_{n_\psi}, & \\ c(t_j,X_j(x_0,w),u_M(t_j;w)) & \leq & 0_{n_c}, & j=0,1,\ldots,N, \\ s(t_j,X_j(x_0,w)) & \leq & 0_{n_s}, & j=0,1,\ldots,N. \end{array}$$

Again, Problem 6.1.2 is a finite dimensional nonlinear optimization problem of type (6.1.7) with the $n_x + M$ optimization variables $z := (x_0, w)^\top$, the objective function

$$F(z) := \varphi(X_0(x_0,w), X_N(x_0,w)), \tag{6.1.12}$$

the $(n_c + n_s) \cdot (N+1)$ inequality constraints

$$G(z) := \begin{pmatrix} c(t_0,X_0(x_0,w),u_M(t_0;w)) \\ \vdots \\ c(t_N,X_N(x_0,w),u_M(t_N;w)) \\ s(t_0,X_0(x_0,w)) \\ \vdots \\ s(t_N,X_N(x_0,w)) \end{pmatrix}, \tag{6.1.13}$$

and the n_ψ equality constraints

$$H(z) := \psi(X_0(x_0,w), X_N(x_0,w)). \tag{6.1.14}$$

Structure of the Optimization Problem:
The size of Problem 6.1.2 is small compared to Problem 6.1.1 since we eliminated most of the

equality constraints as well as the variables $x_2, \ldots, x_N$. Only the initial value x_0 remains as an optimization variable. Unfortunately the derivatives are not sparse anymore and it is more involved to compute them:

$$
\begin{aligned}
G'(z) &= \left(\begin{array}{c|c}
c_x'[t_0] \cdot X_{0,x_0}' & c_x'[t_0] \cdot X_{0,w}' + c_u'[t_0] \cdot u_{M,w}'(t_0; w) \\
c_x'[t_1] \cdot X_{1,x_0}' & c_x'[t_1] \cdot X_{1,w}' + c_u'[t_1] \cdot u_{M,w}'(t_1; w) \\
\vdots & \vdots \\
c_x'[t_N] \cdot X_{N,x_0}' & c_x'[t_N] \cdot X_{N,w}' + c_u'[t_N] \cdot u_{M,w}'(t_N; w) \\
\hline
s_x'[t_0] \cdot X_{0,x_0}' & s_x'[t_0] \cdot X_{0,w}' \\
s_x'[t_1] \cdot X_{1,x_0}' & s_x'[t_1] \cdot X_{1,w}' \\
\vdots & \vdots \\
s_x'[t_N] \cdot X_{N,x_0}' & s_x'[t_N] \cdot X_{N,w}'
\end{array}\right), \qquad (6.1.15) \\
H'(z) &= \left(\psi_{x_0}' \cdot X_{0,x_0}' + \psi_{x_f}' \cdot X_{N,x_0}' \;\middle|\; \psi_{x_0}' \cdot X_{0,w}' + \psi_{x_f}' \cdot X_{N,w}' \right). \qquad (6.1.16)
\end{aligned}
$$

Herein, the sensitivities

$$
X_{i,x_0}'(x_0, w), \quad X_{i,w}'(x_0, w), \quad i = 0, \ldots, N \qquad (6.1.17)
$$

have to be computed. This problem will be addressed later on in detail.

Remark 6.1.3 *Based on the same ideas it is possible and often necessary to extend the reduced discretization method to a reduced multiple shooting method by introducing multiple shooting nodes. Details can be found in Gerdts [Ger01a, Ger03a]. Related methods based on the multiple shooting idea are discussed in, e.g., Deuflhard [Deu74], Bock [BP84], Stoer and Bulirsch [SB90], Leineweber [Lei95], Schulz et al. [SBS98], Hinsberger [Hin97], Oberle and Grimm [OG01].*

6.1.4 Control Discretization

We address the problem of discretizing the control and specify the finite dimensional subspace $L_M^\infty([t_0, t_f], \mathbb{R}^{n_u}) \subset L^\infty([t_0, t_f], \mathbb{R}^{n_u})$. We intend to replace the control variable $u(\cdot) \in L^\infty([t_0, t_f], \mathbb{R}^{n_u})$ by a B-spline representation. Given $k \in \mathbb{N}$ the elementary B-splines $B_i^k(\cdot)$ of order k, $i = 1, \ldots, N + k - 1$, are defined by the recursion

$$
\begin{aligned}
B_i^1(t) &:= \begin{cases} 1, & \text{if } \tau_i \le t < \tau_{i+1} \\ 0, & \text{otherwise} \end{cases}, \\
B_i^k(t) &:= \frac{t - \tau_i}{\tau_{i+k-1} - \tau_i} B_i^{k-1}(t) + \frac{\tau_{i+k} - t}{\tau_{i+k} - \tau_{i+1}} B_{i+1}^{k-1}(t)
\end{aligned} \qquad (6.1.18)
$$

on the auxiliary grid

$$
\mathbb{G}_N^k := \{\tau_i \mid i = 1, \ldots, N + 2k - 1\} \qquad (6.1.19)
$$

with auxiliary grid points

$$
\tau_i := \begin{cases} t_0, & \text{if } 1 \le i \le k, \\ t_{i-k}, & \text{if } k + 1 \le i \le N + k - 1, \\ t_N, & \text{if } N + k \le i \le N + 2k - 1. \end{cases}
$$

Notice, that the convention $0/0 = 0$ is used if auxiliary grid points coincide in the recursion (6.1.18). The evaluation of the recursion (6.1.18) is well-conditioned, cf. Deuflhard and Hohmann [DH91]. The elementary B-splines $B_i^k(\cdot)$, $i = 1, \ldots, N+k-1$ restricted to the intervals $[t_j, t_{j+1}]$, $j = 0, \ldots, N-1$ are polynomials of degree at most $k-1$, i.e.

$$B_i^k(\cdot)\big|_{[t_j,t_{j+1}]} \in \mathcal{P}^{k-1}([t_j, t_{j+1}]),$$

and they form a basis of the space of splines

$$\left\{ s(\cdot) \in C^{k-2}([t_0, t_f], \mathbb{R}) \;\middle|\; s(\cdot)\big|_{[t_j,t_{j+1}]} \in \mathcal{P}^{k-1}([t_j, t_{j+1}]) \text{ for } j = 0, \ldots, N-1 \right\}.$$

Obviously, the elementary B-splines are essentially bounded, i.e. $B_i^k(\cdot) \in L^\infty([t_0, t_f], \mathbb{R})$. For $k \geq 2$ it holds $B_i^k(\cdot) \in C^{k-2}([t_0, t_f], \mathbb{R})$ and for $k \geq 3$ the derivative obeys the recursion

$$\frac{d}{dt} B_i^k(t) = \frac{k-1}{\tau_{i+k-1} - \tau_i} B_i^{k-1}(t) - \frac{k-1}{\tau_{i+k} - \tau_{i+1}} B_{i+1}^{k-1}(t).$$

Furthermore, the elementary B-splines possess local support $\operatorname{supp}(B_i^k) \subset [\tau_i, \tau_{i+k}]$. More precisely, for $k > 1$ it holds

$$B_i^1(t) \begin{cases} > 0, & \text{if } t \in (\tau_i, \tau_{i+k}), \\ = 0, & \text{otherwise.} \end{cases}$$

The cases $k = 1$ or $k = 2$ frequently occur in numerical computations. For $k = 1$ the elementary B-splines are piecewise constant functions, while for $k = 2$ we obtain the continuous and piecewise linear functions

$$B_i^2(t) = \begin{cases} \dfrac{t - \tau_i}{\tau_{i+1} - \tau_i}, & \text{if } \tau_i \leq t < \tau_{i+1}, \\ \dfrac{\tau_{i+2} - t}{\tau_{i+2} - \tau_{i+1}}, & \text{if } \tau_{i+1} \leq t < \tau_{i+2}, \\ 0, & \text{otherwise.} \end{cases}$$

Similarly, in some situations it might be necessary to have a continuously differentiable function as, e.g.,

$$B_i^3(t) = \begin{cases} \dfrac{(t - \tau_i)^2}{(\tau_{i+2} - \tau_i)(\tau_{i+1} - \tau_i)}, & \text{if } t \in [\tau_i, \tau_{i+1}), \\ \dfrac{(t - \tau_i)(\tau_{i+2} - t)}{(\tau_{i+2} - \tau_i)(\tau_{i+2} - \tau_{i+1})} + \dfrac{(\tau_{i+3} - t)(t - \tau_{i+1})}{(\tau_{i+3} - \tau_{i+1})(\tau_{i+2} - \tau_{i+1})}, & \text{if } t \in [\tau_{i+1}, \tau_{i+2}), \\ \dfrac{(\tau_{i+3} - t)^2}{(\tau_{i+3} - \tau_{i+1})(\tau_{i+3} - \tau_{i+2})}, & \text{if } t \in [\tau_{i+2}, \tau_{i+3}), \\ 0, & \text{otherwise.} \end{cases}$$

Figure 6.1 visualizes B-splines of orders $k = 2, 3, 4$.

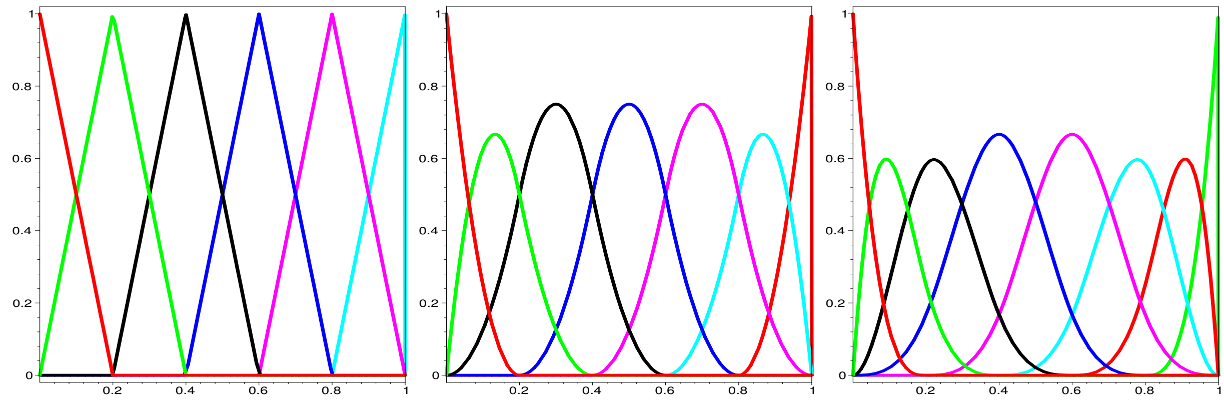


Figure 6.1: B-splines of order $k = 2$ (left), $k = 3$ (middle), and $k = 4$ (right) with $[t_0, t_f] = [0, 1]$ and $N = 5$ on an equidistant grid.

In view of the optimal control problem, we consider discretized controls

$$u_M(\cdot) \in L^\infty_M([t_0,t_f],\mathbb{R}^{n_u}) := \left\{ \sum_{i=1}^{N+k-1} w_i\, B_i^k(\cdot) \;\middle|\; w_i \in \mathbb{R}^{n_u},\ i = 1,\ldots,N+k-1 \right\}. \qquad (6.1.20)$$

Notice, that each $u_M(\cdot) \in L^\infty_M([t_0,t_f],\mathbb{R}^{n_u})$ is determined by the $M := n_u(N+k-1)$ control parameters $w := (w_1,\ldots,w_{N+k+1}) \in \mathbb{R}^{n_u(N+k-1)}$ and hence $L^\infty_M([t_0,t_f],\mathbb{R}^{n_u})$ is a finite dimensional subspace of $L^\infty([t_0,t_f],\mathbb{R}^{n_u})$. The dependence on the vector w is indicated by the notation

$$u_M(t) = u_M(t;w) = u_M(t;w_1,\ldots,w_{N+k-1}).$$

The coefficients w_i are known as *de Boor points*. In the special cases $k = 1$ and $k = 2$ the de Boor points w_i can be interpreted as function values $w_i = u_N(t_i)$. As already mentioned, most often the choices $k = 1$ (piecewise constant approximation) or $k = 2$ (continuous and piecewise linear approximation) are preferred by many authors, cf., e.g., von Stryk [vS94], Büskens [Büs98], and Büskens and Gerdts [BG00].

The choice of B-splines has two major advantages from the numerical point of view. First, it is easy to create approximations with prescribed smoothness properties. Second, the de Boor point w_i influences the function value $u_M(t)$ only in the interval $t \in [\tau_i, \tau_{i+k}]$ due to the local support of B_i^k. This porperty will lead to certain sparsity patterns in the gradient of the objective function and the Jacobian of the constraints. The exploitation of this sparsity reduces the computational effort for the numerical solution considerably, cf. Gerdts [Ger01a, Ger03a].

If instead an interpolating cubic spline is used as control approximation as in Kraft [KE85] the parametrization of the approximate control influences the function values in the whole time interval $[t_0, t_f]$ and the local influence of w_i is lost. Besides the missing sparsity this may lead to numerical instabilities as well. Barclay et al. [BGR97] employ piecewise defined Hermite polynomials as approximations.

Remark 6.1.4 *Of course, it is possible to use individual B-splines of different order for each component of the control vector u_M. For the sake of simplicity, we will not discuss these control approximations and restrict the discussion to the case where B-splines of the same order are used for every component of the control vector.*

The following example shows that the choice of a 'good' control approximation may lead into pitfalls. Suppose we would like to use the modified Euler's scheme with Butcher array

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1/2 & 1/2 & 0 \\ \hline & 0 & 1 \end{array}$$

for the discretization of an ODE. This method requires function evaluations at the time points t_i and $t_i + h/2$. In particular, the method needs the control values $u_i := u_M(t_i)$ and $u_{i+1/2} := u_M(t_i + h/2)$. Instead of choosing a piecewise constant control approximation on $\mathbb{G}_N$ we might have the idea of treating u_i and $u_{i+1/2}$ as independent optimization variables (this would correspond to a piecewise constant control approximation on the grid given by the grid points t_i and $t_{i+1/2}$). The following example shows that this strategy may fail.

Example 6.1.5 (Hager [Hag00], p. 272) *Consider the optimal control problem*

$$\min \quad \frac{1}{2}\int_0^1 u(t)^2 + 2x(t)^2 dt \quad s.t. \quad \dot{x}(t) = \frac{1}{2}x(t) + u(t),\ x(0) = 1.$$

The optimal solution is

$$\hat{x}(t) = \frac{2\exp(3t) + \exp(3)}{\exp(3t/2)(2 + \exp(3))}, \quad \hat{u}(t) = \frac{2(\exp(3t) - \exp(3))}{\exp(3t/2)(2 + \exp(3))}.$$

We consider the modified Euler's method with Butcher array

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1/2 & 1/2 & 0 \\ \hline & 0 & 1 \end{array}$$

and the method of Heun defined by

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline & 1/2 & 1/2 \end{array}$$

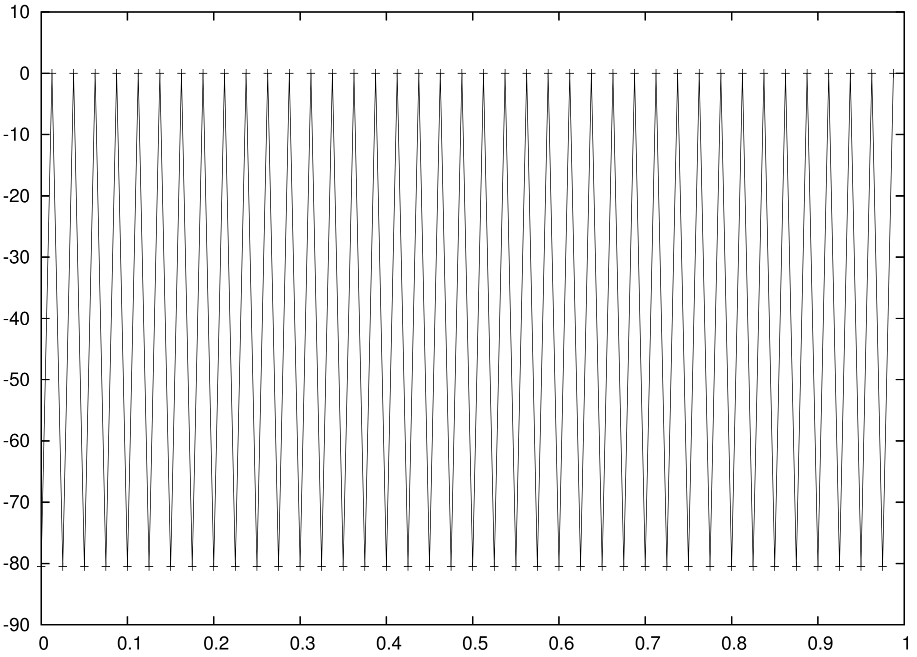
The table shows the error in x in the norm $\|\cdot\|_\infty$ for Heun's method:

N	*Error in x*	*Order*
10	*0.2960507253983891E-02*	–
20	*0.7225108094129906E-03*	*2.0347533*
40	*0.1783364646560370E-03*	*2.0184174*
80	*0.4342336372986644E-04*	*2.0380583*
160	*0.9861920395981549E-05*	*2.138531*
320	*0.2417855093361787E-05*	*2.0281408*

Heun's method converges of second order.
The table shows the error in x in the norm $\|\cdot\|_\infty$ for modified Euler's method. The control is discretized at t_i and $t_i + h/2$ with independent optimization variables u_i and $u_{i+1/2}$:

N	*Error in x*	*Order*
10	*0.4254224673693650E+00*	–
20	*0.4258159920666613E+00*	*-0.0013339*
40	*0.4260329453139864E+00*	*-0.0007349*
80	*0.4260267362368171E+00*	*0.0000210*
160	*0.4261445411996390E+00*	*-0.0003989*
320	*0.4260148465889140E+00*	*0.0004391*

There is no convergence at all. The numerical solution for $N = 40$ shows a control u oscillating strongly between 0 at $t_i + h/2$ and approximately $-1/(2h)$ at t_i:



Now we use a piecewise constant control approximation in the modified Euler's method and obtain the following error in x in the norm $\|\cdot\|_\infty$:

N	*Error in x*	*Order*
10	*0.33588007819529242E-03*	–
20	*0.89303965135845l5E-04*	*1.9111501*
40	*0.22738228194651 99E-04*	*1.9736044*
80	*0.53665001290559 29E-05*	*2.0830664*
160	*0.12507296422992 20E-05*	*2.1012115*
320	*0.78847792728264 92E-06*	*0.6656277*

For this control approximation the modified Euler's method converges of second order!

6.2 Calculation of Gradients

The applicability of the SQP method for solving the discretized optimal control problems 6.1.1 resp. 6.1.2 is straightforward. Nevertheless, the SQP method needs the gradient F' of the objective function and the Jacobians G' and H' of the constraints. In case of the full discretization approach these quantities are easily computed according to (6.1.8)-(6.1.10). In case of the reduced discretization approach it is more involved to compute them. We will see that this can be achieved in different ways:

- The *sensitivity differential equation approach* is advantageous if the number of constraints is (much) larger than the number of variables in the discretized problem.

- The *adjoint equation approach* is preferable if the number of constraints is less than the number of variables in the discretized problem.

- A very powerful and user-friendly tool for the evaluation of gradients is *algorithmic differentiation.* This approach assumes that the evaluation of a function is performed by a Fortran or C procedure (or any other supported programming language). Algorithmic differentiation means that the complete procedure is differentiated step by step using roughly speaking chain and product rules. The result is again a Fortran or C procedure that provides the gradient of the function. Essentially the so-called 'forward mode' in algorithmic differentiation corresponds to the sensitivity equation approach while the 'backward mode' corresponds to the adjoint approach. Details can be found on the web page `www.autodiff.org` and in the book of Griewank [Gri00] and in the review article of Griewank [Gri03].

- The approximation by *finite differences* is straightforward but has the drawback of being computationally expensive and often suffers from low accuracy. Nevertheless, this approach is often used if some solver depending on optimization variables is used as a black box inside an algorithm.

We only concentrate on the first two approaches for calculating derivatives for the reduced discretization approach.

6.2.1 Sensitivity Equation Approach

We restrict the discussion to the one-step method

$$X_{i+1}(z) \;=\; X_i(z) + h_i\Phi(t_i, X_i(z), w, h_i), \qquad i = 0, 1, \ldots, N-1. \tag{6.2.1}$$

Recall that $z = (x_0, w)^\top$ denotes the variables in the optimization problem 6.1.2. We intend to compute the sensitivities

$$S_i := \frac{\partial X_i(z)}{\partial z}, \qquad i = 0, 1, \ldots, N.$$

For $i = 0$ it holds

$$S_0 = \frac{\partial X_0(z)}{\partial z} \in \mathbb{R}^{n_x \times (n_x+M)}, \tag{6.2.2}$$

where $X_0(z)$ is a continuously differentiable function that provides a consistent initial value. Differentiaton of (6.2.1) w.r.t. z yields the relationship

$$S_{i+1} = S_i + h_i\left(\Phi'_x(t_i, X_i(z), w, h_i)\cdot S_i + \Phi'_w(t_i, X_i(z), w, h_i)\cdot\frac{\partial w}{\partial z}\right) \tag{6.2.3}$$

for $i = 0, 1, \ldots, N-1$. This approach is known as internal numerical differentiation (IND), cf. Bock [Boc87]. The IND-approach is based on the differentiation of the discretization scheme (e.g. the one-step method) w.r.t. z. Computing the derivatives for the reduced problem 6.1.2 essentially amounts to solving one initial value problem of size $n_x \cdot (1 + n_x + M)$. It is worth pointing out that the size depends on the number of unknowns in the programming problem, but it does not depend on the number of constraints.

In practice, the computation of the derivatives Φ'_x and Φ'_w is non-trivial. We illustrate the difficulties for the implicit Euler's method and consider the integration step $t_i \to t_{i+1} = t_i + h_i$. Discretization of (6.1) with $X_i \approx x(t_i), X_{i+1} \approx x(t_{i+1})$ leads to the nonlinear equation

$$F\left(t_{i+1}, X_{i+1}, \frac{X_{i+1} - X_i}{h_i}, u_M(t_{i+1}; w)\right) = 0_{n_x}. \tag{6.2.4}$$

Provided that this nonlinear equation possesses a solution and that we may apply the implicit function theorem, this solution will depend on z, i.e. $X_{i+1} = X_{i+1}(z)$. Now, we need the derivative of the numerical solution $X_{i+1}(z)$ w.r.t. to z, i.e. S_{i+1}. According to the implicit function theorem, differentiation of (6.2.4) w.r.t. z yields the linear equation

$$\left(F'_x + \frac{1}{h_i} F'_{\dot{x}}\right)\Big|_{X_{i+1}} \cdot S_{i+1} = -F'_u\Big|_{X_{i+1}} \cdot u'_{M,z}(t_{i+1}; w) + \frac{1}{h_i} F'_{\dot{x}}\Big|_{X_{i+1}} \cdot S_i. \qquad (6.2.5)$$

This formula can be obtained in a different way. Let $x(t;z)$ denote the solution of the DAE (6.1) for given z. If F is sufficiently smooth and x is continuously differentiable w.r.t. t and z and it holds $\frac{\partial}{\partial z}\frac{dx(t;z)}{dt} = \frac{d}{dt}\frac{\partial x(t;z)}{\partial z}$, differentiation of (6.1) w.r.t. z results in a linear matrix DAE – the *sensitivity DAE*

$$F'_x[t] \cdot S(t) + F'_{\dot{x}}[t] \cdot \dot{S}(t) + F'_u[t] \cdot u'_{M,z}(t; w) = 0_{n_x \times n_x}$$

for the *sensitivity matrix* $S(t) := \partial x(t;z)/\partial z$. Discretization of the sensitivity DAE with the same method as in (6.2.4) and for the same time step leads again to the linear equation (6.2.5). Hence, both approaches coincide, provided that X_{i+1} solves (6.2.4) exactly. In practice, this is not the case and (6.2.4) is solved numerically by Newton's method

$$\left(F'_x + \frac{1}{h_i} F'_{\dot{x}}\right)\Big|_{X^{(k)}_{i+1}} \cdot \Delta X^{(k)}_{i+1} = -F\Big|_{X^{(k)}_{i+1}}, \qquad X^{(k+1)}_{i+1} = X^{(k)}_{i+1} + \Delta X^{(k)}_{i+1}, \qquad k = 0, 1, 2, \ldots.$$

Notice, that the iteration matrix $F'_x + \frac{1}{h_i} F'_{\dot{x}}$ at the last iterate can be re-used for computing S_{i+1}. In practice, the iteration matrix within Newton's method is kept constant for several Newton steps to speed up the procedure. But it is important to mention that the iteration matrix has to be re-evaluated prior to the calculation of the sensitivity matrix S_{i+1}. The common strategy of keeping the iteration matrix constant even for some integration steps leads to poor results concerning the accuracy of the sensitivity matrix. Of course, if the iteration matrix is kept constant or if only a fixed number of Newton steps are performed then the resulting S_{i+1} is only an approximation of the correct derivative $\partial X_{i+1}(z)/\partial z$.

Remark 6.2.1

- *The IND approach is applicable for multistep method such as BDF as well.*
- *Similar strategies are discussed in Caracotsios and Stewart [CS85], Maly and Petzold [MP96], Brenan et al. [BCP96], Heim [Hei92], and Kiehl [Kie98]. A comparison of different strategies can be found in Feehery et al. [FTB97].*

6.2.2 Adjoint Equation Approach: The Discrete Case

The adjoint method avoids the calculation of the sensitivities S_i. We demonstrate the method for a prototype function of type

$$\Gamma(z) := \gamma(X_0(z), X_N(z), z).$$

Obviously, φ, ψ and essentially c and s are of this type.
We intend to derive a procedure for calculating $\Gamma'(z)$ subject to the difference equations

$$X_{i+1}(z) - X_i(z) - h_i \Phi(t_i, X_i(z), w, h_i) = 0_{n_x}, \quad i = 0, \ldots, N-1. \qquad (6.2.6)$$

The initial state $X_0(z) \in \mathbb{R}^{n_x}$ is assumed to be sufficiently smooth as a function of z and consistent with the underlying DAE for all z. We consider the auxiliary functional

$$J(z) := \Gamma(z) + \sum_{i=0}^{N-1} \lambda_{i+1}^\top \left(X_{i+1}(z) - X_i(z) - h_i \Phi(t_i, X_i(z), w, h_i)\right)$$

with multipliers λ_i, $i = 1, \ldots, N$. Differentiating J w.r.t. z yields the expression

$$\begin{aligned}
J'(z) &= \gamma'_{x_0} \cdot S_0 + \gamma'_{x_N} \cdot S_N + \gamma'_z \\
&\quad + \sum_{i=0}^{N-1} \lambda_{i+1}^\top \left(S_{i+1} - S_i - h_i \Phi'_x[t_i] \cdot S_i - h_i \Phi'_w[t_i] \cdot \frac{\partial w}{\partial z} \right) \\
&= \gamma'_{x_0} \cdot S_0 + \gamma'_{x_N} \cdot S_N + \gamma'_z \\
&\quad + \sum_{i=1}^{N} \lambda_i^\top S_i - \sum_{i=0}^{N-1} \lambda_{i+1}^\top \left(S_i + h_i \Phi'_x[t_i] \cdot S_i + h_i \Phi'_w[t_i] \cdot \frac{\partial w}{\partial z} \right) \\
&= \left(\gamma'_{x_0} - \lambda_1^\top - h_0 \lambda_1^\top \Phi'_x[t_0] \right) \cdot S_0 + \left(\gamma'_{x_N} + \lambda_N^\top \right) \cdot S_N + \gamma'_z \\
&\quad + \sum_{i=1}^{N-1} \left(\lambda_i^\top - \lambda_{i+1}^\top - h_i \lambda_{i+1}^\top \Phi'_x[t_i] \right) \cdot S_i - \sum_{i=0}^{N-1} h_i \lambda_{i+1}^\top \Phi'_w[t_i] \cdot \frac{\partial w}{\partial z}.
\end{aligned}$$

The terms $S_i = \partial X_i(z)/\partial z$ are just the sensitivities in the sensitivity equation approach which we do not (!) want to compute here. Hence, we have to ensure that the expressions involving S_i are eliminated. This leads to the *adjoint equation*

$$\lambda_i^\top - \lambda_{i+1}^\top - h_i \lambda_{i+1}^\top \Phi'_x[t_i] = 0_{n_x}, \quad i = 0, \ldots, N-1 \tag{6.2.7}$$

and the *transversality condition*

$$\lambda_N^\top = -\gamma'_{x_N}(X_0(z), X_N(z), z). \tag{6.2.8}$$

Notice, that the adjoint equation is solved backwards in time. With these expressions the derivative of J reduces to

$$J'(z) = \left(\gamma'_{x_0} - \lambda_0^\top \right) \cdot S_0 + \gamma'_z - \sum_{i=0}^{N-1} h_i \lambda_{i+1}^\top \Phi'_w[t_i] \cdot \frac{\partial w}{\partial z}. \tag{6.2.9}$$

Herein, the sensitivity matrix S_0 is given by

$$S_0 = X'_0(z). \tag{6.2.10}$$

It remains to show that $J'(z) = \Gamma'(z)$ holds:

Theorem 6.2.2 *It holds*

$$\Gamma'(z) = J'(z) = \left(\gamma'_{x_0} - \lambda_0^\top \right) \cdot S_0 + \gamma'_z - \sum_{i=0}^{N-1} h_i \lambda_{i+1}^\top \Phi'_u[t_i] \cdot \frac{\partial w}{\partial z}. \tag{6.2.11}$$

Proof. By multiplication of the sensitivity equation

$$S_{i+1} - S_i - h_i \Phi'_x[t_i] S_i - h_i \Phi'_w[t_i] \cdot \frac{\partial w}{\partial z} = 0_{n_x}, \quad i = 0, \ldots, N-1$$

with $\lambda_{i+1}^\top$ from the left we obtain

$$-h_i \lambda_{i+1}^\top \Phi'_w[t_i] \cdot \frac{\partial w}{\partial z} = -\lambda_{i+1}^\top S_{i+1} + \lambda_{i+1}^\top S_i + h_i \lambda_{i+1}^\top \Phi'_x[t_i] \cdot S_i, \quad i = 0, \ldots, N-1$$

and hence

$$\begin{aligned}
J'(z) &= \left(\gamma'_{x_0} - \lambda_0^\top\right) \cdot S_0 + \gamma'_z \\
&\quad + \sum_{i=0}^{N-1} \lambda_{i+1}^\top S_i + h_i \lambda_{i+1}^\top \Phi'_x[t_i] S_i - \lambda_{i+1}^\top S_{i+1} \\
&\overset{(6.2.7)}{=} \left(\gamma'_{x_0} - \lambda_0^\top\right) \cdot S_0 + \gamma'_z + \sum_{i=0}^{N-1} \lambda_i^\top S_i - \lambda_{i+1}^\top S_{i+1} \\
&= \left(\gamma'_{x_0} - \lambda_0^\top\right) \cdot S_0 + \gamma'_z + \lambda_0^\top S_0 - \lambda_N^\top S_N \\
&\overset{(6.2.8)}{=} \gamma'_{x_0} S_0 + \gamma'_z + \gamma'_{x_N} S_N \\
&= \Gamma'(z).
\end{aligned}$$

■

With $\Gamma'(z) = J'(z)$ we finally found a formula for the gradient of Γ. Γ itself is a placeholder for the functions $F, G = (G_1, \ldots, G_m), H = (H_1, \ldots, H_p)$ in (6.1.12)-(6.1.14).
In order to compute F', G', H' of (6.1.12)-(6.1.14) for each (!) component of F, G, H an adjoint equation with appropriate transversality condition has to be solved. This essentially corresponds to solving an initial value problem of dimension $n_x \cdot (2 + m + p)$. In addition, the trajectory $(X_i, i = 0, \ldots, N)$ has to be stored. It is important to mention that the effort does not depend on the number M of control parameters! The method is particularly efficient if none or very few constraints are present.

6.2.2.1 Application to Runge-Kutta Methods

The function Φ in the preceding section is specified for Runge-Kutta methods and DAEs of type

$$F(t, x(t), \dot{x}(t), u_M(t; w)) = 0_{n_x}, \qquad x(t_0) = X_0.$$

An implicit Runge-Kutta method with s stages is given by the iteration

$$X_{i+1} = X_i + h\Phi(t_i, X_i, w, h), \quad i = 0, 1, \ldots, N-1,$$

where the increment function Φ is defined by

$$\Phi(t, x, w, h) := \sum_{j=1}^{s} b_j k_j(t, x, w, h) \tag{6.2.12}$$

and the stage derivatives $k = (k_1, \ldots, k_s)$ are implicitly defined by the nonlinear equation

$$G(k, t, x, w, h) := \begin{pmatrix} F\left(t + c_1 h, x + h \sum_{j=1}^{s} a_{1j} k_j, k_1, u_M(t + c_1 h; w)\right) \\ F\left(t + c_2 h, x + h \sum_{j=1}^{s} a_{2j} k_j, k_2, u_M(t + c_2 h; w)\right) \\ \vdots \\ F\left(t + c_s h, x + h \sum_{j=1}^{s} a_{sj} k_j, k_s, u_M(t + c_s h; w)\right) \end{pmatrix} = 0_{s \cdot n_x}. \tag{6.2.13}$$

The nonlinear equation (6.2.13) is solved numerically by Newton's method:

$$\begin{aligned} G_k'(k^{(j)},t,x,w,h)\Delta k^{(j)} &= -G(k^{(j)},t,x,w,h), \\ k^{(j+1)} &= k^{(j)} + \Delta k^{(j)}, \quad j = 0,1,2,\ldots. \end{aligned}$$

Under the assumption that the derivative G_k' is non-singular at a solution k, the implicit function theorem yields the existence of the function $k = k(t,x,w,h)$ satisfying

$$G(k(t,x,w,h),t,x,w,h) = 0_{s\cdot n_x}.$$

By differentiation w.r.t. x and w we find

$$\begin{aligned} k_x'(t,x,w,h) &= -\left(G_k'(k,t,x,w,h)\right)^{-1} G_x'(k,t,x,w,h), \\ k_w'(t,x,w,h) &= -\left(G_k'(k,t,x,w,h)\right)^{-1} G_w'(k,t,x,w,h), \end{aligned}$$

where

$$\begin{aligned} G_k'(k,t,x,w,h) &= \begin{pmatrix} ha_{11}M_1 + T_1 & ha_{12}M_1 & \cdots & ha_{1s}M_1 \\ ha_{21}M_2 & ha_{22}M_2 + T_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & ha_{s-1,s}M_{s-1} \\ ha_{s1}M_s & \cdots & ha_{s,s-1}M_s & ha_{ss}M_s + T_s \end{pmatrix}, \\ M_j &:= F_x'\left(t + c_jh, x + h\sum_{l=1}^{s} a_{jl}k_l, k_j, u_M(t + c_jh; w)\right), \quad j = 1,\ldots,s, \\ T_j &:= F_{\dot{x}}'\left(t + c_jh, x + h\sum_{l=1}^{s} a_{jl}k_l, k_j, u_M(t + c_jh; w)\right), \quad j = 1,\ldots,s, \\ G_x'(k,t,x,w,h) &= \begin{pmatrix} F_x'\left(t + c_1h, x + h\sum_{j=1}^{s} a_{1j}k_j, k_1, u_M(t + c_1h; w)\right) \\ \vdots \\ F_x'\left(t + c_sh, x + h\sum_{j=1}^{s} a_{sj}k_j, k_s, u_M(t + c_sh; w)\right) \end{pmatrix}, \\ G_w'(k,t,x,w,h) &= \begin{pmatrix} F_u'\left(t + c_1h, x + h\sum_{j=1}^{s} a_{1j}k_j, k_1, u_M(t + c_1h; w)\right) \cdot u_{M,w}'(t + c_1; w) \\ \vdots \\ F_u'\left(t + c_sh, x + h\sum_{j=1}^{s} a_{sj}k_j, k_s, u_M(t + c_sh; w)\right) \cdot u_{M,w}'(t + c_s; w) \end{pmatrix}. \end{aligned}$$

Notice, that (6.2.7) and (6.2.9) require the partial derivatives $\Phi_x'(t,x,w,h)$ and $\Phi_w'(t,x,w,h)$. According to (6.2.13) these values are given by

$$\begin{aligned} \Phi_x'(t,x,w,h) &= \sum_{j=1}^{s} b_j k_{j,x}'(t,x,w,h), \\ \Phi_w'(t,x,w,h) &= \sum_{j=1}^{s} b_j k_{j,w}'(t,x,w,h). \end{aligned}$$

In particular, for the implicit Euler's method we have

$$G(k,t,x,w,h)=F\left(t+h,x+hk,k,u_M(t+h;w)\right)$$

and

$$\begin{aligned}\Phi'_x(t,x,w,h) &= k'_x(t,x,w,h)=-\left(hF'_x[t+h]+F'_{\dot{x}}[t+h]\right)^{-1}\cdot F'_x[t+h],\\ \Phi'_w(t,x,w,h) &= k'_w(t,x,w,h)=-\left(hF'_x[t+h]+F'_{\dot{x}}[t+h]\right)^{-1}\cdot F'_u[t+h]\cdot u'_{M,w}(t+h;w).\end{aligned}$$

Example 6.2.3 (Explicit Runge-Kutta Methods)
We briefly discuss the important subclass of ODEs

$$F(t,x(t),\dot{x}(t),u_M(t;w)):=\dot{x}(t)-f(t,x(t),u_M(t;w))$$

in combination with an s-staged explicit Runge-Kutta method

$$X_{i+1}=X_i+h\sum_{j=1}^{s}b_jk_j(t_i,X_i,w,h),\quad i=0,1,\ldots,N-1,$$

where the stage derivatives $k_j=k_j(t,x,w,h)$ are recursively defined by the equation

$$k_j(t,x,w,h):=f\left(t+c_jh,x+h\sum_{\ell=1}^{j-1}a_{j\ell}k_\ell,u_M(t+c_jh;w)\right),\quad j=1,\ldots,s. \tag{6.2.14}$$

By differentiation w.r.t. x and w we find

$$\begin{aligned}k'_{j,x} &= f'_x\left(t+c_jh,x+h\sum_{\ell=1}^{j-1}a_{j\ell}k_\ell,u_M(t+c_jh;w)\right)\left(I+h\sum_{\ell=1}^{j-1}a_{j\ell}k'_{\ell,x}\right),\\ k'_{j,w} &= f'_x\left(t+c_jh,x+h\sum_{\ell=1}^{j-1}a_{j\ell}k_\ell,u_M(t+c_jh;w)\right)\left(h\sum_{\ell=1}^{j-1}a_{j\ell}k'_{\ell,w}\right)\\ &\quad +f'_u\left(t+c_jh,x+h\sum_{\ell=1}^{j-1}a_{j\ell}k_\ell,u_M(t+c_jh;w)\right)\cdot u'_{M,w}(t+c_j;w)\end{aligned}$$

for $j=1,\ldots,s$ and

$$\begin{aligned}\Phi'_x(t,x,w,h) &= \sum_{j=1}^{s}b_jk'_{j,x}(t,x,w,h),\\ \Phi'_w(t,x,w,h) &= \sum_{j=1}^{s}b_jk'_{j,w}(t,x,w,h).\end{aligned}$$

6.2.2.2 Simplecticity

The overall integration scheme consisting of the adjoint equation and the one-step method has a nice additional property – simplecticity, which was also observed by Laurent-Varin et al. [LVBB$^+$04] for ODE optimal control problems. The overall scheme reads as

$$\lambda_{i+1}=\lambda_i-h\Phi'_x(t_i,X_i,w,h)^\top\lambda_{i+1},\qquad X_{i+1}=X_i+h\Phi(t_i,X_i,w,h)$$

Using the auxiliary function $H(t, x, \lambda, w, h) := \lambda^\top \Phi(t, x, w, h)$ the integration scheme can be rewritten in the form

$$\lambda_{i+1} = \lambda_i - hH'_x(t_i, X_i, \lambda_{i+1}, w, h)^\top, \qquad X_{i+1} = X_i + hH'_\lambda(t_i, X_i, \lambda_{i+1}, w, h)^\top.$$

Solving this equation leads to

$$\begin{pmatrix} \lambda_{i+1} \\ X_{i+1} \end{pmatrix} = \begin{pmatrix} \Psi_1(\lambda_i, X_i) \\ \Psi_2(\lambda_i, X_i) \end{pmatrix} = \begin{pmatrix} \left(I + h\Phi'_x(t_i, X_i, w, h)^\top\right)^{-1} \lambda_i \\ X_i + h\Phi(t_i, X_i, w, h) \end{pmatrix}.$$

Hence, we obtain the integration scheme

$$\begin{aligned} \Psi_1(\lambda_i, X_i) + hH'_x(t_i, X_i, \Psi_1(\lambda_i, X_i), w, h)^\top &= \lambda_i, \\ \Psi_2(\lambda_i, X_i) - hH'_\lambda(t_i, X_i, \Psi_1(\lambda_i, X_i), w, h)^\top &= X_i. \end{aligned}$$

Differentiation w.r.t. (λ_i, X_i) leads to

$$\begin{pmatrix} I + h(H''_{x\lambda})^\top & 0 \\ -hH''_{\lambda\lambda} & I \end{pmatrix} \begin{pmatrix} \Psi'_{1,\lambda_i} & \Psi'_{1,x_i} \\ \Psi'_{2,\lambda_i} & \Psi'_{2,x_i} \end{pmatrix} = \begin{pmatrix} I & -hH''_{xx} \\ 0 & I + h(H''_{\lambda x})^\top \end{pmatrix}$$

Exploiting $H''_{\lambda\lambda} = 0$ yields the Jacobian of $\Psi = (\Psi_1, \Psi_2)$ to be

$$\Psi' = \begin{pmatrix} \Psi'_{1,\lambda_i} & \Psi'_{1,x_i} \\ \Psi'_{2,\lambda_i} & \Psi'_{2,x_i} \end{pmatrix} = \begin{pmatrix} \left(I + h(H''_{x\lambda})^\top\right)^{-1} & -h\left(I + h(H''_{x\lambda})^\top\right)^{-1} H''_{xx} \\ 0 & I + hH''_{x\lambda} \end{pmatrix}.$$

It is straightforward to show $(\Psi')^\top J \Psi' = J$, where

$$J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}.$$

Thus, we proved

Theorem 6.2.4 *Ψ is symplectic, i.e. it holds $(\Psi')^\top J \Psi' = J$.*

6.2.3 Adjoint Equation Approach : The Continuous Case

If the DAE is solved by an state-of-the-art integrator with automatic step-size and order selection then the adjoint equation approach as described before becomes very costly in view of the storage needed by the method. Actually, the approximation at every time point generated by the step-size selection algorithm has to be stored. This is a potentially large number. In this case it is more convenient not to consider state approximations obtained by the one-step method. Instead, the continuous DAE (6.1) for the discretized control $u_M(t; w)$ is viewed as a 'continuous' constraint. This will lead to a continuous state $x(t; z)$ depending on z. More precisely, the state x is given by the initial value problem

$$F(t, x(t; z), \dot{x}(t; z), u_M(t; w)) = 0_{n_x}, \qquad x(t_0; z) = X_0(z). \tag{6.2.15}$$

The initial state $X_0(z) \in \mathbb{R}^{n_x}$ is assumed to be sufficiently smooth as a function of z and consistent with the DAE for all z.

