ROBERT BAIER AND MATTHIAS GERDTS

Intensive Course on

Set-Valued Numerical Analysis and Optimal Control

Borovets and Sofia, Bulgaria, 5.9.-16.9.2005

ADDRESSES OF THE AUTHORS:

Dr. Robert Baier
Applied Mathematics V, Departments of Mathematics
University of Bayreuth, D-95440 Bayreuth, Germany

E-Mail: robert.baier@uni-bayreuth.de
WWW: www.uni-bayreuth.de/departments/math/~rbaier/

Prof. Dr. Matthias Gerdts
Fachbereich Mathematik, Schwerpunkt Optimierung und Approximation
Universität Hamburg, D-20146 Hamburg, Germany

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

The authors express their thanks to the participants of the intensive course in Borovets/Sofia, 04.09.–15.09.2005 for their questions, suggestions and error corrections, especially the helpful remarks of Eugene Belogay, Marko Petkovic and Anca Dumitru. Please check the website of Matthias Gerdts for updates. Your comments and suggestions are appreciated.

Preliminary Version: 29th September 2005

Contents

1 Introduction

History

- Variational problems
  - ancient world (Dido's problem)
  - in 1696: *Brachistochrone Problem* (greek: brachistos=shortest, chronos=time) by *Johann Bernoulli (1667-1748)*;
    solved by himself and Sir Isaac Newton (1643–1727), Gottfried Wilhelm Leibniz (1646–1716), Jacob Bernoulli (1654–1705), Guillaume Francois Antoine Marquis de L'Hospital (1661–1704), Ehrenfried Walter von Tschirnhaus (1651–1708)
- Optimal control problems (generalization of variational problems)
  - since approximately 1950
  - initially mainly motivated by military applications
  - $\sim$ **1964**: *Maximum principle* by *Lev S. Pontryagin (1908–1988)* and his students V. G. Boltyanskii, R. V. Gamrelidze and E. F. Mishchenko and independently by *Magnus R. Hestenes*.
  - since then: lots of *theory* (involves different mathematical disciplines like functional analysis, differential equations, optimization, measure theory, numerics) and *applications* in natural sciences, engineering sciences, economical sciences (e.g. simulation of test-drives, robot control, flight path planning, control of plants,...)

History

- Convex Analysis
  milestones:
  - Brunn-Minkowski theory of convex bodies (H. Brunn's thesis 1887, H. Minkowski 1864–1909, important works in 1897, 1903, 1910)
    topics: addition of sets (Minkowski sum); volume of convex bodies; measures of volumes, surfaces, curvature; . . .)
  - books on theory of sets of F. Hausdorff (1914, 1927, English translation)
  - survey of results in book of T. Bonnesen/W. Fenchel (1934), English translation
  - book on convex polyhedra of A. D. Aleksandrov (1938) books on convex polytopes of B. Grünbaum (1967), V. Klee (1963), E. Steinitz/H. Rademacher (1934), H. Weyl (1935/36), . . .
  - book on convex analysis of R. T. Rockafellar (1970)
    focus on convex functions and differential properties, extremal problems like optimization, minimax theory

History

- Set-Valued Analysis
  milestones:
  - limits of sets (P. Painlevé 1906, K. Kuratowski 1958)
  - works on (contingent/tangent) cones (G. Bouligand 1930, generalized gradients by F. C. Clarke 1975) later used for continuity definitions and derivatives of set-valued maps (G. Bouligand, K. Kuratowski in 1932)
  - works on fixed-point theorems of multivalued maps (S. Kakutani 1941, Ky Fan 1969, variation principle of I. Ekeland 1974)
  - book on discontinuous differential equations by A. F. Filippov (1960)
    discontinuity of right-hand side of differential equations w.r.t. state variable, reformulation as differential inclusion solves problem with definition of such solutions
  - works on set-valued integral (since 1965: R. J. Aumann, T. F. Bridgland Jr., G. Debreu, C. Olech, . . .)
  - works on ellipsoidal methods by Schweppe (1968), Chernousko (1980, 1988) and Kurzhanski et al. (1977, 1997)

History

- Set-Valued Analysis (milestones continued):
  - book on differential inclusions by J. P. Aubin/A. Cellina (1984), V. I. Blagodatskikh and A. F. Filippov (1986)
    topics: differential systems with uncertainty, absence of controls, variety of solutions
    earlier works starting from 1966: V. I. Blagodatskih (1973), A. Bressan, C. Castaing (1966), A. Cellina, A. F. Filippov (1967)
    older notions: contingent equations, differential relations, generalized differential equations, multivalued differential equations, ...
  - works on selection theorems of multivalued maps (A. Cellina 1976, C. Castaing/M. Valadier 1976, E. Michael 1956)
  - book on viability theory by J. P. Aubin (1991)
    solutions of differential equations staying in a tube (bounded area)
    J. P. Aubin (1984), V. Křivan (1990), A. B. Kurzhanski/T. F. Filippova (1986), L. Rybiński (1986), N. S. Papageorgiou (1988)
  - book on set-valued analysis by J. P. Aubin/H. Frankowska (1990)
    set-valued maps (svm), selection theorems, measurability of svms, tangent cones, ...

Components of Optimal Control Problems

- time dependent *process* (e.g. population size, chemical process, mechanical system, sales volume of a company)
- $x(t)$: *state* of the process at time t
- $u(t)$: *control* that allows to influence the dynamical behavior of the process (e.g. incorporation of predators to reduce the population size, temperature or feed in a chemical reaction, steering wheel or brakes of a car, change of prices)
- dynamical behavior described by an *ordinary differential equation (ODE)*: $x'(t) = f(t, x(t), u(t))$
- *constraints* due to, e.g., security reasons (altitude of an airplane should be greator or equal zero) or technical limitations
- *objective functional* to be minimized (or maximized)

Optimal Control Problem

Problem 1.1 (Optimal Control Problem). *Minimize*

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

subject to the differential equation

$$x'(t) = f(t, x(t), u(t)), \qquad t_0 \le t \le t_f,$$

the mixed control-state constraints

$$c(t, x(t), u(t)) \le 0, \qquad t_0 \le t \le t_f,$$

the pure state constraints

$$s(t, x(t)) \le 0, \qquad t_0 \le t \le t_f,$$

Optimal Control Problem

Problem 1.1 (continued). *the boundary conditions*

$$\psi(x(t_0), x(t_f)) = 0,$$

and the set constraints

$$u(t) \in \mathcal{U} \subseteq \mathbb{R}^{n_u}, \qquad t_0 \leq t \leq t_f.$$

Questions

- Reachable set:
  - What trajectories are admissible ?
  - How to approximate the set of all admissible trajectories ?
- Optimality:
  - What trajectory is optimal for a given objective functional ?
  - Necessary optimality conditions (minimum principle) ?
  - (Numerical) solution approaches for optimal control problems ?

Set-Valued Maps

A set-valued map is a generalization of an 'ordinary function' $f : X \to Y$:

Definition 1.2. A *set-valued map/mapping* (*multivalued map*) is a map $F : X \Rightarrow Y$, where the images $F(x)$ are subsets of Y for every $x \in X$. often in the following: $X = [t_0, t_f]$, $Y = \mathbb{R}^n$

Special case:
$f : X \to Y$ is the special set-valued map $F : X \Rightarrow Y$ with $F(x) := \{f(x)\}$.

Set-Valued Maps

Questions

- suitable continuity definitions?
- existence of selections of a svm, properties of selections?
- parametrization of set-valued maps (svm)?
- derivative of set-valued maps (svm)?
- integral of set-valued maps (svm)?

Differential Inclusions

Differential inclusions generalize ordinary differential equations:

Problem 1.3. $\mathcal{I} = [t_0, t_f]$, $x : \mathcal{I} \to \mathbb{R}^n$ *is a solution of a differential inclusion (DI), if*

- $x(\cdot)$ *is absolutely continuous, i.e.* $x'(\cdot) \in L_1(\mathcal{I})$ *and*

$$x(t) = x(t_0) + \int_{t_0}^{t} x'(\tau) d\tau \quad (t \in \mathcal{I})$$

- $x'(t) \in F(t, x(t))$ *for a.e.* $t \in \mathcal{I}$, *where* $F : \mathcal{I} \times \mathbb{R}^n \Rightarrow \mathbb{R}^n$ *is a set-valued map*
- $x(t_0) \in X_0$ *with* $X_0 \subset \mathbb{R}^n$ *nonempty*

Special case:
$x' = f(t, x) \Leftrightarrow x' \in F(t, x) := \{f(t, x)\}$.

Differential Inclusions

Remark 1.4. *In general, a differential inclusion has not a unique solution.*
The solution funnel/trajectory tube is defined as

$$\mathcal{X}(t_0, X_0) = \{x(\cdot) \,|\, x(\cdot) \text{ solution of DI with } x(t_0) \in X_0\}$$

The reachable set at time $t \in \mathcal{I}$ *is defined as*

$$\mathcal{R}(t, t_0, X_0) = \{y \in \mathbb{R}^n \,|\, x(\cdot) \in \mathcal{X}(t_0, X_0) \text{ with } y = x(t)\}$$

$\mathcal{R}(t, t_0, X_0)$ = *cross-section of function evaluations of* $\mathcal{X}(t_0, X_0)$ *at time* t

Differential Inclusions

Questions

- existence of solutions of differential inclusions?
- computation of special solutions?
- computation of all solutions, i.e. reachable set?
- set-valued generalization of DE-solver?
- convergence order as in single-valued case?
- appropriate smoothness conditions for convergence analysis?

Connection to Control Problems

Reformulation
control problem:

$$\begin{aligned} x'(t) &= f(t, x(t), u(t)) && \text{for a.e. } t \in \mathcal{I} \\ u(t) &\in \mathcal{U} \subset \mathbb{R}^m && \text{for a.e. } t \in \mathcal{I} \\ x(t_0) &= x_0 \end{aligned}$$

differential inclusion ("drop the control"):

$$\begin{aligned} x'(t) &\in F(t, x(t)) := \bigcup_{u \in U} \{f(t, x(t), u)\} && \text{for a.e. } t \in \mathcal{I} \\ x(t_0) &\in X_0 := \{x_0\} \end{aligned}$$

solution funnel = set of admissible functions $x(\cdot)$ in control problem
reachable sets at time t_f = values of such solutions at time t_f

Connection to Control Problems

State Constraints and Viable Solutions
inital condition:

$$\psi(x(t_0)) = 0 \iff x(t_0) \in X_0 := \{z \in \mathbb{R}^n \,|\, \psi(z) = 0\}$$

A state constraint

$$s(t, x(t)) \leq 0 \quad (t \in \mathcal{I})$$

introduces a viability condition

$$x(t) \in S(t) := \{z \in \mathbb{R}^n \,|\, s(t, z) \leq 0\}.$$

A mixed control-state constraint

$$c(t, x(t), u(t)) \leq 0 \quad \text{for a.e. } t \in \mathcal{I}$$

restricts the choices for $u(t)$ by

$$x'(t) \in f(t, x(t), C(t, x(t)) = \bigcup_{u \in C(t,x(t))} \{f(t, x(t), u)\}$$

with

$$C(t, x) := \{z \in \mathbb{R}^m \,|\, c(t, x, z) \leq 0\}.$$

Shifting the state constraints into a tangent cone $T_{\Theta(t)}(x(t))$ in (DI) without state constraints could destroy Lipschitz properties of the right-hand side:

$$x'(t) \in F(t, x(t)) \cap T_{\Theta(t)}(x(t)) \text{ with } \Theta(t) := \{z \in \mathbb{R}^n \,|\, s(t, z) \leq 0\}.$$

2 Examples and Applications

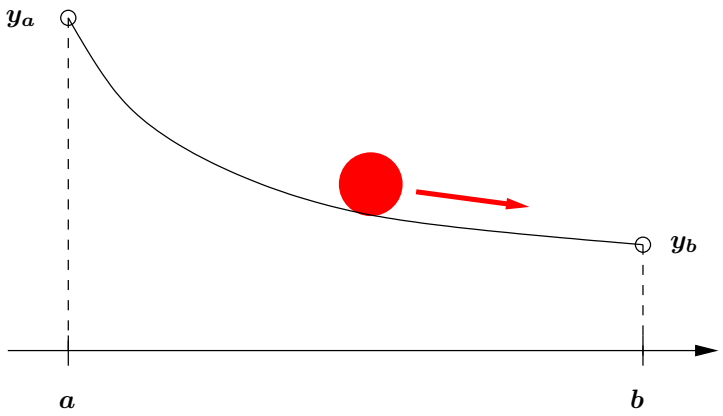
Contents

The Brachistochrone-Problem

In 1696 Johann Bernoulli posed the following problem:

> *Suppose a mass point of mass m is moving along a curve $y(x)$ starting at a point (a, y_a) and ending at (b, y_b) in the (x,y)-plane under the influence of gravity neglecting friction.*
> *Which curve $y(x)$ yields the shortest time?*

The Brachistochrone-Problem



The Brachistochrone-Problem

Bracchistochrone Problem

Minimize

$$J(x, y, \gamma, t_f) = t_f$$

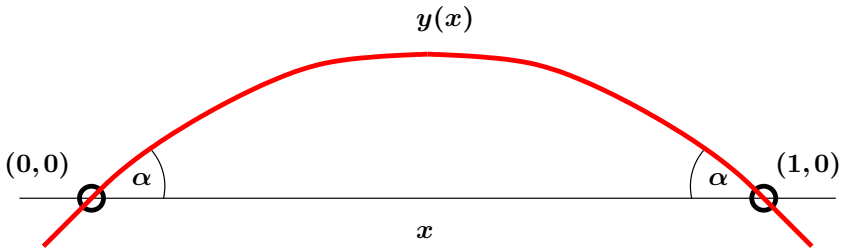
subject to

$$\begin{aligned} x'(t) &= \sqrt{2g(y_a - y(t))}\cos\gamma(t), \qquad x(0) = a,\ x(t_f) = b \\ y'(t) &= \sqrt{2g(y_a - y(t))}\sin\gamma(t), \qquad y(0) = y_a,\ y(t_f) = y_b. \end{aligned}$$

Remark: The solution is neither the direct line nor a segment of a circle!

Minimum-Energy Problem

A rod is fixed at the points $(0,0)$ and $(1,0)$ in the (x,y)-plane in such a way, that it assumes an angle α w.r.t. the x-axis:



What curve yields a minimum of the rod's bending energy?