We discuss the adjoint approach for calculating the gradient of the function

$$\Gamma(z) := \gamma(x(t_0; z), x(t_f; z), z) \tag{6.2.16}$$

w.r.t. the parameter vector $z \in \mathbb{R}^{n_z}$. In order to derive an expression for the derivative $\Gamma'(z)$ the constraint (6.2.15) is coupled by use of the multiplier function λ to the function in (6.2.16) as follows

$$J(z) := \Gamma(z) + \int_{t_0}^{t_f} \lambda(t)^\top F(t, x(t;z), \dot{x}(t;z), u_M(t;w)dt.$$

Differentiating J w.r.t. z yields the expression

$$\begin{aligned} J'(z) &= \Gamma'(z) + \int_{t_0}^{t_f} \lambda(t)^\top \left(F_x'[t] \cdot S(t) + F_{\dot{x}}'[t] \cdot \dot{S}(t) + F_u'[t] \cdot u_{M,z}'(t;w) \right) dt \\ &= \gamma_{x_0}' \cdot S(t_0) + \gamma_{x_f}' \cdot S(t_f) + \gamma_z' \\ &\quad + \int_{t_0}^{t_f} \lambda(t)^\top \cdot F_x'[t] \cdot S(t) + \lambda(t)^\top \cdot F_{\dot{x}}'[t] \cdot \dot{S}(t) + \lambda(t)^\top \cdot F_u'[t] \cdot u_{M,z}'(t;w)dt. \end{aligned}$$

Integration by parts yields

$$\begin{aligned} J'(z) &= \left(\gamma_{x_f}' + \lambda(t_f)^\top \cdot F_{\dot{x}}'[t_f] \right) \cdot S(t_f) + \left(\gamma_{x_0}' - \lambda(t_0)^\top \cdot F_{\dot{x}}'[t_0] \right) \cdot S(t_0) + \gamma_z' \\ &\quad + \int_{t_0}^{t_f} \left(\lambda(t)^\top \cdot F_x'[t] - \frac{d}{dt} \left(\lambda(t)^\top \cdot F_{\dot{x}}'[t] \right) \right) \cdot S(t) + \lambda(t)^\top \cdot F_u'[t] \cdot u_{M,z}'(t;w)dt. \end{aligned}$$

λ is chosen such that the adjoint DAE

$$\lambda(t)^\top \cdot F_x'[t] - \frac{d}{dt} \left(\lambda(t)^\top \cdot F_{\dot{x}}'[t] \right) = 0_{n_x} \tag{6.2.17}$$

is satisfied, provided that such a function λ exists. Then, the derivative of J reduces to

$$\begin{aligned} J'(z) &= \left(\gamma_{x_f}' + \lambda(t_f)^\top \cdot F_{\dot{x}}'[t_f] \right) \cdot S(t_f) + \left(\gamma_{x_0}' - \lambda(t_0)^\top \cdot F_{\dot{x}}'[t_0] \right) \cdot S(t_0) + \gamma_z' \\ &\quad + \int_{t_0}^{t_f} \lambda(t)^\top \cdot F_u'[t] \cdot u_{M,z}'(t;w)dt. \end{aligned}$$

A connection with the sensitivity DAE arises as follows. Almost everywhere it holds

$$\begin{aligned} \frac{d}{dt} \left(\lambda(t)^\top \cdot F_{\dot{x}}'[t] \cdot S(t) \right) &= \frac{d}{dt} \left(\lambda(t)^\top \cdot F_{\dot{x}}'[t] \right) \cdot S(t) + \lambda(t)^\top \cdot F_{\dot{x}}'(t) \cdot \dot{S}(t) \\ &= \lambda(t)^\top F_x'[t] \cdot S(t) + \lambda(t)^\top (-F_x'[t] \cdot S(t) - F_u'[t] \cdot u_{M,z}'(t;w)) \\ &= -\lambda(t)^\top F_u'[t] \cdot u_{M,z}'(t;w). \end{aligned}$$

Using the fundamental theorem of calculus we obtain

$$\left[\lambda(t)^\top \cdot F_{\dot{x}}'[t] \cdot S(t) \right]_{t_0}^{t_f} = -\int_{t_0}^{t_f} \lambda(t)^\top F_u'[t] \cdot u_{M,z}'(t;w)dt. \tag{6.2.18}$$

The gradient then becomes

$$\begin{aligned} J'(z) &= \left(\gamma_{x_f}' + \lambda(t_f)^\top \cdot F_{\dot{x}}'[t_f] \right) \cdot S(t_f) + \left(\gamma_{x_0}' - \lambda(t_0)^\top \cdot F_{\dot{x}}'[t_0] \right) \cdot S(t_0) + \gamma_z' \\ &\quad + \int_{t_0}^{t_f} \lambda(t)^\top \cdot F_u'[t] \cdot u_{M,z}'(t;w)dt \\ &\overset{(6.2.18)}{=} \gamma_{xf}' S(t_f) + \gamma_{x_0}' \cdot S(t_0) + \gamma_z' \\ &= \Gamma'(z). \end{aligned}$$

Hence, $J'(z)$ is identical to $\Gamma'(z)$ provided that λ satisfies (6.2.17).
While it is cheap to compute the sensitivity matrix

$$S(t_0) = X_0'(z), \tag{6.2.19}$$

the sensitivity $S(t_f)$ in the expression for $J'(z)$ has to be eliminated by a proper choice of $\lambda(t_f)$ as we intend to avoid the explicit computation of $S(t_f)$. Moreover, $\lambda(t_f)$ has to be chosen consistently with the adjoint DAE (6.2.17). We discuss the procedure of defining appropriate conditions for $\lambda(t_f)$ for some common cases, cf. Cao et al. [CLPS03]:

(i) Let $F'_{\dot{x}}$ be non-singular a.e. in $[t_0, t_f]$. Then, the DAE has index zero and $\lambda(t_f)$ is obtained by solving the linear equation

$$\gamma'_{x_f} + \lambda(t_f)^\top \cdot F'_{\dot{x}}[t_f] = 0_{n_x} \tag{6.2.20}$$

and we find

$$\lambda(t_f)^\top = -\gamma'_{x_f} \left(F'_{\dot{x}}[t_f]\right)^{-1}$$

and thus

$$J'(z) = \left(\gamma'_{x_0} - \lambda(t_0)^\top \cdot F'_{\dot{x}}[t_0]\right) \cdot S(t_0) + \gamma'_z + \int_{t_0}^{t_f} \lambda(t)^\top \cdot F'_u[t] \cdot u'_{M,z}(t;w)dt. \tag{6.2.21}$$

(ii) Consider a semi-explicit DAE with $x := (x_d, x_a)^\top \in \mathbb{R}^{n_d+n_a}$ and

$$F(t,x,\dot{x},u) := \begin{pmatrix} f(t,x_d,x_a,u) - \dot{x}_d \\ g(t,x_d,x_a,u) \end{pmatrix} \in \mathbb{R}^{n_d+n_a}.$$

The corresponding adjoint equation (6.2.17) for $\lambda := (\lambda_d, \lambda_a)^\top \in \mathbb{R}^{n_d+n_a}$ reads as

$$\begin{aligned} \lambda_d(t)^\top f'_{x_d}[t] + \lambda_a(t)^\top g'_{x_d}[t] + \dot{\lambda}_d(t)^\top &= 0_{n_d}, && (6.2.22) \\ \lambda_d(t)^\top f'_{x_a}[t] + \lambda_a(t)^\top g'_{x_a}[t] &= 0_{n_a}. && (6.2.23) \end{aligned}$$

With $S := (S_d, S_a)^\top$ we find

$$\left(\gamma'_{x_f} + \lambda(t_f)^\top \cdot F'_{\dot{x}}[t_f]\right) \cdot S(t_f) = \left(\gamma'_{x_{d,f}} - \lambda_d(t_f)^\top\right) S_d(t_f) + \gamma'_{x_{a,f}} S_a(t_f). \tag{6.2.24}$$

The task is to seek for a consistent value $\lambda(t_f) = (\lambda_d(t_f), \lambda_a(t_f))^\top$ for (6.2.22)-(6.2.23) such that the expression in (6.2.24) does not depend explicitly on $S(t_f)$.

(a) Let g'_{x_a} be non-singular a.e. in $[t_0, t_f]$. Then, the DAE has index one and

$$\lambda_a(t_f)^\top = -\lambda_d(t_f)^\top f'_{x_a}[t_f](g'_{x_a}[t_f])^{-1}$$

is consistent with the algebraic constraint (6.2.23) of the adjoint system for any $\lambda_d(t_f)$. The expression (6.2.24) vanishes if

$$\begin{aligned} \lambda_d(t_f)^\top &= \gamma'_{x_{d,f}}, \\ \gamma'_{x_{a,f}} &= 0_{n_a}. \end{aligned}$$

Hence, with these settings, $J'(z)$ is given by (6.2.21).
The latter assumption is not as restrictive as it seems as it fits well into the theoretical investigations in Chapter 4. For, the algebraic component x_a is an L^∞-function and thus it makes no sense to allow a pointwise evaluation of x_a at t_0 or t_f. Consequently, it is natural to prohibit γ depending explicitly on x_a.

(b) Let $g'_{x_a} \equiv 0$ and $g'_{x_d} f'_{x_a}$ be non-singular a.e. in $[t_0, t_f]$. Then, the DAE has index two and

$$\lambda_a(t_f)^\top = -\lambda_d(t_f)^\top \left(f'_{x_d}[t_f] f'_{x_a}[t_f] - \frac{d}{dt} f'_{x_a}[t_f] \right) (g'_{x_d}[t_f] f'_{x_a}[t_f])^{-1}$$

is consistent with the derivative of the algebraic constraint (6.2.23) for any $\lambda_d(t_f)$. However, not any $\lambda_f(t_f)$ is consistent with the constraint (6.2.23). We make the ansatz

$$\gamma'_{x_{d,f}} - \lambda_d(t_f)^\top = \xi^\top g'_{x_d}[t_f]$$

with $\xi \in \mathbb{R}^{n_a}$ to be determined such that $\lambda_d(t_f)$ is consistent with (6.2.23), i.e.

$$0_{na} = \lambda_d(t_f)^\top f'_{x_a}[t_f] = \gamma'_{x_{d,f}} f'_{x_a}[t_f] - \xi^\top g'_{x_d}[t_f] f'_{x_a}[t_f],$$

and thus

$$\xi^\top = \gamma'_{x_{d,f}} f'_{x_a}[t_f] \left(g'_{x_d}[t_f] f'_{x_a}[t_f] \right)^{-1}.$$

Moreover, the sensitivity matrix S_d satisfies a.e. the algebraic equation

$$g'_{x_d}[t] S_d(t) + g'_u[t] u'_{M,z}(t; w) = 0_{n_a}$$

and thus the first term of the right handside in (6.2.24) computes to

$$\left(\gamma'_{x_{d,f}} - \lambda_d(t_f)^\top \right) S_d(t_f) = \xi^\top g'_{x_d}[t_f] S_d(t_f) = -\xi^\top g'_u[t_f] u'_{M,z}(t_f; w).$$

Notice, that the expression on the right does not depend on $S(t_f)$ anymore. Finally, if, as above, we assume $\gamma'_{x_{a,f}} = 0_{n_a}$, then $J'(z)$ can be computed without computing $S(t_f)$ according to

$$\begin{aligned} J'(z) &= -\gamma'_{x_{d,f}} f'_{x_a}[t_f] \left(g'_{x_d}[t_f] f'_{x_a}[t_f] \right)^{-1} g'_u[t_f] u'_{M,z}(t_f; w) \\ &\quad + \left(\gamma'_{x_0} - \lambda(t_0)^\top \cdot F'_{\dot{x}}[t_0] \right) \cdot S(t_0) + \gamma'_z + \int_{t_0}^{t_f} \lambda(t)^\top \cdot F'_u[t] \cdot u'_{M,z}(t; w) dt. \end{aligned}$$

Example 6.2.5 *In the special case of an ODE of type*

$$F(t, x(t), \dot{x}(t), u_M(t; w)) := f(t, x(t), u_M(t; w)) - \dot{x}(t)$$

equations (6.2.17) and (6.2.20) reduce to

$$\dot{\lambda}(t)^\top = -\lambda(t)^\top f'_x[t], \qquad \lambda(t_f)^\top = \gamma'_{x_f}.$$

In Cao et al. [CLPS03] stability results for the forward and the adjoint system are derived. It is shown for explicit ODE's and semi-explicit index-1 and index-2 Hessenberg DAE's that stability is preserved for the adjoint DAE. For fully implicit index 0 and index 1 DAE's the augmented adjoint system in (6.2.25) below is stable, if the original DAE was stable, cf. Theorem 4.3 in Cao et. [CLPS03].

6.2.3.1 Numerical Adjoint Approximation by BDF

We discuss the application of the BDF method to the adjoint DAE. Firstly, the implicit DAE

$$F(t, x(t), \dot{x}(t), u_M(t; w)) = 0_{n_x}$$

is transformed to semi-explicit form:

$$\begin{aligned} 0_{n_x} &= \dot{x}(t) - v(t), \\ 0_{n_x} &= F(t, x(t), v(t), u_M(t; w)). \end{aligned}$$

Using the function

$$\begin{aligned} \hat{H}(t, x, \dot{x}, v, u, \lambda_v, \lambda_x) &:= \left(\lambda_v^\top, \lambda_x^\top\right) \begin{pmatrix} \dot{x} - v \\ F(t, x, v, u) \end{pmatrix} \\ &= \lambda_v^\top (\dot{x} - v) + \lambda_x^\top F(t, x, v, u), \end{aligned}$$

the adjoint DAE (6.2.17) for $y = (x, v)^\top$ is given by

$$\hat{H}_y'[t] - \frac{d}{dt}\left(\hat{H}_{\dot{y}}'[t]\right) = 0_{2n_x}.$$

Transposition yields

$$\begin{pmatrix} \left(F_x'[t]\right)^\top \lambda_x(t) - \dot{\lambda}_v(t) \\ \\ -\lambda_v(t) + \left(F_{\dot{x}}'[t]\right)^\top \lambda_x(t) \end{pmatrix} = \begin{pmatrix} 0_{n_x} \\ \\ 0_{n_x} \end{pmatrix}. \tag{6.2.25}$$

This is a linear DAE in $\lambda = (\lambda_v, \lambda_x)^\top$. Employing the BDF discretization method for the integration step $t_{m+k-1} \to t_{m+k}$ and using the approximation

$$\dot{\lambda}_v(t_{m+k}) \approx \frac{1}{h} \sum_{i=0}^{k} \alpha_i \lambda_v(t_{m+i})$$

yields the linear equation

$$\begin{pmatrix} \left(F_x'[t_{m+k}]\right)^\top \lambda_x(t_{m+k}) - \frac{1}{h} \sum_{i=0}^{k} \alpha_i \lambda_v(t_{m+i}) \\ -\lambda_v(t_{m+k}) + \left(F_{\dot{x}}'[t_{m+k}]\right)^\top \lambda_x(t_{m+k}) \end{pmatrix} = \begin{pmatrix} 0_{n_x} \\ 0_{n_x} \end{pmatrix}$$

for $\lambda_x(t_{m+k})$ and $\lambda_v(t_{m+k})$. The size of the equation can be reduced by introducing the second equation

$$\lambda_v(t_{m+k}) = \left(F_{\dot{x}}'[t_{m+k}]\right)^\top \lambda_x(t_{m+k}),$$

into the first equation. Herein, the relations

$$\lambda_v(t_{m+i}) = \left(F_{\dot{x}}'[t_{m+i}]\right)^\top \lambda_x(t_{m+i}), \quad i = 0, 1, \ldots, k$$

are exploited. It remains to solve the linear equation

$$\left(\left(F_x'[t_{m+k}]\right)^\top - \frac{\alpha_k}{h}\left(F_{\dot{x}}'[t_{m+k}]\right)^\top\right) \lambda_x(t_{m+k}) = \frac{1}{h} \sum_{i=0}^{k-1} \alpha_i \left(F_{\dot{x}}'[t_{m+i}]\right)^\top \cdot \lambda_x(t_{m+i}).$$

Recall, that λ_x is the desired adjoint λ of the original system given by $F(\cdot) = 0$, whose Hamilton function is given by

$$H(t, x, \dot{x}, u, \lambda) = \lambda^\top F(t, x, \dot{x}, u).$$

For this, the adjoint DAE is given by

$$\left(F_x'[t]\right)^\top \lambda(t) - \frac{d}{dt}\left[\left(F_{\dot{x}}'[t]\right)^\top \lambda(t)\right] = 0_{n_x}$$

Notice, that this equation arises, if the second equation in (6.2.25) is introduced into the first one.

The numerical calculations introduced for the adjoint equation are very similar to those for the sensitivity equation. The discretized sensitivity equation is given by

$$\left(F_x'[t_{m+k}] + \frac{\alpha_k}{h}F_{\dot{x}}'[t_{m+k}]\right) S(t_{m+k}) = -F_u'[t_{m+k}] \cdot u_z'(t_{m+k}; w) - \frac{1}{h}\sum_{i=0}^{k-1} \alpha_i F_{\dot{x}}'[t_{m+i}] \cdot S(t_{m+i}).$$

Cao et al. [CLPS03] investigate numerical stability for the adjoint system and show for backward Euler's method that the augmented system (6.2.25) preserves asymptotic numerical stability.

6.3 Numerical Example

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. Mayrhofer and Sachs [MS96] and Büskens and Gerdts [BG00]. 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

$$-\left(\frac{\Lambda(t_f) - \Lambda(0)}{\Lambda(0)}\right)^2 - \left(\frac{\Theta(t_f) - \Theta(0)}{\Theta(0)}\right)^2 \qquad (6.3.1)$$

subject to the ODE 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

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

and constants

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

cf Chudej [Chu94]. 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

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

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 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 by the reduced discretization approach. We used the fourth order classical Runge-Kutta method with fixed step-size for time-integration of the ODE. The control is approximated by a continuous and piecewise linear function.

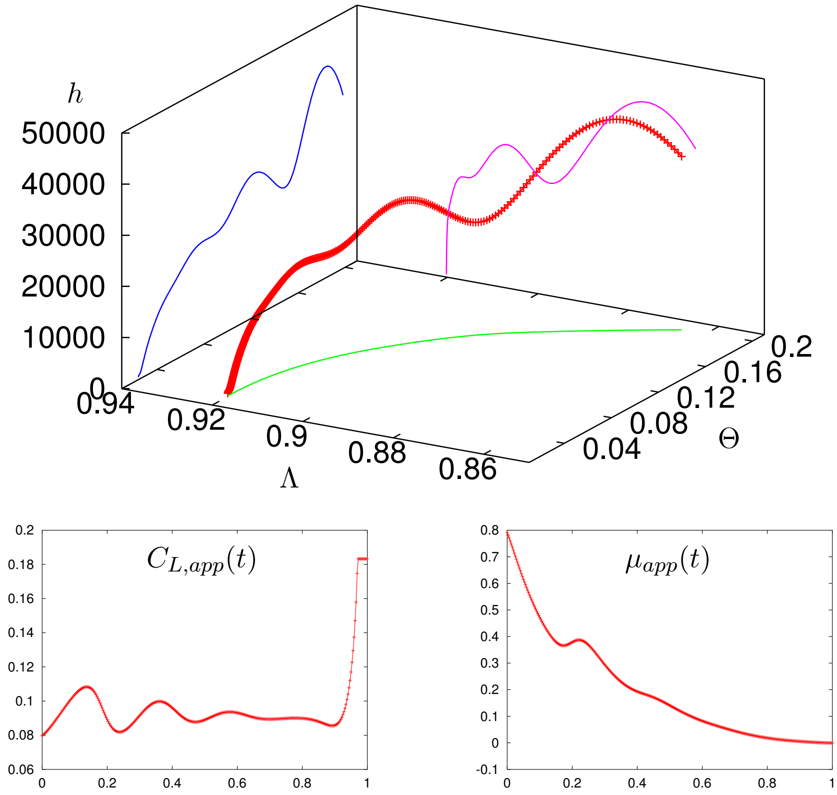


Figure 6.2: Numerical solution of the problem: 3D plot of the state (top) and approximate optimal controls $C_{L,app}$ and μ_{app} (bottom, normalized time scale) for $N = 320$ equidistant intervals.

Figure 6.2 shows the numerical solution for $N = 320$ equidistant intervals. The approximate optimal final time for this highly nonlinear optimization problem is calculated to $t_f = 728.874966$ seconds and the approximate objective function value is -0.77581222.

Figure 6.3 summarizes computational results obtained for the sensitivity equation approach and the adjoint equation approach. While the sensitivity approach grows nonlinearly with the number N of equidistant time intervals in the control grid, the adjoint approach grows at a linear rate. Hence, in this case, the adjoint approach is more efficient than the sensitivity approach. This is the expected behavior since the number of constraints is significantly smaller than the number of optimization variables.

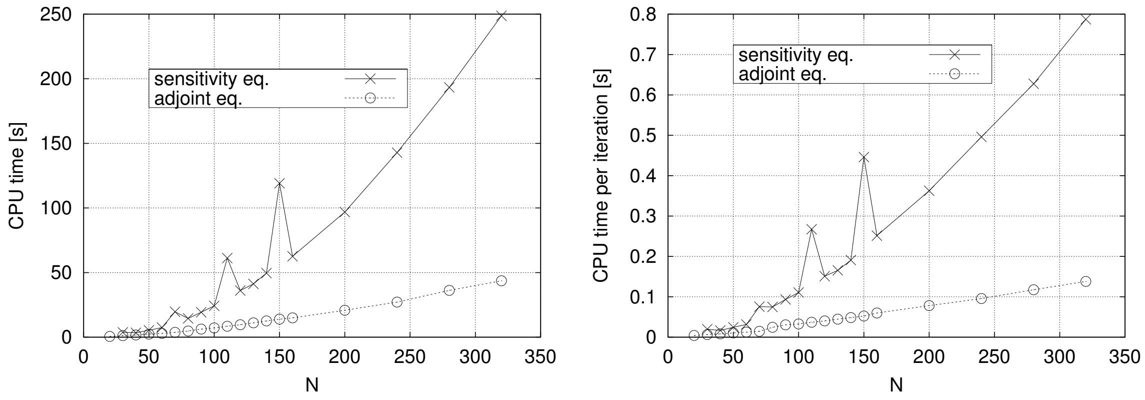


Figure 6.3: Computational results for the emergency landing manoeuvre without dynamic pressure constraint: Performance of the sensitivity equation approach compared to the adjoint equation approach with NPSOL.

6.4 Discrete Minimum Principle and Approximation of Adjoints

The aim of this section is to state a local minimum principle for the full discretization of Problem 4.2.1 using implicit Euler's method. It turns out that this discrete local minimum principle can be interpreted as discretization of the local minimum principles 4.2.4 and 4.2.5, respectively. This observation allows to construct approximations of the multipliers λ_f, λ_g, η, μ, and σ of Section 4.2 to which we refer to as 'continuous' multipliers for brevity (which does not mean that these multipliers are actually continuous). We consider Problem 4.2.1 with $f_0 \equiv 0$:

Problem 6.4.1 (DAE optimal control problem)
Find $x \in W^{1,\infty}([t_0,t_f],\mathbb{R}^{n_x})$, $y \in L^{\infty}([t_0,t_f],\mathbb{R}^{n_y})$, and $u \in L^{\infty}([t_0,t_f],\mathbb{R}^{n_u})$ such that

$$\varphi(x(t_0), x(t_f)) \tag{6.4.1}$$

is minimized subject to the semi-explicit DAE

$$\begin{aligned} \dot{x}(t) &= f(t,x(t),y(t),u(t)) \qquad \textit{a.e. in } [t_0,t_f], & (6.4.2)\\ 0_{n_y} &= g(t,x(t),y(t),u(t)) \qquad \textit{a.e. in } [t_0,t_f], & (6.4.3) \end{aligned}$$

the boundary conditions

$$\psi(x(t_0), x(t_f)) = 0_{n_\psi}, \tag{6.4.4}$$

the mixed control-state constraints

$$c(t,x(t),y(t),u(t)) \le 0_{n_c} \qquad \textit{a.e. in } [t_0,t_f], \tag{6.4.5}$$

the pure state constraints

$$s(t,x(t)) \le 0_{n_s} \qquad \textit{in } [t_0,t_f], \tag{6.4.6}$$

and the set constraints

$$u(t) \in \mathcal{U} \qquad \textit{a.e. in } [t_0,t_f]. \tag{6.4.7}$$

The Hamilton function and the augmented Hamilton function are defined as usual (with $f_0 \equiv 0$):

$$\begin{aligned} \mathcal{H}(t,x,y,u,\lambda_f,\lambda_g) &= \lambda_f^\top f(t,x,y,u) + \lambda_g^\top g(t,x,y,u),\\ \hat{\mathcal{H}}(t,x,y,u,\lambda_f,\lambda_g,\eta) &= \mathcal{H}(t,x,y,u,\lambda_f,\lambda_g) + \eta^\top c(t,x,y,u). \end{aligned}$$

Similar as in Theorem 4.2.5 we assume:

Assumption 6.4.2

(i) Let $g_y'(t,x,y,u)$ be non-singular for every t,x,y,u.

(ii) For every t,x,y,u let rank$(c_u'(t,x,y,u)) = n_c$.

(iii) Let the matrix $g_y' - g_u'(c_u')^+ c_y'$ be non-singular where $(c_u')^+$ denotes the pseudo-inverse of c_u'.

Problem 6.4.1 is discretized by application of the implicit Euler's method on the (not necessarily equidistant) grid $\mathbb{G}_N$ in (6.1.1). The resulting finite dimensional nonlinear program reads as follows.

Problem 6.4.3 (Discretized DAE optimal control problem)
Find grid functions $x_N : \mathbb{G}_N \to \mathbb{R}^{n_x}$, $x_i := x_N(t_i)$, $y_N : \mathbb{G}_N \to \mathbb{R}^{n_y}$, $y_i := y_N(t_i)$, $u_N : \mathbb{G}_N \to \mathbb{R}^{n_u}$, $u_i := u_N(t_i)$, such that

$$\varphi(x_0, x_N) \tag{6.4.8}$$

is minimized subject to the equations

$$\begin{aligned}
f(t_i,x_i,y_i,u_i) - \frac{x_i - x_{i-1}}{h_{i-1}} &= 0_{n_x}, \qquad i = 1,\ldots,N, && (6.4.9)\\
g(t_i,x_i,y_i,u_i) &= 0_{n_y}, \qquad i = 1,\ldots,N, && (6.4.10)\\
\psi(x_0,x_N) &= 0_{n_\psi}, && (6.4.11)\\
c(t_i,x_i,y_i,u_i) &\leq 0_{n_c}, \qquad i = 1,\ldots,N, && (6.4.12)\\
s(t_i,x_i) &\leq 0_{n_s}, \qquad i = 0,\ldots,N, && (6.4.13)\\
u_i &\in \mathcal{U}, \qquad i = 1,\ldots,N. && (6.4.14)
\end{aligned}$$

A few remarks are in order. Equations (6.4.9)-(6.4.10) result from the implicit Euler's method applied to the DAE (6.4.2)-(6.4.3). Notice, that we can dispense with the algebraic constraint (6.4.3) and the mixed control-state constraint (6.4.5) at $t = t_0$ in the discretized problem if Assumption 6.4.2 is valid. This is true because Assumption 6.4.2 guarantees that the equations $g(t_0,x_0,y_0,u_0) = 0_{n_y}$ and $c(t_0,x_0,y_0,u_0) = 0_{n_c}$ can be solved for y_0 and u_0 for arbitrary x_0 provided that a solution exists at all. On the other hand, u_0 and y_0 would not enter the objective function and thus it would be superfluous to impose the constraints $g(t_0,x_0,y_0,u_0) = 0_{n_y}$ and $c(t_0,x_0,y_0,u_0) \leq 0_{n_c}$ in Problem 6.4.3. If Assumption 6.4.2 is not valid these constraints have to be added. The following discrete local minimum principle can be modified accordingly in this case. Unfortunately, then the following interpretations concerning the relationship between the multipliers of Problem 6.4.1 and Problem 6.4.3 do not hold anymore. This lack of consistency could be explained if one could show that the corresponding continuous minimum principle does not hold in the corresponding form. This question is currently under investigation.
Evaluation of the Fritz-John conditions in Theorem 3.6.2 yields the following necessary optimality conditions.

Theorem 6.4.4 (Discrete Local Minimum Principle)

(i) Let the functions $\varphi, f, g, c, s, \psi$ be continuously differentiable w.r.t. x, y, and u.

(ii) Let $\mathcal{U} \subseteq \mathbb{R}^{n_u}$ be a closed and convex set with non-empty interior.

(iii) Let $(\hat{x}_N, \hat{y}_N, \hat{u}_N)$ be a local minimum of Problem 6.4.3.

Then there exist multipliers $\kappa_0 \in \mathbb{R}$, $\kappa \in \mathbb{R}^{n_\psi}$, $\lambda_{f,N} : \mathbb{G}_N \to \mathbb{R}^{n_x}$, $\lambda_{f,i} := \lambda_{f,N}(t_i)$, $\lambda_{g,N} : \mathbb{G}_N \to \mathbb{R}^{n_x}$, $\lambda_{g,i} := \lambda_{g,N}(t_i)$, $\zeta_N : \mathbb{G}_N \to \mathbb{R}^{n_c}$, $\zeta_i := \zeta_N(t_i)$, *and* $\nu_N : \mathbb{G}_N \to \mathbb{R}^{n_s}$, $\nu_i := \nu_N(t_i)$ *such that the following conditions are satisfied:*

(i) $\kappa_0 \geq 0$, $(\kappa_0, \kappa, \lambda_{f,N}, \lambda_{g,N}, \zeta_N, \nu_N) \neq \Theta$,

(ii) Discrete adjoint equations:
For $i = 1, \ldots, N$ *it holds*

$$\begin{aligned}
\lambda_{f,i-1}^\top &= \lambda_{f,i}^\top + h_{i-1}\mathcal{H}'_x\left(t_i, \hat{x}_i, \hat{y}_i, \hat{u}_i, \lambda_{f,i-1}, \lambda_{g,i-1}, \frac{\zeta_i}{h_{i-1}}\right) + \nu_i^\top s'_x(t_i, \hat{x}_i) \\
&= \lambda_{f,N}^\top + \sum_{j=i}^{N} h_{j-1}\mathcal{H}'_x\left(t_j, \hat{x}_j, \hat{y}_j, \hat{u}_j, \lambda_{f,j-1}, \lambda_{g,j-1}, \frac{\zeta_j}{h_{j-1}}\right) \\
&\quad + \sum_{j=i}^{N} \nu_j^\top s'_x(t_j, \hat{x}_j), \qquad (6.4.15) \\
0_{n_y} &= \mathcal{H}'_y\left(t_i, \hat{x}_i, \hat{y}_i, \hat{u}_i, \lambda_{f,i-1}, \lambda_{g,i-1}, \frac{\zeta_i}{h_{i-1}}\right). \qquad (6.4.16)
\end{aligned}$$

(iii) Discrete transversality conditions:

$$\begin{aligned}
\lambda_{f,0}^\top &= -\left(\kappa_0\varphi'_{x0}(\hat{x}_0, \hat{x}_N) + \kappa^\top\psi'_{x0}(\hat{x}_0, \hat{x}_N) + \nu_0^\top s'_x(t_0, \hat{x}_0)\right), \qquad (6.4.17) \\
\lambda_{f,N}^\top &= \kappa_0\varphi'_{x_N}(\hat{x}_0, \hat{x}_N) + \kappa^\top\psi'_{x_N}(\hat{x}_0, \hat{x}_N). \qquad (6.4.18)
\end{aligned}$$

(iv) Discrete optimality conditions:
For all $i = 1, \ldots, N$ *and all* $u \in \mathcal{U}$ *it holds*

$$\mathcal{H}'_u\left(t_i, \hat{x}_i, \hat{y}_i, \hat{u}_i, \lambda_{f,i-1}, \lambda_{g,i-1}, \frac{\zeta_i}{h_{i-1}}\right)(u - \hat{u}_i) \geq 0. \qquad (6.4.19)$$

(v) Discrete complementarity conditions:
It holds

$$\begin{aligned}
\zeta_i &\geq 0_{n_c}, \quad i = 1, \ldots, N, \qquad (6.4.20) \\
\nu_i &\geq 0_{n_s}, \quad i = 0, \ldots, N, \qquad (6.4.21) \\
\zeta_i^\top c(t_i, \hat{x}_i, \hat{y}_i, \hat{u}_i) &= 0, \quad i = 1, \ldots, N, \qquad (6.4.22) \\
\nu_i^\top s(t_i, \hat{x}_i) &= 0, \quad i = 0, \ldots, N. \qquad (6.4.23)
\end{aligned}$$

Proof. Let $\kappa_0 \in \mathbb{R}$, $\hat{x} = (\hat{x}_0, \ldots, \hat{x}_N)^\top \in \mathbb{R}^{n_x(N+1)}$, $\hat{y} = (\hat{y}_1, \ldots, \hat{y}_N)^\top \in \mathbb{R}^{n_y N}$, $\hat{u} = (\hat{u}_1, \ldots, \hat{u}_N)^\top \in \mathbb{R}^{n_u N}$, $\tilde{\lambda}_f = (\tilde{\lambda}_{f,0}, \ldots, \tilde{\lambda}_{f,N-1})^\top \in \mathbb{R}^{n_x N}$, $\tilde{\lambda}_g = (\tilde{\lambda}_{g,0}, \ldots, \tilde{\lambda}_{g,N-1})^\top \in \mathbb{R}^{n_y N}$, $\zeta = (\zeta_1, \ldots, \zeta_N)^\top \in \mathbb{R}^{n_c N}$, $\nu = (\nu_0, \ldots, \nu_N)^\top \in \mathbb{R}^{n_s(N+1)}$, $\kappa \in \mathbb{R}^{n_\psi}$. The Lagrange function of

Problem 6.4.3 is given by

$$\begin{aligned}
L(x,y,u,\tilde{\lambda}_f,\tilde{\lambda}_g,\zeta,\nu,\kappa,\kappa_0) \quad &:= \quad \kappa_0\varphi(x_0,x_N)+\kappa^\top\psi(x_0,x_N)\\
&\qquad +\sum_{i=1}^{N}\tilde{\lambda}_{f,i-1}^\top\left(f(t_i,x_i,y_i,u_i)-\frac{x_i-x_{i-1}}{h_{i-1}}\right)\\
&\qquad +\sum_{i=1}^{N}\tilde{\lambda}_{g,i-1}^\top g(t_i,x_i,y_i,u_i)+\sum_{i=1}^{N}\zeta_i^\top c(t_i,x_i,y_i,u_i)+\sum_{i=0}^{N}\nu_i^\top s(t_i,x_i)\\
&= \quad \kappa_0\varphi(x_0,x_N)+\kappa^\top\psi(x_0,x_N)\\
&\qquad +\sum_{i=1}^{N}\hat{\mathcal{H}}(t,x_i,y_i,u_i,\tilde{\lambda}_{f,i-1},\tilde{\lambda}_{g,i-1},\zeta_i)+\sum_{i=1}^{N}\tilde{\lambda}_{f,i-1}^\top\frac{x_{i-1}-x_i}{h_{i-1}}\\
&\qquad +\sum_{i=0}^{N}\nu_i^\top s(t_i,x_i).
\end{aligned}$$

Application of Theorem 3.6.2 results in the following equations:

1. $L'_{u_i}(u-\hat{u}_i)\geq 0\ \forall u\in\mathcal{U}$: For $i=1,\ldots,N$ it holds

$$\hat{\mathcal{H}}'_u(t_i,\hat{x}_i,\hat{y}_i,\hat{u}_i,\tilde{\lambda}_{f,i-1},\tilde{\lambda}_{g,i-1},\zeta_i)(u-\hat{u}_i)\geq 0 \qquad \forall u\in\mathcal{U}.$$

2. $L'_{y_i}=0$: For $i=1,\ldots,N$ it holds

$$\hat{\mathcal{H}}'_y(t_i,\hat{x}_i,\hat{y}_i,\hat{u}_i,\tilde{\lambda}_{f,i-1},\tilde{\lambda}_{g,i-1},\zeta_i)=0_{n_y}$$

3. $L'_{x_i}=0$: For $i=0$ it holds

$$\left(\kappa_0\varphi'_{x_0}(\hat{x}_0,\hat{x}_N)+\kappa^\top\psi'_{x_0}(\hat{x}_0,\hat{x}_N)+\nu_0^\top s'_x(t_0,\hat{x}_0)\right)+\frac{1}{h_0}\tilde{\lambda}_{f,0}^\top=0_{n_x}.$$

For $i=1,\ldots,N-1$ it holds

$$\hat{\mathcal{H}}'_x(t_i,\hat{x}_i,\hat{y}_i,\hat{u}_i,\tilde{\lambda}_{f,i-1},\tilde{\lambda}_{g,i-1},\zeta_i)-\frac{1}{h_{i-1}}\tilde{\lambda}_{f,i-1}^\top+\frac{1}{h_i}\tilde{\lambda}_{f,i}^\top+\nu_i^\top s'_x(t_i,\hat{x}_i)=0_{n_x}$$

For $i=N$ it holds

$$\begin{aligned}
&\kappa_0\varphi'_{x_N}(\hat{x}_0,\hat{x}_N)+\kappa^\top\psi'_{x_N}(\hat{x}_0,\hat{x}_N)+\nu_N^\top s'_x(t_N,\hat{x}_N)\\
&\quad +\hat{\mathcal{H}}'_x(t_N,\hat{x}_N,\hat{y}_N,\hat{u}_N,\tilde{\lambda}_{f,N-1},\tilde{\lambda}_{g,N-1},\zeta_N)-\frac{1}{h_{N-1}}\tilde{\lambda}_{f,N-1}^\top \quad = \quad 0_{n_x}
\end{aligned}$$

With the definitions

$$\lambda_{f,i}:=\frac{1}{h_i}\tilde{\lambda}_{f,i},\quad \lambda_{g,i}:=\frac{1}{h_i}\tilde{\lambda}_{g,i},\quad i=0,\ldots,N-1,$$

and

$$\lambda_{f,N}^\top:=\kappa_0\varphi'_{x_N}(\hat{x}_0,\hat{x}_N)+\kappa^\top\psi'_{x_N}(\hat{x}_0,\hat{x}_N).$$

we obtain the discrete optimality conditions, the discrete adjoint equations, and the discrete transversality conditions. ■

Remark 6.4.5 *The local minimum principle for ODE optimal control problems can be extended to a global minimum principle. The question arises whether a similar result holds for the discretized optimal control problem. This is not the case in general. However, an approximate minimum principle holds, cf. Mordukhovich [Mor88]. With additional convexity-like conditions a discrete minimum principle holds as well, cf. Ioffe and Tihomirov [IT79], Section 6.4, p. 277.*

We compare the necessary optimality conditions in Theorem 6.4.4 with those in Theorems 4.2.4 and 4.2.5. Our intention is to provide an interpretation of the 'discrete' multipliers in Theorem 6.4.4 and put them into relation to the 'continuous' multipliers in Theorem 4.2.4. Following this interpretation it becomes possible to construct estimates of the continuous multipliers by use of the discrete multipliers only. Of course, the discrete multipliers can be computed numerically by solving the discretized optimal control problem 6.4.3 by SQP. Unfortunately, we are not able to present rigorous convergence proofs which would justify the following interpretations. This is current research. Nevertheless, we will summarize some available convergence results for ODE optimal control problems in the next section.

6.4.1 Problems with Pure State Constraints

In line with the situation of Theorem 4.2.4 we discuss the Problem 6.4.1 without mixed control-state constraints, i.e. $n_c = 0$ and $c \equiv 0$.

The discrete optimality conditions (6.4.19) and the discrete adjoint equations (6.4.16) are easily being recognized as pointwise discretizations of the optimality condition (4.2.12) and the algebraic equation in (4.2.9), respectively, provided we assume for all i (actually there will be an exceptional interpretation for $\lambda_{f,0}$ later on):

$$\hat{x}_i \approx \hat{x}(t_i),\ \hat{y}_i \approx \hat{y}(t_i),\ \hat{u}_i \approx \hat{u}(t_i),\ \lambda_{f,i} \approx \lambda_f(t_i),\ \lambda_{g,i} \approx \lambda_g(t_i).$$

Comparing the discrete transversality condition (6.4.18)

$$\lambda_{f,N}^\top = \kappa_0 \varphi'_{x_N}(\hat{x}_0, \hat{x}_N) + \kappa^\top \psi'_{x_f}(\hat{x}_0, \hat{x}_N)$$

and the transversality condition (4.2.11)

$$\lambda_f(t_f)^\top = l_0 \varphi'_{x_f}(\hat{x}(t_0), \hat{x}(t_f)) + \sigma^\top \psi'_{x_f}(\hat{x}(t_0), \hat{x}(t_f))$$

it is natural to presume

$$\kappa_0 \approx l_0, \qquad \kappa \approx \sigma.$$

The adjoint equation (4.2.8) for $t \in [t_0, t_f]$ reads as

$$\lambda_f(t) = \lambda_f(t_f) + \int_t^{t_f} \mathcal{H}'_x(\tau, \hat{x}(\tau), \hat{y}(\tau), \hat{u}(\tau), \lambda_f(\tau), \lambda_g(\tau))^\top d\tau + \int_t^{t_f} s'_x(\tau, \hat{x}(\tau))^\top d\mu(\tau),$$

where we used the abbreviation

$$\int_t^{t_f} s'_x(\tau, \hat{x}(\tau))^\top d\mu(\tau) := \sum_{i=1}^{n_s} \int_t^{t_f} s'_{i,x}[\tau]^\top d\mu_i(\tau).$$

The discrete adjoint equation (6.4.15) for $i = 1, \ldots, N$ is given by

$$\lambda_{f,i-1} \quad = \quad \lambda_{f,N} + \sum_{j=i}^{N} h_{j-1} \mathcal{H}'_x(t_j, \hat{x}_j, \hat{y}_j, \hat{u}_j, \lambda_{f,j-1}, \lambda_{g,j-1})^\top + \sum_{j=i}^{N} s'_x(t_j, \hat{x}_j)^\top \nu_j.$$

The first sum is easily being recognized to be a Riemann sum on $\mathbb{G}_N$ and thus

$$\sum_{j=i}^{N} h_{j-1}\mathcal{H}'_x\left(t_j,\hat{x}_j,\hat{y}_j,\hat{u}_j,\lambda_{f,j-1},\lambda_{g,j-1}\right)^\top \approx \int_{t_{i-1}}^{t_f} \mathcal{H}'_x(\tau,\hat{x}(\tau),\hat{y}(\tau),\hat{u}(\tau),\lambda_f(\tau),\lambda_g(\tau))^\top d\tau \tag{6.4.24}$$

for $i = 1,\dots,N$. This observation encourages us to expect that for $i = 1,\dots,N$

$$\sum_{j=i}^{N} s'_x(t_j,\hat{x}_j)^\top \nu_j \approx \int_{t_{i-1}}^{t_f} s'_x(\tau,\hat{x}(\tau))^\top d\mu(\tau). \tag{6.4.25}$$

In order to interpret the approximation in (6.4.25) we have to recall the definition of a Riemann-Stieltjes integral, cf. Section 2.4. The Riemann-Stieltjes integral in (6.4.25) is defined to be the limit of the sum

$$\sum_{j=i}^{m} s'_x(\xi_j,\hat{x}(\xi_j))^\top \left(\mu(t_j)-\mu(t_{j-1})\right)$$

where $t_{i-1} < t_i < \ldots < t_m = t_f$ is an arbitrary partition of $[t_{i-1},t_f]$ and $\xi_j \in [t_{j-1},t_j]$ are arbitrary points. If we chose the particular grid $\mathbb{G}_N$ and the points $\xi_j := t_j$ we obtain the approximation

$$\int_{t_{i-1}}^{t_f} s'_x(\tau,\hat{x}(\tau))^\top d\mu(\tau) \approx \sum_{j=i}^{N} s'_x(t_j,\hat{x}(t_j))^\top \left(\mu(t_j)-\mu(t_{j-1})\right).$$

Together with (6.4.25) we draw the conclusion that the relationship of the discrete multipliers ν_i and the continuous counterpart μ must be given by

$$\nu_i \approx \mu(t_i)-\mu(t_{i-1}), \qquad i = 1,\dots,N.$$

Now, we address the remaining transversality condition (4.2.10):

$$\lambda_f(t_0)^\top = -\left(l_0\varphi'_{x_0}(\hat{x}(t_0),\hat{x}(t_f)) + \sigma^\top\psi'_{x_0}(\hat{x}(t_0),\hat{x}(t_f))\right)$$

and the discrete counterpart (6.4.17)

$$\lambda_{f,0}^\top = -\left(\kappa_0\varphi'_{x_0}(\hat{x}_0,\hat{x}_N) + \kappa^\top\psi'_{x_0}(\hat{x}_0,\hat{x}_N) + \nu_0^\top s'_x(t_0,\hat{x}_0)\right).$$

These two conditions can be brought into accordance as follows. Recall, that the multiplier μ is normalized, i.e. $\mu(t_0) = 0_{n_s}$ and μ is continuous from the right in the open interval (t_0,t_f). Hence, μ may jump at t_0 with $0 \le \mu(t_0^+) - \mu(t_0) = \mu(t_0^+)$. Similarly, λ_f may jump at t_0, too. Similar to the derivation of the jump conditions (4.2.21) it follows

$$\lambda_f(t_0^+) - \lambda_f(t_0) = \lim_{\varepsilon\downarrow 0}\lambda_f(t_0+\varepsilon) - \lambda_f(t_0) = -s'_x(t_0,\hat{x}(t_0))^\top(\mu(t_0^+)-\mu(t_0))$$

and thus

$$\begin{aligned}\lambda_f(t_0^+) &= \lambda_f(t_0) - s'_x(t_0,\hat{x}(t_0))^\top(\mu(t_0^+)-\mu(t_0)) \\ &= -\left(l_0\varphi'_{x_0}(\hat{x}(t_0),\hat{x}(t_f)) + \sigma^\top\psi'_{x_0}(\hat{x}(t_0),\hat{x}(t_f)) + s'_x(t_0,\hat{x}(t_0))^\top(\mu(t_0^+)-\mu(t_0))\right).\end{aligned}$$

A comparison with (6.4.17) shows that both are in accordance if ν_0 is interpreted as the jump height of μ at $t = t_0$, i.e.

$$\nu_0 \approx \mu(t_0^+) - \mu(t_0)$$

and if $\lambda_{f,0}$ is interpreted as the right-sided limit of λ_f at t_0, i.e.

$$\lambda_{f,0} \approx \lambda_f(t_0^+).$$

As an approximation of the value $\lambda_f(t_0)$ we then use the value

$$-\left(\kappa_0 \varphi'_{x_0}(\hat{x}_0, \hat{x}_N) + \kappa^\top \psi'_{x_0}(\hat{x}_0, \hat{x}_N)\right).$$

The above interpretations cope well with the complementarity conditions if we recall that ν_i denotes the multiplier for the constraint $s(t_i, x_i) \leq 0_{n_s}$. Then, the complementarity conditions yield

$$0_{n_s} \leq \nu_i \approx \mu(t_i) - \mu(t_{i-1})$$

which reflects the fact that μ is non-decreasing. The condition $\nu_i^\top s(t_i, x_i) = 0$ implies

$$s(t_i, x_i) < 0_{n_s} \qquad \Rightarrow \qquad 0_{n_s} = \nu_i \approx \mu(t_i) - \mu(t_{i-1}),$$

which reflects that μ is constant on inactive arcs.

Remark 6.4.6

- *The above interpretations remain valid for problems with additional mixed control-state constraints and $\mathcal{U} = \mathbb{R}^{n_u}$, cf. Theorem 4.2.5. A comparison of the respective necessary conditions in Theorem 4.2.5 and Theorem 6.4.4 yields the additional relationship*

  $$\frac{\zeta_i}{h_{i-1}} \approx \eta(t_i), \qquad i = 1, \ldots, N$$

  for the multiplier η. The complementarity conditions for ζ_i are discrete versions of the complementarity condition in Theorem 4.2.5.

- *Instead of implicit Euler's method more general Runge-Kutta methods can be investigated provided that these are suitable for DAE optimal control problems. For ODE optimal control problems and explicit Euler's method the analogous results can be found in Gerdts [Ger05b].*

6.4.2 Example

The subsequent example shows that the above interpretations are meaningful. Since the problem is an ODE optimal control problem we use explicit Euler's method for discretization instead of implicit Euler's method. The above interpretations can be adapted accordingly. Though the example at a first glance is very simple it has the nice feature that one of the two state constraints becomes active only at the final time point. This causes the corresponding multiplier to jump only at the final time point. Correspondingly, the adjoint also jumps. It turns out that the above interpretations allow to construct approximations for the multipliers that reflect this behavior for the numerical solution.

Consider a system of two water boxes, where $x(t)$ and $y(t)$ denote the volume of water in the two boxes and $u(t)$ and $v(t)$ denote the outflow rate of water for the respective boxes at time t, cf. Figure 6.4.

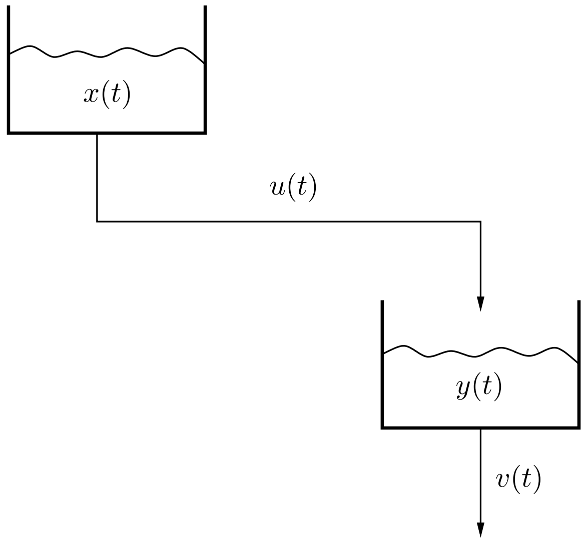


Figure 6.4: System of two water boxes with controllable outflow rates.

Assume, that the outflow rates and the volumes are restricted by $u(t), v(t) \in [0,1]$ and $x(t) \geq 0$, $y(t) \geq 0$, respectively. We consider the following linear optimal control problem.

Problem 6.4.7 *Minimize*

$$-\int_0^{10} (10-t)u(t) + tv(t)dt$$

subject to the ODE

$$\begin{aligned} \dot{x}(t) &= -u(t), \\ \dot{y}(t) &= u(t) - v(t), \end{aligned}$$

the initial conditions

$$x(0) = 4, \quad y(0) = 4,$$

the state constraints

$$x(t) \geq 0, \quad y(t) \geq 0 \quad \textit{in } [0,10],$$

and the control constraints

$$u(t) \in [0,1], \quad v(t) \in [0,1] \quad \textit{a.e. in } [0,10].$$

Evaluation of the local minimum principle 4.2.4 resp. 4.1.10 yields the following result.

Theorem 6.4.8 *The following functions satisfy the local minimum principle 4.2.4 for the linear optimal control problem 6.4.7 and hence are optimal since the problem is convex.*
The optimal control variables are given by

$$\hat{u}(t) = \begin{cases} 1, & \textit{if } 0 \leq t < 4, \\ 0, & \textit{if } 4 \leq t \leq 10, \end{cases}, \qquad \hat{v}(t) = \begin{cases} 0, & \textit{if } 0 \leq t < 2, \\ 1, & \textit{if } 2 \leq t \leq 10. \end{cases}$$

The optimal state variables are given by

$$\hat{x}(t) = \begin{cases} 4-t, & \textit{if } 0 \leq t < 4, \\ 0, & \textit{if } 4 \leq t \leq 10, \end{cases}, \qquad \hat{y}(t) = \begin{cases} 4+t, & \textit{if } 0 \leq t < 2, \\ 6, & \textit{if } 2 \leq t < 4, \\ 10-t, & \textit{if } 4 \leq t \leq 10. \end{cases}$$

The adjoints are given by

$$\hat{\lambda}_x(t) = \mu_x(t) - \mu_x(10), \quad \hat{\lambda}_y(t) = \mu_y(t) - \mu_y(10) = \begin{cases} -2, & \textit{if } 0 \leq t < 10, \\ 0, & \textit{if } t = 10, \end{cases}$$

where the multiplier μ_x is non-decreasing and satisfies $\mu_x(t) = 0$ for $t \in [0,4)$ and

$$\int_4^{10} d\mu_x(\tau) = \mu_x(10) - \mu_x(4) = 8, \quad \mu_x(10) - \mu_x(t) > 12 - t, \quad t \in (4,10).$$

The multiplier μ_y is given by

$$\hat{\mu}_y(t) = \begin{cases} 0, & \textit{if } 0 \leq t < 10, \\ 2, & \textit{if } t = 10. \end{cases}$$

Proof. We apply Theorem 4.2.4. The Hamilton function for Problem 6.4.7 is given by

$$\mathcal{H}(t,x,y,u,v,\lambda_x,\lambda_y,l_0) = -l_0\left((10-t)u + tv\right) - \lambda_x u + \lambda_y(u - v).$$

Let $(\hat{x}, \hat{y}, \hat{u}, \hat{v})$ be a local minimum of Problem 6.4.7. Then, by theorem 4.2.4 there exist multipliers $l_0 \geq 0$, $\sigma = (\sigma_x, \sigma_y)^\top \in \mathbb{R}^2$, $\lambda = (\lambda_x, \lambda_y)^\top \in BV([t_0,t_f],\mathbb{R}^2)$, and $\mu = (\mu_x,\mu_y)^\top \in NBV([t_0,t_f],\mathbb{R}^2)$ with

$$\begin{aligned} \Theta &\neq (l_0, \sigma, \lambda, \mu), \\ \lambda_x(t_f) &= 0, \\ \lambda_y(t_f) &= 0, \\ \lambda_x(t_0) &= -\sigma_x, \\ \lambda_y(t_0) &= -\sigma_y, \\ \lambda_x(t) &= \lambda_x(10) + \int_t^{10} \mathcal{H}'_x[\tau]d\tau - \int_t^{10} d\mu_x(\tau) = \mu_x(t) - \mu_x(10), \\ \lambda_y(t) &= \lambda_y(10) + \int_t^{10} \mathcal{H}'_y[\tau]d\tau - \int_t^{10} d\mu_y(\tau) = \mu_y(t) - \mu_y(10). \end{aligned}$$

Furthermore, a.e. in $[0,10]$ it holds

$$\begin{aligned} 0 &\leq \mathcal{H}'_u[t](u - \hat{u}(t)) = (-l_0(10-t) - \lambda_x(t) + \lambda_y(t))(u - \hat{u}(t)) \\ &= (-l_0(10-t) - \mu_x(t) + \mu_x(10) + \mu_y(t) - \mu_y(10))(u - \hat{u}(t)) \qquad \forall u \in [0,1], \\ 0 &\leq \mathcal{H}'_v[t](v - \hat{v}(t)) = (-l_0 t - \lambda_y(t))(v - \hat{v}(t)) \\ &= (-l_0 t - \mu_y(t) + \mu_y(10))(v - \hat{v}(t)) \qquad \forall v \in [0,1]. \end{aligned}$$

Finally, μ_x, μ_y are non-decreasing functions with $\mu_x(0) = \mu_y(0) = 0$. μ_x and μ_y are constant on intervals with measure greater than zero and $x(t) > 0$ and $y(t) > 0$, respectively.

First, notice that the case $l_0 = 0$ can be excluded, since the Mangasarian-Fromowitz conditions are easily being checked to be satisfied for this problem. Hence, we may set $l_0 = 1$.

From the optimality conditions we conclude

$$\begin{aligned} \hat{u}(t) &= \begin{cases} 0, & \text{if } \lambda_y(t) - \lambda_x(t) > 10 - t, \\ 1, & \text{if } \lambda_y(t) - \lambda_x(t) < 10 - t, \end{cases} \\ \hat{v}(t) &= \begin{cases} 1, & \text{if } -\lambda_y(t) < t, \\ 0, & \text{if } -\lambda_y(t) > t. \end{cases} \end{aligned}$$

The monotonicity properties of μ_x and μ_y imply that $\lambda_x(t) = \mu_x(t) - \mu_y(10) \leq 0$ and $\lambda_y(t) = \mu_y(t) - \mu_y(10) \leq 0$ are non-decreasing functions. Since the function t is strictly increasing and $-\lambda_y(t) \geq 0$ is monotonically decreasing, there exists at most one point $\hat{t} \in [0, 10]$ with $-\lambda_y(\hat{t}) = \hat{t}$. Hence, the control $\hat{v}$ is a bang-bang control with at most one switching point $\hat{t}$. This defines the structure of the control $\hat{v}$.
The first differential equation yields

$$\hat{x}(t) = 4 - \int_0^t \hat{u}(\tau)d\tau, \qquad 0 \leq t \leq 10.$$

Let us assume for a moment, that $\hat{x}(t) > 0$ holds for all $t \in [0, 10]$, i.e. the state constraint $x(t) \geq 0$ is inactive. Then, $\mu_x(t) \equiv 0$ and $\lambda_x(t) = \mu_x(t) - \mu_x(10) \equiv 0$. The optimality condition yields $\hat{u}(t) \equiv 1$, because $\lambda_y(t) - \lambda_x(t) = \lambda_y(t) \leq 0 \leq 10 - t$. But this contradicts $\hat{x}(t) > 0$ for all $t \in [0, 10]$. Hence, there exists a first point $\tilde{t} \in (0, 10]$ with $\hat{x}(\tilde{t}) = 0$. Once $\hat{x}(\tilde{t}) = 0$ is fulfilled, it holds $\hat{x}(t) = 0$ and $\hat{u}(t) = 0$ for all $t \in [\tilde{t}, 10]$ because of

$$0 \leq \hat{x}(t) = \hat{x}(\tilde{t}) - \int_{\tilde{t}}^t \hat{u}(\tau)d\tau = -\int_{\tilde{t}}^t \hat{u}(\tau)d\tau \leq 0.$$

Due to the optimality conditions this implies $\lambda_y(t) - \lambda_x(t) \geq 10 - t$ in $(\tilde{t}, 10]$.
Since $\tilde{t}$ is the first point satisfying $\hat{x}(\tilde{t}) = 0$ it holds $\hat{x}(t) > 0$ in $[0, \tilde{t})$ and thus $\mu_x(t) \equiv 0$ respectively $\lambda_x(t) \equiv -\mu_x(10)$ in $[0, \tilde{t})$. Hence, $\lambda_y(t) - \lambda_x(t)$ is non-decreasing in $[0, \tilde{t})$. Since $10 - t$ is strictly decreasing, there is at most one point $t_1 \in [0, \tilde{t})$ with $\lambda_y(t_1) - \lambda_x(t_1) = 10 - t_1$. Assume $t_1 < \tilde{t}$. Then, necessarily, $\hat{u}(t) = 0$ in $(t_1, \tilde{t})$ and thus $\hat{u}(t) = 0$ in $(t_1, 10]$ and $\hat{x}(t) \equiv \hat{x}(t_1)$ in $[t_1, 10]$. Then, either $\hat{x}(t_1) > 0$, which contradicts the existence of $\tilde{t}$, or $\hat{x}(t_1) = 0$, which contradicts the minimality of $\tilde{t}$. Consequently, there is no point t_1 in $[0, \tilde{t})$ with $\lambda_y(t_1) - \lambda_x(t_1) = 10 - t_1$. Therefore, either $\lambda_y(t) - \lambda_x(t) > 10 - t$ in $[0, \tilde{t})$, which implies $\hat{u}(t) = 0$ and contradicts the existence of $\tilde{t}$, or $\lambda_y(t) - \lambda_x(t) < 10 - t$ in $[0, \tilde{t})$, which implies $\hat{u}(t) = 1$ in $[0, \tilde{t})$. Summarizing, these considerations yield

$$\begin{aligned}
\tilde{t} &= 4, \\
\hat{u}(t) &= \begin{cases} 1, & \text{if } t \in [0, 4), \\ 0, & \text{if } t \in [4, 10], \end{cases} \\
\hat{x}(t) &= \begin{cases} 4 - t, & \text{if } t \in [0, 4), \\ 0, & \text{if } t \in [4, 10], \end{cases} \\
\mu_x(t) &= 0, \quad t \in [0, 4), \\
\lambda_x(t) = \mu_x(t) - \mu_x(10) &= -\mu_x(10), \quad t \in [0, 4), \\
\lambda_y(\tilde{t}) - \lambda_x(\tilde{t}) &= 10 - \tilde{t} = 6.
\end{aligned}$$

Now, we have to determine the switching point $\hat{t}$ of $\hat{v}$. It turns out that $\hat{t} = 2$ is the only possible choice. All other cases ($\hat{t}$ does not occur in $[0, 10]$, $\hat{t} \neq 2$) will lead to contradictions. Hence, we have

$$\hat{t} = 2, \quad \hat{v}(t) = \begin{cases} 0, & \text{if } t \in [0, 2), \\ 1, & \text{if } t \in [2, 10]. \end{cases}$$

Using these controls, it is easy to check that $\hat{y}$ becomes active exactly at $t = 10$. Hence,

$$\mu_y(t) = \begin{cases} 0, & \text{in } [0, 10), \\ \mu_y(10) \geq 0, & \text{if } t = 10, \end{cases} \qquad \lambda_y(t) = \mu_y(t) - \mu_y(10) = \begin{cases} -\mu_y(10), & \text{in } [0, 10), \\ 0, & \text{if } t = 10. \end{cases}$$

On the other hand we know already $\hat{t} = -\lambda_y(\hat{t})$ and hence $\mu_y(10) = 2$.

Moreover, $\mu_x(t) = 0$ in $[0,4)$ and hence

$$\lambda_x(t) = \mu_x(t) - \mu_x(10) = \begin{cases} -\mu_x(10), & \text{if } 0 \le t < 4, \\ \le 0, & \text{if } 4 \le t < 10, \\ 0, & \text{if } t = 10. \end{cases}$$

On the other hand we know already $6 = \lambda_y(\tilde{t}) - \lambda_x(\tilde{t}) = -2 - \lambda_x(\tilde{t})$ and hence $\lambda_x(\tilde{t}) = \mu_x(\tilde{t}) - \mu_x(10) = -\int_{\tilde{t}}^{10} d\mu_x(\tau) = -8$. Finally, since $\hat{u}(t) \equiv 0$ in $[4,10]$ it follows $\lambda_y(t) - \lambda_x(t) = -2 - \lambda_x(t) > 10 - t$ in $(4,10)$, i.e. $\mu_x(10) - \mu_x(t) > 12 - t$ in $(0,4)$. ■

In the sequel, the discretization of Problem 6.4.7 by explicit Euler's method is investigated in detail. Application of explicit Euler's method with constant step size $h = (t_f - t_0)/N$ and equidistant grid points $t_i = t_0 + ih$, $i = 0, \ldots, N$ to Problem 6.4.7 leads to

Problem 6.4.9 (Discretized Problem)
Minimize

$$-h \sum_{i=0}^{N-1} (10 - t_i)u_i + t_i v_i$$

subject to the constraints

$$\begin{array}{rcll} x_0 & = & 4, & \\ x_{i+1} & = & x_i - hu_i, & i = 0,1,\ldots,N-1, \\ y_0 & = & 4, & \\ y_{i+1} & = & y_i + hu_i - hv_i, & i = 0,1,\ldots,N-1, \\ x_i & \ge & 0, & i = 0,1,\ldots,N, \\ y_i & \ge & 0, & i = 0,1,\ldots,N, \\ u_i & \in & [0,1], & i = 0,1,\ldots,N-1, \\ v_i & \in & [0,1], & i = 0,1,\ldots,N-1. \end{array}$$

The Lagrange function of Problem 6.4.9 with $x = (x_1, \ldots, x_N)^\top$, $y = (y_1, \ldots, y_N)^\top$, $u = (u_0, \ldots, u_{N-1})^\top$, $v = (v_0, \ldots, v_{N-1})^\top$, $\lambda^x = (\lambda_1^x, \ldots, \lambda_N^x)^\top$, $\lambda^y = (\lambda_1^y, \ldots, \lambda_N^y)^\top$, $\mu^x = (\mu_1^x, \ldots, \mu_N^x)^\top$, $\mu^y = (\mu_1^y, \ldots, \mu_N^y)^\top$, is given by

$$\begin{aligned} L(x,y,u,v,\lambda^x,\lambda^y,\mu^x,\mu^y) \;=\; & -h \sum_{i=0}^{N-1} \left[(10 - t_i)u_i + t_i v_i\right] \\ & + \sum_{i=0}^{N-1} \lambda_{i+1}^x (x_i - hu_i - x_{i+1}) \\ & + \sum_{i=0}^{N-1} \lambda_{i+1}^y (y_i + hu_i - hv_i - y_{i+1}) \\ & + \sum_{i=0}^{N} \mu_i^x (-x_i) + \sum_{i=0}^{N} \mu_i^y (-y_i). \end{aligned}$$

Notice, that $x_0 = 4$ and $y_0 = 4$ are not considered as constraints in Problem 6.4.9.
We intend to evaluate the Fritz-John conditions in Theorem 3.6.2 for Problem 6.4.9. Notice, that we can choose $l_0 = 1$, since Problem 6.4.9 is linear. Theorem 3.6.2 with $l_0 = 1$ yields the following necessary (and in this case also sufficient) optimality conditions for an optimal solution $\hat{x}, \hat{y}, \hat{u}, \hat{v}$ with multipliers $\lambda^x, \lambda^y, \mu^x, \mu^y$:

(i) For $i = 1, \ldots, N$ it holds $0 = L'_{x_i}(\hat{x}, \hat{y}, \hat{u}, \hat{v}, \lambda^x, \lambda^y, \mu^x, \mu^y)$, i.e.

$$\begin{array}{rcl} 0 & = & \lambda_{i+1}^x - \lambda_i^x - \mu_i^x, \quad i = 1, \ldots, N-1, \\ 0 & = & \lambda_N^x - \mu_N^x. \end{array}$$

Recursive evaluation leads to

$$\lambda_i^x = -\sum_{j=i}^{N} \mu_j^x, \quad i = 1, \ldots, N.$$

(ii) For $i = 1, \ldots, N$ it holds $0 = L'_{y_i}(\hat{x}, \hat{y}, \hat{u}, \hat{v}, \lambda^x, \lambda^y, \mu^x, \mu^y)$, i.e.

$$\begin{array}{rcl} 0 & = & \lambda_{i+1}^y - \lambda_i^y - \mu_i^y, \quad i = 1, \ldots, N-1, \\ 0 & = & \lambda_N^y - \mu_N^y. \end{array}$$

Recursive evaluation leads to

$$\lambda_i^y = -\sum_{j=i}^{N} \mu_j^y, \quad i = 1, \ldots, N.$$

(iii) For $i = 0, \ldots, N$ it holds $L'_{u_i}(\hat{x}, \hat{y}, \hat{u}, \hat{v}, \lambda^x, \lambda^y, \mu^x, \mu^y)(u - \hat{u}_i) \geq 0$ for all $u \in [0, 1]$, i.e.

$$\left[-h(10 - t_i) - h\lambda_{i+1}^x + h\lambda_{i+1}^y\right](u - \hat{u}_i) \geq 0,$$

respectively

$$\left[-h(10 - t_i) + h \sum_{j=i+1}^{N} \left(\mu_j^x - \mu_j^y\right)\right] u \geq \left[-h(10 - t_i) + h \sum_{j=i+1}^{N} \left(\mu_j^x - \mu_j^y\right)\right] \hat{u}_i$$

for all $u \in [0, 1]$. This implies

$$\hat{u}_i = \begin{cases} 1, & \text{if } -(10 - t_i) + \sum_{j=i+1}^{N}(\mu_j^x - \mu_j^y) < 0, \\ 0, & \text{if } -(10 - t_i) + \sum_{j=i+1}^{N}(\mu_j^x - \mu_j^y) > 0, \\ \text{undefined}, & \text{otherwise.} \end{cases} \tag{6.4.26}$$

(iv) For $i = 0, \ldots, N$ it holds $L'_{v_i}(\hat{x}, \hat{y}, \hat{u}, \hat{v}, \lambda^x, \lambda^y, \mu^x, \mu^y)(v - \hat{v}_i) \geq 0$ for all $v \in [0, 1]$, i.e.

$$\left[-ht_i - h\lambda_{i+1}^y\right](v - \hat{v}_i) \geq 0,$$

respectively

$$\left[-ht_i + h \sum_{j=i+1}^{N} \mu_j^y\right] v \geq \left[-ht_i + h \sum_{j=i+1}^{N} \mu_j^y\right] \hat{v}_i$$

for all $v \in [0, 1]$. This implies

$$\hat{v}_i = \begin{cases} 1, & \text{if } -t_i + \sum_{j=i+1}^{N} \mu_j^y < 0, \\ 0, & \text{if } -t_i + \sum_{j=i+1}^{N} \mu_j^y > 0, \\ \text{undefined}, & \text{otherwise.} \end{cases} \tag{6.4.27}$$

(v) It holds $\mu_i^x x_i = 0$, $\mu_i^x \geq 0$, $i = 0, \ldots, N$, and $\mu_i^y y_i = 0$, $\mu_i^y \geq 0$, $i = 0, \ldots, N$.

We will check that

$$\hat{u}_i = \begin{cases} 1, & \text{for } i = 0, \ldots, m-1, \\ 4/h - m, & \text{for } i = m, \\ 0, & \text{for } i = m+1, \ldots, N-1, \end{cases} \tag{6.4.28}$$

$$\hat{v}_i = \begin{cases} 0, & \text{for } i = 0, \ldots, q-1, \\ q+1-2/h, & \text{for } i = q, \\ 1, & \text{for } i = q+1, \ldots, N-1, \end{cases} \tag{6.4.29}$$

with

$$m = \left\lfloor \frac{4}{h} \right\rfloor, \qquad q = \left\lfloor \frac{2}{h} \right\rfloor,$$

is an optimal control for the discretized problem 6.4.9. By application of these controls we find

$$\begin{aligned} \hat{x}_i &= 4 - h\sum_{j=0}^{i-1} \hat{u}_j = 4 - ih > 0, \qquad i = 0, \ldots, m, \\ \hat{x}_{m+1} &= x_m - h\hat{u}_m = 4 - mh - h(4/h - m) = 0, \\ \hat{x}_i &= 0, \quad i = m+2, \ldots, N. \end{aligned}$$

Similarly, we have

$$\begin{aligned} \hat{y}_i &= 4 + h\sum_{j=0}^{i-1}(\hat{u}_j - \hat{v}_j) = 4 + ih, \qquad i = 0, \ldots, q, \\ \hat{y}_{q+1} &= \hat{y}_q + h(\hat{u}_q - \hat{v}_q) = 4 + qh + h(1 - (q+1-2/h)) = 6, \\ \hat{y}_i &= 6, \qquad i = q+2, \ldots, m, \\ \hat{y}_{m+1} &= \hat{y}_m + h(\hat{u}_m - \hat{v}_m) = 6 + h(4/h - m - 1) = 10 - h(m+1), \\ \hat{y}_i &= \hat{y}_{m+1} + h\sum_{j=m+1}^{i-1}(\hat{u}_j - \hat{v}_j) \\ &= 10 - h(m+1) - h(i-1-m) = 10 - ih, \quad i = m+2, \ldots, N. \end{aligned}$$

Notice, that $\hat{y}_i > 0$ for $i = 0, \ldots, N-1$ and $\hat{y}_N = 0$. Hence, according the complementarity conditions (v) the multipliers μ_i^y must satisfy $\mu_i^y = 0$ for $i = 1, \ldots, N-1$. With (6.4.27) and (6.4.29) necessarily it holds

$$\mu_N^y \in (t_{q-1}, t_q).$$

Taking the limit $h \to 0$ we find $t_q = qh = \lfloor 2/h \rfloor h \to 2$ and $t_{q-1} \to 2$ and thus $\mu_N^y \to 2$. According to the complementarity conditions (v) the multipliers μ_i^x must satisfy $\mu_i^x = 0$ for $i = 1, \ldots, m$. With (6.4.26) and (6.4.28) necessarily it holds

$$\begin{aligned} -(10 - t_{m-1}) + \sum_{j=m}^{N}\left(\mu_j^x - \mu_j^y\right) &< 0, \\ -(10 - t_m) + \sum_{j=m+1}^{N}\left(\mu_j^x - \mu_j^y\right) &> 0, \end{aligned}$$

and thus

$$\mu_N^y + (10 - t_m) < \sum_{j=m}^{N} \mu_j^x < \mu_N^y + (10 - t_{m-1}).$$

Taking the limit $h \to 0$ yields $t_m = mh = \lfloor 4/h \rfloor h \to 4$ and

$$\sum_{j=m}^{N} \mu_j^x \to 8.$$

Summarizing, the above considerations showed that (6.4.28), (6.4.29), and the resulting discrete state variables are optimal solutions of the discretized problem provided the multipliers μ^x and μ^y are chosen accordingly. Furthermore, the switching points t_q and t_m converge at a linear rate to the switching points of the continuous problem 6.4.7. The discrete states (viewed as continuous, piecewise linear functions) converge for the norm $\|\cdot\|_{1,\infty}$, whereas the discrete controls (viewed as piecewise constant functions) do not converge to the continuous controls for the norm $\|\cdot\|_\infty$. However, the discrete controls do converge for the norm $\|\cdot\|_1$. Due to the non-uniqueness of the continuous and discrete multipliers it is hard to observe convergence for these quantities.

Figure 6.5 shows the numerical solution for $N = 999$ and piecewise constant control approximation. The numerical solution was computed by the software package SODAS, cf. Gerdts [Ger01b].

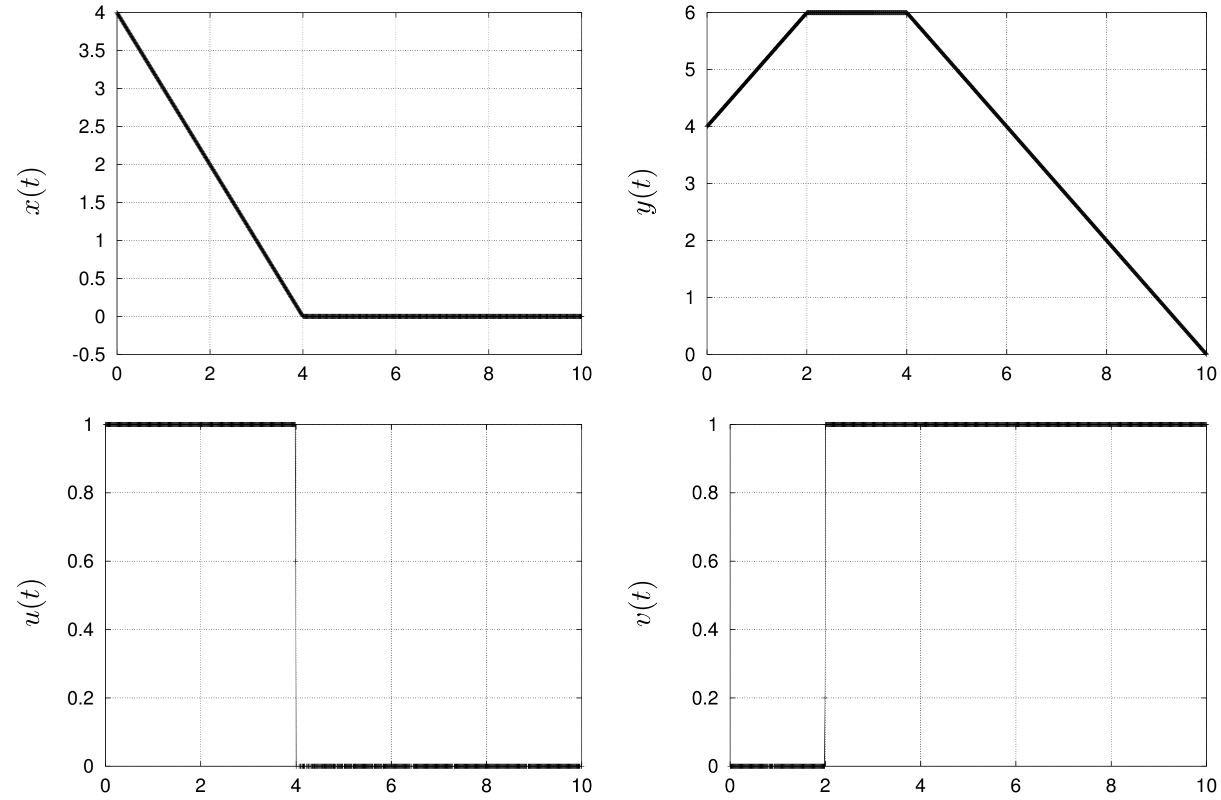


Figure 6.5: Optimal approximate state and control for $N = 999$ grid points and piecewise constant control approximation.