Minimum-Energy Problem

Minimum-Energy Problem

Minimize

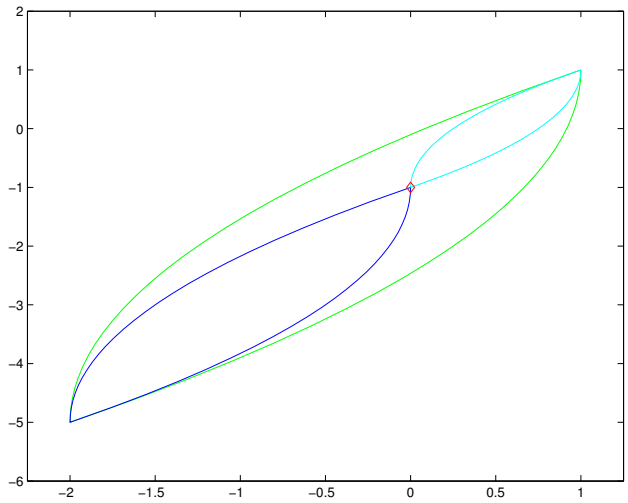
$$J(y_1, y_2, u) = \int_0^1 u(x)^2 dx$$

subject to

$$\begin{aligned} y_1'(x) &= y_2(x), \qquad y_1(0) = y_1(1) = 0, \\ y_2'(x) &= u(x), \qquad y_2(0) = -y_2(1) = \tan\alpha. \end{aligned}$$

Minimum-Energy Problem

Reachable Set for several control sets $\mathcal{U}$



$\diamond\ \mathcal{U} = \{-2\}$

$-\ \mathcal{U} = [-6, -2]$

$-\ \mathcal{U} = [-2, 0]$

$-\ \mathcal{U} = [-6, 0]$

no end conditions,
$u(t) \in \mathcal{U}$!

The reachable sets are calculated by the set-valued combination method "iterated trapezoidal rule/Heun's method" with $N = 100000$ sub-intervals.

Minimum-Energy Problem

Modification: additional constraint $y(x) \leq y_{\max}$

Minimum-Energy Problem

Minimize

$$J(y_1, y_2, u) = \int_0^1 u(x)^2 dx$$

subject to

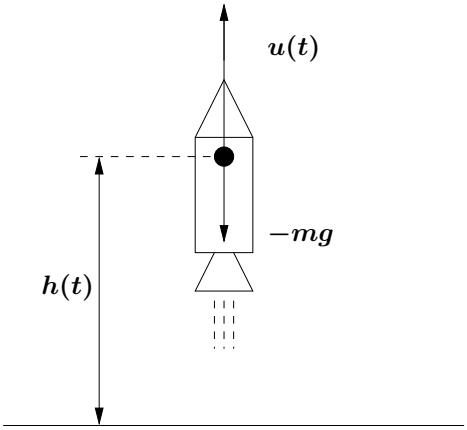
$$\begin{aligned} y_1'(x) &= y_2(x), \qquad y_1(0) = y_1(1) = 0, \\ y_2'(x) &= u(x), \qquad y_2(0) = -y_2(1) = \tan\alpha. \end{aligned}$$

and the state constraint

$$y_1(x) - y_{\max} \leq 0.$$

Vertical Ascent of a Rocket

A rocket of mass m starts at $t = 0$ at rest on earth level $h(0) = 0$ and is accelerated vertically by controlling the thrust $u(t)$:



Vertical Ascent of a Rocket

Task for pilot: For a given amount of fuel η, reach the altitude $H > 0$ in minimal time t_f with bounded thrust $0 \leq u(t) \leq u_{\max}$ (constant mass, no air resistance)!

Vertical Ascent of a Rocket

Minimize

$$J(h, v, u, t_f) = t_f$$

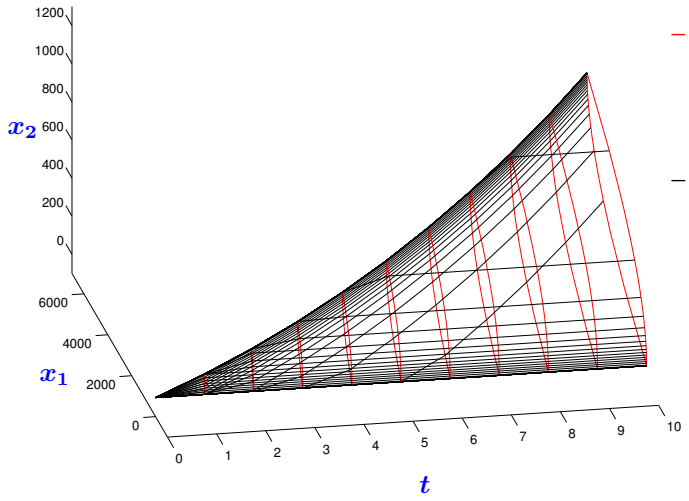
subject to $0 \leq u(t) \leq u_{\max}$,

$$\begin{aligned} h'(t) &= v(t), \qquad h(0) = 0,\ h(t_f) = H, \\ v'(t) &= -g + \frac{u(t)}{m}, \quad v(0) = 0, \end{aligned}$$

$$c \int_0^{t_f} u(t)dt = \eta.$$

Vertical Ascent of a Rocket

Solution Funnel, 2D, Viewpoint (-11, 32)



– reachable sets for $t_i = i$, $i = 0, \ldots, 10$

– corresponding supporting points in same direction

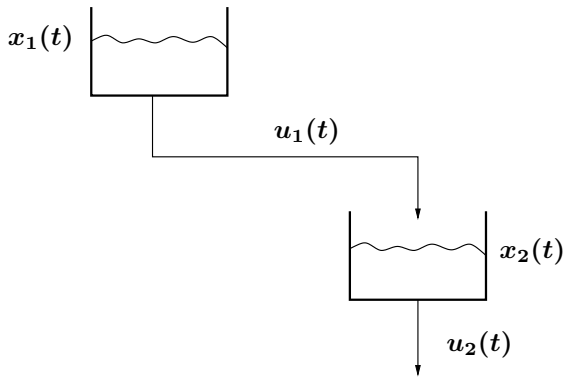
values: $m = 1$, no end condition, no limit on the fuel!

The solution funnel is calculated by the set-valued combination method "iterated trapezoidal rule/Heun's method" with step-size $h = 0.001$ and $u_{\max} = 100$.

System of two Water Boxes

Given:

- 2 water boxes, water level $x_i(t) \geq 0$ at time t in box $i = 1, 2$
- outflow rates $0 \leq u_i(t) \leq 1$, $i = 1, 2$



System of two Water Boxes

Differential equations:

$$\begin{array}{rcll} x_1'(t) & = & -u_1(t), & x_1(0) = x_1^0, \\ x_2'(t) & = & u_1(t) - u_2(t), & x_2(0) = x_2^0. \end{array}$$

Goal: Maximize average outlow

$$\int_0^{10} (10 - t)u_1(t) + tu_2(t)dt$$

System of two Water Boxes

Water Boxes
Maximize

$$J(x_1, x_2, u_1, u_2) = \int_0^{10} (10 - t)u_1(t) + tu_2(t)dt$$

subject to

$$\begin{array}{rcll} x_1'(t) & = & -u_1(t), & x_1(0) = x_1^0, \\ x_2'(t) & = & u_1(t) - u_2(t), & x_2(0) = x_2^0. \end{array}$$

and the state constraints

$$x_i(t) \geq 0, \qquad \forall t \in [0, 10],\ i = 1, 2,$$

and the control constraints $0 \leq u_i(t) \leq 1$, $t \in [0, 10]$, $i = 1, 2$.

Climate Change Model

Model Parameters
WBGU scenario:
(WBGU = German Advisory Council on Global Change) simple model of climate change assessment:

$F(\cdot)$: cumulation of CO_2 emissions caused by humankind

$C(\cdot)$: carbon concentration

$T(\cdot)$: global mean temperature

$E(\cdot)$: CO_2 emission profile controlling the allowed CO_2 emissions

Questions:

- What are the admissible emissions in the year t?
- What are the feasible concentrations $C(t)$ in that year?
- Is it possible to realize a global mean temperature T^* in year t?

Climate Change Model

Reachable Set
control problem:

$$\begin{aligned} F'(t) &= E(t), \\ C'(t) &= B \cdot F(t) + \beta \cdot E(t) - \sigma \cdot (C(t) - C_1), \\ T'(t) &= \mu \cdot \ln\left(\frac{C(t)}{C_1}\right) - \alpha \cdot (T(t) - T_1), \\ E'(t) &= u(t)E(t), \quad |u(t)| \leq u_{\max} \end{aligned}$$

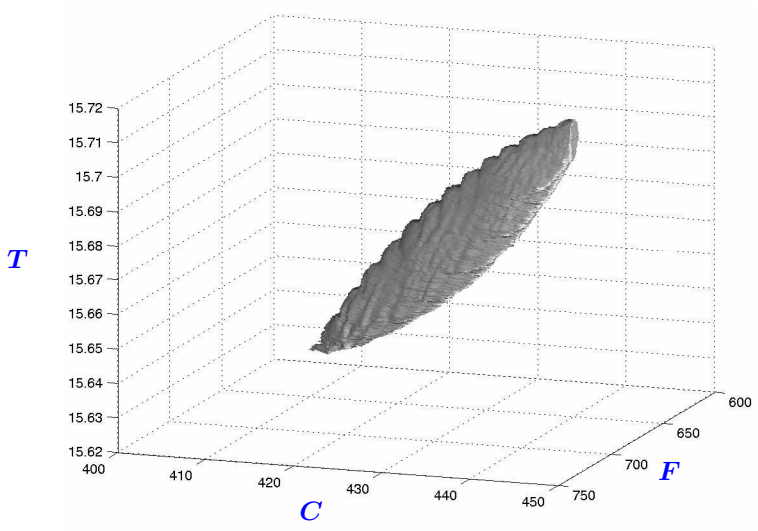
with state constraints

$$\begin{aligned} T_1 \leq\ & T(t) \leq T_{\max}, \\ 0 \leq\ & T'(t) \leq T'_{\text{crit}}(T(t)), \end{aligned}$$

$$T'_{\text{crit}}(T(t)) = \begin{cases} T'_{\max} & \text{if } T_1 \leq T(t) \leq T_{\max} - 1, \\ T'_{\max}\sqrt{T_{\max} - T(t)} & \text{if } T_{\max} - 1 \leq T(t) \leq T_{\max}. \end{cases}$$

Climate Change Model

3D-projection on $F - C - T$-axes from 4D-reachable set, $t_f = 30$



The reachable set for this nonlinear differential inclusion is calculated in [Cha03] with the set-valued Euler's method and step-size $h = 0.5$.

Virtual Test-Drive

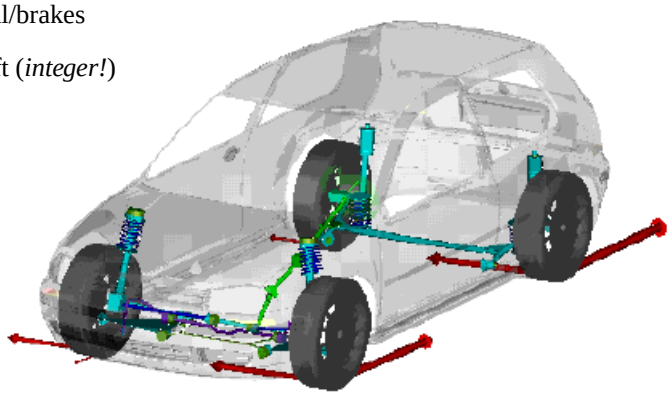
Components:

- mathematical model of the car (⟶ differential equations)
- mathematical model of the test-course
- mathematical model of the driver (⟶ optimal control problem)

VW Golf

Controls:

- Steering wheel
- Gas pedal/brakes
- Gear shift (*integer!*)

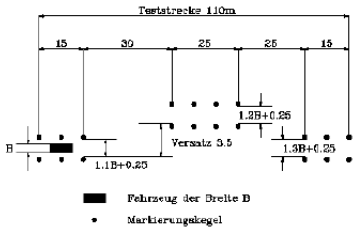


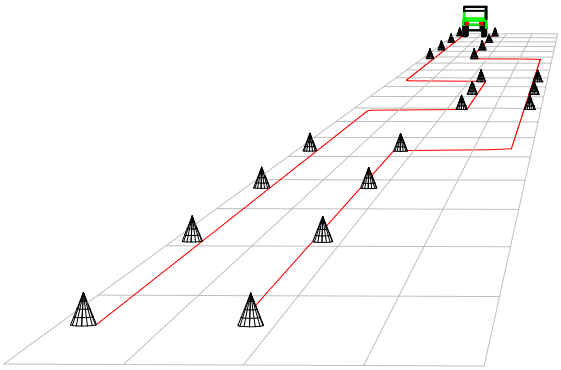
Virtual Test-Drive: Driver

Optimal Control Problem

Min.	$\varphi(x(t_f), t_f, p)$			'steering effort'
Max.				'average velocity'
				'final time'
				'linear combination'
s.t.	$F(x(t), x'(t), u(t), p)$	$=$	0	'car model'
	$C(x(t), u(t), p)$	$\leq$	0	'boundaries of track'
	$\psi(x(t_0), x(t_f), p)$	$=$	0	'initial/final position'
	$u(t)$	$\in$	$\mathcal{U}$	'limitations steering/acceleration'

Virtual Test-Drive: Test-Course

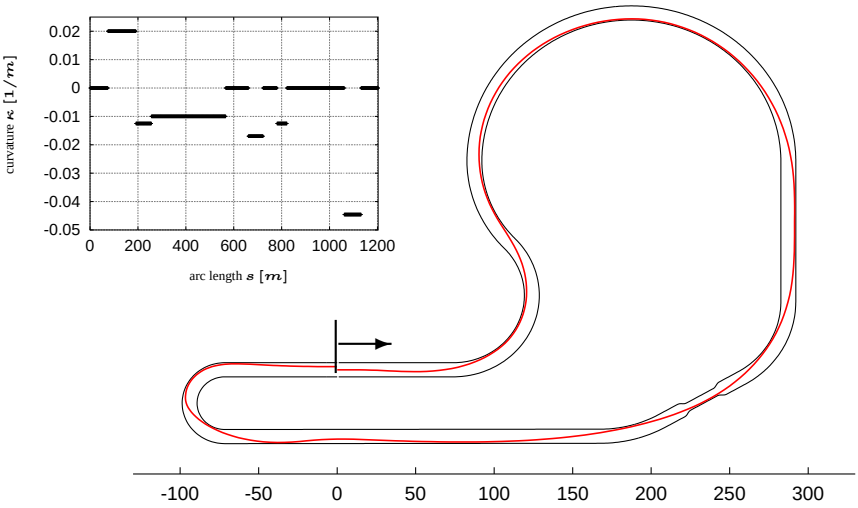




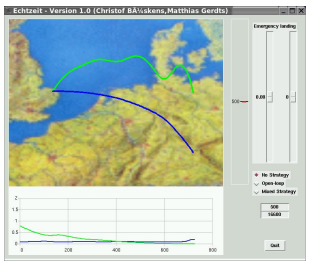
boundary: piecewise def. polynomials

Virtual Test-Drive: Test-Course

middle line: curve $\gamma : [0, L] \to \mathbb{R}^2$, piecewise lin. curv. κ



Emergency Landing Manoeuvre



- Scenario: propulsion system breakdown
- Goal: maximization of range w.r.t. current position
- Controls: lift coefficient, angle of bank
- no thrust available; no fuel consumption (constant mass)

3 Convex Analysis

3.1 Convex Sets

Basic Facts
Why do we deal with convex sets?