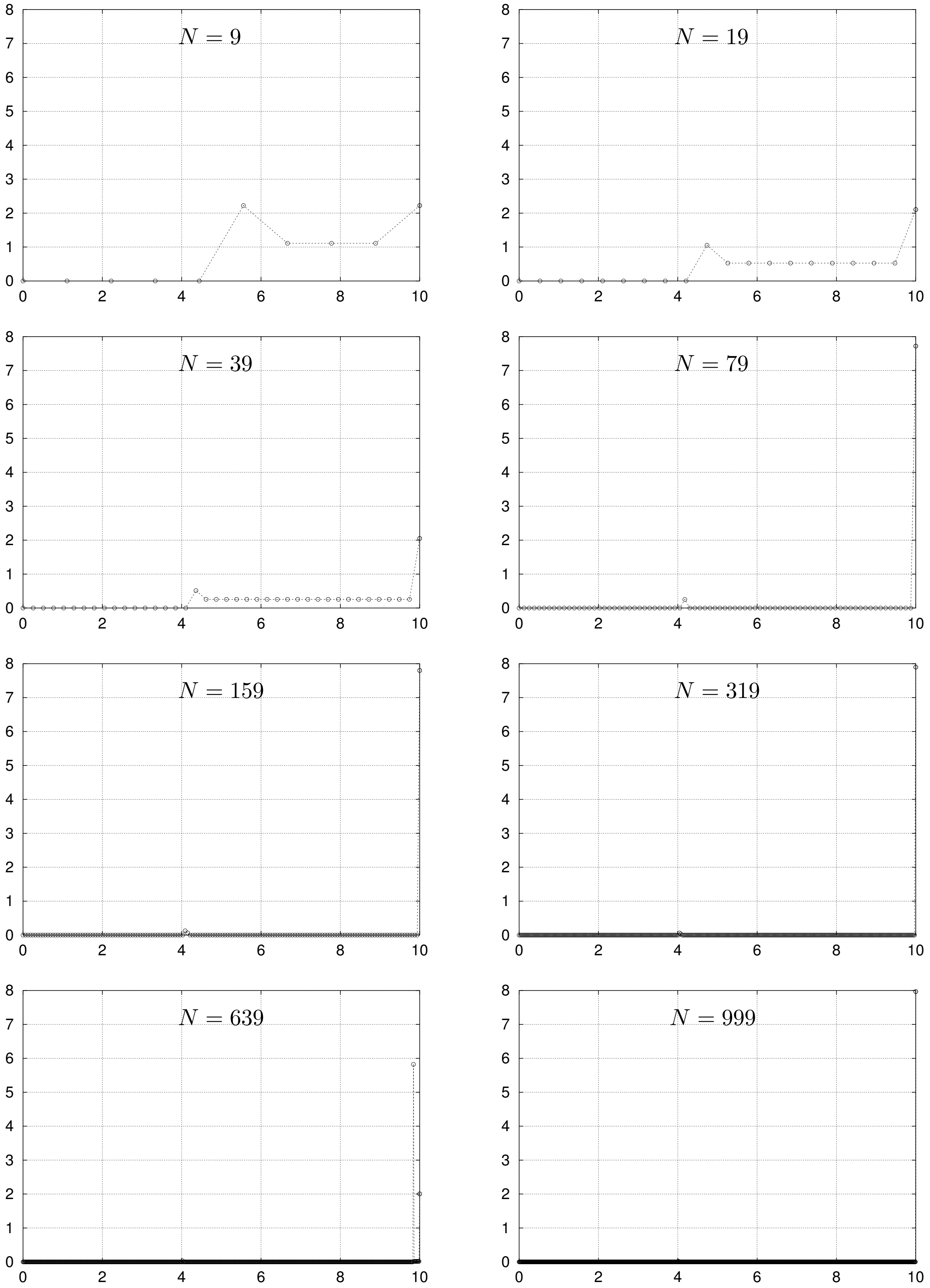


Figure 6.6: Lagrange multiplier μ^x for the discretized state constraint $x_i \geq 0$ in Problem 6.4.9 for $N = 9, 19, 39, 79, 159, 319, 639, 999$.

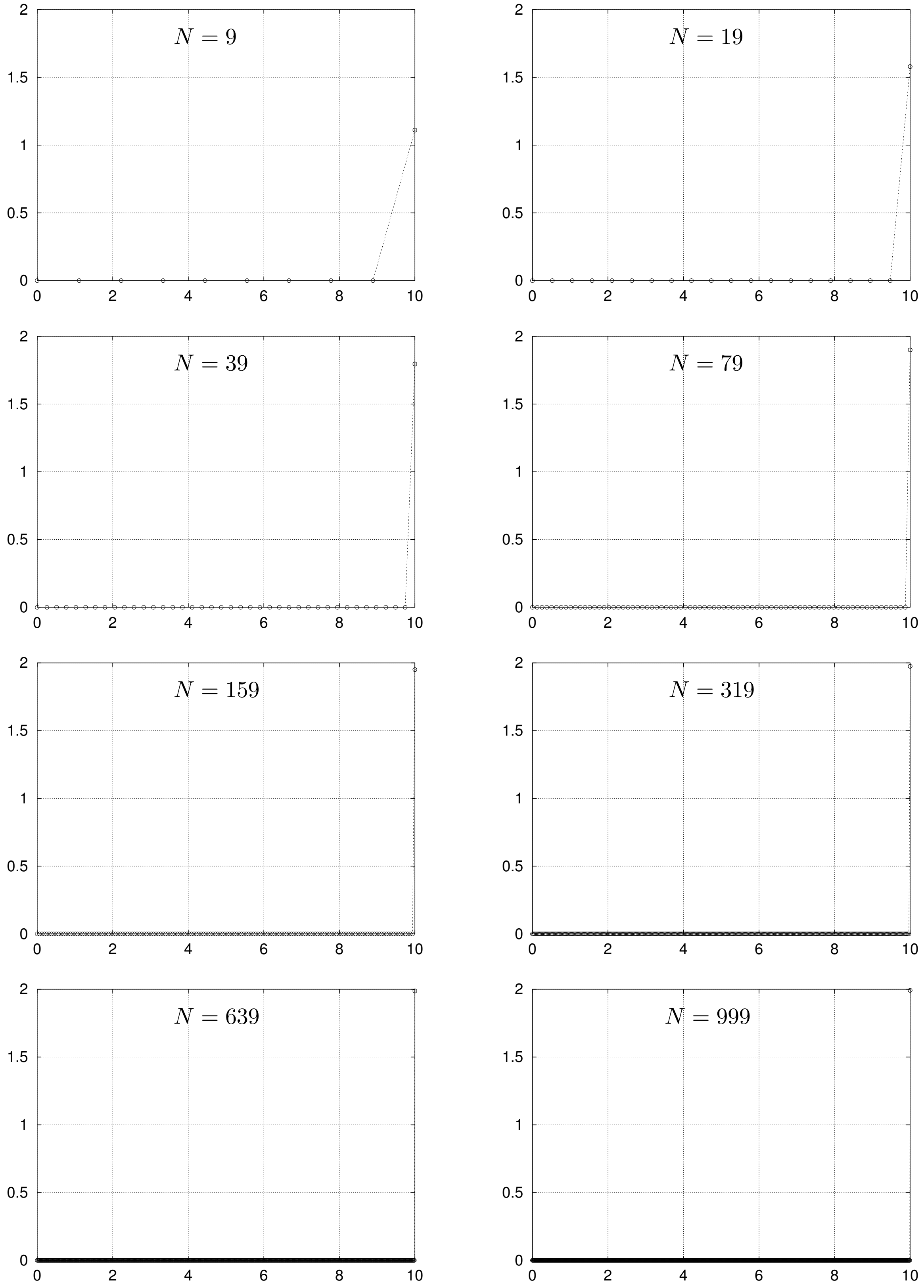


Figure 6.7: Lagrange multiplier μ^y for the discretized state constraint $y_i \geq 0$ in Problem 6.4.9 for $N = 9, 19, 39, 79, 159, 319, 639, 999$.

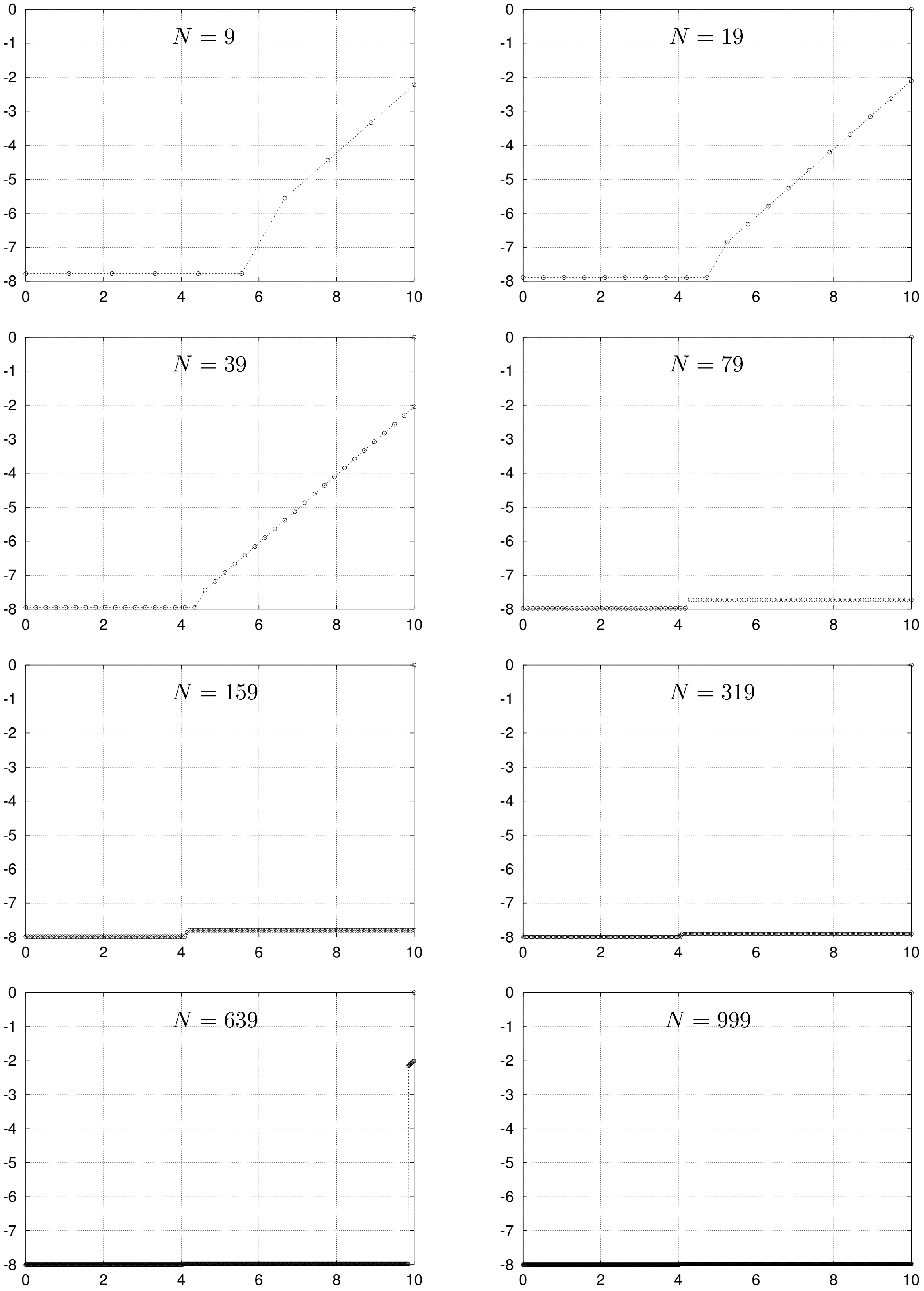


Figure 6.8: Adjoint estimation λ^x for the discretized differential equation in Problem 6.4.9 for $N = 9, 19, 39, 79, 159, 319, 639, 999$.

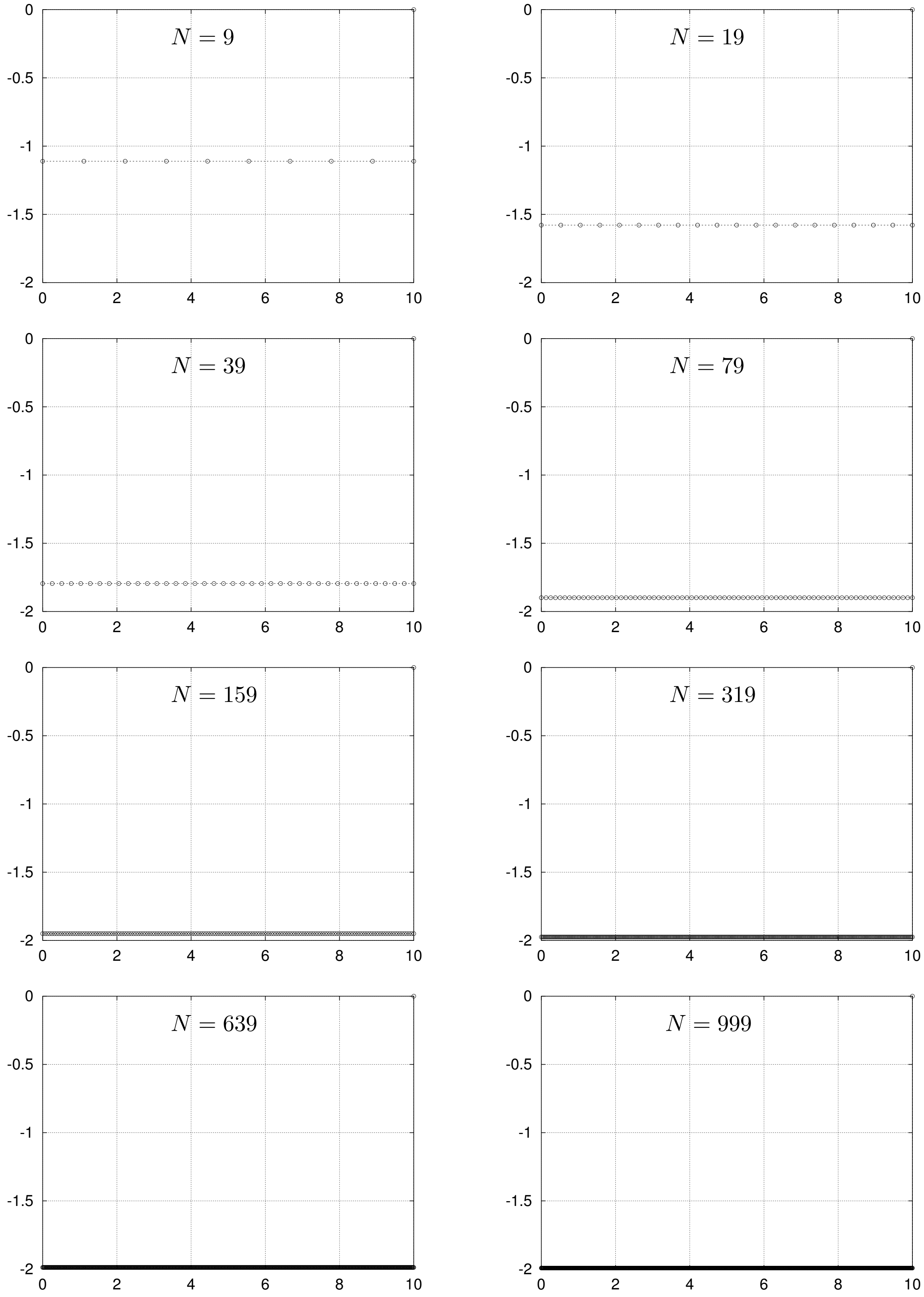


Figure 6.9: Adjoint estimation λ^y for the discretized differential equation in Problem 6.4.9 for $N = 9, 19, 39, 79, 159, 319, 639, 999$.

6.5 Convergence

The convergence of discretized optimal control problems is a current field of research. Only few results are available for ODE optimal control problems. We summarize the existing results without proving them. All results assume that the optimal control is at least continuous. Up to now, the author is not aware of any results considering discontinuous optimal controls or ODE optimal control problems subject to pure state constraints. Hence, a lot of research has to be done in this direction!

6.5.1 Convergence of the Euler Discretization

The proof of convergence for the full Euler discretization can be found in Malanowski et al. [MBM97]. More specifically, the authors investigate the optimal control problem

$$\begin{array}{ll} \min & \varphi(x(t_f)) + \int_0^{t_f} f_0(x(t),u(t))dt \\ \text{s.t.} & \dot{x}(t) - f(x(t),u(t)) = 0_{n_x}, \\ & x(0) - \xi = 0_{n_x}, \quad \psi(x(t_f)) = 0_{n_\psi}, \\ & c(x(t),u(t)) \leq 0_{n_c}, \end{array}$$

and its discretization by Euler's method:

$$\begin{array}{lll} \min & \varphi(x_N) + h\sum_{i=0}^{N-1} f_0(x_i,u_i) & \\ \text{s.t.} & \dfrac{x_{i+1}-x_i}{h} - f(x_i,u_i) = 0_{n_x}, & i = 0,\ldots,N-1, \\ & x_0 - \xi = 0_{n_x}, \quad \psi(x_N) = 0_{n_\psi}, & \\ & c(x_i,u_i) \leq 0_{n_c}, & i = 0,\ldots,N-1. \end{array}$$

We only summarize the assumptions needed to prove a convergence result. As usual $\hat{\mathcal{H}}$ denotes the augmented Hamilton function.

Assumption 6.5.1

(i) *f_0, f, c, φ, ψ are differentiable with locally Lipschitz continuous derivatives.*

(ii) *There exists a local solution $(\hat{x},\hat{u}) \in C^1([0,t_f],\mathbb{R}^{n_x}) \times C([0,t_f],\mathbb{R}^{n_u})$ of the optimal control problem.*

(iii) Uniform rank condition for *c:*
Let $\mathcal{J}(t) := \{i \in \{1,\ldots,n_c\} \mid c_i(\hat{x}(t),\hat{u}(t)) = 0\}$ denote the index set of active mixed control-state constraints at t and $c_{\mathcal{J}(t)}[t]$ the active constraints at t. Let there exists a constant $\alpha > 0$ with

$$\|c'_{\mathcal{J}(t),u}[t]^\top d\| \geq \alpha\|d\| \qquad \forall d \in \mathbb{R}^{|\mathcal{J}(t)|} \text{ a.e. in } [0,t_f].$$

(iv) Surjectivity of the linearized equality constraints:
Let the boundary value problem

$$\dot{y}(t) - \tilde{A}(t)y(t) - \tilde{B}(t)v(t) = 0_{n_x}, \quad y(0) = 0_{n_x}, \quad \psi'_{x_f}(\hat{x}(t_f))y(t_f) = h$$

possess a solution for any $h \in \mathbb{R}^{n_\psi}$, where

$$\begin{aligned}\tilde{A}(t) &= f_x'[t] - f_u'[t] c_{\mathcal{J}(t),u}'[t]^\top \left(c_{\mathcal{J}(t),u}'[t] c_{\mathcal{J}(t),u}'[t]^\top \right)^{-1} c_{\mathcal{J}(t),x}'[t], \\ \tilde{B}(t) &= f_u'[t] \left(I_{n_u} - c_{\mathcal{J}(t),u}'[t]^\top \left(c_{\mathcal{J}(t),u}'[t] c_{\mathcal{J}(t),u}'[t]^\top \right)^{-1} c_{\mathcal{J}(t),x}'[t] \right).\end{aligned}$$

(v) Coercivity:
Define $\mathcal{J}^+(t) := \{i \in \mathcal{J}(t) \mid \eta_i(t) > 0\}$, where η denotes the multiplier for the mixed control-state constraint c.

Let there exist $\beta > 0$ such that

$$d^\top \hat{\mathcal{H}}_{uu}''[t] d \geq \beta \|d\|^2$$

holds for all $d \in \mathbb{R}^{n_u}$ with

$$c_{J^+(t),u}'[t] d = 0_{|\mathcal{J}^+(t)|}.$$

(vi) Riccati equation:
Let the Riccati equation

$$\begin{aligned}\dot{Q}(t) &= -Q(t) f_x'[t] - f_x'[t]^\top Q(t) - \hat{\mathcal{H}}_{xx}''[t] \\ &\quad + \left[\begin{pmatrix} \hat{\mathcal{H}}_{ux}''[t] \\ c_{J^+(t),x}'[t] \end{pmatrix}^\top + Q(t) \begin{pmatrix} f_u'[t]^\top \\ \Theta \end{pmatrix}^\top \right] \begin{pmatrix} \hat{\mathcal{H}}_{uu}''[t] & c_{J^+(t),u}'[t]^\top \\ c_{J^+(t),u}'[t] & \Theta \end{pmatrix} \cdot \\ &\quad \cdot \left[\begin{pmatrix} f_u'[t]^\top \\ \Theta \end{pmatrix} Q(t) + \begin{pmatrix} \hat{\mathcal{H}}_{ux}''[t] \\ c_{J^+(t),x}'[t] \end{pmatrix} \right]\end{aligned}$$

possess a bounded solution Q on $[0, t_f]$ that satisfies the rank condition

$$d^\top (\Gamma - Q(t_f)) d \geq 0 \qquad \forall d \in \mathbb{R}^{n_x} \;:\; \psi_{x_f}'(\hat{x}(t_f)) d = 0_{n_\psi},$$

where

$$\Gamma := \left(\varphi(\hat{x}(t_f)) + \sigma_f^\top \psi(\hat{x}(t_f)) \right)_{xx}''.$$

All function evaluations are at the optimal solution $(\hat{x}, \hat{u})$.

Theorem 6.5.2 *Let the assumptions (i)-(vi) in Assumption 6.5.1 be satisfied. Then, for sufficiently small step-sizes $h > 0$ there exist a locally unique KKT point $(x_h, u_h, \lambda_h, \zeta_h, \kappa_0, \kappa_f)$ of the discretized problem satisfying*

$$\max\{\|x_h - \hat{x}\|_{1,\infty}, \|u_h - \hat{u}\|_\infty, \|\lambda_h - \lambda\|_{1,\infty}, \|\kappa_0 - \sigma_0\|, \|\kappa_f - \sigma_f\|, \|\eta_h - \eta\|_\infty\} = \mathcal{O}(h),$$

where λ_h denotes the discrete adjoint, η_h the discrete multiplier for the mixed control-state constraint, κ_0 the discrete multiplier for the initial condition, and κ_f the discrete multiplier for the final condition.

Proof. The long and complicated proof can be found in Malanowski et al. [MBM97]. ■

Remark 6.5.3

- *The assumptions (v) and (vi) together are sufficient for local optimality of $(\hat{x}, \hat{u})$.*
- *Similar convergence results can be found in Dontchev et al. [DHV00, DHM00].*

6.5.2 Higher Order of Convergence for Runge-Kutta Discretizations

Hager [Hag00] investigates optimal control problems of type

$$\begin{array}{ll} \min & \varphi(x(1)) \\ \text{s.t.} & \dot{x}(t) = f(x(t), u(t)), \quad x(0) = \xi, \quad t \in [0,1], \end{array}$$

and their discretization by s-staged Runge-Kutta methods with fixed step-size $h = 1/N$:

$$\begin{array}{ll} \min & \varphi(x_N) \\ \text{s.t.} & \dfrac{x_{k+1} - x_k}{h} = \displaystyle\sum_{i=1}^{s} b_i f(\eta_i, u_{ki}), \quad k = 0, \ldots, N-1, \\ & \eta_i = x_k + h \displaystyle\sum_{j=1}^{s} a_{ij} f(\eta_j, u_{kj}), \quad i = 1, \ldots, s, \\ & x_0 = \xi. \end{array}$$

Notice, that for each function evaluation of f an independent optimization variable u_{ki} is introduced. Hence, there is no coupling by, e.g., piecewise constant or linear interpolation.
The main result of Hager's article is a convergence result for the discrete solutions assuming the following:

Assumption 6.5.4

(i) Smoothness:
The optimal control problem possesses a solution $(\hat{x}, \hat{u}) \in W^{p,\infty}([0,1], \mathbb{R}^{n_x}) \times W^{p-1,\infty}([0,1], \mathbb{R}^{n_u})$ *with* $p \geq 2$.

The first p *derivatives of* f *and* φ *are supposed to be locally Lipschitz continuous in some neighborhood of* $(\hat{x}, \hat{u})$.

(ii) Coercivity:
There exists some $\alpha > 0$ *with*

$$\mathcal{B}(x,u) \geq \alpha \|u\|_2^2 \qquad \forall (x,u) \in \mathcal{M}$$

where

$$\mathcal{B}(x,u) = \frac{1}{2}\left(x(1)^\top V x(1) + \int_0^1 x(t)^\top Q(t) x(t) + 2x(t)^\top S(t) u(t) + u(t)^\top R(t) u(t) dt \right)$$

and

$$\begin{aligned} A(t) &:= f_x'(\hat{x}(t), \hat{u}(t)), & B(t) &:= f_u'(\hat{x}(t), \hat{u}(t)), \\ V &:= \varphi''(\hat{x}(1)), & Q(t) &:= \mathcal{H}_{xx}''(\hat{x}(t), \hat{u}(t), \lambda(t)), \\ R(t) &:= \mathcal{H}_{uu}''(\hat{x}(t), \hat{u}(t), \lambda(t)), & S(t) &:= \mathcal{H}_{xu}''(\hat{x}(t), \hat{u}(t), \lambda(t)), \end{aligned}$$

and

$$\mathcal{M} = \left\{ (x,u) \in W^{1,2}([0,1], \mathbb{R}^{n_x}) \times L^2([0,1], \mathbb{R}^{n_u}) \;\middle|\; \dot{x} = Ax + Bu, \ x(0) = 0 \right\}.$$

The smoothness property in (i) implies that the optimal control $\hat{u}$ is at least continuous.
It turns out that the well-known order conditions for Runge-Kutta methods for initial value problems are not enough to ensure higher order convergence for discretized optimal control problems. To ensure higher order convergence for the discretized optimal control problem the Runge-Kutta method has to satisfy the stronger conditions in Table 6.1. To distinguish these conditions from the ordinary conditions for IVP's we refer to the conditions in Table 6.1 as *optimal control order conditions*. A closer investigation yields that they are identical with the IVP conditions only up to order $p = 2$.

Table 6.1: Optimal control order conditions for Runge-Kutta methods up to order 4 (higher order conditions require the validity of all conditions of lower order)

Order	Conditions ($c_i = \sum a_{ij}$, $d_j = \sum b_i a_{ij}$)
$p = 1$	$\sum b_i = 1$
$p = 2$	$\sum d_i = \frac{1}{2}$
$p = 3$	$\sum c_i d_i = \frac{1}{6}$, $\sum b_i c_i^2 = \frac{1}{3}$, $\sum \frac{d_i^2}{b_i} = \frac{1}{3}$
$p = 4$	$\sum b_i c_i^3 = \frac{1}{4}$, $\sum b_i c_i a_{ij} c_j = \frac{1}{8}$, $\sum d_i c_i^2 = \frac{1}{12}$, $\sum d_i a_{ij} c_j = \frac{1}{24}$,
	$\sum \frac{c_i d_i^2}{b_i} = \frac{1}{12}$, $\sum \frac{d_i^3}{b_i^2} = \frac{1}{4}$, $\sum \frac{b_i c_i a_{ij} d_j}{b_j} = \frac{5}{24}$, $\sum \frac{d_i a_{ij} d_j}{b_j} = \frac{1}{8}$

Theorem 6.5.5 *Let the assumptions (i)–(ii) in Assumption 6.5.4 be satisfied. Suppose $b_i > 0$, $i = 1, \ldots, s$ holds for the coefficients of the Runge-Kutta method. Let the Runge-Kutta method be of optimal control order p, cf. Table 6.1.*
Then, for any sufficiently small step-size $h > 0$ there exists a strict local minimum of the discretized optimal control problem.
If $\frac{d^{p-1}\hat{u}}{dt^{p-1}}$ is of bounded variation then

$$\max_{0 \leq k \leq N} \{\|x_k - \hat{x}(t_k)\| + \|\lambda_k - \lambda(t_k)\| + \|u^*(x_k, \lambda_k) - \hat{u}(t_k)\|\} = \mathcal{O}(h^p).$$

If $\frac{d^{p-1}\hat{u}}{dt^{p-1}}$ is Riemann-integrable then

$$\max_{0 \leq k \leq N} \{\|x_k - \hat{x}(t_k)\| + \|\lambda_k - \lambda(t_k)\| + \|u^*(x_k, \lambda_k) - \hat{u}(t_k)\|\} = o(h^{p-1}).$$

Herein, $u^*(x_k, \lambda_k)$ denotes a local minimum of the Hamilton function $\mathcal{H}(x_k, u, \lambda_k)$ w.r.t. u.

Proof. The long and complicated proof can be found in Hager [Hag00]. ■

Remark 6.5.6 *The assumption $b_i > 0$ is essential as Example 6.1.5 shows.*

Chapter 7

Selected Applications and Extensions

This chapter shows in an exemplified way several applications and extensions of the direct discretization method of Chapter 6. Within each application it is necessary to solve dynamic optimization problems.

7.1 Mixed-Integer Optimal Control

Many technical or economical processes lead to optimal control problems involving so-called 'continuous' controls $u \in \mathcal{U}$ as well as 'discrete' controls $v \in \mathcal{V}$. Herein, the control set $\mathcal{U}$ is assumed to be a closed convex subset of $\mathbb{R}^{n_u}$ with nonempty interior whereas the control set $\mathcal{V}$ is a discrete set with finitely many elements defined by

$$\mathcal{V} = \{v_1, \dots, v_M\} \qquad v_i \in \mathbb{R}^{n_v},\ M \in \mathbb{N}.$$

We consider

Problem 7.1.1 (Mixed-Integer Optimal Control Problem (MIOCP))

$$\begin{array}{lrcll} \textit{Minimize} & \multicolumn{4}{l}{\varphi(x(t_0), x(t_f))} \\ \textit{w.r.t.} & \multicolumn{4}{l}{x \in W^{1,\infty}([t_0,t_f],\mathbb{R}^{n_x}), u \in L^{\infty}([t_0,t_f],\mathbb{R}^{n_u}), v \in L^{\infty}([t_0,t_f],\mathbb{R}^{n_v})} \\ \textit{s.t.} & \dot{x}(t) - f(t,x(t),u(t),v(t)) & = & 0 & \textit{a.e. in } [t_0,t_f], \\ & \psi(x(t_0),x(t_f)) & = & 0, & \\ & s(t,x(t)) & \leq & 0 & \textit{for all } t \in [t_0,t_f], \\ & u(t) & \in & \mathcal{U} & \textit{for a.e. } t \in [t_0,t_f], \\ & v(t) & \in & \mathcal{V} & \textit{for a.e. } t \in [t_0,t_f]. \end{array}$$

The functions $\varphi : \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \to \mathbb{R}$, $f : [t_0,t_f] \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \times \mathcal{V} \to \mathbb{R}^{n_x}$, $\psi : \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \to \mathbb{R}^{n_\psi}$, $s : [t_0,t_f] \times \mathbb{R}^{n_x} \to \mathbb{R}^{n_s}$ are supposed to be sufficiently smooth w.r.t. to all arguments. Without loss of generality, it is assumed, that the initial time t_0 and the final time t_f are fixed. Furthermore, the restriction to ODEs is not essential. The following methods can be extended directly to DAEs.

One way to solve the above optimal control problem is to formulate and solve necessary conditions provided by the well-known global minimum principle, cf. Ioffe and Tihomirov [IT79], Theorem 1, p. 234 and Girsanov [Gir72]. Unfortunately, for more complicated problems this approach usually turns out to be cumbersome and complex. Hence, we will concentrate on alternative approaches.

7.1.1 Branch&Bound

Another way, which was followed in Gerdts [Ger05e], is to apply the direct discretization approach of Chapter 6 to MIOCP. This approach leads to a large scale finite dimensional nonlinear mixed-integer programming problem of type 3.10.1. This mixed-integer programming problem is solved by a Branch&Bound method as described in Section 3.10. The Branch&Bound method itself requires to solve discretized optimal control problems with relaxed control sets as subproblems.

7.1.2 Variable Time Transformation Method

The variable time transformation to be considered in this section was used by Dubovitskii and Milyutin in a very elegant and comparatively easy way to derive a global minimum principle for arbitrary control sets $\mathcal{U}$ from a local one being valid only for convex control sets with nonempty interior, cf. Ioffe and Tihomirov [IT79], p. 148. Interestingly, the same variable time transformation can be used numerically to solve optimal control problems with purely discrete control set, cf. Teo et al. [TJL99], Lee et al. [LTRJ97, LTRJ99], and Siburian [Sib04]. In Lee et al. [LTRJ97] a similar approach was used to solve time optimal control problems. Siburian and Rehbock [SR04] discussed a procedure for solving singular optimal control problems. In Lee et al. [LTC98] a method for solving nonlinear mixed-integer programming problems based on a suitable formulation of an equivalent optimal control problem was introduced. We will extend their results to problems with mixed-integer control sets, cf. Gerdts [Ger06].

The variable time transformation method is based on a discretization. Therefore we introduce a main grid

$$\mathbb{G}_h \ : \ t_i = t_0 + ih, \qquad i = 0,\ldots,N, \ h = \frac{t_f - t_0}{N}$$

with $N \in \mathbb{N}$ intervals and a minor grid

$$\mathbb{G}_{h,M} \ : \ \tau_{i,j} = t_i + j\frac{h}{M}, \qquad j = 0,\ldots,M, \ i = 0,\ldots,N-1.$$

Recall that M is the number of values in the discrete control set $\mathcal{V}$. For simplicity only equally spaced main grids are discussed. However, a generalization towards non-equidistant grids is straightforward. In contrast to the main grid the number of minor grid intervals is determined by the number M of elements in the discrete control set $\mathcal{V}$. For the minor grid points it will be sufficient to consider only equally spaced grids.

On the minor grid we define the fixed and piecewise constant function

$$v_{\mathbb{G}_{h,M}}(\tau) := v_j \quad \text{for } \tau \in [\tau_{i,j-1}, \tau_{i,j}), \ i = 0,\ldots,N-1, \ j = 1,\ldots,M, \qquad (7.1.1)$$

cf. Figure 7.1.

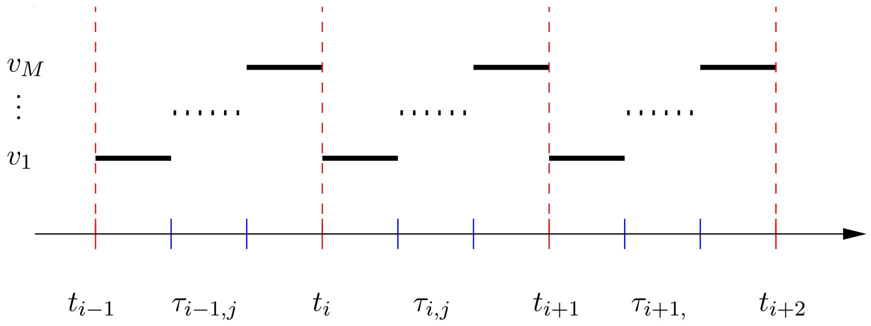


Figure 7.1: Piecewise constant staircase function $v_{\mathbb{G}_{h,M}}$ used for re-parametrization of the discrete controls.

The idea is to define an appropriate variable time transformation $t = t(\tau)$ and to control the length of the transformed minor grid interval $[t(\tau_{i,j}), t(\tau_{i,j+1})]$ by an additional function w.

By doing so, non-optimal values of the discrete control $v_{\mathbb{G}_{h,M}}$ within each main grid interval $[t(t_i), t(t_{i+1})]$ will be 'wiped out'. As in Ioffe and Tihomirov [IT79], p. 148 and in Lee et al. [LTRJ99] we define the variable time transformation $t = t(\tau)$ by setting

$$t(\tau) := t_0 + \int_{t_0}^{\tau} w(s)ds, \qquad \tau \in [t_0, t_f]$$

with

$$\int_{t_0}^{t_f} w(s)ds = t_f - t_0,$$

i.e $t(t_f) = t_f$. This transformation maps $[t_0, t_f]$ onto itself but changes the speed of running through this interval. In particular, it holds

$$\frac{dt}{d\tau} = w(\tau), \qquad \tau \in [t_0, t_f].$$

The function values $w(\tau)$ with $\tau \in [\tau_{i,j}, \tau_{i,j+1})$ are related to the length of the interval $[t(\tau_{i,j}), t(\tau_{i,j+1})]$ according to

$$\int_{\tau_{i,j}}^{\tau_{i,j+1}} w(\tau)d\tau = t(\tau_{i,j+1}) - t(\tau_{i,j}) \qquad \text{in} \qquad [\tau_{i,j}, \tau_{i,j+1}).$$

As a consequence the interval $[t(\tau_{i,j}), t(\tau_{i,j+1})]$ shrinks to the point $\{t(\tau_{i,j})\}$ if

$$w(\tau) = 0 \qquad \text{in} \qquad [\tau_{i,j}, \tau_{i,j+1}).$$

Hence, w is restricted to the class of piecewise constant functions on the minor grid as depicted in Figure 7.2.

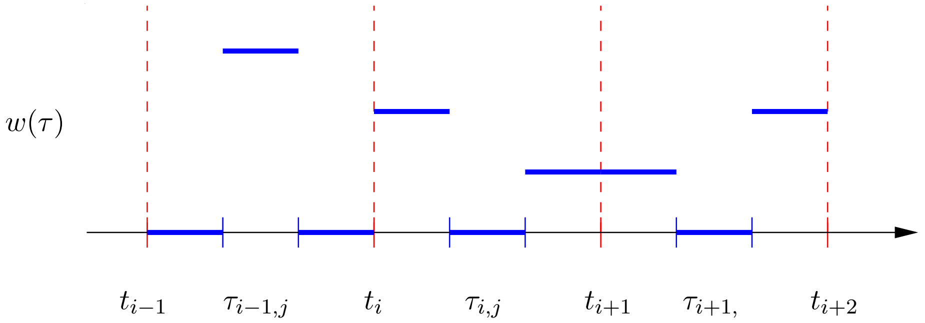


Figure 7.2: Example of a function w controlling the length of every minor grid interval.

Motivated by the previous consideration concerning the interpretation of w the function w is subject to the following restrictions:

(i) $w(\tau) \geq 0$ for all τ;

(ii) $w(\tau)$ is piecewise constant on the minor grid $\mathbb{G}_{h,M}$;

(iii) It holds

$$\int_{t_i}^{t_{i+1}} w(\tau)d\tau = t_{i+1} - t_i = h, \qquad i = 0, \ldots, N-1.$$

The first condition prohibits to run back in time and yields a non-decreasing function $t(\tau)$. The latter condition is necessary to ensure that the parametrized time actually passes through the entire main grid interval $[t_i, t_{i+1}]$. The time points t_i, $i = 0, \ldots, N$ serve as fixed time points in the original time t. Condition (iii) is essentially needed for mixed-integer optimal control problems, i.e. in the case that 'continuous' controls u and 'discrete' controls v are present simultaneously. The condition (iii) was not imposed in the papers [LTRJ97, LTRJ99, TJL99, Sib04, SR04] because only problems with a purely discrete control v and without additional 'continuous' controls u were considered. In this case, condition (iii) is not needed. The reason behind it is that we need a sufficiently dense grid for the 'continuous' controls u in order to get a good approximation to MIOCP on $\mathbb{G}_h$. If condition (iii) is not present then it will often happen that after optimizing the discretized problem many main grid intervals will be shrunken to zero. In particular, for the numerical example from virtual test-driving at the end of this section we will get solutions where the car is tunneling through the state constraints in the middle of the test-course because all main grid points will be moved close to the start or the end of the test-course. Of course such a behavior is not desired and leads to wrong results.

Figure 7.3 illustrates the variable time transformation.

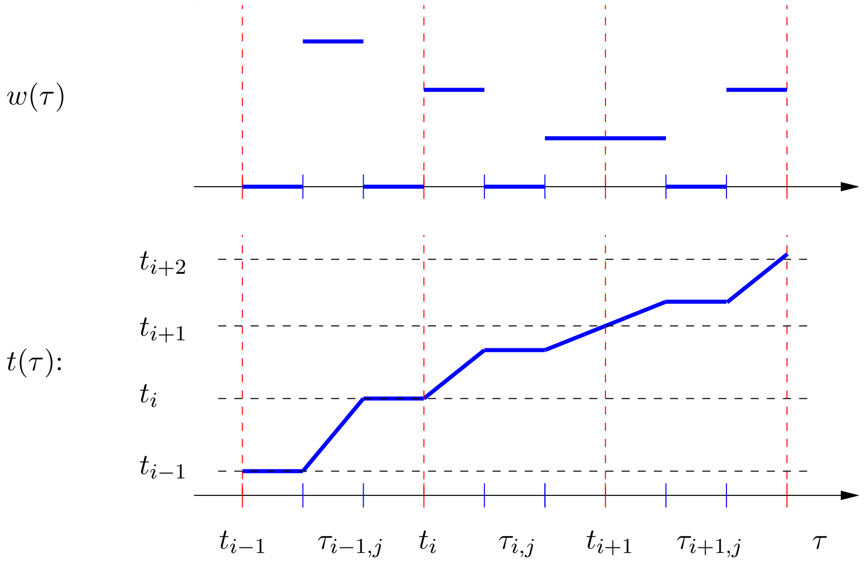


Figure 7.3: Illustration of variable time transformation: Parametrization w (top) and corresponding time $t = t(\tau)$ (bottom).

Notice, that the mapping $\tau \mapsto t(\tau)$ is not invertible. But with the convention

$$t^{-1}(s) := \inf\{\tau \mid s = t(\tau)\} \tag{7.1.2}$$

it becomes invertible. The function $v_{\mathbb{G}_{h,M}}$ from (7.1.1) together with any w satisfying the conditions (i)–(iii), e.g. w in Figure 7.2, correspond to a feasible discrete control $v(t) \in \mathcal{V}$ defined by

$$v(s) := v_{\mathbb{G}_{h,M}}(t^{-1}(s)), \qquad s \in [t_0, t_f]$$

cf. Figure 7.4. Notice, that minor intervals with $w(\tau) = 0$ do not contribute to $v(s)$.

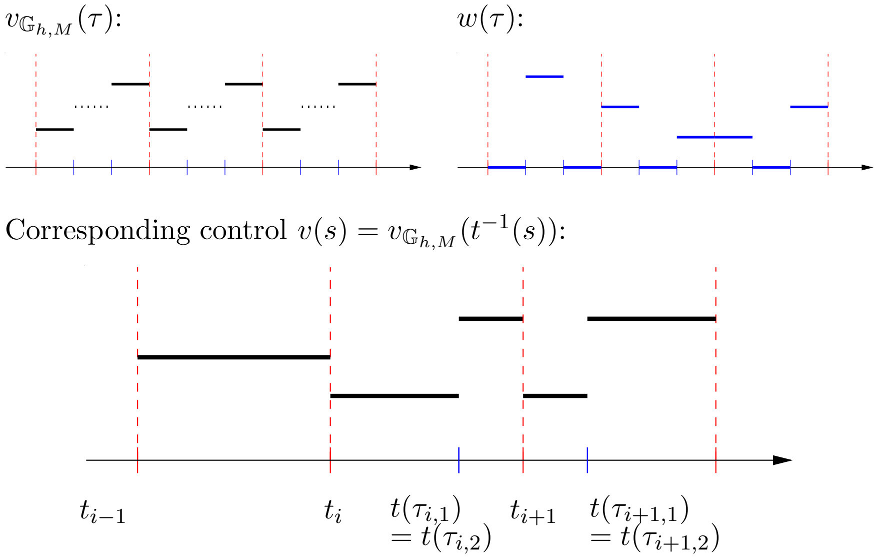


Figure 7.4: Back-transformation v (bottom) of variable time transformation for given w and fixed $v_{\mathbb{G}_{h,M}}$ (top).

Vice versa, every piecewise constant discrete control v on the main grid $\mathbb{G}_h$ can be described by $v_{\mathbb{G}_{h,M}}$ and some suitable w. Actually, the set of all functions that can be constructed by $v_{\mathbb{G}_{h,M}}$ and w is larger than the set of all piecewise constant discrete-valued functions with values in $\mathcal{V}$ on the grid $\mathbb{G}_h$. But it is smaller than the set of all such functions on the minor grid $\mathbb{G}_{h,M}$. For instance, a function that switches from the value v_M to v_1 in some main grid interval can not be represented. This is due to the preference of values given by the definition of the fixed function $v_{\mathbb{G}_{h,M}}$. If we would have defined $v_{\mathbb{G}_{h,M}}$ in each main grid interval starting with the value v_M and ending with v_1 we would get a different set of representable functions. However, for $h \to 0$ any discrete-valued function with finitely many jumps in (t_0, t_f) can be approximated arbitrarily close (in the L^1-norm) by $v_{\mathbb{G}_{h,M}}$ and a suitable w. Thus, the suggested transformation for sufficiently small h will yield good approximations for mixed-integer optimal control problems with the optimal v having only finitely many jumps.

Since our intention is to create an optimal v we introduce w as a new control subject to

$$w \in \mathcal{W} := \left\{ w \in L^\infty([t_0,t_f],\mathbb{R}^{n_v}) \;\middle|\; \begin{array}{l} w(\tau) \geq 0, \\ w \text{ piecewise constant on } \mathbb{G}_{h,M}, \\ \displaystyle\int_{t_i}^{t_{i+1}} w(\tau) d\tau = t_{i+1} - t_i \quad \forall i \end{array} \right\}.$$

The transformed mixed-integer optimal control problem (TMIOCP) reads as

$$\begin{array}{lrcll}
\text{Minimize} & \multicolumn{4}{l}{\varphi(x(t_0), x(t_f))} \\
\text{w.r.t.} & \multicolumn{4}{l}{x \in W^{1,\infty}([t_0,t_f],\mathbb{R}^{n_x}), u \in L^\infty([t_0,t_f],\mathbb{R}^{n_u}), w \in L^\infty([t_0,t_f],\mathbb{R}^{n_v})} \\
\text{s.t.} & \dot{x}(\tau) - w(\tau) f(\tau, x(\tau), u(\tau), v_{\mathbb{G}_{h,M}}(\tau)) & = & 0 & \text{a.e. in } [t_0,t_f], \\
& s(\tau, x(\tau)) & \leq & 0 & \text{in } [t_0,t_f], \\
& \psi(x(t_0), x(t_f)) & = & 0 & \\
& u(\tau) & \in & \mathcal{U} & \text{a.e. in } [t_0,t_f] \\
& w & \in & \mathcal{W} &
\end{array}$$

Notice, that TMIOCP only has ‘continuous’ controls u and w since $v_{\mathbb{G}_{h,M}}$ is a fixed function. Furthermore, if $w(\tau) \equiv 0$ in $[\tau_{i,j}, \tau_{i,j+1}]$ then x remains constant in that interval. As explained before, any piecewise constant function v on $\mathbb{G}_h$ with values in $\mathcal{V}$ can be represented by $v_{\mathbb{G}_{h,M}}$ and w. This justifies the replacement of v in MIOCP by $v_{\mathbb{G}_{h,M}}$ in TMIOCP.
Solving TMIOCP numerically by a direct discretization method using the grids $\mathbb{G}_h$ and $\mathbb{G}_{h,M}$ yields the solutions x^*, u^*, w^*. Approximate solutions of MIOCP are then given by back-transformation

$$x(s) := x^*(t^{-1}(s)), \quad u(s) := u^*(t^{-1}(s)), \quad v(s) := v_{G_{h,M}}(t^{-1}(s))$$

where

$$t(\tau) = t_0 + \int_{t_0}^{\tau} w^*(s)ds$$

and $t^{-1}(s)$ is defined according to (7.1.2).

Remark 7.1.2 *Notice, that this approach is not limited to v being a scalar function. For instance, consider the case of $n_v = 2$ discrete controls each assuming values in $\{0, 1\}$. Then the control set $\mathcal{V}$ is given by all possible combinations of values:*

$$\mathcal{V} = \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\}.$$

TMIOCP will be solved numerically by a direct discretization approach. The ‘continuous’ control u is discretized on the main grid $\mathbb{G}_h$ by the piecewise constant function

$$u_h(\tau) = u_i \qquad \text{for } \tau \in [t_i, t_{i+1}),\ i = 0, \ldots, N-1.$$

Any $w \in \mathcal{W}$ can be parametrized by the values $w_{i,j}$, $i = 0, \ldots, N-1$, $j = 0, \ldots, M-1$ according to

$$w(\tau) = w_{i,j} \qquad \text{for } \tau \in [\tau_{i,j}, \tau_{i,j+1}),\ i = 0, \ldots, N-1,\ j = 0, \ldots, M-1.$$

The differential equation is discretized by, e.g., a s-staged Runge-Kutta method with coefficients $b_\nu, c_\nu, a_{\nu\mu}$, $1 \leq \nu, \mu \leq s$ on the minor grid $\mathbb{G}_{h,M}$:

$$\begin{aligned} x_{i,j+1} &= x_{i,j} + \frac{h}{M} \sum_{\nu=1}^{s} b_\nu k_\nu^{i,j}, \qquad j = 0, \ldots, M-1, && (7.1.3) \\ k_\nu^{i,j} &= w_{i,j} f\left(\tau_{i,j} + c_\nu \tfrac{h}{M}, x_{i,j} + \tfrac{h}{M} \sum_{\mu=1}^{s} a_{\nu\mu} k_\mu^{i,j}, u_i, v_{\mathbb{G}_{h,M}}(\tau_{i,j}) \right), && (7.1.4) \\ x_{i+1,0} &= x_{i,M}, \qquad i = 0, \ldots, N-1. && (7.1.5) \end{aligned}$$

The first index always refers to the main grid interval, the second index to the minor grid interval. Notice, that $v_{\mathbb{G}_{h,M}}(\tau_{i,j}) = v_{j+1}$ in (7.1.4). Furthermore, it holds

$$\int_{t_i}^{t_{i+1}} w(\tau) d\tau = \frac{h}{M} \sum_{j=0}^{M-1} w_{i,j}.$$

A discretization of TMIOCP is thus given by the subsequent finite dimensional nonlinear programming problem:

Problem 7.1.3

$$
\begin{array}{lrcll}
\textit{Minimize} & \varphi(x_{0,0}, x_{N,0}) & & & \\
\textit{w.r.t.} & x_{i,j}, u_i, w_{i,j}, & & & \\
\textit{s.t.} & \textit{equations } (7.1.3) - (7.1.5), & & & \\
 & s(t_i, x_{i,0}) & \leq & 0 & i = 0, \ldots, N, \\
 & \psi(x_{0,0}, x_{N,0}) & = & 0, & \\
 & u_i & \in & \mathcal{U} & i = 0, \ldots, N-1, \\
 & w_{i,j} & \geq & 0, & i = 0, \ldots, N-1,\ j = 0, \ldots, M-1, \\
 & \displaystyle\sum_{j=0}^{M-1} w_{i,j} & = & M, & i = 0, \ldots, N-1.
\end{array}
$$

This problem can be solved numerically by a SQP method.

Remark 7.1.4 *The state constraint $s(t, x(t)) \leq 0$ is only evaluated on the main grid and not on the minor grid. This keeps the number of constraints small. Likewise, u_h is kept piecewise constant on the main grid and not on the minor grid. This keeps the number of variables small. This approach should be sufficient in view of a possible convergence result as $h \to 0$. Nevertheless, it is straightforward to construct an alternative method discretizing everything on the minor grid.*

7.1.3 Numerical Results

We will consider a particular optimal control problem arising from automobile test-driving with gear shifts. Herein, one component of the control is discrete.

7.1.3.1 Car Model

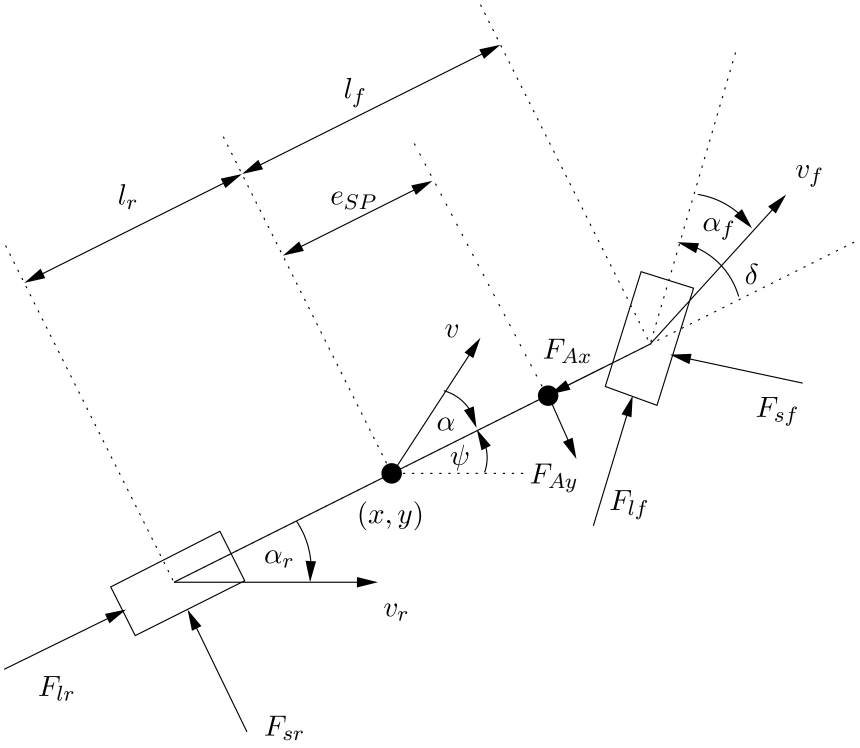


Figure 7.5: Geometrical description of the single-track car model.

We use the single-track car model to be described below in detail. It is a simplified car model, which is commonly used in automobile industry for basic investigations of the dynamical behavior of cars, cf., e.g., Mayr [May91], Risse [Ris91], Neculau [Nec92], Moder [Mod94].

The simplifying assumption made to derive the single-track car model is that the rolling and pitching behavior of the car body can be neglected, that is, the roll angle and the pitch angle are small. This allows to replace the two wheels on the front and rear axis by a virtual wheel located in the center of the respective axis. Furthermore, due to the simplifying assumptions it can be assumed that the car's center of gravity is located on the roadway and therefore, it is sufficient to consider the motion of the car solely in the horizontal plane. The upcoming car model includes four control variables for the driver: the steering angle velocity $|w_\delta| \leq 0.5$ $[rad/s]$, the total braking force $0 \leq F_B \leq 15000$ $[N]$, the gear $\mu \in \{1, 2, 3, 4, 5\}$, and the accelerator pedal position $\phi \in [0, 1]$. The configuration of the car is depicted in Figure 7.5.

Herein, (x, y) denotes the center of gravity in a reference coordinate system, v, v_f, v_r denote the velocity of the car and the velocities of the front and rear wheel, respectively, δ, β, ψ denote the steering angle, the side slip angle, and the yaw angle, respectively, α_f, α_r denote the slip angles at front and rear wheel, respectively, F_{sf}, F_{sr} denote the lateral tire forces (side forces) at front and rear wheel, respectively, F_{lf}, F_{lr} denote the longitudinal tire forces at front and rear wheel, respectively, l_f, l_r, e_{SP} denote the distances from the center of gravity to the front and rear wheel, and to the drag mount point, respectively, F_{Ax}, F_{Ay} denote the drag according to air resistance and side wind, respectively, and m denotes the mass of the car.

The equations of motion are given by the following system of ordinary differential equations

$$\begin{aligned}
\dot{x} &= v\cos(\psi - \beta), && (7.1.6)\\
\dot{y} &= v\sin(\psi - \beta), && (7.1.7)\\
\dot{v} &= \frac{1}{m}\Big[(F_{lr} - F_{Ax})\cos\beta + F_{lf}\cos(\delta + \beta) - (F_{sr} - F_{Ay})\sin\beta \\
&\qquad -F_{sf}\sin(\delta + \beta)\Big], && (7.1.8)\\
\dot{\beta} &= w_z - \frac{1}{m \cdot v}\Big[(F_{lr} - F_{Ax})\sin\beta + F_{lf}\sin(\delta + \beta) \\
&\qquad + (F_{sr} - F_{Ay})\cos\beta + F_{sf}\cos(\delta + \beta)\Big], && (7.1.9)\\
\dot{\psi} &= w_z, && (7.1.10)\\
\dot{w}_z &= \frac{1}{I_{zz}}\Big[F_{sf} \cdot l_f \cdot \cos\delta - F_{sr} \cdot l_r - F_{Ay} \cdot e_{SP} + F_{lf} \cdot l_f \cdot \sin\delta\Big], && (7.1.11)\\
\dot{\delta} &= w_\delta. && (7.1.12)
\end{aligned}$$

The lateral tire forces are functions of the respective slip angles (and the tire loads, which are constant in our model). A famous model for the lateral tire forces is the 'magic formula' of Pacejka [PB93]:

$$\begin{aligned}
F_{sf}(\alpha_f) &= D_f \sin\left(C_f \arctan\left(B_f\alpha_f - E_f\left(B_f\alpha_f - \arctan(B_f\alpha_f)\right)\right)\right),\\
F_{sr}(\alpha_r) &= D_r \sin\left(C_r \arctan\left(B_r\alpha_r - E_r\left(B_r\alpha_r - \arctan(B_r\alpha_r)\right)\right)\right),
\end{aligned}$$

compare Figure 7.6. Herein, $B_f, B_r, C_f, C_r, D_f, D_r, E_f, E_r$ are constants depending on the tire.

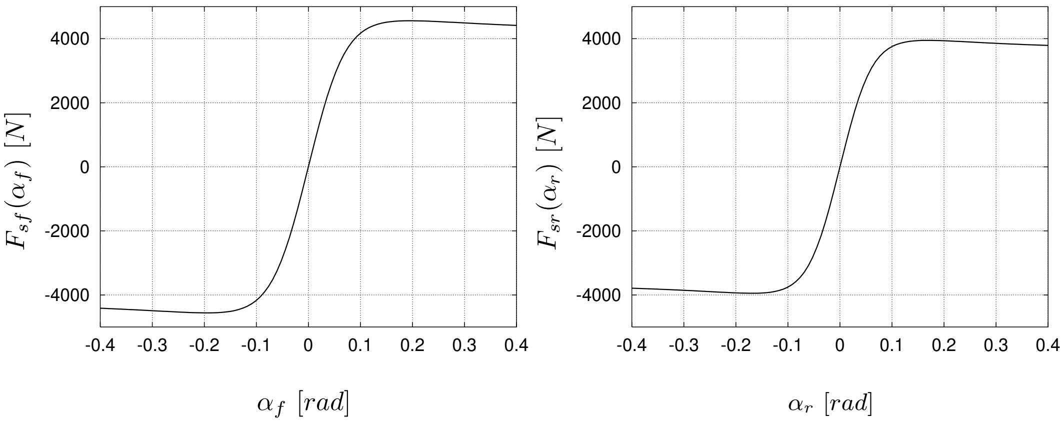


Figure 7.6: Lateral tire forces at front (left) and rear (right) wheel as functions of the slip angle computed by the 'magic formula' of Pacejka [PB93].

According to Mitschke [Mit90], p. 23, (notice the opposite sign in the definition of β) the slip angles are given by

$$\alpha_f = \delta - \arctan\left(\frac{l_f\dot{\psi} - v\sin\beta}{v\cos\beta}\right), \qquad \alpha_r = \arctan\left(\frac{l_r\dot{\psi} + v\sin\beta}{v\cos\beta}\right).$$

The drag due to air resistance is modeled by

$$F_{Ax} = \frac{1}{2} \cdot c_w \cdot \rho \cdot A \cdot v^2,$$

where c_w is the air drag coefficient, ρ is the air density, and A is the effective flow surface. In this article, it is assumed that there is no side wind, i.e. $F_{Ay} = 0$.
In the sequel, we assume that the car has rear wheel drive. Then, the longitudinal tire force at the front wheel is given by

$$F_{lf} = -F_{Bf} - F_{Rf},$$

where F_{Bf} is the braking force at the front wheel and F_{Rf} denotes the rolling resistance at the front wheel. The longitudinal tire force at the rear wheel is given by

$$F_{lr} = \frac{M_{wheel}(\phi,\mu)}{R} - F_{Br} - F_{Rr}$$

where $M_{wheel}(\phi,\mu)$ is the torque resulting from the drive-train applied to the rear wheel. Again, F_{Br} is the braking force at the rear wheel and F_{Rr} denotes the rolling resistance at the rear wheel. According to Neculau [Nec92] it holds

$$M_{wheel}(\phi,\mu) = i_g(\mu) \cdot i_t \cdot M_{mot}(\phi,\mu), \tag{7.1.13}$$

where

$$M_{mot}(\phi,\mu) = f_1(\phi) \cdot f_2(w_{mot}(\mu)) + (1 - f_1(\phi))f_3(w_{mot}(\mu))$$

denotes the motor torque and

$$w_{mot}(\mu) = \frac{v \cdot i_g(\mu) \cdot i_t}{R} \tag{7.1.14}$$

denotes the rotary frequency of the motor depending on the gear μ. Notice, that this relation is based on the assumption that the longitudinal slip can be neglected. The functions f_1, f_2 and f_3 are given by

$$\begin{aligned} f_1(\phi) &= 1 - \exp(-3\phi), \\ f_2(w_{mot}) &= -37.8 + 1.54 \cdot w_{mot} - 0.0019 \cdot w_{mot}^2, \\ f_3(w_{mot}) &= -34.9 - 0.04775 \cdot w_{mot}. \end{aligned}$$

The total braking force F_B controlled by the driver is distributed on the front and rear wheels such that $F_{Bf} + F_{Br} = F_B$ holds. Here, we chose

$$F_{Bf} = \frac{2}{3} F_B, \qquad F_{Br} = \frac{1}{3} F_B.$$

Finally, the rolling resistance forces are given by

$$F_{Rf} = f_R(v) \cdot F_{zf}, \qquad F_{Rr} = f_R(v) \cdot F_{zr},$$

where

$$f_R(v) = f_{R0} + f_{R1} \frac{v}{100} + f_{R4} \left(\frac{v}{100}\right)^4 \qquad (v \text{ in [km/h]}),$$

is the friction coefficient and

$$F_{zf} = \frac{m \cdot l_r \cdot g}{l_f + l_r}, \qquad F_{zr} = \frac{m \cdot l_f \cdot g}{l_f + l_r}$$

denote the static tire loads at the front and rear wheel, respectively, cf. Risse [Ris91].

For the upcoming numerical computations we used the parameters summarized in Table 7.1.

Table 7.1: Parameters for the single-track car model (partly taken from Risse [Ris91]).

Parameter	Value	Description
m	1239 $[kg]$	mass of car
g	9.81 $[m/s^2]$	acceleration due to gravity
l_f/l_r	1.19016/1.37484 $[m]$	dist. center of gravity to front/rear wheel
e_{SP}	0.5 $[m]$	dist. center of gravity to drag mount point
R	0.302 $[m]$	wheel radius
I_{zz}	1752 $[kgm^2]$	moment of inertia
c_w	0.3	air drag coefficient
ρ	1.249512 $[N/m^2]$	air density
A	1.4378946874 $[m^2]$	effective flow surface
$i_g(1)$	3.91	first gear
$i_g(2)$	2.002	second gear
$i_g(3)$	1.33	third gear
$i_g(4)$	1.0	fourth gear
$i_g(5)$	0.805	fifth gear
i_t	3.91	motor torque transmission
B_f/B_r	10.96/12.67	Pacejka-model (stiffness factor)
$C_f = C_r$	1.3	Pacejka-model (shape factor)
D_f/D_r	4560.40/3947.81	Pacejka-model (peak value)
$E_f = E_r$	-0.5	Pacejka-model (curvature factor)
$f_{R0}/f_{R1}/f_{R4}$	0.009/0.002/0.0003	coefficients

7.1.3.2 Test-Course

The discussion is restricted to the double-lane-change manoeuvre being a standardized manoeuvre in automobile industry. The driver has to manage an offset of 3.5 $[m]$ and afterwards he has to reach the original track, see Figure 7.7. The double-lane-change manoeuvre can be understood as a model for a jink caused by a suddenly occurring obstacle on the road.

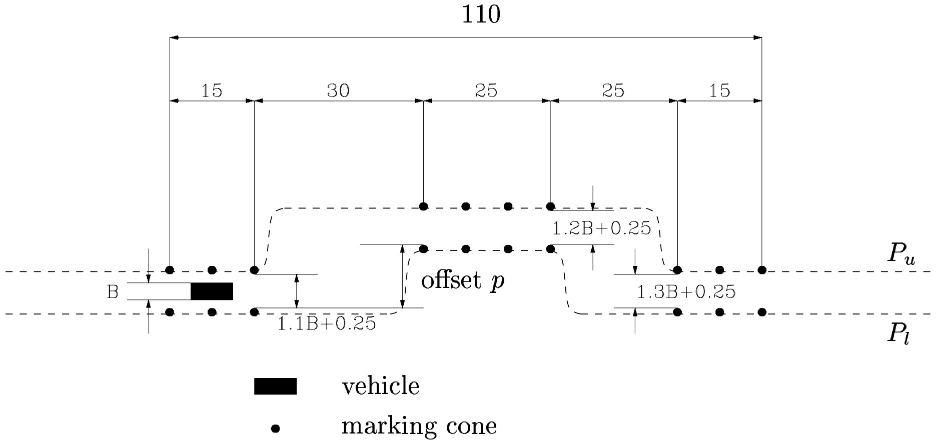


Figure 7.7: Measurements of the track and boundaries P_l and P_u (dashed), cf. Zomotor [Zom91]

The boundaries of the test-course will play the role of state constraints in a suitable optimal control problem later on and are described by continuously differentiable functions $P_l(x)$ (lower boundary) and $P_u(x)$ (upper boundary), which are piecewise defined cubic polynomials. For a car with width $B = 1.5$ $[m]$ we find, cf. Moder [Mod94], Gerdts [Ger01a]:

$$P_l(x) = \begin{cases} 0, & \text{if } x \leq 44 \ , \\ 4 \cdot h_2 \cdot (x-44)^3, & \text{if } 44 < x \leq 44.5 \ , \\ 4 \cdot h_2 \cdot (x-45)^3 + h_2, & \text{if } 44.5 < x \leq 45 \ , \\ h_2, & \text{if } 45 < x \leq 70 \ , \\ 4 \cdot h_2 \cdot (70-x)^3 + h_2, & \text{if } 70 < x \leq 70.5 \ , \\ 4 \cdot h_2 \cdot (71-x)^3, & \text{if } 70.5 < x \leq 71 \ , \\ 0, & \text{if } x > 71 \ , \end{cases} \quad , \tag{7.1.15}$$

$$P_u(x) = \begin{cases} h_1, & \text{if } x \leq 15 \ , \\ 4 \cdot (h_3 - h_1) \cdot (x-15)^3 + h_1, & \text{if } 15 < x \leq 15.5 \ , \\ 4 \cdot (h_3 - h_1) \cdot (x-16)^3 + h_3, & \text{if } 15.5 < x \leq 16 \ , \\ h_3, & \text{if } 16 < x \leq 94 \ , \\ 4 \cdot (h_3 - h_4) \cdot (94-x)^3 + h_3, & \text{if } 94 < x \leq 94.5 \ , \\ 4 \cdot (h_3 - h_4) \cdot (95-x)^3 + h_4, & \text{if } 94.5 < x \leq 95 \ , \\ h_4, & \text{if } x > 95 \ , \end{cases} \tag{7.1.16}$$

where $h_1 = 1.1 \cdot B + 0.25$, $h_2 = 3.5$, $h_3 = 1.2 \cdot B + 3.75$, $h_4 = 1.3 \cdot B + 0.25$.

7.1.3.3 Driver Model

The driver is modeled by formulation of a suitable state constrained optimal control problem with free final time. The required observation of the marking cones results in two state constraints

$$y(t) \leq P_u(x(t)) - B/2, \qquad y(t) \geq P_l(x(t)) + B/2, \tag{7.1.17}$$

where $(y(t), x(t))$ denotes the position of the car's center of gravity at time t and B the car's width. Let the initial state of the car at time $t_0 = 0$ on the track be prescribed with exception of the initial y-position:

$$(x(0), y(0), v(0), \beta(0), \psi(0), w_z(0), \delta(0)) = (-30, free, 10, 0, 0, 0, 0). \tag{7.1.18}$$

To ensure, that the car completes the course, we impose additional boundary conditions at the final time t_f:

$$x(t_f) = 140, \qquad \psi(t_f) = 0. \tag{7.1.19}$$

The latter condition ensures, that the longitudinal axis of the car is parallel to the boundary of the test-course at t_f. The intention of this condition is to ensure that the driver could continue his drive at least for a short time period after t_f has been reached without violating the boundaries of the course too severely.

Finally, the driver is modeled by minimizing a linear combination of the steering effort and the final time. Roughly speaking, this is a compromise between driving fast (minimizing time) and driving safely (minimizing steering effort).

Summarizing, the double-lane-change manoeuvre is modeled by the following optimal control problem:

$$\begin{array}{ll} \text{Minimize} & t_f + \displaystyle\int_0^{t_f} w_\delta(t)^2 dt \\ \text{s.t.} & \text{differential equations (7.1.6)-(7.1.12) a.e. in } [0, t_f], \\ & \text{boundary conditions (7.1.18) and (7.1.19)}, \\ & \text{state constraints (7.1.17) for all } t \in [0, t_f], \\ & \text{control constraints } w_\delta(t) \in [-0.5, 0.5], F_B(t) \in [0, 15000], \\ & \phi \in [0, 1], \mu(t) \in \{1, 2, 3, 4, 5\} \text{ for all } t \in [0, t_f]. \end{array}$$

Remark 7.1.5 *For the applicability of the Branch&Bound method it is necessary to relax the discrete function $i_g(\mu)$. This can be done by constructing an interpolating natural cubic spline as it was done in Gerdts [Ger05e].*

7.1.4 Results

We compare the Branch&Bound method and the variable time transformation method. For the numerical calculations we used the software package SODAS, cf. Gerdts [Ger01a, Ger03a], with the classical Runge-Kutta discretization scheme of order four with piecewise constant control approximations on an equidistant grid. All computations concerning the Branch&Bound method were performed on a Pentium 3 processor with 750 MHz. All computations concerning the variable time transformation method were performed on a Pentium mobile processor with 1.6 GHz processing speed.

In all cases the optimal braking forces and the optimal gas pedal positions compute to $F_B \equiv 0$ and $\varphi \equiv 1$, respectively.

Table 7.2 summarizes the numerical results of the Branch&Bound method for $N = 20$ grid points. Herein, the Branch&Bound algorithm is stopped after #NLP nodes of the search tree are generated. Recall, that each node corresponds to solving one relaxed discretized optimal control problem. At termination the lower and upper bound (columns LOWER and UPPER) are computed in order to obtain an estimate for the exact optimal objective function value. With these bounds the accuracy of the solution obtained at termination is given by column GAP, where GAP=(UPPER-LOWER)/LOWER. This procedure allows to investigate how good the currently best solutions are after a fixed number of iterations. Without artificial stopping, for

$N = 20$ the Branch&Bound algorithm terminates with the solution after 1119 nodes have been generated. Herein, the optimal discrete gear shift $\mu = (\mu_0, \mu_1, \dots, \mu_{N-1})^\top$ is computed to

$$\mu_i = \begin{cases} 1, & \text{if } i = 0, \\ 2, & \text{if } 1 \leq i \leq 7, \\ 3, & \text{if } 8 \leq i \leq 18, \\ 4, & \text{if } i = 19. \end{cases}$$

Table 7.2: Numerical results for $N = 20$. The Branch&Bound algorithm is stopped after #NLP nodes of the search tree have been generated.

N	#NLP	LOWER	UPPER	GAP	CPU [h:m,s]
20	100	6.635221	6.781922	2.2 %	00:02,19
20	500	6.664889	6.781922	1.8 %	00:10,41
20	1000	6.674979	6.781922	1.6 %	00:21,42
20	1119	6.781922	6.781922	0.0 %	00:23,52

It is remarkable, that the first feasible solution found by the Branch&Bound method is already the optimal one. But this is just a coincidence as the results for $N = 40$ show.
The final time is $t_f = 6.779751$ $[s]$ and the optimal objective function value is 6.781922. Compared to the number of possible combinations for the discrete control μ given by 5^N, the Branch&Bound algorithm requires comparatively few evaluations, namely 1119. But, one has to keep in mind, that this corresponds to solving 1119 nonlinear, state constrained optimal control problems numerically, which in general may be a demanding task for itself. However, in our case the solution of a single optimal control problem corresponding to a node in the search tree can be accelerated substantially, if, e.g., the optimal solution of the father's node is taken as an initial guess for the successors. Usually, this initial guess is very close to the optimal solution, since according to the branching rule only one additional component of the control is restricted.
It was also possible to solve the problem by Branch&Bound for $N = 40$ points, but with much more computational effort. The Branch&Bound algorithm terminates after approximately 232 hours CPU time and 146941 nodes generated. Herein, the optimal discrete gear shift $\mu = (\mu_0, \mu_1, \dots, \mu_{N-1})^\top$ is computed to

$$\mu_i = \begin{cases} 1, & \text{if } 0 \leq i \leq 2, \\ 2, & \text{if } 3 \leq i \leq 15, \\ 3, & \text{if } 16 \leq i \leq 38, \\ 4, & \text{if } i = 39. \end{cases}$$

Table 7.3 summarizes the numerical results for $N = 40$ grid points. The final time is $t_f = 6.786781$ $[s]$ and the optimal objective function value is 6.791374.

Table 7.3: Numerical results for $N = 40$. The Branch&Bound algorithm is stopped after #NLP nodes of the search tree have been generated.

N	#NLP	LOWER	UPPER	GAP	CPU [h:m,s]
40	1000	6.643892	6.795074	2.3 %	1:32,23
40	10000	6.643892	6.791484	2.2 %	14:40,54
40	50000	6.643892	6.791375	2.2 %	77:31,44
40	146941	6.791374	6.791374	0.0 %	232:25,31

Figure 7.8 shows the numerical solution for the variable time transformation method for $N = 20$. The variable time transformation method only needs 2 minutes and 1.154 seconds to solve the problem with objective function value 6.774669 and final time $t_f = 6.772516$ $[s]$. Recall, that the Branch&Bound method needed 23 minutes and 52 seconds! The lower objective function value for the variable time transformation method is due to additional degrees of freedom in Problem 7.1.3, since the method allows to switch the discrete control even within some main grid interval. The switching points of the discrete control in the Branch&Bound method were restricted to the main grid points. Nevertheless, the qualitative switching structure of the optimal discrete control μ is identical for both approaches.

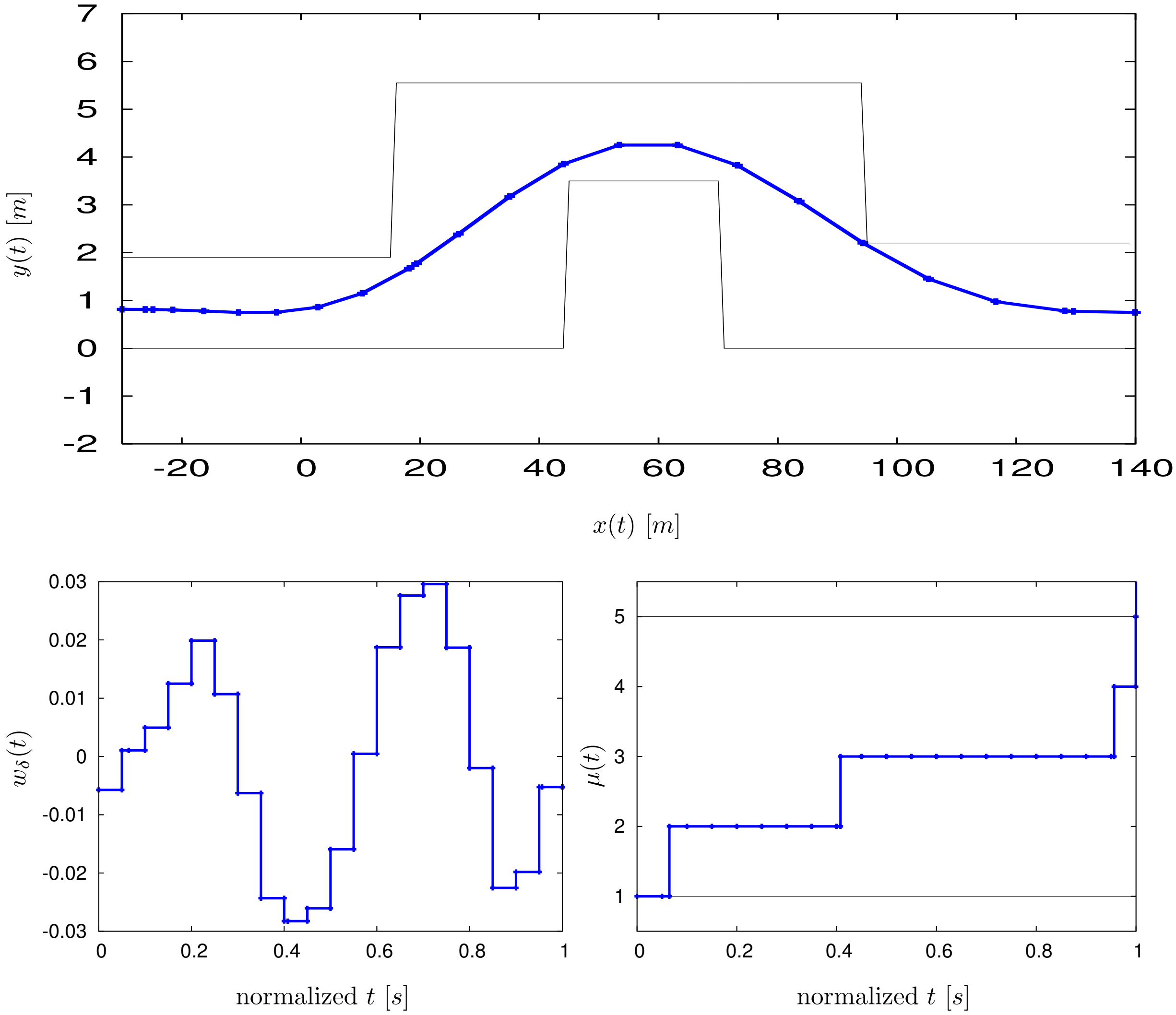


Figure 7.8: Numerical result for $N = 20$: Path $(x(t), y(t))$ of the car's center of gravity (top), steering angle velocity $w_\delta(t)$ (bottom, left), and gear shift $\mu(t)$ (bottom, right).

Figure 7.9 shows the numerical solution for the variable time transformation method for $N = 40$. The Branch&Bound method needed 232 hours, 25 minutes and 31 seconds of CPU time to solve the problem and yields the optimal objective function value 6.791374 and final time $t_f = 6.786781$ $[s]$. The variable time transformation method only needs 9 minutes and 39.664 seconds to solve the problem with objective function value 6.787982 and final time $t_f = 6.783380$ $[s]$.

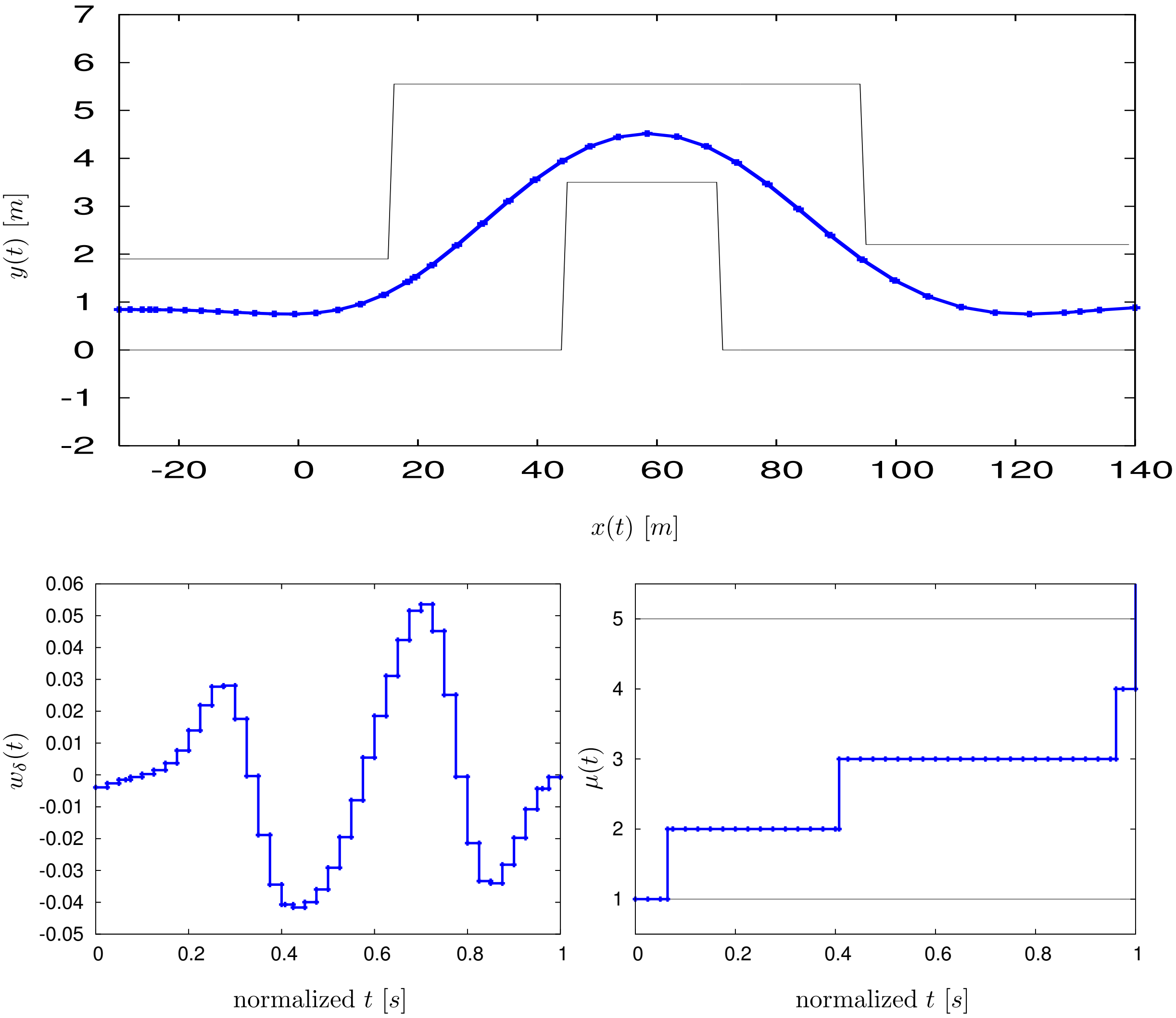


Figure 7.9: Numerical result for $N = 40$: Path $(x(t), y(t))$ of the car's center of gravity (top), steering angle velocity $w_\delta(t)$ (bottom, left), and gear shift $\mu(t)$ (bottom, right).