- reachable sets are convex for linear control problems
- properties of S remain valid for its convex hull $\mathbf{co}(S)$ (compactness, boundedness)
- uniqueness of best approximation of a point to a set
- convex sets can be easily described by support functions or supporting points
- convex sets can be easily stored in computer (only store its boundary, extreme points, exposed points, supporting points, support functions)

Important Tools for (Later) Proofs

- Caratheodory's theorem (convex combination of max. $n+1$ elements necessary for convex hull)
- separation theorems (separate two "almost" disjoint sets by a hyperplane)
- representation theorems of convex sets:
  Theorem of Minkowski (resp. Krein/Milman) on convex hull of extreme points,
  Theorem of Straszewicz on closed convex hull of exposed points

3.1.1 Basic Definitions and Properties

Notation 3.1. *Let $x, y \in \mathbb{R}^n$. Then, $\|x\|$ stands for the Euclidean norm of x and $\langle x, y\rangle$ means the scalar product of x and y.*

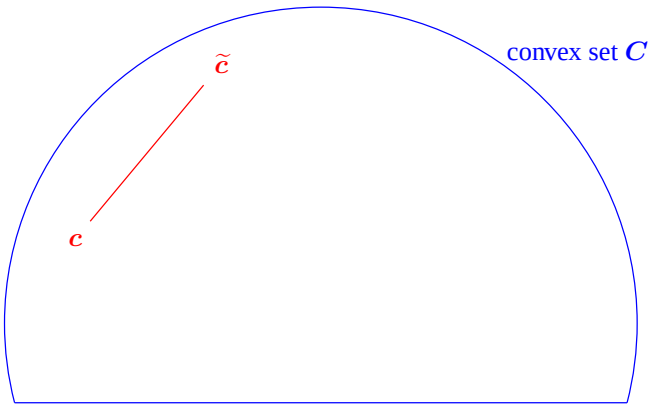
$B_1(0)$ is the closed Euclidean unit ball in $\mathbb{R}^n$ and S_{n-1} its boundary, the unit sphere.

Definition 3.2 (convex set). A set $C \subset \mathbb{R}^n$ is *convex*, if

$$(1-\lambda)c + \lambda\tilde{c} \in C \quad \text{for all } c, \tilde{c} \in C \text{ and for all } \lambda \in [0,1].$$

Denote by $\mathcal{C}(\mathbb{R}^n)$ the *set of all nonempty, convex, compact sets in* $\mathbb{R}^n$ and by $\mathcal{K}(\mathbb{R}^n)$ the *set of all nonempty, compact sets in* $\mathbb{R}^n$.

convex set



Example 3.3 *(convex and nonconvex sets).*

(i) All closed and bounded (or unbounded) intervals in $\mathbb{R}$ are convex.

(ii) All open and bounded (or unbounded) intervals in $\mathbb{R}$ are convex.

(iii) The half-open/half-closed interval $[a, b)$ resp. $(a, b]$ are also convex.

(iv) The unit ball $B_1(0) \subset \mathbb{R}^n$ is convex.

(v) The unit square $[-1, 1]^n$ is convex.

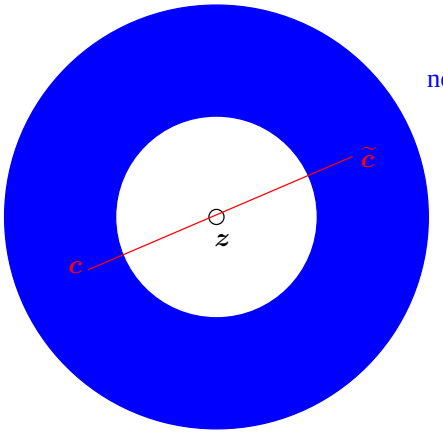
(vi) A point set $\{x\}$ with $x \in \mathbb{R}^n$ is convex.

(vii) The set $\{x, y\}$ with $x, y \in \mathbb{R}^n$ and $x \neq y$ is not convex.

(viii) Let $r \in [0, 1)$. Then, $M := B_1(0) \setminus B_r(0_{\mathbb{R}^n})$ is not convex.

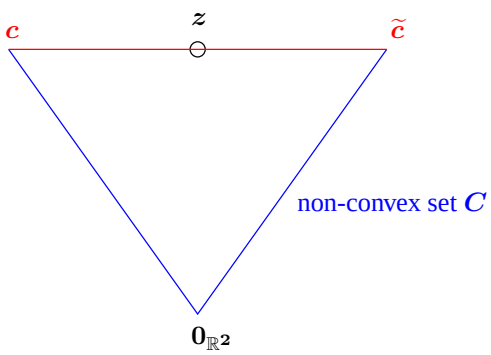
(ix) The set $M := \operatorname{co}\{-e^1 + e^2, 0_{\mathbb{R}^n}\} \cup \operatorname{co}\{e^1 + e^2, 0_{\mathbb{R}^n}\}$ is not convex.

(viii) non-convex set



$$z = \lambda c + (1-\lambda)\widetilde{c} \notin C$$

(ix) non-convex set (union of two convex line segments)



$$c = -e^1 + e^2$$
$$\widetilde{c} = e^1 + e^2$$
$$\lambda = \tfrac{1}{2}$$
$$z = \lambda c + (1-\lambda)\widetilde{c} = e^2 \notin C$$

Proposition 3.4. *Let $C \subset \mathbb{R}^n$ be convex. Then, $\mathbf{int}(C)$ and $\overline{C}$ is also convex.*

Proof. Let $x, y \in \mathbf{int}(C)$ and $\lambda \in [0,1]$.
Then, there exists $\varepsilon_1, \varepsilon_2 > 0$ with $B_{\varepsilon_i}(x) \subset C$, $i = 1,2$. Consider an arbitrary $\eta \in B_1(0)$ for which $x + \varepsilon_1\eta, y + \varepsilon_2\eta \in C$. Hence,

$$\begin{aligned}&\lambda(x+\varepsilon_1\eta) + (1-\lambda)(y+\varepsilon_2\eta)\\ =&(\lambda x + (1-\lambda)y) + \underbrace{(\lambda\varepsilon_1 + (1-\lambda)\varepsilon_2)}_{=:\varepsilon>0}\eta \in C.\end{aligned}$$

Therefore, $B_\varepsilon(z) \subset C$ with $z := \lambda x + (1-\lambda)y$, such that $z \in \mathbf{int}(C)$.

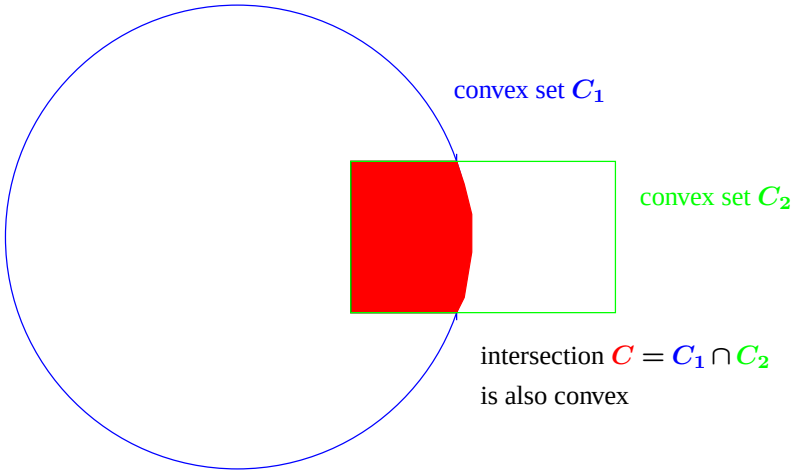
Let $x, y \in \overline{C}$ and $\lambda \in [0,1]$ and choose $(x^m)_m, (y^m)_m \subset C$ with $x^m \xrightarrow[m\to\infty]{} x$ and $y^m \xrightarrow[m\to\infty]{} y$. Since C is convex, we have $\lambda x^m + (1-\lambda)y^m \in C$. Hence,

$$\lambda x^m + (1-\lambda)y^m \xrightarrow[m\to\infty]{} \lambda x + (1-\lambda)y \in \overline{C}.$$

□

Proposition 3.5. *Let I be an arbitrary index set and $C_i \subset \mathbb{R}^n$ be convex for every $i \in I$. Then, $C := \bigcap_{i \in I} C_i$ is also convex.*

intersection of two convex sets



Remark 3.6. *In general, the union of convex sets need not be convex any longer. See Example 3.3(ix) for an example which consists of a nonconvex union of two (convex) line segments in $\mathbb{R}^2$.*

Proposition 3.7. *Let $C_i \in \mathbb{R}^{n_i}$, $i = 1, 2$, be convex sets and $n_1 + n_2 = n$.*
Then, $C = C_1 \times C_2 \subset \mathbb{R}^n$ with $C_1 \times C_2 = \{(c^1, c^2) \in \mathbb{R}^n \,|\, c^i \in C_i, i = 1, 2\}$ is convex.

Example 3.8 *(Further examples).*

(i) Clearly, the complete space $\mathbb{R}^n$ and the empty set are convex.

(ii) Each linear subspace of $\mathbb{R}^n$ is convex.

(iii) Each affine set of $\mathbb{R}^n$ is convex.

(iv) Each line in $\mathbb{R}^n$ is convex.

(v) Each half-space of $\mathbb{R}^n$ is convex.

(vi) Let $A \in \mathbb{R}^{m \times n}$ be a matrix and $b \in \mathbb{R}^m$. Then, the set of admissible points $M := \{x \in \mathbb{R}^n \,|\, Ax = b\}$ and $\widetilde{M} := \{x \in \mathbb{R}^n \,|\, Ax \leq b\}$ are convex.

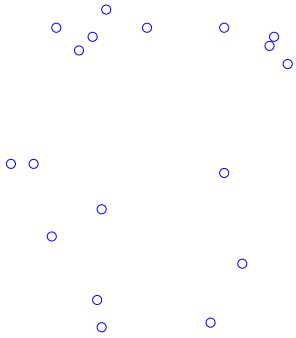
Definition 3.9. Let $C \subset \mathbb{R}^n$, $k \in \mathbb{N}$ and $c^i \in C$, $i = 1, \ldots, k$. A *convex combination* of $(c^i)_{i=1,\ldots,k}$ is a sum $\sum_{i=1}^k \lambda_i c^i$ with $\lambda_i \geq 0$, $i = 1, \ldots, k$ and $\sum_{i=1}^k \lambda_i = 1$.
The *convex hull* of C is defined as

$$\mathrm{co}(C) = \bigcap_{D \supset C \text{ convex}} D.$$

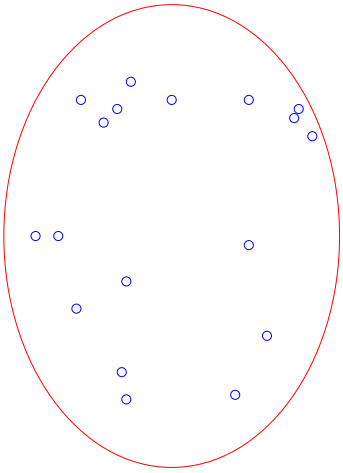
Similarly, the *affine hull* of C, denoted by $\mathbf{aff}(C)$, could be defined.

Lemma 3.10. *If $C \subset \mathbb{R}^n$ is convex, each convex combination is an element of C.*

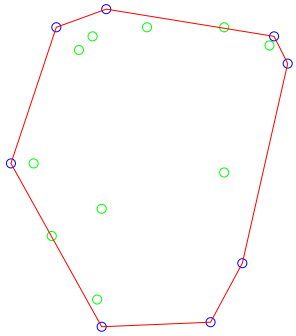
Building a Convex Hull from Discrete Points



Imagine a Rubber Band which Includes all Nails = Points



Rubber Band Forms the Convex Hull



– boundary of convex hull = polytope

o vertices of convex hull

o non-extreme points or points
in the interior of the convex hull

Remark 3.11. *Algorithms for the computation of convex hulls can be found, cf. e.g. [dvOS97-ICSCG, O'R98-ICSCG, BY98-ICSCG, PS88-ICSCG] and the citations on computational geometry at the end of this subsubsection. Graham's algorithm in $\mathbb{R}^2$ for N points and Divide-and-Conquer in $\mathbb{R}^3$ achieve $\mathcal{O}(N\log(N))$ complexity. The lower bound in $\mathbb{R}^n$ is $\mathcal{O}(N\log(N)+N^{\lfloor n/2\rfloor})$ (cf. [BY98-ICSCG, Theorem 8.2.1]).*

Example 3.12. (i) Let $a,b\in\mathbb{R}$ with $a<b$. Then, $\mathbf{co}\{a,b\}=[a,b]$.

(ii) Let $x,y\in\mathbb{R}^n$. Then, $\mathbf{co}\{x,y\}=\{\lambda x+(1-\lambda)y\,|\,\lambda\in[0,1]\}$ is the line segment spanned by x,y.

(iii) Let $\{e^1,e^2\}$ be the standard basis of $\mathbb{R}^2$. Then,

$$\mathbf{co}\{e^1+e^2,-e^1+e^2,-e^1-e^2,e^1-e^2\}=[-1,1]^2.$$

(iv) $\mathbf{co}(S_{n-1})=B_1(0)$

Proposition 3.13. *Let $C\subset\mathbb{R}^n$. Then,*

$$\mathbf{co}(C)=\{\sum_{i=1}^{k}\lambda_i c^i\,|\,c^i\in C,\ \lambda_i\geq 0,\ \sum_{i=1}^{k}\lambda_i=1,\ k\in\mathbb{N}\}.$$

Proof. "$\supset$": For any convex set $D\supset C$, we have for $c^i\in C\subset D$ that $\sum_{i=1}^{k}\lambda_i c^i\in D$. Hence, $\sum_{i=1}^{k}\lambda_i c^i\in\mathbf{co}(C)$.

"$\subset$": Let $\sum_{i=1}^{k}\lambda_i c^i$ and $\sum_{i=1}^{m}\widetilde{\lambda}_i\widetilde{c}^i$ be elements from $\mathbf{co}(C)$ and $\mu\in[0,1]$. Then,

$$\mu\sum_{i=1}^{k}\lambda_i c^i+(1-\mu)\sum_{i=1}^{m}\widetilde{\lambda}_i\widetilde{c}^i$$

is again a convex combination of elements in C and therefore the right-hand side is convex. Therefore, $\mathbf{co}(C)$ is included in the right-hand side per definition. □

Theorem 3.14 (Carathéodory). *Let $M\subset\mathbb{R}^n$. Then,*

$$\mathbf{co}(M)=\{\sum_{i=1}^{n+1}\lambda_i v^i\,|\,v^i\in M,\lambda_i\geq 0,\ \sum_{i=1}^{n+1}\lambda_i=1\}.$$

Proof. Take a convex combination $z=\sum_{i=1}^{k}\lambda_i v^i$ with $k\in\mathbb{N}$ and $k>n+1$.
Let us consider the k vectors $\binom{v^i}{1}\in\mathbb{R}^{n+1}$ which have to be linear dependent. Hence, there exists non-zero vector $(\alpha_1,\ldots,\alpha_k)^\top\in\mathbb{R}^k$ with

$$\sum_{i=1}^{k}\alpha_i\begin{pmatrix}v^i\\1\end{pmatrix}=0_{\mathbb{R}^{n+1}}.$$

Since $\sum_{i=1}^{k}\alpha_i=0$, not all α_i could be negative or could be equal to zero. Set

$$\widehat{t}:=\frac{\lambda_{i_0}}{\alpha_{i_0}}:=\min_{\alpha_i>0}\frac{\lambda_i}{\alpha_i}\geq 0.$$

For all i with $\alpha_i>0$:

$$\frac{\lambda_i}{\alpha_i}\geq\widehat{t}\quad\text{and therefore,}\quad\lambda_i-\widehat{t}\cdot\alpha_i\geq 0$$

For all i with $\alpha_i\leq 0$:

$$\lambda_i-\underbrace{\widehat{t}}_{\geq 0}\cdot\alpha_i\geq\lambda_i\geq 0$$

With the help of $\widehat{t}$ a new convex combination could be defined: Set

$$\widetilde{\lambda}_i := \lambda_i - \widehat{t} \cdot \alpha_i \geq 0 \quad \text{(i=1,...,k)}.$$

The new convex combination yields

$$\sum_{i=1}^{k} \widetilde{\lambda}_i = \underbrace{\sum_{i=1}^{k} \lambda_i}_{=1} - \widehat{t} \cdot \underbrace{\sum_{i=1}^{k} \alpha_i}_{=0} = 1,$$
$$\sum_{i=1}^{k} \widetilde{\lambda}_i v^i = \underbrace{\sum_{i=1}^{k} \lambda_i v^i}_{=z} - \widehat{t} \cdot \underbrace{\sum_{i=1}^{k} \alpha_i v^i}_{=0_{\mathbb{R}^n}} = z.$$

The weight λ_{i_0} yields

$$\lambda_{i_0} = \alpha_{i_0} \widehat{t} \quad \text{and therefore,} \quad \widetilde{\lambda}_{i_0} = \lambda_{i_0} - \alpha_{i_0} \widehat{t} = 0.$$

Now, z can be represented as

$$z = \sum_{\substack{i=1,\dots,k \\ i \neq i_0}} \widetilde{\lambda}_i v^i,$$

a convex combination with $k-1$ elements of M.
If $k-1 = n+1$, then the proof is finished. Otherwise, repeat this method starting with $\sum_{\substack{i=1,\dots,k \\ i \neq i_0}} \widetilde{\lambda}_i v^i$ until only $n+1$ elements remain in the convex combination. These $n+1$ elements may happen to be linear independent. □

Remark 3.15. *Let $M = \{v^1, v^2, \dots, v^m\} \subset \mathbb{R}^n$ with m different points.*
Assume first dimension $n = 2$.

$m = 1$: $\mathrm{co}(M)$ *is a point (and consists of convex combinations of 1 element).*

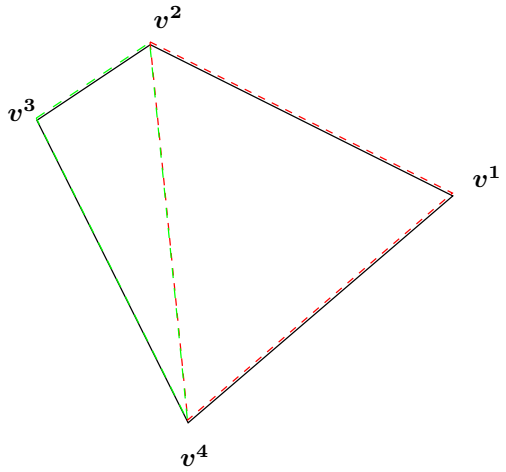
$m = 2$: $\mathrm{co}(M)$ *is a line segment and consists of convex combinations of 2 elements.*

$m = 3$: $\mathrm{co}(M)$ *is a triangle and consists of convex combinations of 3 elements (see Carathéodory's Theorem 3.14).*

$m = 4$: $\mathrm{co}(M)$ *is a general quadrangle, but consists of convex combinations of 3 elements (see Carathéodory's Theorem 3.14), since the quadrangle could be partitioned into two triangles.*

For $n = 3$ a general polytope mit m vertices could be partitioned with the help of tetrahedrons (convex combinations of 4 elements according to Carathéodory's Theorem 3.14).

Carathéodory's Theorem ($m = 4$)



convex hull of set

$M = \{v^i \mid i = 1, \dots, 4\}$

A point $z \in \mathbf{co}(M)$

with $z = \sum_{i=1}^{4} \lambda_i v^i$ lies

either in the green triangle

or in the red one.

Proposition 3.16. *Let $C \subset \mathbb{R}^n$.*

(i) If C is convex, then $\mathbf{co}(C) = C$.

(ii) If C is bounded, i.e. $C \subset B_r(m)$ with $r > 0$, then $\mathbf{co}(C)$ is also bounded by $B_r(m)$.

(iii) If C is compact, then $\mathbf{co}(C)$ is also compact.

Proof.
(i) Clearly, $C \subset \mathbf{co}\, C$. From the definition the convex set $D := C$ fulfills $D \supset C$ and therefore, $\mathbf{co}\, C \subset C$.
(ii) Let $z = \sum_{i=1}^{k} \lambda_i c^i \in \mathbf{co}\, C$. Then,

$$\|z - m\| = \|\sum_{i=1}^{k} \lambda_i (c^i - m)\| \leq \sum_{i=1}^{k} \lambda_i \|c^i - m\|$$
$$\leq \sum_{i=1}^{k} \lambda_i r = r.$$

(iii) $\mathbf{co}(C)$ is bounded, since C is bounded.
To show the closedness, let $(z^m)_m \subset \mathbf{co}\, C$ with $z^m \xrightarrow[m\to\infty]{} z$. Caratheodory's theorem shows that

$$z^m = \sum_{i=1}^{n+1} \lambda_{i,m} c^{(i,m)}.$$

Since $(\lambda_{i,m})_m \subset [0,1]$ and $(c^{i,m_k})_k \subset C$ are bounded sequences, they contain convergent (w.l.o.g. common) subsequences

$$(\lambda_{i,m_k})_k \subset (\lambda_{i,m})_m \quad \text{with} \quad \lambda_{i,m_k} \xrightarrow[k\to\infty]{} \lambda_i \in [0,1],$$
$$(c^{i,m_k})_k \subset (c^{i,m})_m \quad \text{with} \quad c^{i,m_k} \xrightarrow[k\to\infty]{} c^i \in C.$$

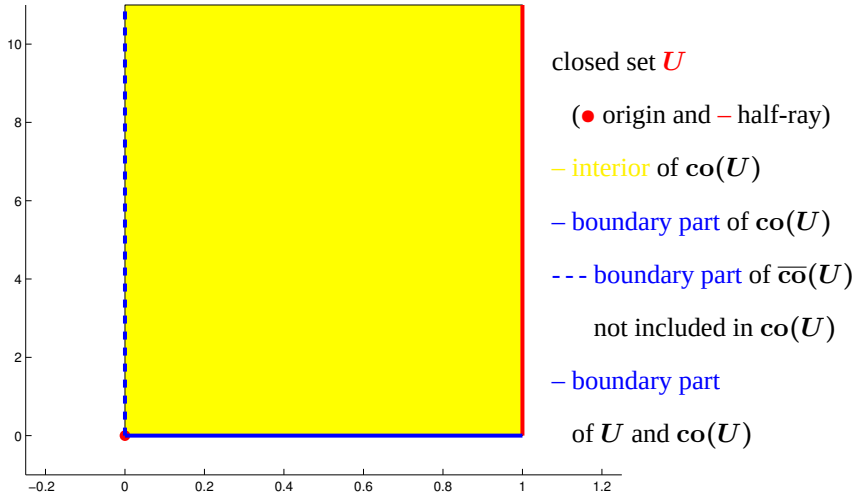
Altogether, we have

$$z^{m_k} = \sum_{i=1}^{n+1} \lambda_{i,m_k} c^{(i,m_k)} \xrightarrow[k\to\infty]{} \sum_{i=1}^{n+1} \lambda_i c^i \in \mathbf{co}(C).$$

Since the complete sequence converges to z, the subsequence converges also to z which shows that $z = \sum_{i=1}^{n+1} \lambda_i c^i \in \mathbf{co}(C)$. □

Example 3.17. If $C \subset \mathbb{R}^n$ is only closed, then $\mathbf{co}(C)$ is not necessarily closed. Take $C := \{\binom{0}{0}\} \cup \{\binom{1}{y} \mid y \geq 0\}$. The convex hull of C is $\{\binom{0}{0}\} \cup \{\binom{t}{y} \mid t \in (0,1],\ y \geq 0\}$ which is not closed.

Non-Closed Convex Hull of Closed Set



Definition 3.18. Let $p^i \in \mathbb{R}^n$, $i = 1, \ldots, k$. Then, $P := \mathbf{co}\{p^i \mid i = 1, \ldots, k\}$ is called (convex) *polytope* with vertices p^i, $i = 1, \ldots, k$.

Remark 3.19. *A rich literature exists dedicated only to polytopes,*
cf. e.g. [Grü03-ICSP, Zie98-ICSP, Brø83-ICSP] or sections in [Sch93-ICSP] and the citations on polytopes at the end of this subsubsection.

Example 3.20. (i) A convex polytope with 1 vertex is just the point itself.

(ii) Let $v^1, v^2 \in \mathbb{R}^n$ be two different points. Then, the convex polytope with these two vertices is the line segment connecting both points.

(iii) Let $w^1, w^2 \in \mathbb{R}^2$ be two independent vectors. Then, the convex polytope with the three vertices $v^1, v^1 + w^1, v^1 + w^2$ is a triangle for every $v^1 \in \mathbb{R}^n$.

(iv) Let v^i, $i = 1, 2, 3, 4$, be four points such that every of the four points is not in the convex hull of the other three. Then, the convex polytope with these four vertices is a (convex) quadrangle.

(v) Let $\{e^1, \ldots, e^n\}$ be the standard basis of $\mathbb{R}^n$. Then, $\mathbf{co}\{0_{\mathbb{R}^n}, e^1, e^2, \ldots, e^n\}$ is the n-dimensional unit simplex.

For Further Reading on Convex Sets

References

[HUL93-ICS] J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms I*, volume 305 of *Grundlehren der mathematischen Wissenschaften*. Springer-Verlag, Berlin–Heidelberg–New York–London–Paris–Tokyo–Hong Kong–Barcelona–Budapest, 1993.

[Sch93-ICS] R. Schneider. *Convex Bodies: The Brunn-Minkowski Theory*, volume 44 of *Encyclopedia of Mathematics and Applications*. Cambridge University Press, Cambridge, 1993.

[Web94-ICS] R. Webster. *Convexity*. Oxford Science Publications. Oxford University Press, Oxford–New York–Tokyo, 1994.

[Mar77-ICS] J. T. Marti. *Konvexe Analysis*, volume 54 of *Lehrbücher und Monographien aus dem Gebiet der Exakten Wissenschaften, Mathematische Reihe*. Birkhäuser, Basel–Stuttgart, 1977.

[Roc70-ICS] R. T. Rockafellar. *Convex Analysis*, volume 28 of *Princeton Mathematical Series*. Princeton University Press, Princeton, NJ, 2nd edition, 1972.

[Val64-ICS] F. A. Valentine. *Convex Sets*. Robert E. Krieger Publishing Company, Huntington, N. Y., 1964. reprint.

[Lei85-ICS] K. Leichtweiß. *Vypuklye mnozhestva*. Glavnaya Redaktsiya Fiziko-Matematicheskoj Literatury. Nauka, Moskva, 1985. Russian translation by V. A. Zalgaller and T. V. Khachaturova.

[Lei80-ICS] K. Leichtweiß. *Konvexe Mengen*. Hochschultext. Springer-Verlag, Berlin–Heidelberg–New York, 1980.

[GW83-ICS] P. M. Gruber and J. M. Wills. *Convexity and Its Application*, volume XVI of *Pure and Applied Mathematics*. Birkhäuser Verlag, Basel–Boston–Stuttgart, 1983.

[GW93-ICS] P. M. Gruber and J. M. Wills. *Handbook of Convex Geometry. Volume A*. North-Holland, Amsterdam, 1993.

For Further Reading on Computational Geometry

References

[dvOS97-ICSCG] M. de Berg, M. van Kreveld, M. Overmars, and O. Schwarzkopf. *Computational Geometry. Algorithms and Applications*. Springer, Berlin–Heidelberg–New York–Barcelona–Budapest–Hong Kong–London–Milan–Paris–Santa Clara–Singapore–Tokyo, 1997.

[O'R98-ICSCG] J. O'Rourke. *Computational Geometry in C*. Cambridge University Press, Cambridge, 1998. 2nd ed.

[BY98-ICSCG] J.-D. Boissonnat and M. Yvinec. *Algorithmic Geometry*. Cambridge University Press, Cambridge, 1998.

[PS88-ICSCG] F. P. Preparata and M. I. Shamos. *Computational Geometry. An Introduction*. Texts and Monographs in Computer Science. Springer, New York et al., 1988. corr. and expanded 2. print.

[GO97-ICSCG] J. E. Goodman and J. O'Rourke, editors. *Handbook of Discrete and Computational Geometry*. CRC Press Series on Discrete Mathematics and Its Applications. CRC Press, Boca Raton, FL–New York, 1997.