Figure 7.10 shows the numerical solution for $N = 80$. The Branch&Bound method did not terminate in a reasonable time. The variable time transformation method only needs 65 minutes and 3.496 seconds to solve the problem with objective function value 6.795366 and final time $t_f = 6.789325$ $[s]$.

It has to be mentioned that the calculation of gradients in the SQP method involved in the variable time transformation method for simplicity was done by finite differences. Using a more sophisticated approach as in Chapter 6 would even lead to a substantial reduction of CPU time for the variable time transformation method. These computational results show that the variable time transformation method is particularly well-suited for mixed-integer optimal control problems and is much less expensive than the Branch&Bound method.

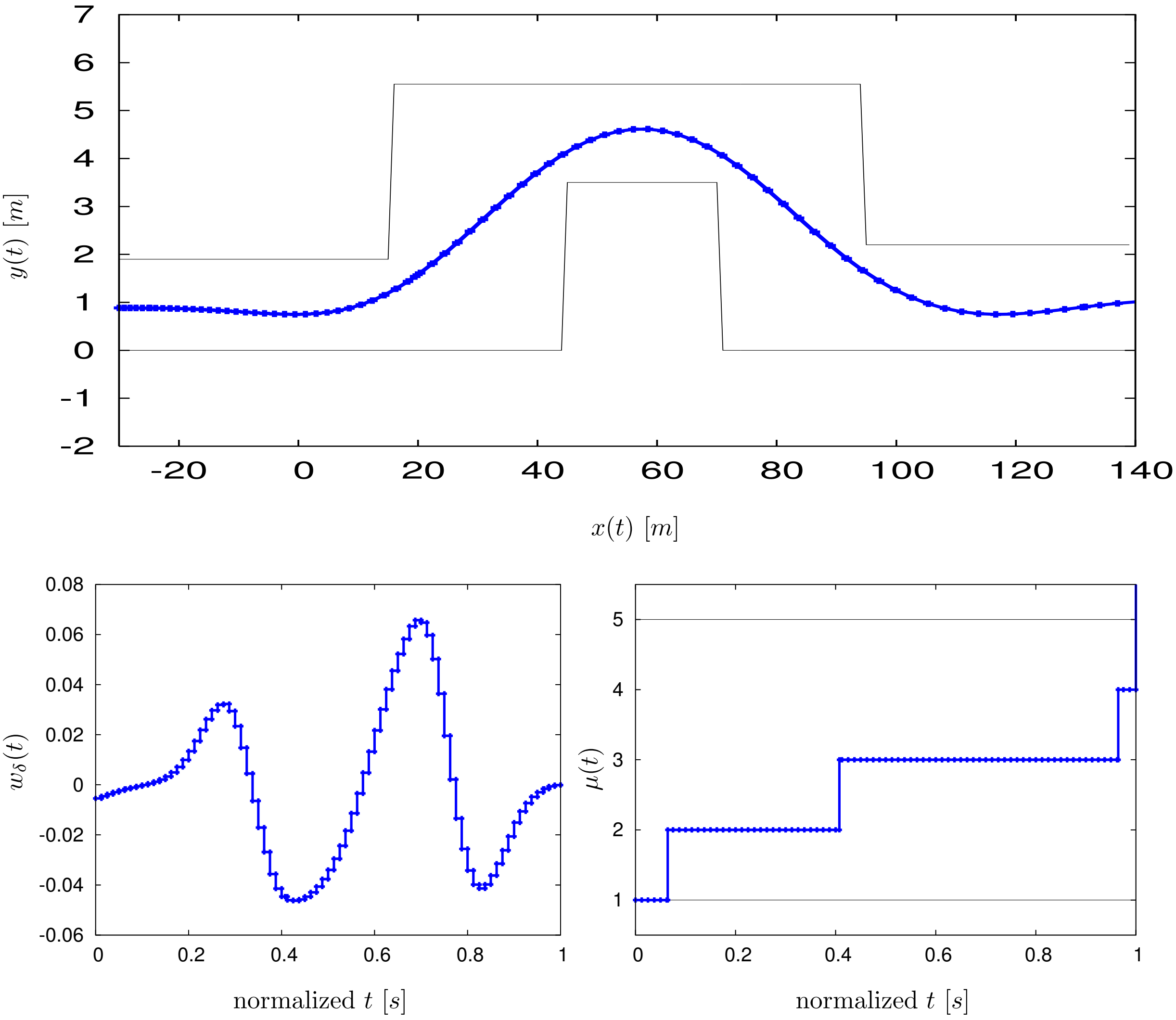


Figure 7.10: Numerical result for $N = 80$: Path $(x(t), y(t))$ of the car's center of gravity (top), steering angle velocity $w_\delta(t)$ (bottom, left), and gear shift $\mu(t)$ (bottom, right).

7.2 Open-Loop-Real-Time Control

In practical applications the optimal control problem depends on system parameters. For example, for the computation of optimal flight paths of a space shuttle mission a common model for the air density depends on the altitude and certain parameters describing a standard atmosphere. Of course, in reality the air density differs from the model and the actual air density can be viewed as a perturbation of the model. This in turn implies that the computed optimal flight path for the standard atmosphere is not optimal anymore for the actual air density. Accordingly, the flight path has to be adapted to the perturbed situation in an optimal way. One way to achieve this would be to re-compute the optimal flight path for the perturbed model. However, for time critical processes the re-computation of an optimal solution for the perturbed problem, e.g. by the direct discretization method, may be not fast enough to provide an optimal solution in real-time. Therefore, an alternative method is needed which is capable of providing at least an approximation of the optimal solution for the perturbed problem in real-time. In the sequel a method based on the parametric sensitivity analysis of the underlying optimal control problem is suggested to calculate such real-time approximations.
We consider

Problem 7.2.1 (Perturbed DAE Optimal Control Problem $OCP(p)$)

$$\begin{array}{lrcll}
\textit{Minimize} & \varphi(x(t_0), x(t_f), p) & & & \\
\textit{s.t.} & F(t, x(t), \dot{x}(t), u(t), p) & = & 0_{n_x} & \textit{a.e. in } [t_0, t_f], \\
 & \psi(x(t_0), x(t_f), p) & = & 0_{n_\psi}, & \\
 & c(t, x(t), u(t), p) & \leq & 0_{n_c} & \textit{a.e. in } [t_0, t_f], \\
 & s(t, x(t), p) & \leq & 0_{n_s} & \textit{in } [t_0, t_f].
\end{array}$$

Notice, that $p \in \mathbb{R}^{n_p}$ is not an optimization variable. Let $\hat{p} \in \mathbb{R}^{n_p}$ denote a fixed *nominal parameter.* The corresponding optimal control problem $OCP(\hat{p})$ is called *nominal problem.*
For a given parameter p we may apply the reduced discretization approach to solve the problem numerically.

Problem 7.2.2 (Reduced Discretization $DOCP(p)$)
Find $z = (x_0, w) \in \mathbb{R}^{n_x+M}$ such that the objective function

$$\varphi(X_0(z, p), X_N(z, p), p)$$

is minimized subject to

$$\begin{array}{rcll}
\psi(X_0(z, p), X_N(z, p), p) & = & 0_{n_\psi}, & \\
c(t_j, X_j(z, p), u_M(t_j; w), p) & \leq & 0_{n_c}, & j = 0, 1, \ldots, N, \\
s(t_j, X_j(z, p), p) & \leq & 0_{n_s}, & j = 0, 1, \ldots, N.
\end{array}$$

Herein, X_0 is a function that provides a consistent initial value for x_0, control parametrization w, and parameter p. The values X_j, $j = 1, \ldots, N$ are given by a suitable integration scheme, e.g. a one-step method:

$$X_{j+1}(z, p) = X_j(z, p) + h_j \Phi(t_j, X_j(z, p), w, p, h_j), \qquad j = 0, 1, \ldots, N-1.$$

According to (6.1.2) the discretized control is given by

$$u_M(\cdot) = \sum_{i=1}^{M} w_i B_i(\cdot)$$

with basis functions B_i, e.g. B-Splines.
Suppose that $DOCP(p)$ is solvable for any $p \in \mathbb{R}^n$, let $z(p) = (x_0(p), w(p))$ denote the optimal solution and $z(\hat{p})$ the *nominal solution.* In particular, the optimal discretized control for $DOCP(p)$ is given by

$$u_M(\cdot) = \sum_{i=1}^{M} w_i(p) B_i(\cdot).$$

The underlying assumption to obtain a real-time update formula, i.e. a formula or method that is capable of providing an approximation for the optimal solution $z(p)$ in real-time, is that there exists a neighborhood $\emptyset \neq \mathcal{N}(\hat{p}) \subseteq \mathbb{R}^{n_p}$ of the nominal parameter $\hat{p}$, such that the perturbed problem $DOCP(p)$ possesses an optimal solution $z(p)$ for all $p \in \mathcal{N}(\hat{p})$, which is continuously differentiable w.r.t. p in $\mathcal{N}(\hat{p})$. Sufficient conditions for differentiability are stated in Theorem 3.7.3. In case of differentiability, Taylor expansion yields the *real-time update formula*

$$z(p) \approx z(\hat{p}) + \frac{dz(\hat{p})}{dp} \cdot (p - \hat{p}), \tag{7.2.1}$$

cf. equation (3.7.4). Application of this formula to the discretized control yields a real-time approximation of the optimal perturbed control according to

$$u_M(\cdot) \approx \sum_{i=1}^{M} \left(w_i(\hat{p}) + \frac{dw_i(\hat{p})}{dp}(p - \hat{p}) \right) B_i(\cdot). \tag{7.2.2}$$

Please notice, that only matrix-vector multiplications and vector-vector summations are needed to evaluate the real-time update formulae. In the case of an one-dimensional parameter, these operations can be performed within micro seconds or even below on standard computers. Hence, the real-time update formulae provide an extremely fast approximation of the optimal solution of the perturbed optimal control problem. It remains to compute the sensitivity differential $dz(\hat{p})/dp$. Two approaches for the calculation of the sensitivity differential are described in the sequel: the brute-force method and the approach via sensitivity analysis. In either case, the calculation of the sensitivity differential takes place *offline* and may be very time consuming. Nevertheless, if the sensitivity differential is obtained, the *on-line* evaluation of the real-time update formulas (7.2.1) and (7.2.2), respectively, is extremely cheap as mentioned above.

Remark 7.2.3 *Conditions for the solution differentiability of the infinite dimensional optimal control problem subject to ODEs in appropriate Banach spaces can be found in Maurer and Pesch [MP94a, MP94b, MP95a], Malanowski and Maurer [MM96, MM98, MM01], and Maurer and Augustin [AM01, MA01]. Numerical approaches based on a linearization of the necessary optimality conditions of the optimal control problem are discussed in Pesch [Pes78, Pes79, Pes89a, Pes89b].*

7.2.1 Brute-force method

The idea of the brute-force method is very simple. Let p_i be the $i-$th component of the parameter vector $p = (p_1, \ldots, p_{n_p})^\top$ and $e_i \in \mathbb{R}^{n_p}$ the $i-$th unit vector. The sensitivity differential $dz(\hat{p})/dp_i$

is approximated by, e.g., the central finite difference scheme

$$\frac{dz(\hat{p})}{dp_i} \approx \frac{z(\hat{p} + he_i) - z(\hat{p} - he_i)}{2h}, \tag{7.2.3}$$

where $z(\hat{p} \pm he_i)$ denotes the optimal solution of $DOCP(\hat{p} \pm he_i)$. Thus, the calculation of the sensitivity differential $dz(\hat{p})/dp$ requires the solution of $2n_p$ discretized optimal control problems. Depending on n_p this is a potentially large number of nonlinear programming problems to be solved. In addition, the choice of the step size h is crucial. Despite this disadvantages, one has the advantage, that a very good initial estimate for the SQP method is known: the nominal solution $z(\hat{p})$. In addition, this approach may be preferable to the subsequent method, if the optimal solution can not be calculated very accurately, which is often the case for large scale problems.

7.2.2 Real-time approximation via sensitivity analysis

Büskens and Maurer [BM96, BM01b, BM01a] established a method for ODE optimal control problems, which allows to calculate an approximation of the optimal solution of a perturbed optimal control problem in real-time. This method is based on a sensitivity analysis of the discretized optimal control problem w.r.t. to the perturbation parameters p. Büskens and Gerdts [GB01] extended this method to optimal control problems subject to semi explicit index-1 DAEs.

The discretized optimal control problem $DOCP(\hat{p})$ is a parametric nonlinear program as in Problem 3.7.1 with objective function and constraints according to (6.1.12)-(6.1.14).

Suppose that the SQP method applied to $DOCP(\hat{p})$ converges to a local minimum $\hat{z} = z(\hat{p})$ with Lagrange multiplier $\hat{\mu}$ and that the assumptions of Theorem 3.7.3 hold at $\hat{z}$. Then, Theorem 3.7.3 guarantees the differentiability of the mapping $p \mapsto z(p)$ in some neighborhood of $\hat{p}$. Furtheremore, a solution of the linear equation (3.7.1) yields the sensitivity differential $\frac{dz}{dp}(\hat{p})$ at the solution $\hat{z}$.

The solution of (3.7.1) requires the second derivatives L''_{zz} and L''_{zp} of the Lagrange function, the derivatives H'_z, G'_z in (6.1.15)-(6.1.16), and the corresponding derivatives H'_p, G'_p at $\hat{z}$. More precisely, only the derivatives of the active constraints are neeeded. Due to high computational costs, most implementations of SQP methods, e.g. NPSOL of Gill et. al. [GMSW98], do not use the Hessian of the Lagrangian within the QP subproblem. Instead, the Hessian is replaced by a matrix computed with the modified BFGS update formula (3.8.7). It is well known, that even in the optimal solution the matrix computed by the BFGS update formula may differ substantially from the exact Hessian and therefore can not be used for sensitivity computations. Boggs et. al. [BTW82] stated a necessary and sufficient condition for Q-superlinear convergence of quasi-Newton methods for equality constrained nonlinear optimization problems: Under certain assumptions, the sequence $\{z^{(k)}\}$ converges Q-superlinearly to $\hat{z}$, if and only if

$$\frac{\|P(z^{(k)}, \hat{p})\,(B_k - L''_{zz}(\hat{z}, \hat{\mu}, \hat{p}))\,d^{(k)}\|}{\|d^{(k)}\|} \to 0,$$

where B_k is the BFGS update matrix and P is a projector on the tangent space of the constraints. This means, that the BFGS update is only a good approximation of the Hessian when projected on the linearized constraints at iterate $z^{(k)}$ in direction $d^{(k)}$.

Consequently, the Hessian L''_{zz} as well as the quantities L''_{zp}, H'_z, G'_z, H'_p, and G'_p have to be reevaluated after the SQP method converged.

The computation of the derivatives H'_z, G'_z, H'_p, and G'_p can be done efficiently by either the sensitivity equation approach of Section 6.2.1 or the adjoint equation approach of Sections 6.2.2

and 6.2.3. Notice that these approaches usually are less expensive and more reliable than finite difference approximations.

One idea to approximate the derivatives L''_{zz} and L''_{zp} is to use the finite difference approximations

$$\begin{aligned} L''_{z_iz_j}(\hat{z},\hat{\mu},\hat{p}) &\approx \frac{1}{4h^2}\left(L(\hat{z}+he_i+he_j,\hat{\mu},\hat{p}) - L(\hat{z}-he_i+he_j,\hat{\mu},\hat{p})\right.\\ &\qquad \left.-L(\hat{z}+he_i-he_j,\hat{\mu},\hat{p}) + L(\hat{z}-he_i-he_j,\hat{\mu},\hat{p})\right),\\ L''_{z_ip_j}(\hat{z},\hat{\mu},\hat{p}) &\approx \frac{1}{4h^2}\left(L(\hat{z}+he_i,\hat{\mu},\hat{p}+he_j) - L(\hat{z}-he_i,\hat{\mu},\hat{p}+he_j)\right.\\ &\qquad \left.-L(\hat{z}+he_i,\hat{\mu},\hat{p}-he_j) + L(\hat{z}-he_i,\hat{\mu},\hat{p}-he_j)\right) \end{aligned} \tag{7.2.4}$$

with appropriate unity vectors e_i, e_j. For each evaluation of the Lagrange function L one DAE has to be solved numerically. Hence, this approach leads to a total number of $4\frac{n_z(n_z+1)}{2}+4n_zn_p$ DAE evaluations if the symmetry of the Hessian L''_{zz} is exploited.

An alternative approach is to exploit the information obtained by the sensitivity equation. Recall that one evaluation of the DAE and the corresponding sensitivity equation for the NLP variables z provides the gradient of the objective function and the Jacobian of the constraints and with this information the gradient of the Lagrangian is obtained easily. Again, a finite difference approximation with these gradients yields the second order derivatives according to

$$\begin{aligned} L''_{zz_i}(\hat{z},\hat{\mu},\hat{p}) &\approx \frac{L'_z(\hat{z}+he_i,\hat{\mu},\hat{p}) - L'_z(\hat{z}-he_i,\hat{\mu},\hat{p})}{2h},\\ L''_{zp_j}(\hat{z},\hat{\mu},\hat{p}) &\approx \frac{L'_z(\hat{z},\hat{\mu},\hat{p}+he_j) - L'_z(\hat{z},\hat{\mu},\hat{p}-he_j)}{2h}. \end{aligned}$$

This time, only $2n_z+2n_p$ DAEs including the corresponding sensitivity equations (6.2.3) w.r.t. z have to be solved. Since the evaluation of n_z nonlinear DAEs in combination with the corresponding linear sensitivity DAE with n_z columns is usually much cheaper than the evaluation of ${n_z}^2$ nonlinear DAEs, the second approach is found to be much faster than a pure finite difference approximation (7.2.4).

The main drawback of this method based on a sensitivity analysis of ($DOCP(p)$ is, that a very accurate solution of $DOCP(\hat{p})$ is needed in view of getting good results for the sensitivity differentials. As a rule of thumb, the feasibility and optimality tolerance of the SQP method, compare Gill et al. [GMSW98], should be smaller than 10^{-9}, if the derivatives are computed by finite differences.

7.2.3 Numerical Example: Emergency Landing

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. Mayrhofer and Sachs [MS96], Büskens and Gerdts [BG00, BG03]. 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

$$\Phi(C_L,\mu,t_f) = -\left(\frac{\Lambda(t_f)-\Lambda(0)}{\Lambda(0)}\right)^2 - \left(\frac{\Theta(t_f)-\Theta(0)}{\Theta(0)}\right)^2 \tag{7.2.5}$$

subject to the ODE for the velocity v, the inclination γ, the azimuth angle χ, the altitude h, the latitude Λ, and the longitude Θ

$$\begin{aligned}
\dot{v} &= -D(v,h;C_L)\frac{1}{m} - g(h)\sin\gamma + \\
&\quad + \omega^2\cos\Lambda(\sin\gamma\cos\Lambda - \cos\gamma\sin\chi\sin\Lambda)R(h), && (7.2.6)\\
\dot{\gamma} &= L(v,h;C_L)\frac{\cos\mu}{mv} - \left(\frac{g(h)}{v} - \frac{v}{R(h)}\right)\cos\gamma + \\
&\quad + 2\omega\cos\chi\cos\Lambda + \omega^2\cos\Lambda(\sin\gamma\sin\chi\sin\Lambda + \cos\gamma\cos\Lambda)\frac{R(h)}{v}, && (7.2.7)\\
\dot{\chi} &= L(v,h;C_L)\frac{\sin\mu}{mv\cos\gamma} - \cos\gamma\cos\chi\tan\Lambda\frac{v}{R(h)} + \\
&\quad + 2\omega(\sin\chi\cos\Lambda\tan\gamma - \sin\Lambda) - \omega^2\cos\Lambda\sin\Lambda\cos\chi\frac{R(h)}{v\cos\gamma}, && (7.2.8)\\
\dot{h} &= v\sin\gamma, && (7.2.9)\\
\dot{\Lambda} &= \cos\gamma\sin\chi\frac{v}{R(h)}, && (7.2.10)\\
\dot{\Theta} &= \cos\gamma\cos\chi\frac{v}{R(h)\cos\Lambda}, && (7.2.11)
\end{aligned}$$

with functions

$$\begin{array}{rcllrcl}
L(v,h,C_L) &=& q(v,h)\ F\ C_L, & & \rho(h) &=& \rho_0\exp(-\beta h),\\
D(v,h,C_L) &=& q(v,h)\ F\ C_D(C_L), & & R(h) &=& r_0 + h,\\
C_D(C_L) &=& C_{D_0} + k\ {C_L}^2, & & g(h) &=& g_0(r_0/R(h))^2,\\
q(v,h) &=& \frac{1}{2}\rho(h)v^2 & & & &
\end{array} \qquad (7.2.12)$$

and constants

$$\begin{array}{rclrclrcl}
F &=& 305, & r_0 &=& 6.371\cdot 10^6, & C_{D_0} &=& 0.017,\\
k &=& 2, & \rho_0 &=& 1.249512\cdot(1+p), & \beta &=& 1/6900,\\
g_0 &=& 9.80665, & \omega &=& 7.27\cdot 10^{-5}, & m &=& 115000.
\end{array} \qquad (7.2.13)$$

The parameter p will be used to model perturbations in the air density $\rho(h)$ in (7.2.12). For the unperturbed problem it holds $p = \hat{p} = 0$.
Since the propulsion system is damaged, the mass m remains constant. Box constraints for the two control functions lift coefficient C_L and angle of bank μ are given by

$$\begin{array}{rcccl}
0.01 &\leq& C_L &\leq& 0.18326,\\
-\frac{\pi}{2} &\leq& \mu &\leq& \frac{\pi}{2}.
\end{array} \qquad (7.2.14)$$

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

$$\begin{pmatrix} v(0)\\ \gamma(0)\\ \chi(0)\\ h(0)\\ \Lambda(0)\\ \Theta(0) \end{pmatrix} = \begin{pmatrix} 2150.5452900\\ 0.1520181770\\ 2.2689279889\\ 33900.000000\\ 0.8651597102\\ 0.1980948701 \end{pmatrix}. \qquad (7.2.15)$$

An additional restriction is given by the terminal condition

$$h(t_f) = 500 - \varepsilon. \tag{7.2.16}$$

The parameter ε will be used to model perturbations in the final altitude $h(t_f)$. For the unperturbed problem it holds $\varepsilon = \varepsilon_0 = 0$.

The final time t_f is assumed to be free and thus t_f is an additional optimization variable. Finally, the dynamic pressure constraint

$$q(v, h) \leq q_{max} \tag{7.2.17}$$

has to be obeyed. Here, we used the value $q_{max} = 60000$ $[N/m^2]$.

The infinite dimensional optimal control problem is discretized by the reduced discretization approach. We used the classical fourth order Runge-Kutta scheme for time integration. The control is approximated by a continuous and piecewise linear function.

Figure 7.11 shows the numerical solution for 151 discretization points of the unperturbed problem with $p = \hat{p} = 0$ and $\varepsilon = \varepsilon_0 = 0$. The dynamic pressure constraint is active in the normalized time interval $[0.18, 0.186666666]$.

Table 7.4: Results for the emergency landing manoeuvre with and without dynamic pressure constraint for different SQP methods.

method	dyn. press.	t_f in $[s]$	objective	CPU nom. in $[s]$	CPU total in $[s]$	eigenvalue red. Hessian	eval
NPSOL	no	728.8493380	-0.7758097	60.66	93.40	$3.97 \cdot 10^{-5}$	500
NPSOL	yes	727.1079063	-0.7649252	60.04	94.54	$4.31 \cdot 10^{-5}$	382

Table 7.4 summarizes data for the emergency landing manoeuvre with and without dynamic pressure constraint. The columns 't_f' and 'objective' denote the final time and the objective function value, respectively, of the nominal solution of the discretized optimal control problems. The computations were performed on a personal computer with Pentium 4 processor and 2.66 GHz. The column 'CPU nom.' indicates the CPU time needed for the nominal problem, whereas 'CPU total' includes in addition the sensitivity analysis of the nonlinear programming problem. The column 'eigenvalue red. Hessian' denotes the smallest eigenvalue of the reduced Hessian. If this entry is positive then the second order sufficient conditions in Theorem 3.6.8 are satisfied for the nonlinear optimization problem. In particular the Hessian of the Lagrangian is positive definite on the kernel of the linearized active constraints. The column 'eval' denotes the number of objective function and constraint evaluations of the employed SQP method. The optimality tolerance and feasibility tolerance for NPSOL were both set to 10^{-12}.

Table 7.5 summarizes data for the emergency landing manoeuvre with and without dynamic pressure constraint.

Table 7.5: Sensitivities of the nominal final time t_f w.r.t. perturbation parameters p and ε for the emergency landing manoeuvre with and without dynamic pressure constraint.

method	dyn. press.	dt_f/dp	$dt_f/d\varepsilon$
NPSOL	no	91.8538973	0.0145585106
NPSOL	yes	92.3851543	0.0141449020

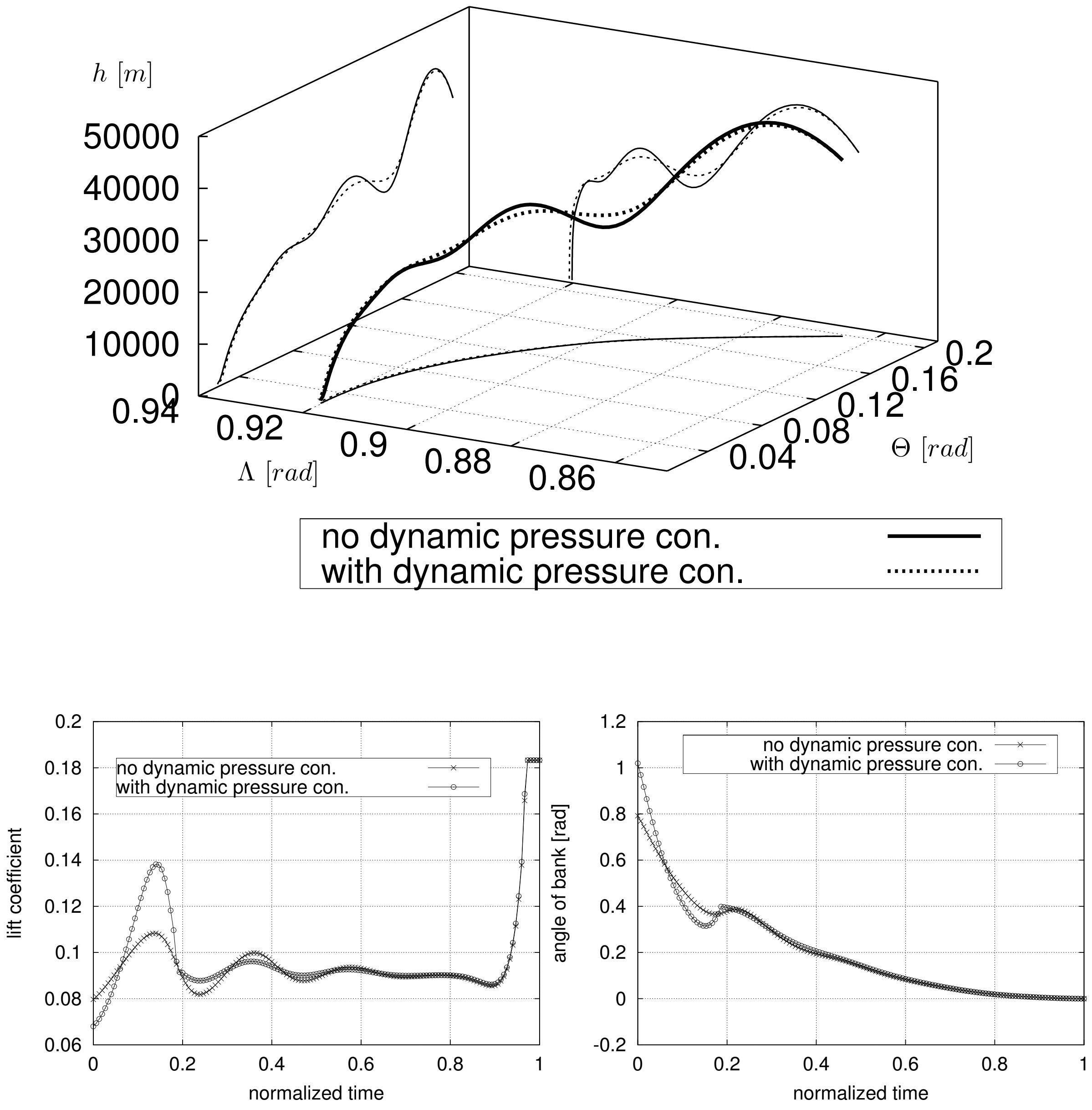


Figure 7.11: Comparison of the numerical nominal solutions for the problem with and without dynamic pressure constraint: 3D plot of the flight path (top) and the approximate optimal controls lift coefficient C_L and angle of bank μ (bottom, normalized time scale) for 151 grid points.

Figures 7.12-7.15 show the sensitivities of the nominal controls C_L and μ w.r.t. the perturbation parameters p and ε. For the problem with dynamic pressure constraint the sensitivities jump at points where the state constraint becomes active resp. inactive.

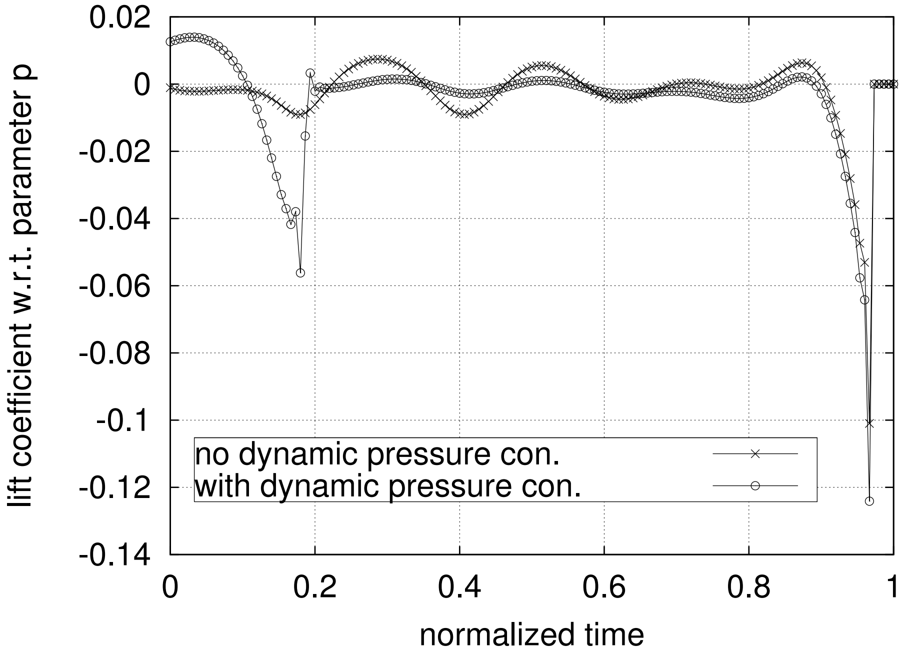


Figure 7.12: Comparison of the sensitivities dC_L/dp of the lift coefficient C_L w.r.t. perturbation parameter p for the problem with and without dynamic pressure constraint (normalized time scale) for 151 grid points.

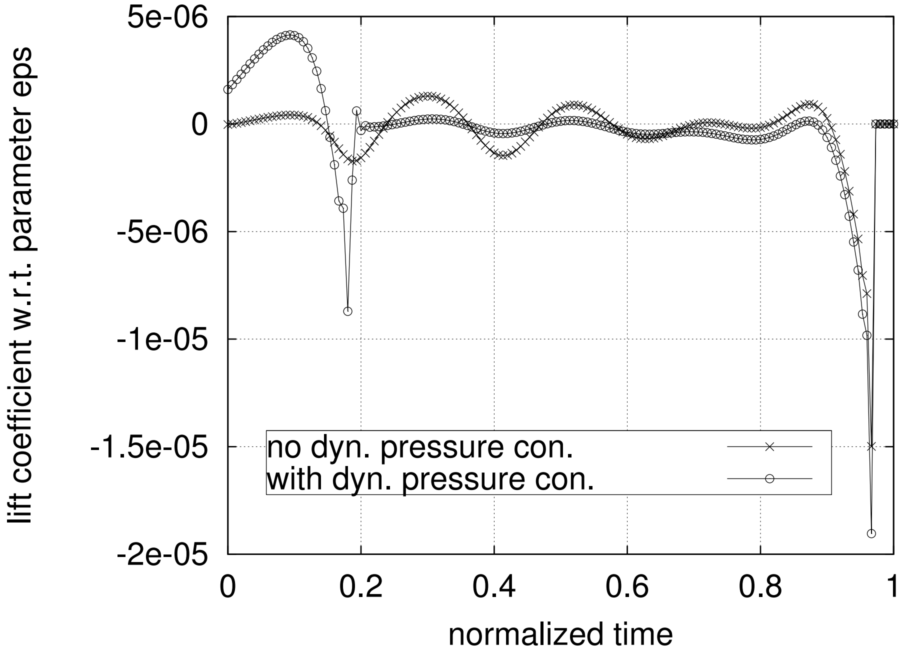


Figure 7.13: Comparison of the sensitivities $dC_L/d\varepsilon$ of the lift coefficient C_L w.r.t. perturbation parameter ε for the problem with and without dynamic pressure constraint (normalized time scale) for 151 grid points.

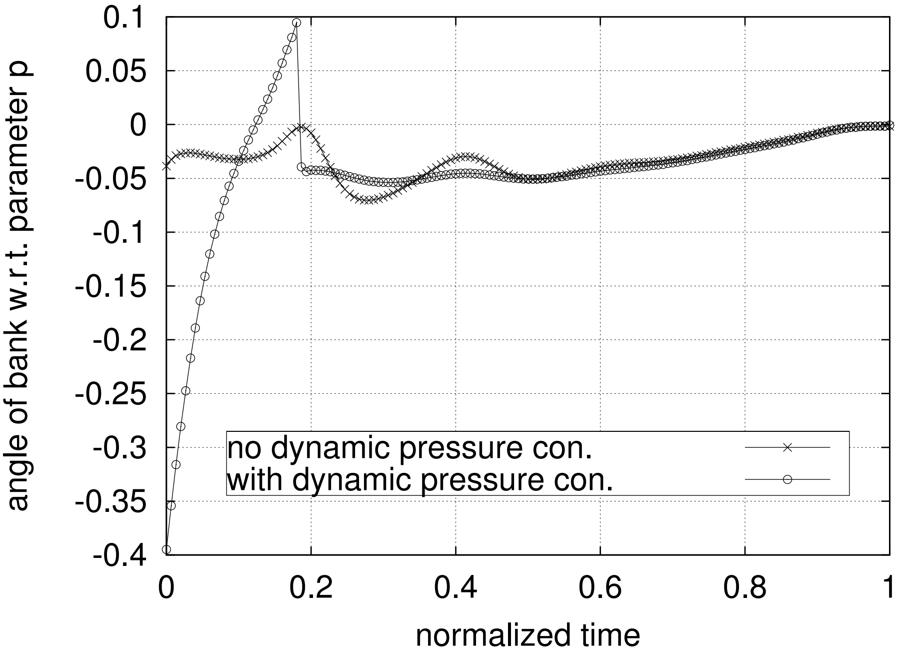


Figure 7.14: Comparison of the sensitivities $d\mu/dp$ of the angle of bank μ w.r.t. perturbation parameter p for the problem with and without dynamic pressure constraint (normalized time scale) for 151 grid points.

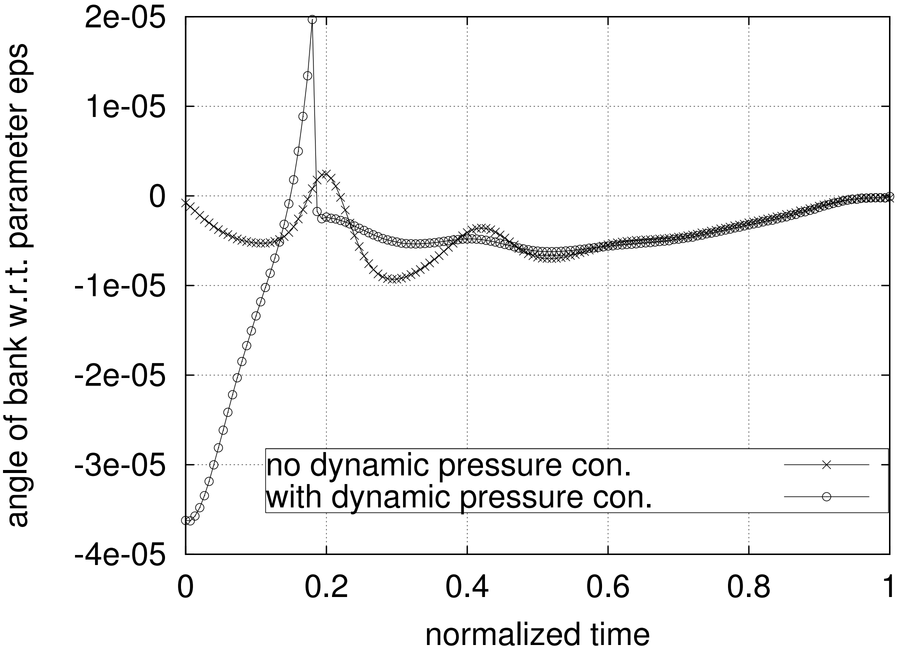


Figure 7.15: Comparison of the sensitivities $d\mu/d\varepsilon$ of the angle of bank μ w.r.t. perturbation parameter ε for the problem with and without dynamic pressure constraint (normalized time scale) for 151 grid points.

Remark 7.2.4 *A corrector iteration method for the reduction of constraint violations which* *unavoidedly occur due to the linearization in (3.7.4) was developed in Büskens [Büs01]. The application to the emergency problem is discussed in Büskens and Gerdts [BG03] as well.*

7.2.4 Numerical Example: Test-Drives

The test-drive to be simulated is again the double-lane change manoeuvre discussed already in Section 7.1.3. Instead of the single-track model a more sophisticated full car model of a BMW 1800/2000 is used. The equations of motion form a semi-explicit index 1 DAE (1.2)-(1.2) with $n_x = 40$ differential equations and $n_y = 4$ algebraic equations. The detailed model can be found in von Heydenaber [vH80] and Gerdts [Ger01a, Ger03b, Ger03c]. We assume that the double-lane change manoeuvre is driven at (almost) constant velocity, i.e. the braking force is set to zero, the acceleration is chosen such that it compensates the effect of rolling resistance, and the gear is fixed. The remaining control u denotes the steering wheel velocity. There are two real-time parameters p_1 and p_2 involved in the problem. The first one p_1 denotes the offset of the test-course with nominal value 3.5 $[m]$, cf. Figure 7.7. The second one p_2 denotes the height of the car's center of gravity with nominal value 0.56 $[m]$. While p_2 influences the dynamics of the car, p_1 influences only the state constraints of the optimal control problem. A similar problem with a different car model was already investigated in Büskens and Gerdts [GB01].

Again, the driver is modeled by formulation of an appropriate optimal control problem with free final time t_f. The car's initial position on the course is fixed. At final time t_f boundary conditions are given by prescribed x-position (140 $[m]$) and yaw angle (0 $[rad]$). In addition, the steering capability of the driver is restricted by $|u(t)| \leq 3$ $[rad/s]$. The objective is to minimize a linear combination of final time and steering effort, i.e.

$$40\,t_f + \int_{t_0}^{t_f} u(t)^2 dt \to \min .$$

Figure 7.16 shows the nominal control and the sensitivity differentials for the control at 101 grid points obtained with a piecewise linear approximation of the control, i.e. a B-spline representation with $k = 2$, and the linearized RADAUIIA method of Section 5.4 for time integration.