[Ede87-ICSCG] H. Edelsbrunner. *Algorithms in combinatorial geometry*, volume 10 of *European Association for Theoretical Computer Science: EATCS monographs on theoretical computer science*. Springer, Berlin et al., 1987.

For Further Reading on Polytopes and Polyhedra

References

[Grü03-ICSP] B. Grünbaum. *Convex Polytopes*, volume 221 of *Graduate Texts in Mathematics*. Springer, New York–Berlin–Heidelberg–Hong Kong–London–Milan–Paris–Tokyo, 2nd edition, 2003.

[Zie98-ICSP] G. M. Ziegler. *Lectures on Polytopes*, volume 152 of *Graduate Texts in Mathematics*. Springer, New York–Berlin–Heidelberg–London–Paris–Tokyo–Hong Kong–Barcelone–Budapest, 1998. revised 1st ed., corr. 2. print.

[Brø83-ICSP] A. Brønsted. *An Introduction to Convex Polytopes*, volume 90 of *Graduate Texts in Mathematics*. Springer, New York–Heidelberg–Berlin, 1983.

[Kle63-ICS] V. Klee, editor. *Convexity. Proceedings of symposia in pure mathematics. Vol. VIII. Held at the University of Washington Scattle, Washington June 13-15, 1961*, Providence, Rhode Island, 1963. AMS.

[GW83-ICSP] P. M. Gruber and J. M. Wills. *Convexity and Its Application*, volume XVI of *Pure and Applied Mathematics*. Birkhäuser Verlag, Basel–Boston–Stuttgart, 1983.

[Ale05-ICSP] A. D. Aleksandrov. *Convex Polyhedra*. Springer Monographs in Mathematics. Springer, Berlin–Heidelberg–New York, 2005. revised edition and English translation by V. A. Zalgaller.

[Ale50-ICSP] A. D. Aleksandrov. *Vypuklye mnogogranniki*. Gosudarstv. Izdat. Tekhn.-Teor. Lit. Nauka, Moskau–Leningrad, 1950.

[Sch93-ICSP] R. Schneider. *Convex Bodies: The Brunn-Minkowski Theory*, volume 44 of *Encyclopedia of Mathematics and Applications*. Cambridge University Press, Cambridge, 1993.

3.1.2 Extreme Sets

The following definition of an extreme set does not require the convexity of the set.

Definition 3.21. Let $M \subset \mathbb{R}^n$ be an arbitrary set and $E \subset M$.
E is called *extreme set* of M (or *extreme* in M), if for all $x, y \in M$ and all $\lambda \in (0,1)$ with $\lambda x + (1-\lambda)y \in E$ always follows that $x, y \in E$.
If $E = \{e\}$ is an extreme set of M with one element, then e is called *extreme point* of M.
The set of extreme points of M is denoted by $\mathbf{ext}(M)$.

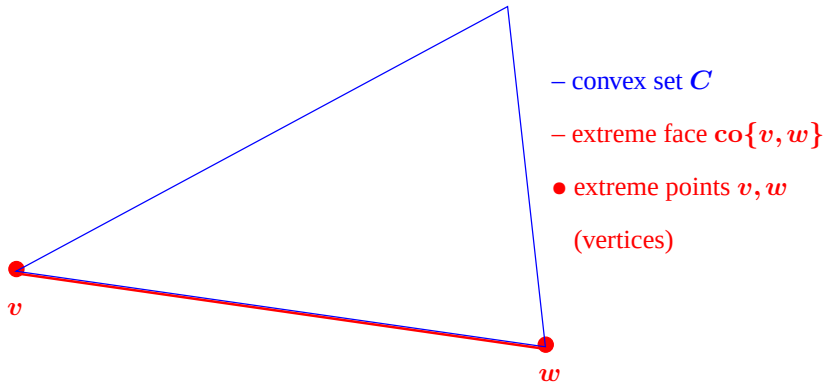
If a point z lies in the open line segment connecting x, y and is contained in E, then the end-points of the line segment x and y themselves must lie in E.

Applied to convex sets, the notion of faces could be introduced by additionally requiring the convexity of the extreme set.

Definition 3.22. Let $C \subset \mathbb{R}^n$ be convex and $E \subset C$ be nonempty.
E is called a *face* of C, if E is extreme in C and E is itself convex.
0-dimensional faces of C are called *(extreme) vertices*,
1-dimensional faces of C are called *(extreme) edges*,
$(k-1)$-dimensional faces of C are called *(extreme) facet*, if $\dim(C) = \dim \operatorname{aff}(C) = k$ with $k \in \{1, \dots, n\}$.

C itself is a face of C!

Extreme Points and Faces of a Triangle



Example 3.23.

(i) $C = B_1(0) \subset \mathbb{R}^n$.
Then, $\mathbf{ext}(C) = \partial C$ and C has infinitely many extreme points.

(ii) $C = \mathbf{int}(B_1(0)) \subset \mathbb{R}^n$.
Then, $\mathbf{ext}(C) = \emptyset$.

(iii) $C = \mathbb{R}^n$ has $\mathbf{ext}(C) = \emptyset$ and $C = \{x \in \mathbb{R}^n \mid x \geq 0_{\mathbb{R}^n}\}$ has $\mathbf{ext}(C) = \{0_{\mathbb{R}^n}\}$.

(iv) $C = [-1,1]^2 \subset \mathbb{R}^2$.
Then, $\mathbf{ext}(C) = \{\binom{1}{1}, \binom{-1}{1}, \binom{1}{-1}, \binom{-1}{-1}\}$ are the extreme points (vertices = 0-dimensional faces) and $\mathbf{co}\{\binom{1}{1}, \binom{-1}{1}\}, \mathbf{co}\{\binom{-1}{1}, \binom{-1}{-1}\}, \mathbf{co}\{\binom{-1}{-1}, \binom{1}{-1}\}, \mathbf{co}\{\binom{1}{-1}, \binom{1}{1}\}$ are extreme sets (edges = 1-dim. face, in $\mathbb{R}^2$ edges are the facets).
C is the 2-dim. face.

Remark 3.24. *A definition for an extreme point $e \in M$ equivalent to Definition 3.21 could be given as follows:*

(i) *For all* $x, y \in M$ *and all* $\lambda \in (0,1)$ *with* $\lambda x + (1-\lambda)y = e$ *always follows that* $x = y = e$.
(ii) *There does not exist points* $x, y \in M \setminus \{e\}$ *with* $e \in \mathbf{co}\{x, y\}$.
(iii) $C \setminus \{e\}$ *is convex.*
(iv) *For all* $x, y \in M$ *with* $\frac{1}{2}(x+y) = e$ *always follows that* $x = y = e$.
(v) *There does not exist points* $x, y \in M \setminus \{e\}$ *with* $e = \frac{1}{2}(x+y)$.
(vi) *For all convex combinations* $e = \sum_{i=1}^{k} \lambda_i x^i$ *with* $x^i \in M$, $\lambda_i > 0$ *for all* $i = 1, \dots, k$ *follows that* $x^i = e$, $i = 1, \dots, k$.

In cases (i)–(v), it is sufficient to prove this for two different points $x \neq y$, *but (i) could not be satisfied for all* $\lambda \in [0,1]$*!*

An extreme point lies always on the boundary of the set.

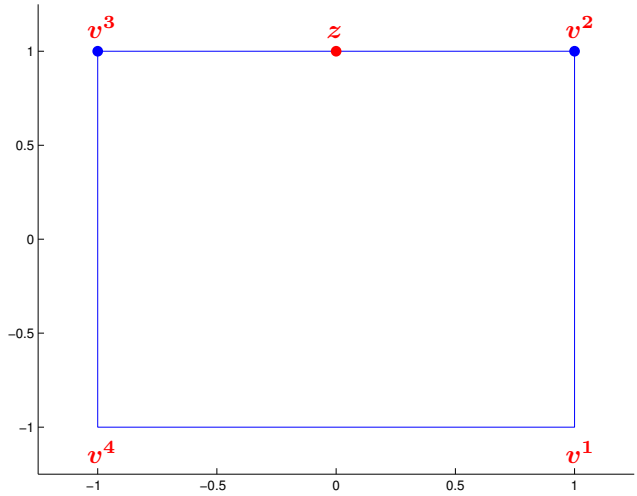
Proposition 3.25. *Let* $M \subset \mathbb{R}^n$ *be an arbitrary set. Then,* $\mathbf{ext}(M) \subset \partial M$.

Proof. Let $x \in \mathbf{ext}(M)$ be an extreme point and assume that it lies in the interior of M. Then, there exists a ball $B_\varepsilon(x)$ around x contained fully in M. For any $\eta \in S_{n-1}$ follows:

$$x + \varepsilon\eta, x - \varepsilon\eta \in B_\varepsilon(x) \subset M,$$
$$\frac{1}{2}(x+\varepsilon\eta) + \frac{1}{2}(x-\varepsilon\eta) = x.$$

Hence, x could not be an extreme point which is a contradiction. □

Not every Boundary Point is Extreme



– convex set C
• extreme points v^i, $i = 1, \dots, 4$
• boundary point z with $z = \frac{1}{2}(v^2 + v^3)$ is not extreme

Lemma 3.27. *Let* $M \subset \mathbb{R}^n$. *Then,* $\mathbf{ext}(\mathbf{co}(M)) \subset M$.

Proof. Let $z \in \mathbf{ext}(\mathbf{co}(M))$. Then, $z \in \mathbf{co}(M)$, i.e. $z = \sum_{i=1}^{k} \lambda_i x^i$ with $x^i \in M$, $i = 1, \dots, k$. W.l.o.g. we may assume that all $\lambda_i > 0$, $i = 1, \dots, k$ (otherwise reduce the sum to those, clearly there must be some $\lambda_i > 0$).
If $k = 1$, then $z = x^1 \in M$.

Otherwise $k \geq 2$ and $\lambda_1 \in (0,1)$. Then,

$$z = \lambda_1 x^1 + \sum_{i=2}^{k} \lambda_i x^i = \lambda_1 x^1 + (1-\lambda_1) \sum_{i=2}^{k} \frac{\lambda_i}{1-\lambda_1} x^i.$$

Since $\frac{\lambda_i}{1-\lambda_1} > 0$ for all $i = 2, \dots, k$ and

$$\sum_{i=2}^{k} \frac{\lambda_i}{1-\lambda_1} = \frac{1}{1-\lambda_1} \sum_{i=2}^{k} \lambda_i = \frac{1}{1-\lambda_1}(1-\lambda_1) = 1,$$

z is written as a convex combination of x^1 and $\sum_{i=2}^{k} \frac{\lambda_i}{1-\lambda_1} x^i$ which are both elements of $\mathbf{co}(M)$. The extremality of z in $\mathbf{co}(M)$ yields $z = x^1$ and hence $z \in M$. □

Corollary 3.28. *Let $P = \mathbf{co}\{p^i \mid i = 1, \ldots, r\} \subset \mathbb{R}^n$ with $r \geq 0$. Then, all extreme points (vertices) are contained in the set $\{p^i \mid i = 1, \ldots, r\}$.*

Proof. Apply Lemma 3.27 to $M := \{p^i \mid i = 1, \ldots, r\}$. □

Proposition 3.30. *Let $M \subset \mathbb{R}^n$ be an arbitrary set and $\widetilde{E} \subset E \subset M$.*
If $\widetilde{E}$ is extreme in E and E is extreme in M, then $\widetilde{E}$ is also extreme in M.

Proposition 3.31. *If $M \subset \mathbb{R}^n$ is compact and nonempty, there there exists an extreme point, i.e. $\mathbf{ext}(M) \neq \emptyset$.*

Proof. Since M is compact and nonempty, there exists a maximum $\widehat{x}$ of the continuous function $x \mapsto \|x\|_2$. We will show that $\widehat{x}$ is an extreme point. To prove this, consider two points $x, y \in M$ with $\frac{1}{2}(x+y) = \widehat{x}$ (this is sufficient by Remark 3.24(iv)). The equation

$$\frac{1}{2}\|x+y\|^2 + \frac{1}{2}\|x-y\|^2 = \|x\|^2 + \|y\|^2 \tag{7}$$

in the parallelogram $\mathbf{co}\{x, y, x+y, x-y\}$ shows that

$$\|\widehat{x}\|^2 = \|\frac{1}{2}(x+y)\|^2 \underbrace{\leq}_{(*)} \frac{1}{2}(\|x\|^2 + \|y\|^2) \underbrace{\leq}_{(**)} \frac{1}{2}(\|\widehat{x}\|^2 + \|\widehat{x}\|^2) = \|\widehat{x}\|^2$$

(*) and (**) must be equalities to avoid a contradiction. Multiply (7) by $\frac{1}{2}$, resolve for $\frac{1}{2}\|x+y\|^2$ and insert this in equality (*) gives us $x = y$, hence $\widehat{x} = x = y$. □

Remark 3.32. *Let $C \subset \mathbb{R}^n$ be convex, closed. Then, $\mathbf{ext}(C)$ is not closed in general for $\dim C = \dim \mathbf{aff}(C) > 2$.*
Consider (cf. [Web94, remarks after Theorem 2.6.16] or [Lei80, Lei85])

$$\begin{aligned}
C^1 &:= \{x \in \mathbb{R}^3 \mid x_1^2 + x_2^2 \leq 1,\ x_3 = 0\},\\
C^2 &:= \{x \in \mathbb{R}^3 \mid x_1 = 1,\ x_2 = 0,\ x_3 \in [-1, 1]\},\\
C &:= \mathbf{co}(C^1 \cup C^2),\\
B^1 &:= \{x \in \mathbb{R}^3 \mid x_1^2 + x_2^2 = 1,\ x_3 = 0\} \quad (\textit{relative boundary of } C^2).
\end{aligned}$$

Then, $C \in \mathcal{C}(\mathbb{R}^3)$ and

$$\mathbf{ext}(C) = (B^1 \setminus \{(1,0,0)^\top\}) \cup \{(1,0,1)^\top, (1,0,-1)^\top\}$$

which is not closed ($(1,0,0)^\top \notin \mathbf{ext}(C)$).
Nevertheless, $\mathbf{ext}(C)$ is closed for $\dim C = 2$, cf. [Web94, exercise 2.6, 2.] or [Lei80, Lei85].

3.1.3 Separation Theorems

Notation 3.33. *Let $\eta \in \mathbb{R}^n$, $\eta \neq 0_{\mathbb{R}^n}$ and $\alpha \in \mathbb{R}$. Then,*

$$H := H(\eta, \alpha) \ := \{x \in \mathbb{R}^n \mid \langle \eta, x \rangle = \alpha\}$$

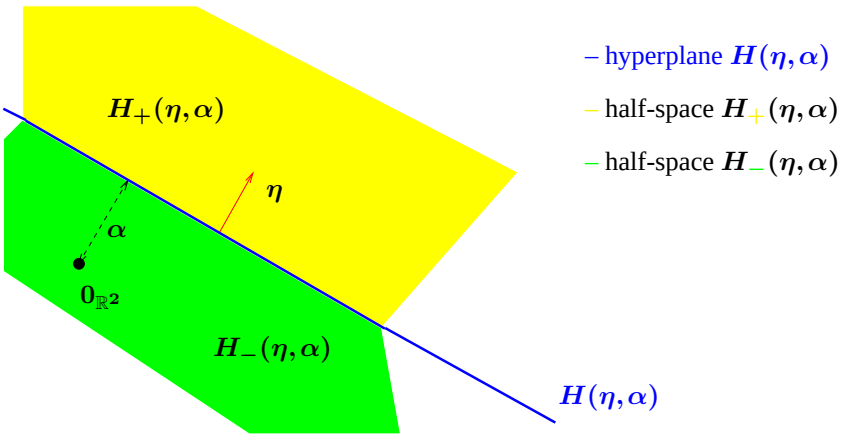
denotes the hyperplane with the normal η and offset α.

$$H_+ := H_+(\eta, \alpha) := \{x \in \mathbb{R}^n \mid \langle \eta, x \rangle > \alpha\},$$
$$H_- := H_-(\eta, \alpha) := \{x \in \mathbb{R}^n \mid \langle \eta, x \rangle < \alpha\}$$

denote the corresponding (open) half-spaces, $\overline{H}_+$ and $\overline{H}_-$ the (closed) half-spaces.
If $x^0 \in H$, then clearly $\alpha = \langle \eta, x^0 \rangle$.

Clearly, $H(\eta, \alpha) = H(\frac{1}{\|\eta\|}\eta, \frac{\alpha}{\|\eta\|})$ so that $\eta \in S_{n-1}$ could always be assumed.

Partition of $\mathbb{R}^n$ in 3 Parts by Hyperplanes



Remark 3.34. *Let $H(\eta, \alpha) \subset \mathbb{R}^n$ be a hyperplane. Then,*

$$\mathbb{R}^n = H_-(\eta, \alpha) \cup H(\eta, \alpha) \cup H_+(\eta, \alpha) = \overline{H}_-(\eta, \alpha) \cup H_+(\eta, \alpha)$$

is a disjoint union of half-spaces and the hyperplane resp. closed and open half-spaces.

Definition 3.35. Let $C, D \subset \mathbb{R}^n$ be convex sets and H be a hyperplane.

(i) H *separates* C and D, if $C \subset \overline{H}_+$, $D \subset \overline{H}_-$ or vice versa, if $C \subset \overline{H}_-$, $D \subset \overline{H}_+$.

(ii) H *separates* C and D *properly*, if H separates C and D and additionally, $C \not\subset H$ or $D \not\subset H$.

(iii) H *separates* C and D *strictly*, if $C \subset H_+$, $D \subset H_-$ or vice versa.

(iv) H *separates* C and D *strongly*, if there exists $\varepsilon > 0$ with $C + \varepsilon B_1(0) \subset \overline{H}_+$, $D + \varepsilon B_1(0) \subset \overline{H}_-$ or vice versa.

Remark 3.36. *Let $C, D \subset \mathbb{R}^n$ be convex and $H(\eta, \alpha)$ be a hyperplane with $C \subset H_-(\eta, \alpha)$. Then, $C \subset H_+(-\eta, -\alpha)$.*
If $H(\eta, \alpha)$ separates C and D, then $H(-\eta, -\alpha)$ is also a separating hyperplane.
The same is valid with $C \subset H(\eta, \alpha)$ and $C \subset H(-\eta, -\alpha)$ resp. $C \subset \overline{H}_-(\eta, \alpha)$ and $C \subset \overline{H}_+(-\eta, -\alpha)$.

Sketch of Proof. If $c \in C$, then $\langle \eta, c \rangle \leq \alpha$. Hence, $\langle -\eta, c \rangle \geq -\alpha$, i.e. $c \in H_+(-\eta, -\alpha)$. □

Proposition 3.37. *Let $C, D \subset \mathbb{R}^n$ be convex, nonempty sets and $H(\eta, \alpha)$ be a hyperplane.*

(i) $\boldsymbol{H}$ *separates* $\boldsymbol{C}$ *and* $\boldsymbol{D}$*, if and only if* $\sup_{y\in D}\langle\eta,y\rangle\leq\alpha\leq\inf_{x\in C}\langle\eta,x\rangle$ *or vice versa, i.e.* $\sup_{x\in C}\langle\eta,x\rangle\leq\alpha\leq\inf_{y\in D}\langle\eta,y\rangle$.

(ii) $\boldsymbol{H}$ *separates* $\boldsymbol{C}$ *and* $\boldsymbol{D}$ *properly, if and only if* $\sup_{y\in D}\langle\eta,y\rangle\leq\alpha\leq\inf_{x\in C}\langle\eta,x\rangle$ *and* $\inf_{y\in D}\langle\eta,y\rangle<\sup_{x\in C}\langle\eta,x\rangle$ *or the roles of* $\boldsymbol{C}$ *and* $\boldsymbol{D}$ *are interchanged.*

(iii) $\boldsymbol{H}$ *separates* $\boldsymbol{C}$ *and* $\boldsymbol{D}$ *strictly, if and only if* $\langle\eta,y\rangle<\alpha<\langle\eta,x\rangle$ *for all* $x\in C$, $y\in D$.

(iv) $\boldsymbol{H}$ *separates* $\boldsymbol{C}$ *and* $\boldsymbol{D}$ *strongly, if and only if* $\sup_{y\in D}\langle\eta,y\rangle\leq\alpha-\varepsilon<\alpha<\alpha+\varepsilon\leq\inf_{x\in C}\langle\eta,x\rangle$ *or vice versa.*

Proof. (i) "⇒": Since Remark 3.36 is valid, we study only the first case "$D\subset\overline{H}_-(\eta,\alpha)$, $C\subset\overline{H}_+(\eta,\alpha)$".

$$D\subset\overline{H}_-(\eta,\alpha)\Rightarrow\sup_{y\in D}\langle\eta,y\rangle\leq\sup_{y\in\overline{H}_-(\eta,\alpha)}\langle\eta,y\rangle=\alpha,$$
$$C\subset\overline{H}_+(\eta,\alpha)\Rightarrow\inf_{x\in C}\langle\eta,x\rangle\geq\inf_{x\in\overline{H}_+(\eta,\alpha)}\langle\eta,x\rangle=\alpha$$

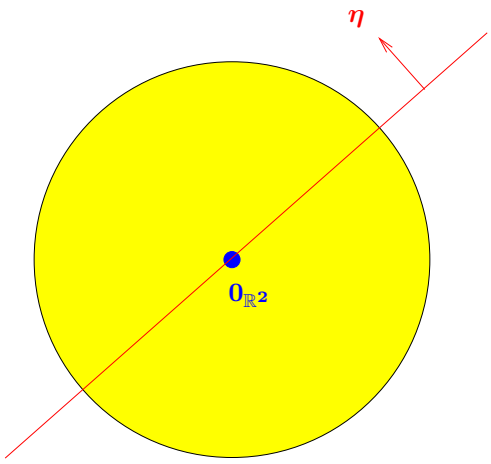
which proves the inequality.
"⇐": If the inequalities are fulfilled, then

$$\langle\eta,y\rangle\leq\sup_{y\in D}\langle\eta,y\rangle\leq\alpha,\quad\langle\eta,x\rangle\geq\inf_{x\in C}\langle\eta,x\rangle\geq\alpha$$

for all $x\in C$, $y\in D$.
By the definition of the half-spaces, we have $D\subset\overline{H}_-(\eta,\alpha)$ and $C\subset\overline{H}_+(\eta,\alpha)$.
The other proofs are similar. □

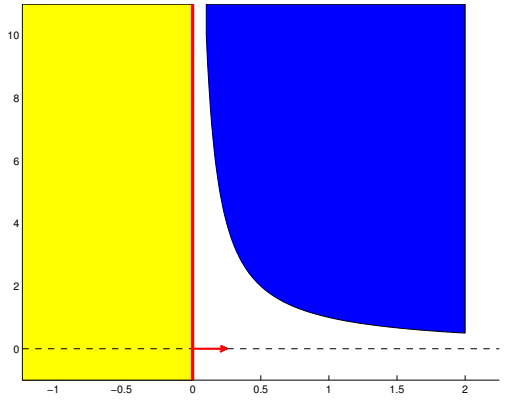
Remark 3.38. *strong separation* ⇒ *strict separation* ⇒ *proper separation* ⇒ *separation for any two convex, nonempty subsets of* $\mathbb{R}^n$

Example (i): non-separable sets



– $D=B_1(0)$
• $C=\{0_{\mathbb{R}^2}\}$
– non-separating hyperplane $H(\eta,0)$

Example (ii): properly, non-strictly separable sets



– $D = \{\binom{x}{y} \in \mathbb{R}^2 \,|\, x \leq 0\}$

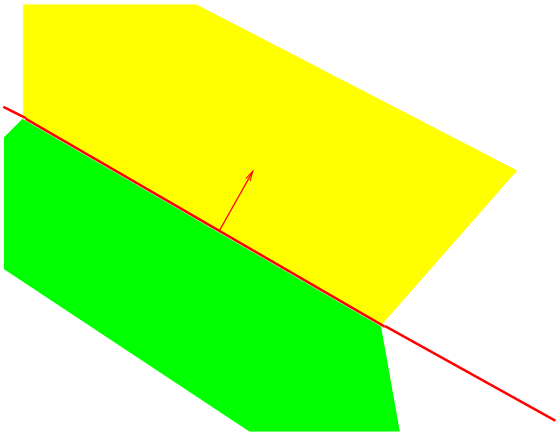
– $C = \{\binom{x}{y} \in \,|\, x > 0, y \geq \frac{1}{x}\}$

– separating hyperplane

$H(\eta, 0)$ with $\eta = \binom{1}{0}$

(non-strictly separation)

Example (iii): strictly, non-strongly separated half-spaces



– $D = H_-(\eta, \alpha)$

– $C = H_+(\eta, \alpha)$

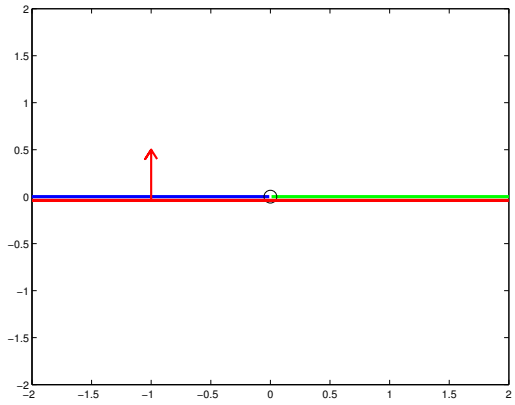
– strictly separating

hyperplane $H(\eta, \alpha)$

for C and D

$H(\eta, \alpha)$ also separates $H_+(\eta, \alpha)$
and $H(\eta, \alpha)$ properly
and even (non-properly)
$H(\eta, \alpha)$ and $H(\eta, \alpha)$

Example (iv): non-properly separation



– $D = \{\binom{x}{0} \in \mathbb{R}^2 \,|\, x < 0\}$

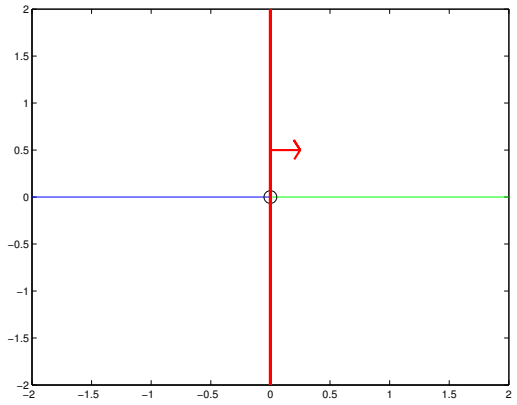
– $C = \{\binom{x}{0} \in \,|\, x > 0\}$

– separating hyperplane

$H(\eta, 0)$ with $\eta = \binom{0}{1}$

(not a clever choice)

Example (iv): better choice of hyperplane → strictly separation

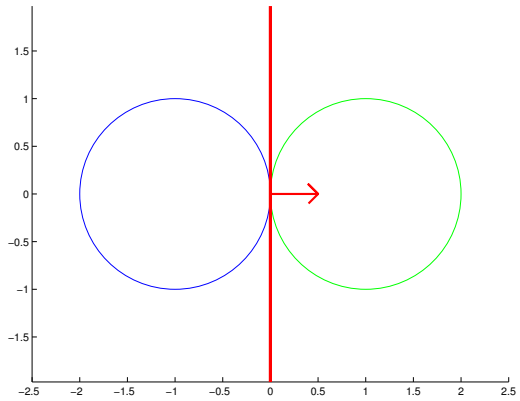


– $D = \{\binom{x}{0} \in \mathbb{R}^2 \,|\, x < 0\}$

– $C = \{\binom{x}{0} \in \,|\, x > 0\}$

– strictly separating hyperplane $H(\eta, 0)$ with $\eta = \binom{1}{0}$ (better choice)

Example (v): properly, non-strictly separable sets



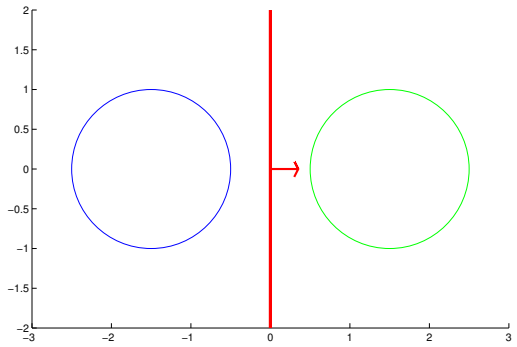
– $D = B_1(\binom{-1}{0})$

– $C = B_1(\binom{1}{0})$

– properly, but non-strictly separating hyperplane $H(\eta, 0)$ with $\eta = \binom{1}{0}$

$H(\eta, 0)$ separates $\mathbf{int}(C)$ and $\mathbf{int}(C)$ strictly, but not strongly