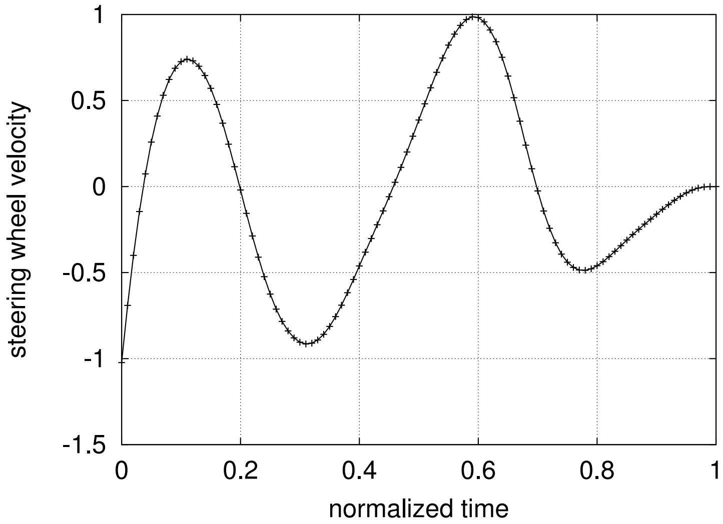


Figure 7.16: Nominal control for $p_1 = 3.5$ $[m]$ and $p_2 = 0.56$ $[m]$ for 101 grid points.

The nominal objective function value is 273.7460. The optimal final time is $t_f = 6.7930367$ $[s]$

and its sensitivity differentials are

$$\frac{dt_f}{dp_1} = 0.042409, \qquad \frac{dt_f}{dp_2} = 0.022685.$$

The second order sufficient conditions in Theorem 3.6.8 are satisfied. The minimal eigenvalue of the reduced Hessian is 0.034. The CPU time for the nominal solution amounts to 318.53 $[s]$, the overall CPU time including sensitivity analysis is 463.78 $[s]$.
The sensitivities are obtained by a sensitivity analysis of the discretized optimal control problem and are depicted in Figures 7.17 and 7.18.

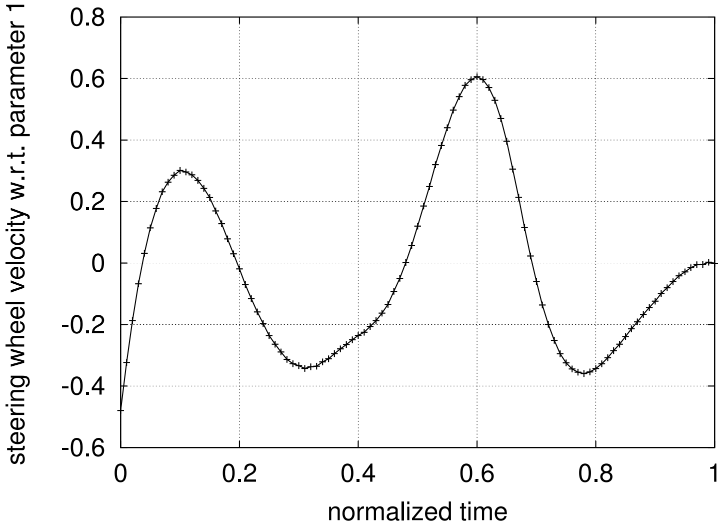


Figure 7.17: Sensitivity of the steering wheel velocity w.r.t. to the offset in the state constraints p_1 for $N = 100$ control grid points.

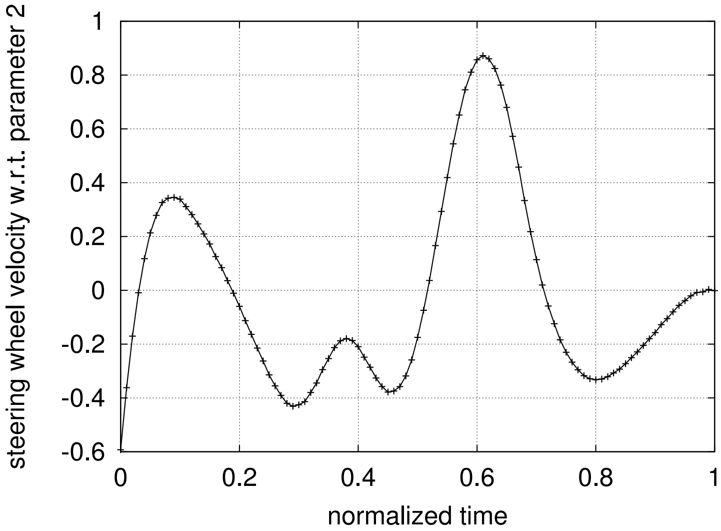


Figure 7.18: Sensitivity of the steering wheel velocity w.r.t. to the height of the center of gravity p_2 for $N = 100$ control grid points.

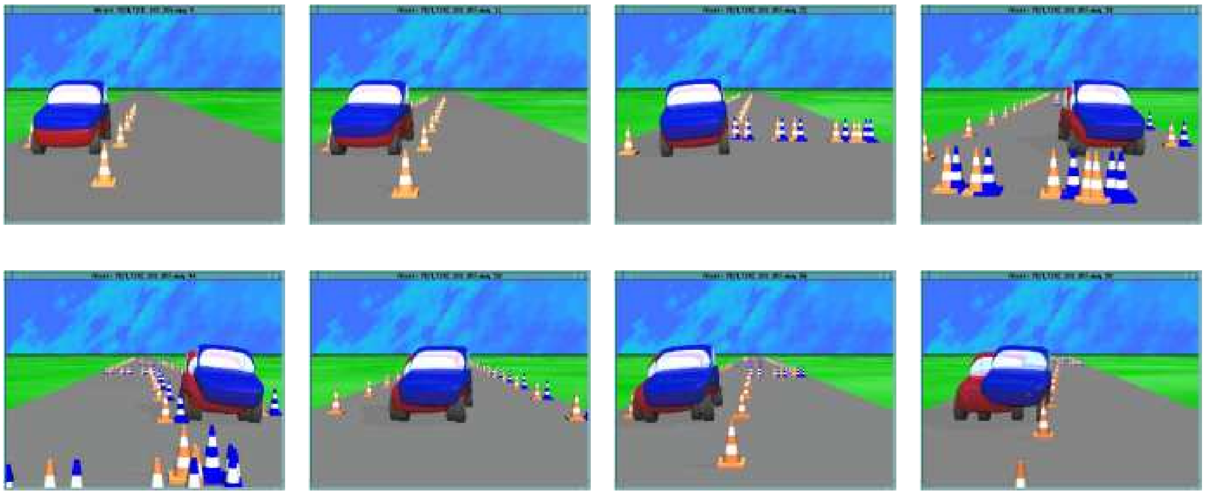
Table 7.6 points out the accuracy of the real-time approximation when compared to the optimal solution for that particular perturbation. For small perturbations, i.e. ±1%, the real-time approximation is very accurate.

Table 7.6: Errors in objective function, boundary conditions, path constraints and control. The errors denote the difference between the real-time approximation and the corresponding optimal solution for the particular perturbations in p_1 and p_2. For a perturbation of $\geq +2\%$ or $\leq -3\%$ in both parameters, the active set changes and Theorem 3.7.3 is not applicable anymore.

Perturb. %	objective abs./rel.	bound. cond. abs.	state constr. max. abs.	control max. abs./abs. L_2	structure of control
$+20$	$2.2 \cdot 10^{-2}/7.8 \cdot 10^{-3}$	$8.4 \cdot 10^{-4}/2.5 \cdot 10^{-3}$	$3.5 \cdot 10^{-1}$	$4.3 \cdot 10^{-1}/1.6 \cdot 10^{-2}$	diff.
$+10$	$6.0 \cdot 10^{-3}/2.2 \cdot 10^{-3}$	$1.5 \cdot 10^{-4}/1.1 \cdot 10^{-3}$	$1.6 \cdot 10^{-1}$	$2.6 \cdot 10^{-1}/8.4 \cdot 10^{-2}$	diff.
$+5$	$2.2 \cdot 10^{-3}/7.9 \cdot 10^{-4}$	$2.9 \cdot 10^{-5}/3.3 \cdot 10^{-4}$	$8.1 \cdot 10^{-2}$	$2.2 \cdot 10^{-1}/6.0 \cdot 10^{-2}$	diff.
$+4$	$1.5 \cdot 10^{-3}/5.4 \cdot 10^{-4}$	$1.7 \cdot 10^{-5}/2.2 \cdot 10^{-4}$	$6.2 \cdot 10^{-2}$	$1.9 \cdot 10^{-1}/5.5 \cdot 10^{-2}$	diff.
$+3$	$8.7 \cdot 10^{-4}/3.2 \cdot 10^{-4}$	$9.0 \cdot 10^{-6}/1.2 \cdot 10^{-4}$	$4.2 \cdot 10^{-2}$	$1.7 \cdot 10^{-1}/5.0 \cdot 10^{-2}$	diff.
$+2$	$3.0 \cdot 10^{-4}/1.1 \cdot 10^{-4}$	$3.7 \cdot 10^{-6}/5.6 \cdot 10^{-5}$	$2.1 \cdot 10^{-2}$	$1.4 \cdot 10^{-1}/4.7 \cdot 10^{-2}$	diff.
$+1$	$1.9 \cdot 10^{-5}/6.8 \cdot 10^{-6}$	$8.4 \cdot 10^{-7}/1.4 \cdot 10^{-5}$	$8.0 \cdot 10^{-4}$	$4.7 \cdot 10^{-4}/1.8 \cdot 10^{-4}$	eq.
-1	$1.8 \cdot 10^{-5}/6.4 \cdot 10^{-6}$	$6.6 \cdot 10^{-7}/1.5 \cdot 10^{-5}$	$8.1 \cdot 10^{-4}$	$4.4 \cdot 10^{-4}/1.7 \cdot 10^{-4}$	eq.
-2	$6.8 \cdot 10^{-5}/2.5 \cdot 10^{-5}$	$2.3 \cdot 10^{-6}/6.0 \cdot 10^{-5}$	$3.3 \cdot 10^{-3}$	$1.7 \cdot 10^{-3}/6.6 \cdot 10^{-4}$	eq.
-3	$1.5 \cdot 10^{-4}/5.5 \cdot 10^{-5}$	$4.4 \cdot 10^{-6}/1.4 \cdot 10^{-4}$	$7.4 \cdot 10^{-3}$	$1.2 \cdot 10^{-2}/4.8 \cdot 10^{-3}$	diff.
-4	$2.7 \cdot 10^{-4}/9.9 \cdot 10^{-5}$	$6.4 \cdot 10^{-6}/2.5 \cdot 10^{-4}$	$1.3 \cdot 10^{-2}$	$2.6 \cdot 10^{-2}/1.0 \cdot 10^{-2}$	diff.
-5	$4.2 \cdot 10^{-4}/1.5 \cdot 10^{-4}$	$7.8 \cdot 10^{-6}/3.9 \cdot 10^{-4}$	$2.1 \cdot 10^{-2}$	$2.7 \cdot 10^{-2}/1.1 \cdot 10^{-2}$	diff.
-10	$1.6 \cdot 10^{-3}/5.9 \cdot 10^{-4}$	$1.5 \cdot 10^{-5}/1.7 \cdot 10^{-3}$	$8.5 \cdot 10^{-2}$	$5.3 \cdot 10^{-2}/2.4 \cdot 10^{-2}$	diff.
-20	$5.0 \cdot 10^{-3}/1.8 \cdot 10^{-3}$	$4.2 \cdot 10^{-4}/6.8 \cdot 10^{-3}$	$3.3 \cdot 10^{-1}$	$1.9 \cdot 10^{-1}/8.2 \cdot 10^{-2}$	diff.

But even for large perturbations $p_1 = 3.85$ $[m]$ and $p_2 = 0.728$ $[m]$ the real-time approximation provides reasonable results as depicted in Figure 7.19, since the car still stays within the boundaries.

$t_0 = 0$ $[s]$



$t_f = 6.793$ $[s]$

Figure 7.19: Snapshots of the nominal solution and the real-time approximation for perturbations $p_1 = 3.85$ $[m]$ and $p_2 = 0.728$ $[m]$.

7.3 Dynamic Parameter Identification

Suppose that a time dependent technical process within some fixed time interval $[t_0, t_f]$ can be modeled mathematically in terms of the parametric DAE

$$F(t, x(t), \dot{x}(t), p) = 0_{n_x}, \quad x(t_0) = X_0(p) \tag{7.3.1}$$

with sufficiently smooth functions $F : [t_0, t_f] \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_p} \to \mathbb{R}^{n_x}$ and $X_0 : \mathbb{R}^{n_p} \to \mathbb{R}^{n_x}$. Again, X_0 is a function that provides a consistent initial value for a given parameter p. The unknown parameters $p \in \mathbb{R}^{n_p}$ in it have to be identified out of measured data. Suppose that a measurement of the technical process at the grid points

$$\mathbb{G} := \{t_1, t_2, \ldots, t_N\}, \quad N \in \mathbb{N}, \quad t_0 \le t_1 < \ldots < t_N \le t_f \tag{7.3.2}$$

yields the N measurement vectors

$$y_i = (y_{i1}, \ldots, y_{iM})^\top \in \mathbb{R}^M, \quad i = 1, \ldots, N.$$

The measurements are related to the mathematical model by a sufficiently smooth *output function* $h = (h_1, \ldots, h_M)^\top : [t_0, t_f] \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_p} \to \mathbb{R}^M$ according to

$$\varepsilon_{ij} = y_{ij} - h_j(t_i, x(t_i), p), \quad i = 1, \ldots, N, \; j = 1, \ldots, M. \tag{7.3.3}$$

Herein, $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iM})^\top \in \mathbb{R}^M$ denote measurement errors at the grid points. It is assumed that the time points t_i, $i = 1, \ldots, N$ are deterministic, while the measurement vectors y_i, $i = 1, \ldots, N$ underly measurement errors.

The errors ε_{ij} are interpreted as realizations of the N random vectors $\xi_i = (\xi_{i1}, \ldots, \xi_{iM})^\top$ with expectations $\mu_i = 0_M$ and covariance matrices $V_i \in \mathbb{R}^{M \times M}$, $i = 1, \ldots, N$. Furthermore, it is assumed that the variables ξ_i, $i = 1, \ldots, N$ are Gaussian variables with probability density

$$\begin{aligned} f_i(\xi_i) &= \frac{1}{\sqrt{2\pi}^M \cdot \sqrt{det(V_i)}} \exp\left(-\frac{1}{2}(\xi_i - \mu_i)^\top V_i^{-1}(\xi_i - \mu_i)\right) \\ &= \frac{1}{\sqrt{2\pi}^M \cdot \sqrt{det(V_i)}} \exp\left(-\frac{1}{2}\xi_i^\top V_i^{-1}\xi_i\right). \end{aligned}$$

If ξ_i, $i = 1, \ldots, N$ are independent, then the probability density of the random matrix $\xi = (\xi_1 \mid \ldots \mid \xi_N) \in \mathbb{R}^{M \times N}$ is given by

$$\begin{aligned} f(\xi) &= \prod_{i=1}^N f_i(\xi_i) \\ &= \prod_{i=1}^N \frac{1}{\sqrt{2\pi}^M \cdot \sqrt{det(V_i)}} \exp\left(-\frac{1}{2}\xi_i^\top V_i^{-1}\xi_i\right) \\ &= \frac{1}{\sqrt{2\pi}^{N \cdot M}} \cdot \left(\prod_{i=1}^N \frac{1}{\sqrt{det(V_i)}}\right) \exp\left(-\frac{1}{2}\sum_{i=1}^N \xi_i^\top V_i^{-1}\xi_i\right). \end{aligned}$$

The likelihood function $\mathcal{L}(p)$ arises, if the random variable ξ_i is replaced by $y_i - h(t_i, x(t_i), p)$:

$$\begin{aligned} \mathcal{L}(p) &:= f(Y - H(t, x, p)) \\ &= \frac{1}{\sqrt{2\pi}^{N \cdot M}} \left(\prod_{i=1}^N \frac{1}{\sqrt{det(V_i)}}\right) \exp\left(-\frac{1}{2}\sum_{i=1}^N (y_i - h(t_i, x(t_i), p))^\top V_i^{-1}(y_i - h(t_i, x(t_i), p))\right), \end{aligned}$$

where $Y = (y_1 \mid \dots \mid y_N) \in \mathbb{R}^{M\times N}$, $t = (t_1,\dots,t_N)^\top \in \mathbb{R}^N$, $x = (x(t_1),\dots,x(t_N))^\top \in \mathbb{R}^{N\cdot n_x}$, and $H = (h(t_1,x(t_1),p) \mid \dots \mid h(t_N,x(t_N),p)) \in \mathbb{R}^{M\times N}$. Herein, $\mathcal{L}(p)$ denotes the probability that $\xi_{ij} = \varepsilon_{ij} = y_{ij} - h_j(t_i,x(t_i),p)$, $i = 1,\dots,N$ hold. The aim of the maximum-likelihood method is to maximize the likelihood function $\mathcal{L}(p)$ w.r.t. p in order to obtain the largest probability that the parameter p reproduces the measurements. Application of the natural logarithm yields the equivalent problem

$$\begin{aligned} \ln(\mathcal{L}(p)) &= -\frac{N\cdot M}{2}\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{N}\ln(det(V_i)) \\ &\quad -\frac{1}{2}\sum_{i=1}^{M}(y_i - h(t_i,x(t_i),p))^\top V_i^{-1}(y_i - h(t_i,x(t_i),p)) \\ &\to \max. \end{aligned}$$

If the covariance matrices are known, then this problem is equivalent with the weighted least squares problem

$$\frac{1}{2}\sum_{i=1}^{N}(y_i - h(t_i,x(t_i),p))^\top V_i^{-1}(y_i - h(t_i,x(t_i),p)) \to \min. \tag{7.3.4}$$

As a special case we consider independent Gaussian random variables ξ_{ij} with variances σ_{ij}^2. Then, it follows $V_i = diag(\sigma_{i1}^2,\dots,\sigma_{im}^2)$ and the least squares problem reduces to

$$\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{(y_{ij} - h_j(t_i,x(t_i),p))^2}{\sigma_{ij}^2} \to \min. \tag{7.3.5}$$

We obtained a maximum likelihood estimator for the parameters, cf. von Schwerin [vS99] and Grupp [Gru96]. Moreover, in practical applications also inequality and equality constraints depending on the parameter vector p have to be obeyed. The problems (7.3.4) and (7.3.5) can be subsumed in

Problem 7.3.1 (Least-Squares-Problem)
Find $p \in \mathbb{R}^{n_p}$ such that (7.3.4) is minimized subject to (7.3.1) and

$$G(p) \le 0_{n_G}, \qquad H(p) = 0_{n_H}.$$

Herein, all functions are assumed to be sufficiently smooth. Problem 7.3.1 is a nonlinear program and thus can be solved by general purpose SQP methods similar as in the reduced discretization approach of Chapter 6. In each iteration of the SQP method the DAE has to be solved numerically and a sensitivity analysis has to be performed. Often the Gauss-Newton method is preferable for least-squares problems, since, under suitable assumptions it converges quadratically, provided that the optimal objective function is close to zero, cf. Bock [Boc87]. Schittkowski [Sch94] shows how to simulate a Gauss-Newton method by use of a general purpose SQP method.

7.3.1 Numerical Example: Parameter Identification in a Truck Model

A planar model of a truck driving on an uneven road at constant speed 30 $[m/s]$ is described in detail in Simeon et al. [SGFR94]. The road excitation is modeled by a Fourier series. The mechanical multi-body model, cf. Example 1.9, consists of seven bodies with the generalized position vector $q = (q_1, \ldots, q_{11})^\top \in \mathbb{R}^{11}$ and the components

q_1	:	vertical motion of rear tire
q_2	:	vertical motion of front tire
q_3	:	vertical motion of truck chassis
q_4	:	rotation about y-axis of truck chassis
q_5	:	vertical motion of engine
q_6	:	rotation about y-axis of engine
q_7	:	vertical motion of driver cabin
q_8	:	rotation about y-axis of driver cabin
q_9	:	vertical motion of driver seat
q_{10}	:	vertical motion of loading area
q_{11}	:	rotation about y-axis of loading area

The equations of motion are given by the stabilized index 2 DAE

$$\begin{aligned}
m_1\ddot{q}_1 &= -F_{10} + F_{13} - m_1 g,\\
m_2\ddot{q}_2 &= -F_{20} + F_{23} - m_2 g,\\
m_3\ddot{q}_3 &= -F_{13} - F_{23} + F_{35} + F_{34} + F_{43} + F_{53} + F_{37} - m_3 g + \lambda,\\
l_3\ddot{q}_4 &= (a_{23}F_{23} - a_{13}F_{13} - a_{37}F_{37} - a_{34}F_{34} - a_{35}F_{35} - a_{43}F_{43}\\
&\qquad\qquad -a_{53}F_{53}) \cos q_4 - (-a_{z_1}\cos q_4 + a_{z_2}\sin q_4)\lambda,\\
m_4\ddot{q}_5 &= -F_{43} - F_{34} - m_4 g,\\
l_4\ddot{q}_6 &= (b_{43}F_{43} - b_{34}F_{34}) \cos q_6,\\
m_5\ddot{q}_7 &= -F_{53} - F_{35} + F_{56} - m_5 g,\\
l_5\ddot{q}_8 &= (c_{53}F_{53} - c_{35}F_{35} - c_{56}F_{56}) \cos q_8,\\
m_6\ddot{q}_9 &= -F_{56} - m_6 g,\\
m_7\ddot{q}_{10} &= -F_{37} - m_7 g - \lambda,\\
l_7\ddot{q}_{11} &= e_{37}\cos(q_{11})F_{37} - (e_{z_1}\cos q_{11} + e_{z_2}\sin q_{11})\lambda,\\
0 &= C(q_3, q_4, q_{10}, q_{11}, \dot{q}_3, \dot{q}_4, \dot{q}_{10}, \dot{q}_{11}),\\
0 &= c(q_3, q_4, q_{10}, q_{11})
\end{aligned}$$

with the algebraic constraint on velocity level (index 2 constraint)

$$\begin{aligned}
&C(q_3, q_4, q_{10}, q_{11}, \dot{q}_3, \dot{q}_4, \dot{q}_{10}, \dot{q}_{11}) =\\
&\quad = -\dot{q}_3 + \dot{q}_4(-a_{z_1}\cos q_4 + a_{z_2}\sin q_4) + \dot{q}_{10} + \dot{q}_{11}(e_{z_1}\cos q_{11} + e_{z_2}\sin q_{11})
\end{aligned}$$

and the algebraic constraint on position level (stabilizing constraint)

$$\begin{aligned}
&c(q_3, q_4, q_{10}, q_{11}) =\\
&\quad = -q_3 - a_{z_1}\sin q_4 - a_{z_2}\cos q_4 + q_{10} + e_{z_1}\sin q_{11} - e_{z_2}\cos q_{11} + h_{eq}.
\end{aligned}$$

The model depends on a variety of parameters, cf. Simeon et al. [SGFR94] for a detailed description. In this example, we focus in particular on the damping coefficients d_{13} and d_{23}, the coefficient κ influencing the pneumatic spring force law, and the geometric constant h_{eq}. The forces F_{13} and F_{23} depend on the parameters d_{13}, d_{23}, and κ, whereas the position constraint c depends on h_{eq}. Three different parameter identification problems are to be developed in the sequel.

First the output function h in (7.3.4) and the sourcing of measured data is described according to Heim [Hei92]. For simplicity $N = 15$ equally spaced measure points

$$t_i = t_0 + (i-1)\cdot h, \quad i = 1,\ldots,N, \quad h = (t_f - t_0)/(N-1)$$

with initial time $t_0 = 3$ $[s]$ and final time $t_f = 6.5$ $[s]$ are used. The $n_h = 7$ dimensional output function $h = (h_1,\ldots,h_7)^\top$ is given by

$$h(q_1,q_2,q_3,q_5,q_7,q_9,q_{10}) = (q_1,q_2,q_3,q_5,q_7,q_9,q_{10})^\top \in \mathbb{R}^7.$$

The measured data y_i at measure point t_i is obtained by a perturbation $\varepsilon_i \in \mathbb{R}^{n_h}$ of the nominal solution $q_j(t_i)$, $j = 1,2,3,5,7,9,10$ of the DAE, which is given for

$$\kappa = 1.4,\ d_{13} = 21593\ [Ns/m],\ d_{23} = 38537\ [Ns/m],\ h_{eq} = 0.9\ [m],\ q(t_0) = \dot{q}(t_0) = 0_{11}.$$

Consistent initial values for $\lambda(t_0)$ are obtained as in Section 6.1.1 and in Gerdts [Ger01a, Ger05d, Ger03a]. The measurement errors ε_i are assumed to be normally distributed with expected value 0 $[m]$. In addition, it is assumed that the standard deviation of the measurement errors ε_i at every measure point t_i, $i = 1,\ldots,N$ for every component h_j, $j = 1,\ldots,n_h$ is equal, i.e. $\sigma = \sigma_{ij}$, $i = 1,\ldots,N$, $j = 1,\ldots,n_h$. In the nominal solution, the positions $q_j(t_i)$, $j = 1,2,3,5,7,9,10$ approximately range from -0.015 $[m]$ to 0.015 $[m]$.

The objective function is given by the least squares objective function

$$\Phi = \frac{1}{2}\sum_{i=1}^{N} w\|y_i - h(q_1(t_i),q_2(t_i),q_3(t_i),q_5(t_i),q_7(t_i),q_9(t_i),q_{10}(t_i))\|_2^2$$

with Euclidian norm $\|\cdot\|_2$ and weight w.

According to Heim [Hei92] two identification problems can be formulated:

(i) For given κ and h_{eq} the parameters d_{13} and d_{23} are to be identified out of measured data.

(ii) For given coefficients d_{13}, d_{23} and h_{eq} the parameter κ is to be identified out of measured data.

Table 7.7: Identification of d_{13} and d_{23}: Dependency of the results with respect to the standard deviation σ of normally distributed measurement errors with expected value 0 $[m]$.

Parameter	Nominal	σ $[m]$	Result	Abs. Error	Rel. Error	CPU $[s]$
d_{13}	21593	10^{-3}	22209.7229	616.7229	$2.86\cdot10^{-2}$	18.02
		10^{-4}	21655.3010	62.3010	$2.89\cdot10^{-3}$	19.65
		10^{-5}	21599.3308	6.3308	$2.93\cdot10^{-4}$	18.05
		10^{-6}	21593.6391	0.6391	$2.96\cdot10^{-5}$	19.72
d_{23}	38537	10^{-3}	39297.3527	760.3527	$1.97\cdot10^{-2}$	cf. d_{13}
		10^{-4}	38614.1861	77.1861	$2.00\cdot10^{-3}$	cf. d_{13}
		10^{-5}	38545.7403	8.7403	$2.27\cdot10^{-4}$	cf. d_{13}
		10^{-6}	38537.7730	0.7730	$2.01\cdot10^{-5}$	cf. d_{13}

In Table 7.7 the results of the first identification problem for d_{13} and d_{23} for given κ are summarized. The initial guess for d_{13} was 20000 $[Ns/m]$, that for d_{23} was 30000 $[Ns/m]$. The optimality and feasibility tolerance within the SQP method was set to 10^{-7}. The objective function in the resulting nonlinear program was scaled by a factor of $w = 10^6$. The code DASSL of Petzold [Pet82a] was used for time integration.

In the second identification problem the parameter κ for given d_{13} and d_{23} is to be identified out of measured data. In Table 7.8 the results of the second identification problem are summarized. The initial guess for κ was 1.3. The optimality and feasibility tolerance within the SQP method was set to 10^{-7}. The objective function in the resulting nonlinear program was not scaled, i.e. $w = 1$.

Table 7.8: Identification of κ: Dependency of the results with respect to the standard deviation σ of normally distributed measurement errors with expected value 0 $[m]$.

Parameter	Nominal	σ $[m]$	Result	Abs. Error	Rel. Error	CPU $[s]$
κ	1.4	10^{-3}	1.4129911	$1.29911 \cdot 10^{-2}$	$9.28 \cdot 10^{-3}$	12.36
		10^{-4}	1.4012414	$1.2414 \cdot 10^{-3}$	$8.87 \cdot 10^{-4}$	12.21
		10^{-5}	1.4001233	$1.233 \cdot 10^{-4}$	$8.81 \cdot 10^{-5}$	12.24
		10^{-6}	1.4000121	$1.21 \cdot 10^{-5}$	$8.64 \cdot 10^{-6}$	12.21

The parameter h_{eq} influences the algebraic constraint on position level. Hence, a third identification problem is given by

(iii) The parameter h_{eq} as well as the unknown initial vertical positions $q_3(t_0)$ and $q_{10}(t_0)$ are to be identified out of the measured data for given d_{13}, d_{23} and κ.

This problem is more complicated than the previous problems, since the sought values occur in the stabilizing position constraint. During the optimization process the current iterates in general are inconsistent with the position constraint. Hence, the projection method described in Section 6.1.1 is essentially needed to compute consistent values.

Table 7.9 summarizes the results of the third identification problem. As before the optimality and feasibility tolerance within the SQP method was set to 10^{-7}. The initial guesses for h_{eq}, $q_3(t_0)$ and $q_{10}(t_0)$ were 0.7 $[m]$, 0 $[m]$ and 0 $[m]$, respectively. The objective function in the resulting nonlinear program was scaled with $w = 2$. A multiple shooting approach similar to the reduced approach was used with three equidistant multiple shooting nodes. Consistent values at the multiple shooting nodes are obtained by the projection method as in Section 6.1.1.

Table 7.9: Identification of h_{eq}, $q_3(t_0)$ and $q_{10}(t_0)$: Dependency of the results with respect to the standard deviation σ of normally distributed measurement errors with expected value 0 $[m]$.

Parameter	Nominal	σ $[m]$	Result	Abs. Error	Rel. Error	CPU $[s]$
h_{eq}	0.9	10^{-3}	0.8993478	$6.522 \cdot 10^{-4}$	$7.25 \cdot 10^{-4}$	69.12
		10^{-4}	0.8999344	$6.56 \cdot 10^{-5}$	$7.29 \cdot 10^{-5}$	83.00
		10^{-5}	0.8999935	$6.5 \cdot 10^{-6}$	$7.22 \cdot 10^{-6}$	83.18
		10^{-6}	0.8999994	$6.0 \cdot 10^{-7}$	$6.67 \cdot 10^{-7}$	82.46
$q_3(t_0)$	0	10^{-3}	$9.99 \cdot 10^{-5}$	$9.99 \cdot 10^{-5}$		cf. h_{eq}
		10^{-4}	$9.91 \cdot 10^{-6}$	$9.91 \cdot 10^{-6}$		cf. h_{eq}
		10^{-5}	$9.93 \cdot 10^{-7}$	$9.93 \cdot 10^{-7}$		cf. h_{eq}
		10^{-6}	$1.06 \cdot 10^{-7}$	$1.06 \cdot 10^{-7}$		cf. h_{eq}
$q_{10}(t_0)$	0	10^{-3}	$7.52 \cdot 10^{-4}$	$7.52 \cdot 10^{-4}$		cf. h_{eq}
		10^{-4}	$7.55 \cdot 10^{-5}$	$7.55 \cdot 10^{-5}$		cf. h_{eq}
		10^{-5}	$7.54 \cdot 10^{-6}$	$7.54 \cdot 10^{-6}$		cf. h_{eq}
		10^{-6}	$7.50 \cdot 10^{-7}$	$7.50 \cdot 10^{-7}$		cf. h_{eq}

All computations are performed on a personal computer with 750 MHz processing speed.

Bibliography

[AF96] Adjiman, C. S. and Floudas, C. A. *Rigorous convex underestimators for general twice-differentiable problems*. Journal of Global Optimization, 9 (1); 23–40, 1996.

[AF01] Adjiman, C. S. and Floudas, C. A. *The αBB global optimization algorithm for nonconvex problems: An overview*. In *From local to global optimization. Papers from the conference dedicated to Professor Hoang Tuy on the occasion of his 70th birthday, Rimforsa, Sweden, August 1997* (A. M. et al., editor), volume 53 of *Nonconvex Optimization and Applications*, pp. 155–186. Kluwer Academic Publishers, Dordrecht, 2001.

[AF04a] Akrotirianakis, I. G. and Floudas, C. A. *Computational experience with a new class of convex underestimators: Box-constrained NLP problems*. Journal of Global Optimization, 29 (3); 249–264, 2004.

[AF04b] Akrotirianakis, I. G. and Floudas, C. A. *A new class of improved convex underestimators for twice continuously differentiable constrained NLPs*. Journal of Global Optimization, 30 (4); 367–390, 2004.

[Alt91] Alt, W. *Sequential Quadratic Programming in Banach Spaces*. In *Advances in Optimization* (W. Oettli and D. Pallaschke, editors), pp. 281–301. Springer, Berlin, 1991.

[Alt02] Alt, W. *Nichtlineare Optimierung: Eine Einführung in Theorie, Verfahren und Anwendungen*. Vieweg, Braunschweig/Wiesbaden, 2002.

[AM98] Arnold, M. and Murua, A. *Non-stiff integrators for differential-algebraic systems of index 2*. Numer. Algorithms, 19 (1-4); 25–41, 1998.

[AM01] Augustin, D. and Maurer, H. *Computational sensitivity analysis for state constrained optimal control problems*. Annals of Operations Research, 101; 75–99, 2001.

[AMF95] Androulakis, I., Maranas, C. and Floudas, C. *αBB: A global optimization method for general constrained nonconvex problems*. Journal of Global Optimization, 7 (4); 337–363, 1995.

[AP91] Ascher, U. M. and Petzold, L. R. *Projected implicit Runge-Kutta methods for differential-algebraic equations*. SIAM Journal on Numerical Analysis, 28 (4); 1097–1120, 1991.

[Arn95] Arnold, M. *A perturbation analysis for the dynamical simulation of mechnical multibody systems*. Applied Numerical Mathematics, 18 (1); 37–56, 1995.

[Arn98] Arnold, M. *Half-explicit Runge-Kutta methods with explicit stages for differential-algebraic systems of index 2*. BIT, 38 (3); 415–438, 1998.

[Bac06] Backes, A. *Extremalbedingungen für Optimierungs-Probleme mit Algebro-Differentialgleichungen.* Ph.D. thesis, Mathematisch-Naturwissenschaftliche Fakultät, Humboldt-Universität Berlin, Berlin, Germany, 2006.

[BCP96] Brenan, K. E., Campbell, S. L. and Petzold, L. R. *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations*, volume 14 of *Classics In Applied Mathematics.* SIAM, Philadelphia, 1996.

[BE88] Brenan, K. E. and Engquist, B. E. *Backward Differentiation Approximations of Nonlinear Differential/Algebraic Systems.* Mathematics of Computations, 51 (184); 659–676, 1988.

[Bet90] Betts, J. T. *Sparse Jacobian Updates in the Collocation Method for Optimal Control Problems.* Journal of Guidance, Control and Dynamics, 13 (3); 409–415, 1990.

[BF97] Billups, S. C. and Ferris, M. C. *QPCOMP: A quadratic programming based solver for mixed complementarity problems.* Mathematical Programming, 76 (3); 533–562, 1997.

[BG00] Büskens, C. and Gerdts, M. *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*, pp. 27–38. München, 2000.

[BG03] Büskens, C. and Gerdts, M. *Emergency Landing of a Hypersonic Flight System: A Corrector Iteration Method for Admissible Real–Time Optimal Control Approximations.* In *Optimalsteuerungsprobleme in der Luft- und Raumfahrt, Workshop in Greifswald des Sonderforschungsbereichs 255: Transatmospärische Flugsysteme*, pp. 51–60. München, 2003.

[BG05] Büskens, C. and Gerdts, M. *Differentiability of Consistency Functions for DAE Systems.* Journal of Optimization Theory and Applications, 125 (1); 37–61, 2005.

[BGK$^+$83] Bank, B., Guddat, J., Klatte, D., Kummer, B. and Tammer, K. *Non-linear parametric optimization.* Birkhäuser, Basel, 1983.

[BGR97] Barcley, A., Gill, P. E. and Rosen, J. B. *SQP methods and their application to numerical optimal control.* Report NA 97-3, Dep. of Mathematics, University of California, San Diego, 1997.

[BH75] Bryson, A. E. and Ho, Y.-C. *Applied Optimal Control.* Hemisphere Publishing Corporation, Washington, 1975.

[BH92] Betts, J. T. and Huffman, W. P. *Application of Sparse Nonlinear Programming to Trajectory Optimization.* Journal of Guidance, Control and Dynamics, 15 (1); 198–206, 1992.

[BH99] Betts, J. T. and Huffman, W. P. *Exploiting Sparsity in the Direct Transcription Method for Optimal Control.* Computational Optimization and Applications, 14 (2); 179–201, 1999.

[BM96] Büskens, C. and Maurer, H. *Sensitivity Analysis and Real-Time Control of Nonlinear Optimal Control Systems via Nonlinear Programming Methods.* Proceedings of the 12th Conference on Calculus of Variations, Optimal Control and Applications. Trassenheide, 1996.

[BM01a] Büskens, C. and Maurer, H. *Sensitivity Analysis and Real-Time Control of Parametric Optimal Control Problems Using Nonlinear Programming Methods*. In *Online Optimization of Large Scale Systems* (M. Grötschel, S. O. Krumke and J. Rambau, editors), pp. 56–68. Springer, 2001.

[BM01b] Büskens, C. and Maurer, H. *Sensitivity Analysis and Real-Time Optimization of Parametric Nonlinear Programming Problems*. In *Online Optimization of Large Scale Systems* (M. Grötschel, S. O. Krumke and J. Rambau, editors), pp. 3–16. Springer, 2001.

[BO75] Blum, E. and Oettli, W. *Mathematische Optimierung*. volume 20 of *Ökonometrie und Unternehmensforschung*. Springer-Verlag Berlin Heidelberg New York, Berlin, 1975.

[Boc87] Bock, H. G. *Randwertproblemmethoden zur Parameteridentifizierung in Systemen nichtlinearer Differentialgleichungen*. volume 183 of *Bonner Mathematische Schriften*. Bonn, 1987.

[BP84] Bock, H. G. and Plitt, K. J. *A Multiple Shooting Algorithm for Direct Solution of Optimal Control Problems*. Proceedings of the 9th IFAC Worldcongress, Budapest, Hungary. 1984.

[BP89] Brenan, K. E. and Petzold, L. R. *The numerical solution of higher index differential/algebraic equations by implicit methods*. SIAM Journal on Numerical Analysis, 26 (4); 976–996, 1989.

[BS79] Bazaraa, M. S. and Shetty, C. M. *Nonlinear Programming: Theory and Algorithms*. John Wiley & Sons, 1979.

[BS00] Bonnans, J. F. and Shapiro, A. *Perturbation Analysis of Optimization Problems*. Springer Series in Operations Research. Springer, New York, 2000.

[BSS93] Bazaraa, M. S., Sherali, H. D. and Shetty, C. M. *Nonlinear Programming: Theory and Algorithms*. John Wiley & Sons, 2nd edition, 1993.

[BTW82] Boggs, P. T., Tolle, J. W. and Wang, P. *On the local convergence of quasi-newton methods for constrained optimization*. SIAM Journal on Control and Optimization, 20 (2); 161–171, 1982.

[Bul71] Bulirsch, R. *Die Mehrzielmethode zur numerischen Lösung von nichtlinearen Randwertproblemen und Aufgaben der optimalen Steuerung*. Report der Carl-Cranz-Gesellschaft, 1971.

[Büs98] Büskens, C. *Optimierungsmethoden und Sensitivitätsanalyse für optimale Steuerprozesse mit Steuer- und Zustandsbeschränkungen*. Ph.D. thesis, Fachbereich Mathematik, Westfälische Wilhems-Universität Münster, 1998.

[Büs01] Büskens, C. *Real-Time Solutions for Perturbed Optimal Control Problems by a Mixed Open- and Closed-Loop Strategy*. In *Online Optimization of Large Scale Systems* (M. Grötschel, S. O. Krumke and J. Rambau, editors), pp. 105–116. Springer, 2001.

[CG95] Campbell, S. L. and Gear, C. W. *The index of general nonlinear DAEs*. Numerische Mathematik, 72; 173–196, 1995.

[CH52] Curtiss, C. F. and Hirschfelder, J. O. *Integration of stiff equations*. Proceedings of the National Academy of Sciences of the United States of America, 38; 235–243, 1952.

[Chu94] Chudej, K. *Optimale Steuerung des Aufstiegs eines zweistufigen Hyperschall-Raumtransporters*. Ph.D. thesis, Mathematisches Institut, Technische Universität München, 1994.

[CL82] Chernousko, F. L. and Lyubushin, A. A. *Method of successive approximations for solution of optimal control problems*. Optimal Control Applications and Methods, 3; 101–114, 1982.

[Cla83] Clarke, F. H. *Optimization and Nonsmooth Analysis*. John Wiley & Sons, New York, 1983.

[CLPS03] Cao, Y., Li, S., Petzold, L. R. and Serban, R. *Adjoint sensitivity analysis for differential-algebraic equations: The adjoint DAE system and its numerical solution*. SIAM Journal on Scientific Computing, 24 (3); 1076–1089, 2003.

[CS85] Caracotsios, M. and Stewart, W. E. *Sensitivity analysis of initial-boundary-value problems with mixed PDEs and algebraic equations*. Computers chem. Engng, 19 (9); 1019–1030, 1985.

[DB02] Deuflhard, P. and Bornemann, F. *Scientific Computing with Ordinary Differential Equations*. volume 42 of *Texts in Applied Mathematics*. Springer-Verlag New York, New York, 2002.

[Deu74] Deuflhard, P. *A modified Newton method for the solution of ill-conditioned systems of nonlinear equations with apllication to multiple shooting*. Numerische Mathematik, 22; 289–315, 1974.

[Deu79] Deuflhard, P. *A Stepsize Control for Continuation Methods and its Special Application to Multiple Shooting Techniques*. Numerische Mathematik, 33; 115–146, 1979.

[DG86a] Duff, I. S. and Gear, C. W. *Computing the structural index*. SIAM Journal on Algebraic Discrete Methods, 7 (4); 594–603, 1986.

[DG86b] Duran, M. A. and Grossmann, I. E. *An outer-approximation algorithm for a class of mixed-integer nonlinear programs*. Mathematical Programming, 36; 307–339, 1986.

[DH91] Deuflhard, P. and Hohmann, A. *Numerische Mathematik*. de Gruyter, Berlin, 1991.

[DHM00] Dontchev, A. L., Hager, W. W. and Malanowski, K. *Error Bounds for Euler Approximation of a State and Control Constrained Optimal Control Problem*. Numerical Functional Analysis and Optimization, 21 (5 & 6); 653–682, 2000.

[DHV00] Dontchev, A. L., Hager, W. W. and Veliov, V. M. *Second-Order Runge-Kutta Approximations in Control Constrained Optimal Control*. SIAM Journal on Numerical Analysis, 38 (1); 202–226, 2000.

[DL99] Devdariani, E. N. and Ledyaev, Y. S. *Maximum Principle for Implicit Control Systems*. Applied Mathematics and Optimization, 40; 79–103, 1999.

[DLO$^+$77] Diekhoff, H.-J., Lory, P., Oberle, H., Pesch, H.-J., Rentrop, P. and Seydel, R. *Comparing routines for the numerical solution of initial value problems of ordinary differential equations in multiple shooting*. Numerische Mathematik, 27; 449–469, 1977.

[DPR76] Deuflhard, P., Pesch, H. J. and Rentrop, P. *A modified continuation method for the numerical solution of nonlinear two-point boundary value problems by shooting techniques*. Numerische Mathematik, 26; 327–343, 1976.

[dPV97] de Pinho, M. and Vinter, R. B. *Necessary Conditions for Optimal Control Problems Involving Nonlinear Differential Algebraic Equations*. Journal of Mathematical Analysis and Applications, 212; 493–516, 1997.

[DW02] Dingguo, P. and Weiwen, T. *Globally convergent inexact generalized Newton's methods for nonsmooth equations*. Journal of Computational and Applied Mathematics, 138; 37–49, 2002.

[Eic93] Eich, E. *Convergence results for a coordinate projection method applied to mechanical systems with algebraic constraints*. SIAM Journal on Numerical Analysis, 30 (5); 1467–1482, 1993.

[EKKvS99] Engl, G., Kröner, A., Kronseder, T. and von Stryk, O. *Numerical Simulation and Optimal Control of Air Separation Plants*. In *High Performance Scientific and Engineering Computing* (H.-J. Bungartz, F. Durst and C. Zenger, editors), volume 8 of *Lecture Notes in Computational Science and Engineering*, pp. 221–231. Springer, 1999.

[Fia83] Fiacco, A. V. *Introduction to Sensitivity and Stability Analysis in Nonlinear Programming*, volume 165 of *Mathematics in Science and Engineering*. Academic Press, New York, 1983.

[Fis97] Fischer, A. *Solution of monotone complementarity problems with locally Lipschitzian functions*. Mathematical Programming, 76 (3); 513–532, 1997.

[FK97] Facchinei, F. and Kanzow, C. *A nonsmooth inexact Newton method for the solution of large-scale nonlinear complementarity problems*. Mathematical Programming, 76 (3); 493–512, 1997.

[FL91] Führer, C. and Leimkuhler, B. J. *Numerical solution of differential-algebraic equations for constraint mechanical motion*. Numerische Mathematik, 59; 55–69, 1991.

[FL94] Fletcher, R. and Leyffer, S. *Solving mixed integer nonlinear programs by outer approximation*. Mathematical Programming, 66; 327–349, 1994.

[FM90] Fiacco, A. V. and McCormick, G. P. *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, volume 4 of *Classics In Applied Mathematics*. SIAM, Philadelphia, 1990.

[FTB97] Feehery, W. F., Tolsma, J. E. and Barton, P. I. *Efficient sensitivity analysis of large-scale differential-algebraic systems*. Applied Numerical Mathematics, 25; 41–54, 1997.

[Füh88] Führer, C. *Differential-algebraische Gleichungssysteme in mechanischen Mehrkörpersystemen: Theorie, numerische Ansätze und Anwendungen*. Ph.D. thesis, Fakultät für Mathematik und Informatik, Technische Universität München, 1988.

[GB01] Gerdts, M. and Büskens, C. *Computation of Consistent Initial Values for Optimal Control Problems with DAE Systems of Higher Index*. ZAMM, 81 S2; 249–250, 2001.

[GB02] Gerdts, M. and Büskens, C. *Consistent Initialization of Sensitivity Matrices for a Class of Parametric DAE Systems*. BIT Numerical Mathematics, 42 (4); 796–813, 2002.

[Gea71] Gear, C. W. *Simultaneous Numerical Solution of Differential-Algebraic Equations*. IEEE Transactions on Circuit Theory, 18 (1); 89–95, 1971.

[Gea88] Gear, C. W. *Differential-algebraic equation index transformations*. SIAM Journal on Scientific and Statistical Computing, 9; 39–47, 1988.

[Gea90] Gear, C. W. *Differential algebraic equations, indices, and integral algebraic equations*. SIAM Journal on Numerical Analysis, 27 (6); 1527–1534, 1990.

[Ger01a] Gerdts, M. *Numerische Methoden optimaler Steuerprozesse mit differential-algebraischen Gleichungssystemen höheren Indexes und ihre Anwendungen in der Kraftfahrzeugsimulation und Mechanik*. volume 61 of *Bayreuther Mathematische Schriften*. Bayreuth, 2001.

[Ger01b] Gerdts, M. *SODAS – Software for Optimal Control Problems with Differential-Algebraic Systems: User's guide*. Technical report, Institut für Mathematik, Universität Bayreuth, 2001.

[Ger03a] Gerdts, M. *Direct Shooting Method for the Numerical Solution of Higher Index DAE Optimal Control Problems*. Journal of Optimization Theory and Applications, 117 (2); 267–294, 2003.

[Ger03b] Gerdts, M. *A Moving Horizon Technique for the Simulation of Automobile Test-Drives*. ZAMM, 83 (3); 147–162, 2003.

[Ger03c] Gerdts, M. *Optimal Control and Real-Time Optimization of Mechanical Multi-Body Systems*. ZAMM, 83 (10); 705–719, 2003.

[Ger04] Gerdts, M. *Parameter Optimization in Mechanical Multibody Systems and Linearized Runge-Kutta Methods*. In *Progress in Industrial Mathematics at ECMI 2002* (A. Buikis, R. Ciegis and A. D. Flitt, editors), volume 5 of *Mathematics in Industry*, pp. 121–126. Springer, 2004.

[Ger05a] Gerdts, M. *Local minimum principle for optimal control problems subject to index one differential-algebraic equations*. Technical report, Department of Mathematics, University of Hamburg, http://www.math.uni-hamburg.de/home/gerdts/Report_index1.pdf, 2005.

[Ger05b] Gerdts, M. *Numerische Lösungsverfahren für Optimalsteuerungsprobleme*. Lecture Note, University of Hamburg, Department of Mathematics, 2005.

[Ger05c] Gerdts, M. *On the Convergence of Linearized Implicit Runge-Kutta Methods and their Use in Parameter Optimization.* Mathematica Balkanica (New Series), 19; 75–83, 2005.

[Ger05d] Gerdts, M. *Parameter Identification in Higher Index DAE Systems.* Technical report, Department of Mathematics, Universität Hamburg, 2005.

[Ger05e] Gerdts, M. *Solving Mixed-Integer Optimal Control Problems by Branch&Bound: A Case Study from Automobile Test-Driving with Gear Shift.* Optimal Control, Applications and Methods, 26 (1); 1–18, 2005.

[Ger06] Gerdts, M. *A variable time transformation method for mixed-integer optimal control problems.* Optimal Control, Applications and Methods, 27 (3); 169–182, 2006.

[GI83] Goldfarb, D. and Idnani, A. *A numerically stable dual method for solving strictly convex quadratic programs.* Mathematical Programming, 27; 1–33, 1983.

[Gir72] Girsanov, I. V. *Lectures on Mathematical Theory of Extremum Problems.* volume 67 of *Lecture Notes in Economics and Mathematical Systems.* Springer, Berlin-Heidelberg-New York, 1972.

[GK97] Grossmann, I. and Kravanja, Z. *Mixed-Integer Nonlinear Programming: A Survey of Algorithms and Applications.* volume 93 of *The IMA Volumes in Mathematics and its Applications*, pp. 73–100. Springer, New York,Berlin,Heidelberg, 1997.

[GK99] Geiger, C. and Kanzow, C. *Numerische Verfahren zur Lösung unrestringierter Optimierungsaufgaben.* Springer, Berlin-Heidelberg-New York, 1999.

[GK02] Geiger, C. and Kanzow, C. *Theorie und Numerik restringierter Optimierungsaufgaben.* Springer, Berlin-Heidelberg-New York, 2002.

[GLG85] Gear, C. W., Leimkuhler, B. and Gupta, G. K. *Automatic integration of Euler-Lagrange equations with constraints.* Journal of Computational and Applied Mathematics, 12/13; 77–90, 1985.

[GM78] Gill, P. E. and Murray, W. *Numerically stable methods for quadratic programming.* Mathematical Programming, 14; 349–372, 1978.

[GMS94] Gill, P. E., Murray, W. and Saunders, M. A. *Large-scale SQP Methods and their Application in Trajectory Optimization*, volume 115 of *International Series of Numerical Mathematics*, pp. 29–42. Birkhäuser, Basel, 1994.

[GMS02] Gill, P. E., Murray, W. and Saunders, M. A. *SNOPT: An SQP algorithm for large-scale constrained optimization.* SIAM Journal on Optimization, 12 (4); 979–1006, 2002.

[GMSW91] Gill, P. E., Murray, W., Saunders, M. A. and Wright, M. H. *Inertia-controlling methods for general quadratic programming.* SIAM Review, 33 (1); 1–36, 1991.

[GMSW98] Gill, P. E., Murray, W., Saunders, M. A. and Wright, M. H. *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.

[GMW81] Gill, P. E., Murray, W. and Wright, M. H. *Practical Optimization*. Academic Press, London, 1981.

[GP84] Gear, C. W. and Petzold, L. *ODE methods for the solution of differential/algebraic systems*. SIAM Journal on Numerical Analysis, 21 (4); 716–728, 1984.

[GR80] Göpfert, A. and Riedrich, T. *Funktionalanalysis*. Teubner, Leipzig, 1980.

[Gri00] Griewank, A. *Evaluating derivatives. Principles and techniques of algorithmic differentiation*, volume 19 of *Frontiers in Applied Mathematics*. SIAM, Philadelphia, 2000.

[Gri03] Griewank, A. *A mathematical view of automatic differentiation*. Acta Numerica, 12; 321–398, 2003.

[Gru96] Grupp, F. *Parameteridentifizierung nichtlinearer mechanischer Deskriptorsysteme mit Anwendungen in der Rad–Schiene–Dynamik*. Fortschritt-Berichte VDI Reihe 8, Nr. 550. VDI–Verlag, Düsseldorf, 1996.

[GS01] Ghildyal, V. and Sahinidis, N. V. *Solving global optimization problems with Baron*. In *From local to global optimization. Papers from the conference dedicated to Professor Hoang Tuy on the occasion of his 70th birthday, Rimforsa, Sweden, August 1997* (A. M. et al., editor), volume 53 of *Nonconvex Optimization and Applications*, pp. 205–230. Kluwer Academic Publishers, Dordrecht, 2001.

[Gün95] Günther, M. *Ladungsorientierte Rosenbrock-Wanner-Methoden zur numerischen Simulation digitaler Schaltungen*, volume 168 of *VDI Fortschrittberichte Reihe 20: Rechnergestützte Verfahren*. VDI-Verlag, 1995.

[Hag00] Hager, W. W. *Runge-Kutta methods in optimal control and the transformed adjoint system*. Numerische Mathematik, 87 (2); 247–282, 2000.

[Han77] Han, S. P. *A Globally Convergent Method for Nonlinear Programming*. Journal of Optimization Theory and Applications, 22 (3); 297–309, 1977.

[HCB93] Hiltmann, P., Chudej, K. and Breitner, M. H. *Eine modifizierte Mehrzielmethode zur Lösung von Mehrpunkt-Randwertproblemen*. Technical Report 14, Sonderforschungsbereich 255 der Deutschen Forschungsgemeinschaft: Transatmosphärische Flugsysteme, Lehrstuhl für Höhere und Numerische Mathematik, Technische Universität München, 1993.

[Hei92] Heim, A. *Parameteridentifizierung in differential-algebraischen Gleichungssystemen*. Master's thesis, Mathematisches Institut, Technische Universität München, 1992.

[Hes66] Hestenes, M. R. *Calculus of variations and optimal control theory*. John Wiley & Sons, New York, 1966.

[Hin97] Hinsberger, H. *Ein direktes Mehrzielverfahren zur Lösung von Optimalsteuerungsproblemen mit großen, differential-algebraischen Gleichungssystemen und Anwendungen aus der Verfahrenstechnik*. Ph.D. thesis, Institut für Mathematik, Technische Universität Clausthal, 1997.

[HL69] Hermes, H. and Lasalle, J. P. *Functional Analysis and Time Optimal Control*, volume 56 of *Mathematics in Science and Engineering*. Academic Press, New York, 1969.

[HLR89] Hairer, E., Lubich, C. and Roche, M. *The numerical solution of differential-algebraic systems by Runge-Kutta methods*. volume 1409 of *Lecture Notes in Mathematics*. Springer, Berlin-Heidelberg-New York, 1989.

[HP87] Hargraves, C. R. and Paris, S. W. *Direct trajectory optimization using nonlinear programming and collocation*. Journal of Guidance, Control and Dynamics, 10; 338–342, 1987.

[HPR92] Han, S. P., Pang, J. S. and Rangaraj, N. *Globally convergent Newton methods for nonsmooth equations*. Mathematics of Operations Research, 17 (3); 586–607, 1992.

[HSV95] Hartl, R. F., Sethi, S. P. and Vickson, G. *A Survey of the Maximum Principles for Optimal Control Problems with State Constraints*. SIAM Review, 37 (2); 181–218, 1995.

[HW96] Hairer, E. and Wanner, G. *Solving ordinary differential equations II: Stiff and differential-algebraic problems*, volume 14. Springer Series in Computational Mathematics, Berlin-Heidelberg-New York, 2nd edition, 1996.

[IT79] Ioffe, A. D. and Tihomirov, V. M. *Theory of extremal problems*. volume 6 of *Studies in Mathematics and its Applications*. North-Holland Publishing Company, Amsterdam, New York, Oxford, 1979.

[Jay93] Jay, L. *Collocation methods for differential-algebraic equations of index 3*. Numerische Mathematik, 65; 407–421, 1993.

[Jay95] Jay, L. *Convergence of Runge-Kutta methods for differential-algebraic systems of index 3*. Applied Numerical Mathematics, 17; 97–118, 1995.

[Jia99] Jiang, H. *Global convergence analysis of the generalized Newton and Gauss-Newton methods of the Fischer-Burmeister equation for the complementarity problem*. Mathematics of Operations Research, 24 (3); 529–543, 1999.

[JLS71] Jacobson, D. H., Lele, M. M. and Speyer, J. L. *New Necessary Conditions of Optimality for Constrained Problems with State-Variable Inequality Constraints*. Journal of Mathematical Analysis and Applications, 35; 255–284, 1971.

[JQ97] Jiang, H. and Qi, L. *A new nonsmooth equations approach to nonlinear complementarity problems*. SIAM Journal on Control and Optimization, 35 (1); 178–193, 1997.

[KE85] Krämer-Eis, P. *Ein Mehrzielverfahren zur numerischen Berechnung optimaler Feedback-Steuerungen bei beschränkten nichtlinearen Steuerungsproblemen*. volume 166 of *Bonner Mathematische Schriften*. Bonn, 1985.

[Kie98] Kiehl, M. *Sensitivity Analysis of ODEs and DAEs - Theory and Implementation Guide*. Technische Universität München, TUM-M5004, 1998.

[Kla90] Klatte, D. *Nonlinear optimization problems under data perturbations*. volume 378 of *Lecture Notes in Economics and Mathematical Systems*, pp. 204–235. Springer, Berlin-Heidelberg-New York, 1990.

[KM97] Kunkel, P. and Mehrmann, V. *The linear quadratic optimal control problem for linear descriptor systems with variable coefficients*. Mathematics of Control, Signals, and Systems, 10 (3); 247–264, 1997.

[KM04] Kurina, G. A. and März, R. *On linear-quadratic optimal control problems for time-varying descriptor systems*. SIAM Journal on Control and Optimization, 42 (6); 2062–2077, 2004.

[Kno75] Knobloch, H. W. *Das Pontryaginsche Maximumprinzip für Probleme mit Zustandsbeschränkungen I und II*. Zeitschrift für Angewandte Mathematik und Mechanik, 55; 545–556, 621–634, 1975.

[KOS77] Kufner, A., Oldrich, J. and Svatopluk, F. *Function Spaces*. Noordhoff International Publishing, Leyden, 1977.

[Kow69] Kowalsky, H.-J. *Lineare Algebra*. Walter de Gruyter & Co, Berlin, 4th edition, 1969.

[Kra88] Kraft, D. *A Software Package for Sequential Quadratic Programming*. DFVLR-FB-88-28, Oberpfaffenhofen, 1988.

[Kre82] Kreindler, E. *Additional Necessary Conditions for Optimal Control with State-Variable Inequality Constraints*. Journal of Optimization Theory and Applications, 38 (2); 241–250, 1982.

[Kur76] Kurcyusz, S. *On the Existence and Nonexistence of Lagrange Multipliers in Banach Spaces*. Journal of Optimization Theory and Applications, 20 (1); 81–110, 1976.

[KWW78] Kirsch, A., Warth, W. and Werner, J. *Notwendige Optimalitätsbedingungen und ihre Anwendung*. volume 152 of *Lecture Notes in Economics and Mathematical Systems*. Springer, Berlin-Heidelberg-New York, 1978.

[Lei95] Leineweber, D. B. *Analyse und Restrukturierung eines Verfahrens zur direkten Lösung von Optimal-Steuerungsproblemen*. Master's thesis, Interdisziplinäres Zentrum für Wissenschaftliches Rechnen, Universität Heidelberg, 1995.

[Lem62] Lemke, C. E. *A method of solution for quadratic programs*. Management Science, 8; 442–453, 1962.

[Lem71a] Lempio, F. *Lineare Optimierung in unendlichdimensionalen Vektorräumen*. Computing, 8; 284–290, 1971.

[Lem71b] Lempio, F. *Separation und Optimierung in linearen Räumen*. Ph.D. thesis, Universität Hamburg, Hamburg, 1971.

[Lem72] Lempio, F. *Tangentialmannigfaltigkeiten und infinite Optimierung*. Habilitationsschrift, Universität Hamburg, Hamburg, 1972.

[Ley01] Leyffer, S. *Integrating SQP and Branch-and-Bound for Mixed Integer Nonlinear Programming*. Computational Optimization and Applications, 18; 295–309, 2001.

[LM80] Lempio, F. and Maurer, H. *Differential stability in infinite-dimensional nonlinear programming*. Applied Mathematics and Optimization, 6; 139–152, 1980.

[LP86] Lötstedt, P. and Petzold, L. R. *Numerical Solution of Nonlinear Differential Equations with Algebraic Constraints I: Convergence Results for Backward Differentiation Formulas*. Mathematics of Computation, 46; 491–516, 1986.

[LPG91] Leimkuhler, B., Petzold, L. R. and Gear, C. W. *Approximation methods for the consistent initialization of differential-algebraic equations*. SIAM Journal on Numerical Analysis, 28 (1); 205–226, 1991.

[LS76] Ljusternik, L. A. and Sobolew, W. I. *Elemente der Funktionalanalysis*. Verlag Harri Deutsch, Zürich-Frankfurt/Main-Thun, 1976.

[LTC98] Lee, H. W. J., Teo, K. L. and Cai, X. Q. *An Optimal Control Approach to Nonlinear Mixed Integer Programming Problems*. Computers & Mathematics with Applications, 36 (3); 87–105, 1998.

[LTRJ97] Lee, H. W. J., Teo, K. L., Rehbock, V. and Jennings, L. S. *Control parameterization enhancing technique for time optimal control problems*. Dynamic Systems and Applications, 6 (2); 243–262, 1997.

[LTRJ99] Lee, H. W. J., Teo, K. L., Rehbock, V. and Jennings, L. S. *Control parametrization enhancing technique for optimal discrete-valued control problems*. Automatica, 35 (8); 1401–1407, 1999.

[Lue69] Luenberger, D. G. *Optimization by Vector Space Methods*. John Wiley & Sons, New York-London-Sydney-Toronto, 1969.

[LVBB$^+$04] Laurent-Varin, J., Bonnans, F., Berend, N., Talbot, C. and Haddou, M. *On the refinement of discretization for optimal control problems*. IFAC Symposium on Automatic Control in Aerospace, St. Petersburg, 2004.

[MA01] Maurer, H. and Augustin, D. *Sensitivity Analysis and Real-Time Control of Parametric Optimal Control Problems Using Boundary Value Methods*. In *Online Optimization of Large Scale Systems* (M. Grötschel, S. O. Krumke and J. Rambau, editors), pp. 17–55. Springer, 2001.

[Mac88] Machielsen, K. C. P. *Numerical Solution of Optimal Control Problems with State Constraints by Sequential Quadratic Programming in Function Space*. volume 53 of *CWI Tract*. Centrum voor Wiskunde en Informatica, Amsterdam, 1988.

[Mal97] Malanowski, K. *Sufficient Optimality Conditions for Optimal Control subject to State Constraints*. SIAM Journal on Control and Optimization, 35 (1); 205–227, 1997.

[Man94] Mangasarian, O. L. *Nonlinear Programming*, volume 10 of *Classics In Applied Mathematics*. SIAM, Philadelphia, 1994.

[Mär95] März, R. *On linear differential-algebraic equations and linerizations*. Applied Numerical Mathematics, 18 (1); 267–292, 1995.

[Mär98a] März, R. *Criteria of the trivial solution of differential algebraic equations with small nonlinearities to be asymptotically stable*. Journal of Mathematical Analysis and Applications, 225 (2); 587–607, 1998.

[Mär98b] März, R. *EXTRA-ordinary differential equations: Attempts to an analysis of differential-algebraic systems*. In *European congress of mathematics* (A. Balog, editor), volume 1 of *Prog. Math. 168*, pp. 313–334. Birkhäuser, Basel, 1998.

[Mau77] Maurer, H. *On Optimal Control Problems with Boundary State Variables and Control Appearing Linearly*. SIAM Journal on Control and Optimization, 15 (3); 345–362, 1977.

[Mau79] Maurer, H. *On the Minimum Principle for Optimal Control Problems with State Constraints*. Schriftenreihe des Rechenzentrums der Universität Münster, 41, 1979.

[Mau81] Maurer, H. *First and Second Order Sufficient Optimality Conditions in Mathematical Programming and Optimal Control*. Mathematical Programming Study, 14; 163–177, 1981.

[May91] Mayr, R. *Verfahren zur Bahnfolgeregelung für ein automatisch geführtes Fahrzeug*. Ph.D. thesis, Fakultät für Elektrotechnik, Universität Dortmund, 1991.

[MBM97] Malanowski, K., Büskens, C. and Maurer, H. *Convergence of Approximations to Nonlinear Optimal Control Problems*. In *Mathematical programming with data perturbations* (A. Fiacco, editor), volume 195, pp. 253–284. Dekker. Lecture Notes in Pure and Applied Mathematics, 1997.

[Meh91] Mehrmann, V. *The autonomous linear quadratic control problem. Theory and numerical solution*. volume 163 of *Lecture Notes in Control and Information Sciences*. Springer, Berlin, 1991.

[MF05] Meyer, C. A. and Floudas, C. A. *Convex underestimation of twice continuously differentiable functions by piecewise quadratic perturbation: Spline α BB underestimators*. Journal of Global Optimization, 32 (2); 221–258, 2005.

[Mit90] Mitschke, M. *Dynamik der Kraftfahrzeuge, Band C: Fahrverhalten*. Springer, Berlin-Heidelberg-New York, 2nd edition, 1990.

[MM96] Malanowski, K. and Maurer, H. *Sensitivity analysis for parametric control problems with control-state constraints*. Computational Optimization and Applications, 5 (3); 253–283, 1996.

[MM98] Malanowski, K. and Maurer, H. *Sensitivity analysis for state constrained optimal control problems*. Discrete and Continuous Dynamical Systems, 4 (2); 3–14, 1998.

[MM01] Malanowski, K. and Maurer, H. *Sensitivity analysis for optimal control problems subject to higher order state constraints*. Annals of Operations Research, 101; 43–73, 2001.

[MMP04] Malanowski, K., Maurer, H. and Pickenhain, S. *Second-order Sufficient Conditions for State-constrained Optimal Control Problems*. Journal of Optimization Theory and Applications, 123 (3); 595–617, 2004.

[MO02] Maurer, H. and Oberle, H. J. *Second order sufficient conditions for optimal control problems with free final time: The Riccati approach*. SIAM Journal on Control Optimization, 41 (2); 380–403, 2002.

[Mod94] Moder, T. *Optimale Steuerung eines KFZ im fahrdynamischen Grenzbereich*. Master's thesis, Mathematisches Institut, Technische Universität München, 1994.

[Mor88] Mordukhovich, B. S. *An approximate maximum principle for finite-difference control systems*. U.S.S.R. Computational Mathematics and Mathematical Physics, 28 (1); 106–114, 1988.

[MP94a] Maurer, H. and Pesch, H. J. *Solution differentiability for nonlinear parametric control problems*. SIAM Journal on Control and Optimization, 32 (6); 1542–1554, 1994.

[MP94b] Maurer, H. and Pesch, H. J. *Solution differentiability for parametric nonlinear control problems with control-state constraints*. Control and Cybernetics, 23 (1-2); 201–227, 1994.

[MP95a] Maurer, H. and Pesch, H. J. *Solution differentiability for parametric nonlinear control problems with control-state constraints*. Journal of Optimization Theory and Applications, 86 (2); 285–309, 1995.

[MP95b] Maurer, H. and Pickenhain, S. *Second-order Sufficient Conditions for Control Problems with Mixed Control-state Constraints*. Journal of Optimization Theory and Applications, 86 (3); 649–667, 1995.

[MP96] Maly, T. and Petzold, L. R. *Numerical Methods and Software for Sensitivity Analysis of Differential-Algebraic Systems*. Applied Numerical Mathematics, 20 (1); 57–79, 1996.

[MS96] Mayrhofer, M. and Sachs, G. *Notflugbahnen eines zweistufigen Hyperschall-Flugsystems ausgehend vom Trennmanöver*. Seminarbericht des Sonderforschungsbereichs 255: Transatmosphärische Flugsysteme, TU München, pp. 109–118, 1996.

[MT97] März, R. and Tischendorf, C. *Recent results in solving index-2 differential-algebraic equations in circuit simulation*. SIAM Journal on Scientific Computing, 18 (1); 139–159, 1997.

[Mül03] Müller, P. C. *Optimal control of proper and nonproper descriptor systems*. Archive of Applied Mechanics, 72; 875–884, 2003.

[MZ79] Maurer, H. and Zowe, J. *First and Second-Order Necessary and Sufficient Optimality Conditions for Infinite-Dimensional Programming Problems*. Mathematical Programming, 16; 98–110, 1979.

[Nat75] Natanson, I. P. *Theorie der Funktionen einer reellen Veränderlichen*. Verlag Harri Deutsch, Zürich-Frankfurt-Thun, 1975.

[Nec92] Neculau, M. *Modellierung des Fahrverhaltens: Informationsaufnahme, Regel- und Steuerstrategien in Experiment und Simulation*. Ph.D. thesis, Fachbereich 12: Verkehrswesen, Technische Universität Berlin, 1992.

[Neu76] Neustadt, L. W. *Optimization: A Theory of Necessary Conditions*. Princeton, New Jersey, 1976.

[Obe86] Oberle, H. *Numerical solution of minimax optimal control problems by multiple shooting technique*. Journal of Optimization Theory and Applications, 50; 331–357, 1986.

[OG01] Oberle, H. J. and Grimm, W. *BNDSCO – A Program for the Numerical Solution of Optimal Control Problems*. Technical Report Reihe B, Bericht 36, Hamburger Beiträge zur Angewandten Mathematik, Department of Mathematics, University of Hamburg, http://www.math.uni-hamburg.de/home/oberle/software.html, 2001.

[PB93] Pacejka, H. and Bakker, E. *The Magic Formula Tyre Model*. Vehicle System Dynamics, 21 supplement; 1–18, 1993.

[PB94] Pesch, H. J. and Bulirsch, R. *The Maximum Principle, Bellman's Equation and Caratheodorys Work*. Journal of Optimization Theory and Applications, 80 (2); 199–225, 1994.

[PBGM64] Pontryagin, L. S., Boltyanskij, V., Gamkrelidze, R. and Mishchenko, E. *Mathematische Theorie optimaler Prozesse*. Oldenbourg, München, 1964.

[Pes78] Pesch, H. J. *Numerische Berechnung optimaler Flugbahnkorrekturen in Echtzeit-Rechnung*. Ph.D. thesis, Institut für Mathematik, Technische Universität München, 1978.

[Pes79] Pesch, H. J. *Numerical computation of neighboring optimum feedback control schemes in real-time*. Applied Mathematics and Optimization, 5; 231–252, 1979.

[Pes89a] Pesch, H. J. *Real-time computation of feedback controls for constrained optimal control problems. I: Neighbouring extremals*. Optimal Control Applications and Methods, 10 (2); 129–145, 1989.

[Pes89b] Pesch, H. J. *Real-time computation of feedback controls for constrained optimal control problems. II: A correction method based on multiple shooting*. Optimal Control Applications and Methods, 10 (2); 147–171, 1989.

[Pes02] Pesch, H. J. *Schlüsseltechnologie Mathematik: Einblicke in aktuelle Anwendungen der Mathematik*. B. G. Teubner, Stuttgart – Leipzig – Wiesbaden, 2002.

[Pet82a] Petzold, L. R. *A description of DASSL: a differential/algebraic system solver*. Rep. Sand 82-8637, Sandia National Laboratory, Livermore, 1982.

[Pet82b] Petzold, L. R. *Differential/algebraic equations are not ODE's*. SIAM Journal on Scientific and Statistical Computing, 3 (3); 367–384, 1982.

[Pet89] Petzold, L. R. *Recent developments in the numerical solution of differential/algebraic systems*. Computer Methods in Applied Mechanics and Engineering, 75; 77–89, 1989.

[Pow78] Powell, M. J. D. *A fast algorithm for nonlinearily constrained optimization calculation*. In *Numerical Analysis* (G. Watson, editor), volume 630 of *Lecture Notes in Mathematics*. Springer, Berlin-Heidelberg-New York, 1978.

[Pyt98] Pytlak, R. *Runge-Kutta Based Procedure for the Optimal Control of Differential-Algebraic Equations.* Journal of Optimization Theory and Applications, 97 (3); 675–705, 1998.

[Qi93] Qi, L. *Convergence analysis of some algorithms for solving nonsmooth equations.* Mathematics of Operations Research, 18 (1); 227–244, 1993.

[QS93] Qi, L. and Sun, J. *A nonsmooth version of Newton's method.* Mathematical Programming, 58 (3); 353–367, 1993.

[Ral94] Ralph, D. *Global convergence of damped Newton's method for nonsmooth equations via the path search.* Mathematics of Operations Research, 19 (2); 352–389, 1994.

[Ris91] Risse, H.-J. *Das Fahrverhalten bei normaler Fahrzeugführung*, volume 160 of *VDI Fortschrittberichte Reihe 12: Verkehrstechnik/Fahrzeugtechnik.* VDI-Verlag, 1991.

[Rob76] Robinson, S. M. *Stability Theory for Systems of Inequalities, Part II: Differentiable Nonlinear Systems.* SIAM Journal on Numerical Analysis, 13 (4); 487–513, 1976.

[Roc70] Rockafellar, R. T. *Convex Analysis.* Princeton University Press, New Jersey, 1970.

[RS96] Ryoo, H. S. and Sahinidis, N. V. *A branch-and-reduce approach to global optimization.* Journal of Global Optimization, 8 (2); 107–138, 1996.

[RV02] Roubicek, T. and Valásek, M. *Optimal control of causal differential-algebraic systems.* Journal of Mathematical Analysis and Applications, 269 (2); 616–641, 2002.

[Sah96] Sahinidis, N. V. *BARON: A general purpose global optimization software package.* Journal of Global Optimization, 8 (2); 201–205, 1996.

[SB90] Stoer, J. and Bulirsch, R. *Numerische Mathematik II.* Springer, Berlin-Heidelberg-New York, 3rd edition, 1990.

[SBS98] Schulz, V. H., Bock, H. G. and Steinbach, M. C. *Exploiting invariants in the numerical solution of multipoint boundary value problems for DAE.* SIAM Journal on Scientific Computing, 19 (2); 440–467, 1998.

[Sch81] Schittkowski, K. *The Nonlinear Programming Method of Wilson, Han, and Powell with an Augmented Lagrangean 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.

[Sch83] Schittkowski, K. *On the Convergence of a Sequential Quadratic Programming Method with an Augmented Lagrangean Line Search Function.* Mathematische Operationsforschung und Statistik, Series Optimization, 14 (2); 197–216, 1983.

[Sch85] Schittkowski, K. *NLPQL: A Fortran subroutine for solving constrained nonlinear programming problems.* Annals of Operations Research, 5; 484–500, 1985.

[Sch94] Schittkowski, K. *Parameter estimation in systems of nonlinear equations.* Numerische Mathematik, 68; 129–142, 1994.

[Sch96] Schulz, V. H. *Reduced SQP Methods for Large-Scale Optimal Control Problems in DAE with Application to Path Planning Problems for Satellite Mounted Robots.* Ph.D. thesis, Interdisziplinäres Zentrum für Wissenschaftliches Rechnen, Universität Heidelberg, 1996.

[SGFR94] Simeon, B., Grupp, F., Führer, C. and Rentrop, P. *A nonlinear truck model and its treatment as a multibody system.* Journal of Computational and Applied Mathematics, 50; 523–532, 1994.

[Sib04] Siburian, A. *Numerical Methods for Robust, Singular and Discrete Valued Optimal Control Problems.* Ph.D. thesis, Curtin University of Technology, Perth, Australia, 2004.

[Sim94] Simeon, B. *Numerische Integration mechanischer Mehrkörpersysteme: Projizierende Deskriptorformen, Algorithmen und Rechenprogramme*, volume 130 of *VDI Fortschrittberichte Reihe 20: Rechnergestützte Verfahren.* VDI-Verlag, 1994.

[Spe93] Spellucci, P. *Numerische Verfahren der nichtlinearen Optimierung.* Birkhäuser, Basel, 1993.

[SR04] Siburian, A. and Rehbock, V. *Numerical procedure for solving a class of singular optimal control problems.* Optimization Methods and Software, 19 (3–4); 413–426, 2004.

[Ste73] Stetter, H. J. *Analysis of Discretization Methods for Ordinary Differential Equations.* volume 23 of *Springer Tracts in Natural Philosophy.* Springer-Verlag Berlin Heidelberg New York, 1973.

[Ste95] Steinbach, M. C. *Fast Recursive SQP Methods for Large-Scale Optimal Control Problems.* Ph.D. thesis, Interdisziplinäres Zentrum für Wissenschaftliches Rechnen, Universität Heidelberg, 1995.

[Sto85] Stoer, J. *Principles of sequential quadratic programming methods for solving nonlinear programs.* In *Computational Mathematical Programming* (K. Schittkowski, editor), volume F15 of *NATO ASI Series*, pp. 165–207. Springer, Berlin-Heidelberg-New York, 1985.

[SW95] Strehmel, K. and Weiner, R. *Numerik gewöhnlicher Differentialgleichungen.* Teubner, Stuttgart, 1995.

[TG87] Teo, K. L. and Goh, C. J. *MISER: An optimal control software.* Applied Research Corporation, National University of Singapore, Kent Ridge, Singapore, 1987.

[TJL99] Teo, K. L., Jennings, L. S. and Lee, H. W. J. Rehbock, V. *The control parameterization enhancing transform for constrained optimal control problems.* Journal of the Australian Mathematics Society, 40 (3); 314–335, 1999.

[Trö05] Tröltzsch, F. *Optimale Steuerung partieller Differentialgleichungen.* Vieweg, Wiesbaden, 2005.

[vH80] von Heydenaber, T. *Simulation der Fahrdynamik von Kraftfahrzeugen.* Master's thesis, Institut für Mathematik, Technische Universität München, 1980.

[vS94] von Stryk, O. *Numerische Lösung optimaler Steuerungsprobleme: Diskretisierung, Parameteroptimierung und Berechnung der adjungierten Variablen*, volume 441 of *VDI Fortschrittberichte Reihe 8: Meß-, Steuerungs- und Regeleungstechnik*. VDI-Verlag, 1994.

[vS99] von Schwerin, R. *Multibody System Simulation: Numerical Methods, Algorithms, and Software*. volume 7 of *Lecture Notes in Computational Science and Engineering*. Springer, Berlin-Heidelberg-New York, 1999.

[Wer95] Werner, D. *Funktionalanalysis*. Springer, Berlin-Heidelberg-New York, 1995.

[Wid46] Widder, D. V. *The Laplace Transform*. Princeton University Press, Princeton, 1946.

[XC97] Xu, H. and Chang, X. W. *Approximate Newton methods for nonsmooth equations*. Journal of Optimization Theory and Applications, 93 (2); 373–394, 1997.

[XG97] Xu, H. and Glover, B. M. *New version of the Newton method for nonsmooth equations*. Journal of Optimization Theory and Applications, 93 (2); 395–415, 1997.

[YF97] Yamashita, N. and Fukushima, M. *Modified Newton methods for solving a semismooth reformulation of monotone complementarity problems*. Mathematical Programming, 76 (3); 469–491, 1997.

[Zei94] Zeidan, V. *The Riccati Equation for Optimal Control Problems with Mixed State-Control Constraints: Necessity and Sufficiency*. SIAM Journal on Control and Optimization, 32 (5); 1297–1321, 1994.

[ZK79] Zowe, J. and Kurcyusz, S. *Regularity and Stability of the Mathematical Programming Problem in Banach Spaces*. Applied Mathematics and Optimization, 5; 49–62, 1979.

[Zom91] Zomotor, A. *Fahrwerktechnik: Fahrverhalten*. Vogel Buchverlag, Stuttgart, 1991.