Example (vi): strongly separable sets



– $D = B_1(\binom{-1.5}{0})$

– $C = B_1(\binom{1.5}{0})$

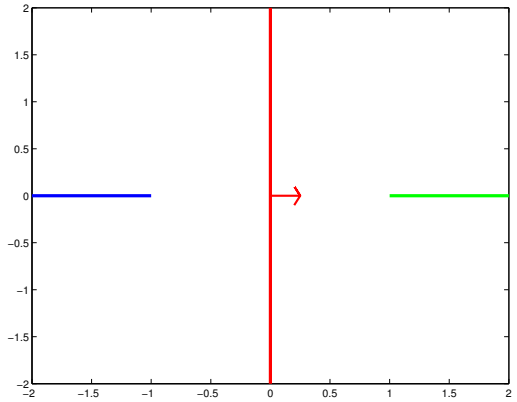
– strongly

separating hyperplane

$H(\eta, 0)$ with $\eta = \binom{1}{0}$

$H(\eta, 0)$ separates $\mathbf{int}(C)$ and $\mathbf{int}(C)$ strongly

Example (vii): strongly separable sets



– $D = \{\binom{x}{0} \in \mathbb{R}^2 \,|\, x < -1\}$

– $C = \{\binom{x}{0} \in \, | \, x > 1\}$

– strongly separating

hyperplane $H(\eta, 0)$

with $\eta = \binom{1}{0}$

Lemma 3.40. *Let $C, D \subset \mathbb{R}^n$ be convex, nonempty. Then, $H(\eta, \alpha)$ separates C and D in one of the four ways, if and only if this hyperplane separates $C - D$ and $0_{\mathbb{R}^n}$ in the same way.*

Proof. Let us study only strong separation. Then, Proposition 3.37 shows that for any $x \in C$, $y \in D$ yields

$$\langle \eta, y \rangle \leq \sup_{y \in D} \langle \eta, y \rangle \leq \alpha - \varepsilon \quad \text{and} \quad \alpha + \varepsilon \leq \inf_{x \in C} \langle \eta, x \rangle \leq \langle \eta, x \rangle.$$

Using the two equations

$$\alpha + \varepsilon \leq \langle \eta, x \rangle \quad \text{and} \quad \langle \eta, y \rangle \leq \alpha - \varepsilon,$$
$$\langle \eta, x - y \rangle = \langle \eta, x \rangle - \langle \eta, y \rangle \geq (\alpha + \varepsilon) - (\alpha - \varepsilon) = 2\varepsilon > \varepsilon > 0$$

yields finally $\inf_{z \in C - D} \langle \eta, z \rangle \geq 2\varepsilon > \varepsilon \geq \sup_{z = 0_{\mathbb{R}^n}} \langle \eta, z \rangle = 0$.

Taking $\varepsilon \searrow 0$, we have the result for the separation only. For proper and strict separation the reasoning is similar and evident. □

Hence, the separation of two sets can be reduced to the separation of one (convex) set and a point.

Lemma 3.41. *Let $C, D \subset \mathbb{R}^n$ be convex, $H(\eta, \alpha)$ be a separating hyperplane and $x^0 \in H$. Then, $H(\eta, \alpha - \langle \eta, x^0 \rangle)$ separates $C - x^0$ and $D - x^0$.*

Proposition 3.46 (Separation). *Let $C, D \subset \mathbb{R}^n$ be convex and disjoint and let C have nonempty interior. Then, C and D could be separated by a hyperplane H.*

Proposition 3.47 (strong separation). *Let $C, D \subset \mathbb{R}^n$ be convex, disjoint sets and let C be closed and D compact. Then, there exists a hyperplane which separates C and D strongly.*

Proof. In Subsection 3.2 Propositions 3.86(iv), 3.106(v) and Remark 3.74 show the convexity of $D - C$. The closedness of this set follows from Proposition 3.106(iii) and Lemma 3.84. From the assumption follows $0_{\mathbb{R}^n} \notin D - C$ (otherwise, $C \cap D \neq \emptyset$).
We will prove that $\operatorname{dist}(0_{\mathbb{R}^n}, D - C) =: \delta > 0$. Assume the contrary and consider $(x^m)_m \subset D - C$ with

$$\|x^m\| = \|0_{\mathbb{R}^n} - x^m\| \underset{m\to\infty}{\searrow} \delta = 0.$$

Hence, $(x^m)_m$ is bounded and contains a convergent subsequence $(x^{m_k})_k$ with $x^{m_k} \xrightarrow[m\to\infty]{} x$. The continuity of $\|\cdot\|$ shows $\|x\| = \lim_{k\to\infty} \|x^{m_k}\| = 0$. The closedness of $D - C$ shows that $x = 0_{\mathbb{R}^n} \in D - C$. But C and D are disjoint which shows that δ must be positive.

$V := B_{\frac{\delta}{2}}(0)$ is an convex neighborhood of the origin with $V \cap (D - C) = \emptyset$. Since $0_{\mathbb{R}^n}$ is an inner point of V, Proposition 3.46 shows the existence of a separating hyperplane $H(\eta, \alpha)$ for V and $D - C$.
$\eta \neq 0_{\mathbb{R}^n}$, so w.l.o.g. $\eta \in S_{n-1}$. Since V has nonempty interior, we set $\varepsilon := \frac{\delta}{2} > 0$. Then,

$$\begin{aligned}\sup_{x\in\{0_{\mathbb{R}^n}\}} \langle \eta, x\rangle = 0 < \varepsilon = \varepsilon\langle \eta, \eta\rangle = \langle \eta, \varepsilon\eta\rangle \\ \leq \sup_{x\in V} \langle \eta, x\rangle \leq \alpha \leq \inf_{z\in D-C} \langle \eta, z\rangle\end{aligned}$$

which shows that $H(\eta, \frac{\varepsilon}{2})$ separates $0_{\mathbb{R}^n}$ and $D - C$ strongly by Proposition 3.37(iv). Lemma 3.40 shows finally the assertion. □

3.1.4 Support Function, Supporting Faces, Exposed Sets

Proposition 3.48. *Let $C \subset \mathbb{R}^n$ be convex, closed and $x \notin C$. Then, there exists $\eta \in S_{n-1}$ with*

$$\sup_{c \in C} \langle \eta, c \rangle < \langle \eta, x \rangle.$$

Proof. Set $D := \{x\}$, then Proposition 3.47 shows the existence of a hyperplane $H(\eta, \alpha)$ which separates C and $\{x\}$ strongly, hence strictly. □

Definition 3.49. Let $C \subset \mathbb{R}^n$, $\eta \in \mathbb{R}^n$. The function

$$\begin{aligned} \delta^\star(\cdot, C) : \mathbb{R}^n &\to \mathbb{R} \cup \{\pm\infty\} \\ \eta &\mapsto \delta^\star(\eta, C) := \sup_{c \in C} \langle \eta, c \rangle \end{aligned}$$

is called the *support function* of C.

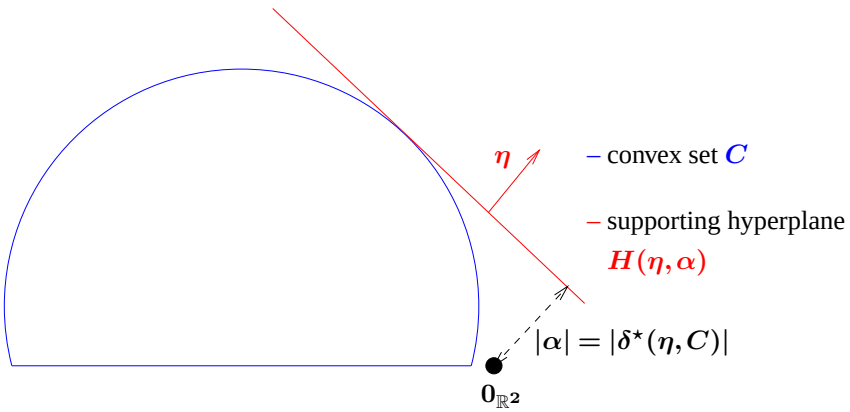
Remark 3.50. *Let $C \subset \mathbb{R}^n$, $\eta \in \mathbb{R}^n$.*

- *If $C \neq \emptyset$, then $\delta^\star(\eta, C) \in \mathbb{R} \cup \{+\infty\}$.*
- *If $0_{\mathbb{R}^n} \in C$, then $\delta^\star(\eta, C) \geq 0$.*
- *If $\delta^\star(\eta, C) < 0$ for all $\eta \in \mathbb{R}^n$, then $\delta^\star(\eta, C) = -\infty$ for all $\eta \in \mathbb{R}^n$ and $C = \emptyset$.*
- *There exists $C \neq \emptyset$ and some $\eta \in \mathbb{R}^n$ with $\delta^\star(\eta, C) < 0$, e.g. from $C = \mathrm{co}\{-e^1 + e^2, e^1 + e^2\}$ (e^k the k-th unit vector of $\mathbb{R}^n$) and $\eta = -e^2$ follows that*

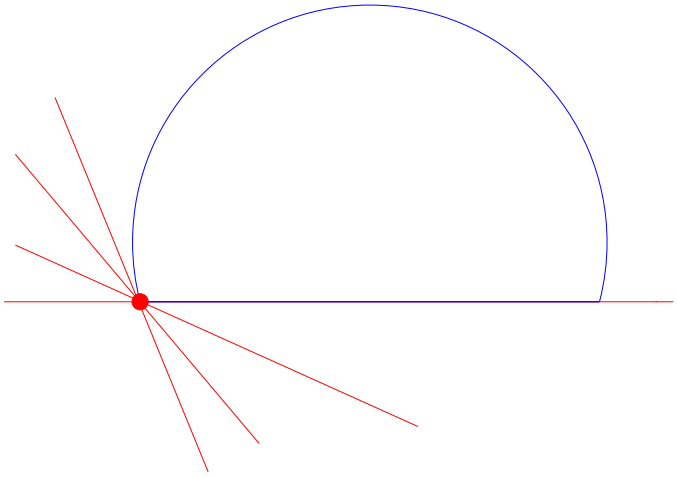
$$\delta^\star(\eta, C) = \max\{\langle \eta, -e^1 + e^2 \rangle, \langle \eta, e^1 + e^2 \rangle\} = -\langle e^2, e^2 \rangle = -1.$$

The following picture shows that the support function in direction η is the (signed) distance of the origin to the supporting hyperplane H, if $\|\eta\| = 1$. $\delta^\star(\eta, C) \cdot \eta \in H$, but in general $\delta^\star(\eta, C) \cdot \eta \notin C$.

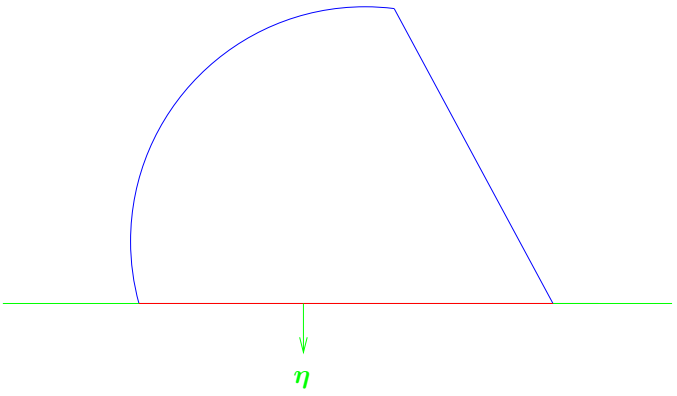
support function in a direction $\eta \in S_{n-1}$ with unique touching point



unique touching point in various directions



non-unique touching points in one direction



Lemma 3.51. *Let $M \subset \mathbb{R}^n$ be a set, then for all $l \in \mathbb{R}^n$*

$$\delta^\star(l, M) = \delta^\star(l, \overline{M}),$$
$$\delta^\star(l, M) = \delta^\star(l, \mathrm{co}(M))$$

Proof. For $M = \emptyset$ nothing is to prove.
Otherwise, let $x \in \overline{M}$ and $(x^k)_k \subset M$ be a sequence with

$$x^k \xrightarrow[k\to\infty]{} x.$$

The continuity of the scalar product yields

$$\langle l, x^k \rangle \xrightarrow[k\to\infty]{} \langle l, x \rangle.$$

On the other hand,

$$\langle l, x^k \rangle \le \delta^\star(l, M) \quad \text{and} \quad \langle l, x \rangle \le \delta^\star(l, M),$$
$$\delta^\star(l, \overline{M}) \le \delta^\star(l, M).$$

Obviously, $\delta^\star(l,M) \le \delta^\star(l,\overline{M})$.
Let $z = \sum_{i=1}^{n+1} \alpha_i x^i \in \mathbf{co}(M)$. Then,

$$\langle l,z\rangle = \sum_{i=1}^{n+1} \underbrace{\alpha_i}_{\ge 0} \langle l,x^i\rangle \le \underbrace{\sum_{i=1}^{n+1} \alpha_i}_{=1} \delta^\star(l,M) = \delta^\star(l,M)$$

which proves $\delta^\star(l,\mathbf{co}(M)) \le \delta^\star(l,M)$. The converse inequality is obvious. □

Lemma 3.52. *Let $C \subset \mathbb{R}^n$ be convex, nonempty and $x \in \partial C$. Then, there exists a hyperplane $H(\eta,\alpha)$ which supports C in x, i.e.*

$$\begin{aligned} \langle \eta, x\rangle &\le \alpha \quad \text{for all } c \in C,\\ \langle \eta, x\rangle &= \alpha. \end{aligned}$$

Shortly, this means that $C \subset \overline{H}_-(\eta,\alpha)$ and $x \in H(\eta,\alpha)$.

Proof. Since $\partial C = \partial(\mathbb{R}^n \setminus \overline{C})$, there exists a sequence $(x^k)_k \subset \mathbb{R}^n \setminus \overline{C}$ with $x^k \xrightarrow[k\to\infty]{} x$.
Proposition 3.48 states the existence of separating hyperplanes $H(\eta^k,\alpha^k)$ for $\overline{C}$ and $\{x^k\}$ strictly, i.e.

$$\sup_{c\in\overline{C}} \langle \eta^k, c\rangle < \alpha_k < \langle \eta^k, x^k\rangle.$$

Since $(\eta^k)_k \subset S_{n-1}$ is bounded, it contains a subsequence $(\eta^{k_\nu})_\nu$ converging to $\eta \in S_{n-1}$ yielding

$$\begin{aligned} \langle \eta^{k_\nu}, c\rangle &< \langle \eta^{k_\nu}, x^{k_\nu}\rangle \quad \text{for all } \nu \in \mathbb{N},\\ \langle \eta, c\rangle &\le \langle \eta, x\rangle \end{aligned}$$

for all $c \in \overline{C}$. Altogether we reached to prove

$$\sup_{c\in C} \langle \eta, c\rangle = \sup_{c\in\overline{C}} \langle \eta, c\rangle \le \langle \eta, x\rangle =: \alpha$$

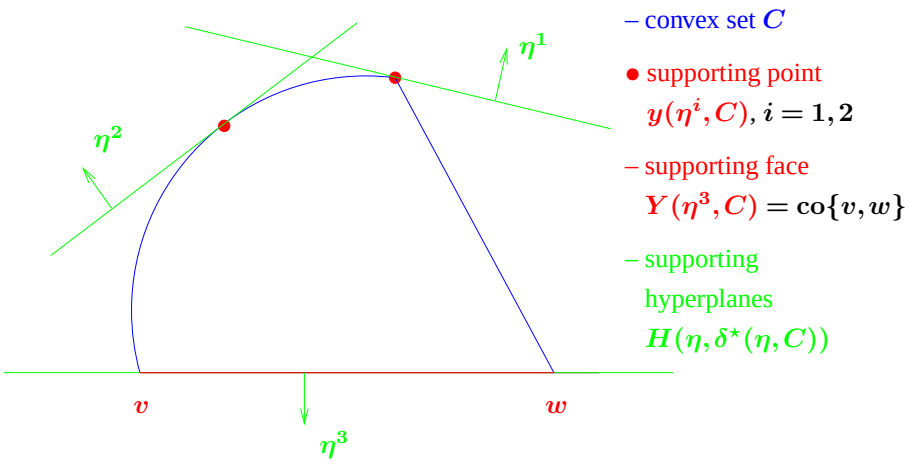
so that $H(\eta,\alpha)$ is the wanted separating hyperplane. □

Definition 3.53. Let $C \subset \mathbb{R}^n$ be convex, nonempty. Each hyperplane $H(\eta,\alpha)$ for a boundary point $x \in \partial C \cap C$ in Lemma 3.52 is called *supporting hyperplane* of C in direction η.
x is called *supporting point* of C in this direction, if there exists such a hyperplane and additionally $x \in C$.
Formally,

$$Y(\eta,C) := H(\eta,\alpha) \cap C = \{c \in C \mid \langle \eta, c\rangle = \delta^\star(\eta,C)\}$$

is called the *set of supporting points*/*supporting face* of C in direction η.
We denote $Y(\eta,C) = \{y(\eta,C)\}$ in the unique case.
$Y(\eta,C)$ is also called *exposed face* of C, in the unique case $y(\eta,C)$ is called *exposed point*.
The set of all exposed points is denoted by $\mathbf{exp}(C)$.

Exposed Faces



The following fact could also be justified by Proposition 3.59.

Proposition 3.54. *An exposed point x of a convex, nonempty set C is a boundary point of C.*

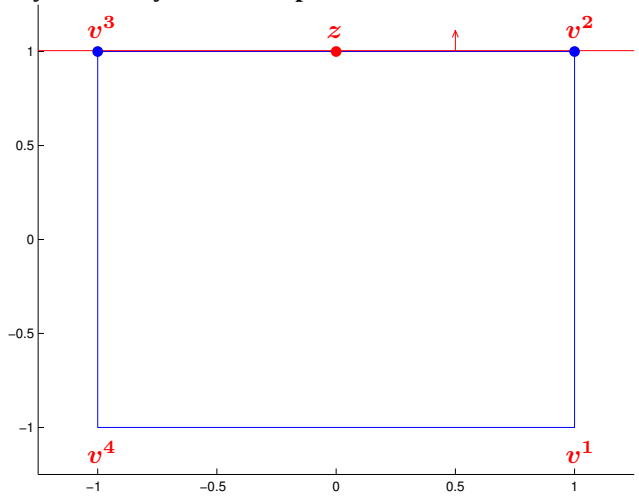
Proof. Let $H(\eta, \alpha)$ be a supporting hyperplane with $\eta \in S_{n-1}$, i.e. $C \subset \overline{H}_-(\eta, \alpha)$, $x \in H(\eta, \alpha)$. Assume that $x \in \mathrm{int}(C)$ and let $\delta > 0$ such that $B_\delta(x) \subset C$. Then, $x + \delta\eta \in B_\delta(x) \subset C$ and

$$\langle \eta, x + \delta\eta \rangle = \langle \eta, x \rangle + \delta \|\eta\|^2 = \alpha + \delta > \alpha,$$

but this is a contradiction to $x + \delta\eta \in C \subset \overline{H}_-(\eta, \alpha)$. □

Corollary 3.55. *Under the assumptions of Lemma 3.52 there exists an element of a supporting face (namely $x \in Y(\eta, C)$), if there exists a boundary point of C which belongs to C.*
Especially, this condition is fulfilled, if C is closed.
A boundary point is an element of a supporting face $Y(\eta, C)$, but it need not be exposed. It is only exposed, if $Y(\eta, C)$ consists of only one element.

Not every Boundary Point is Exposed



Proposition 3.59. *An exposed face F of a convex, nonempty set C is an (extreme) face, but not vice versa in general.*

Compare Proposition 3.59 for **0**-dimensional faces (namely, the exposed points) with Theorem 3.69.

Proposition 3.60. *Let $C \subset \mathbb{R}^n$ be convex and $F \subset C$ be an exposed face. If $z \in \mathrm{ext}(F)$, then $z \in \mathrm{ext}(C)$.*

3.1.5 Representation of Convex Sets

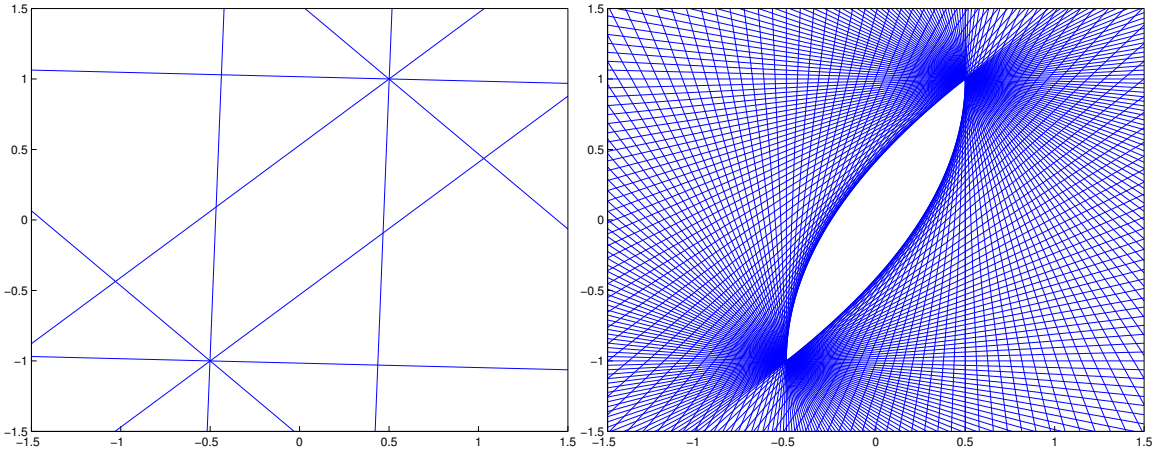
Proposition 3.61. *Let $C \subset \mathbb{R}^n$ be closed, convex. Then,*

$$C = \bigcap_{\eta \in S_{n-1}} \{x \in \mathbb{R}^n \,|\, \langle \eta, x \rangle \leq \delta^\star(\eta, C)\}.$$

If you know the support function of a closed, convex set for every normed direction, you can recover the set itself by Proposition 3.61.

representation of a convex sets via support(ing) functions/hyperplanes

intersection based on 8 and 200 supporting hyperplanes



Proof. "$\subset$": Let $c \in C$ and $\eta \in S_{n-1}$. Then,

$$\langle \eta, c \rangle \leq \sup_{c \in C} \langle \eta, c \rangle \leq \delta^\star(\eta, C).$$

"$\supset$": If $x \in \mathbb{R}^n \setminus C$, then Proposition 3.48 shows the existence of $H(\eta, \alpha)$ (w.l.o.g. $\eta \in S_{n-1}$) which separates C and $\{x\}$, i.e.

$$\delta^\star(\eta, C) = \sup_{c \in C} \langle \eta, c \rangle < \langle \eta, x \rangle.$$

Clearly, x lies in the complement of the right-hand side $D := \bigcap_{\eta \in S_{n-1}} \{x \in \mathbb{R}^n \,|\, \langle \eta, x \rangle \leq \delta^\star(\eta, C)\}$ and therefore, $C \supset D$. □

Remark 3.62. *Propositions 3.61 and 3.16(iii) state that for $S \in \mathcal{K}(\mathbb{R}^n)$ the representation*

$$\operatorname{co}(S) = \bigcap_{\substack{D \supset S \\ D \text{ convex}}} D$$

of the convex hull could be simplified as infinite intersection of halfspaces:

$$\operatorname{co}(S) = \bigcap_{\substack{\eta \in S_{n-1} \\ \alpha \in \mathbb{R}: \\ \overline{H}_-(\eta, \alpha) \supset S}} \overline{H}_-(\eta, \alpha)$$

α can be specified as $\delta^\star(\eta, \operatorname{co}(S))$.

Proposition 3.63. *Let $C \subset \mathbb{R}^n$ be closed, convex, nonempty, $x \in \mathbb{R}^n$ and $U \subset \mathbb{R}^n$. Then,*

$$\begin{aligned}
x \in \overline{\mathrm{co}}(U) &\iff \forall\, \eta \in S_{n-1}: && \langle \eta, x\rangle \le \delta^\star(\eta, U),\\
x \in C &\iff \forall\, \eta \in S_{n-1}: && \langle \eta, x\rangle \le \delta^\star(\eta, C),\\
x \in \mathrm{int}(C) &\iff \forall\, \eta \in S_{n-1}: && \langle \eta, x\rangle < \delta^\star(\eta, C),\\
x \in \partial C &\iff \forall\, \eta \in S_{n-1}: && \langle \eta, x\rangle \le \delta^\star(\eta, C),\\
&\quad \text{and } \exists \eta^0 \in S_{n-1} \text{ with} && \langle \eta^0, x\rangle = \delta^\star(\eta^0, C),\\
x \in \mathrm{aff}(C) &\iff \forall\, \eta \in S_{n-1} \text{ with } \delta^\star(\eta, C) = -\delta^\star(-\eta, C): && \langle \eta, x\rangle = \delta^\star(\eta, C),\\
x \in \mathrm{ri}(C) &\iff \forall\, \eta \in S_{n-1} \text{ with } \delta^\star(\eta, C) > -\delta^\star(-\eta, C): && \langle \eta, x\rangle < \delta^\star(\eta, C)
\end{aligned}$$

Hereby, $\mathrm{ri}(C)$ is the relative interior of C (the interior of C w.r.t. $\mathrm{aff}(C)$). For details on the relative interior see e.g. [HUL93, III, 2.1].

Proof. cf. [Roc72, Theorem 13.1] or [HUL93, V., 2.2. and Theorem 2.2.3], Proposition 3.61 and Lemma 3.52 □

Proposition 3.64. *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{-\infty, \infty\}$ be a positive homogeneous, convex function which is not identically $+\infty$.*
Then, the closure $\mathrm{cl}(f)$ of f defined as

$$\mathrm{cl}(f) := \inf\{\mu \,|\, (x, \mu) \in \mathrm{epi}(f)\}$$

is support function of the closed, convex set

$$C = \bigcap_{l \in \mathbb{R}^n} \{x \in \mathbb{R}^n \,|\, \langle l, x\rangle \le f(l)\}.$$

If $f(\cdot)$ is finite, then $f(\cdot)$ itself is support function of C.

Proof. cf. [Roc72, Corollary 13.2.1] If $f(\cdot)$ is finite, then the closedness of $f(\cdot)$ follows by [Roc72, Corollary 7.4.2], i.e. $\mathrm{cl}(f) = f$. □

Definition 3.65. $P \subset \mathbb{R}^n$ is called a *polyhedral set*, if there exists finitely many hyperplanes $H(\eta^i, \alpha_i)$, $i = 1, \dots, k$ with

$$P = \bigcap_{i=1,\dots,k} \overline{H}_-(\eta^i, \alpha_i).$$

Remark 3.66. *Clearly, a polyhedral set is convex by Proposition 3.5.*
Furthermore, each convex polytope $P \subset \mathbb{R}^n$ is a polyhedral set, but not vice versa (since P could be unbounded which could not happen for polytopes).

Proposition 3.67. *Let $C \subset \mathbb{R}^n$ be compact and convex. Then, C is the convex hull of its boundary, i.e. $C = \mathrm{co}(\partial C)$.*

Proof. Clearly, $C = \mathrm{int}(C) \cup \partial C$. Let $x \in \mathrm{int}(C)$. To prove that $x \in \mathrm{co}(\partial C)$ we choose $\eta \in S_{n-1}$. Then, the half-rays

$$\{x + \lambda\eta \,|\, \lambda \ge 0\} \quad \text{and} \quad \{x - \lambda\eta \,|\, \lambda \ge 0\}$$

each meet only once the boundary of C (this follows by [Lei98, Lemma 2.8] or from [Sch93, Lemma 1.1.8]). The resulting values are denoted by λ_1 resp. λ_2. The function $\varphi(\lambda) := \mathrm{dist}(x + \lambda\eta, \mathbb{R}^n \setminus \mathrm{int}(C))$ is continuous on $\mathbb{R}$ by Proposition 3.145. Since C is bounded by $B_r(x)$ with suitable $r \ge 0$, the function $\varphi(\cdot)$ must attain its maximum on $[0, r]$, hence both values exist. Clearly, $x + \lambda_1\eta, x - \lambda_2\eta \in \partial C$.

Finally, let us show that x is a convex combination of these two boundary points.
Set $\lambda := \frac{\lambda_2}{\lambda_1 + \lambda_2} \in (0, 1)$ and consider

$$\lambda(x + \lambda_1\eta) + (1 - \lambda)(x - \lambda_2\eta) = x + (\lambda\lambda_1 - (1 - \lambda)\lambda_2)\eta = x.$$

Hence, $x \in \mathrm{co}(\partial C)$. □

Theorem 3.68 (H. Minkowski resp. M. Krein/D. Milman). *Let $C \subset \mathbb{R}^n$ be compact and convex. Then, C is the convex hull of its extreme points, i.e. $C = \mathrm{co}(\mathrm{ext}(C))$.*
Krein/Milman (1940): If $C \subset X$ compact, convex and X locally convex vector space, then $C = \overline{\mathrm{co}}(\mathrm{ext}(C))$.

Proof. "$\supset$" is clear, the proof for "$\subset$" uses induction on n.
The start with $n = 1$ is easy:
$C = [a, b]$ with $a \leq b$ is a typical element of $\mathcal{C}(\mathbb{R})$.
Let us prove that $\mathbf{ext}([a, b]) = \{a, b\}$.
Consider $x, y \in [a, b]$ and $\lambda \in (0, 1)$ with $a = \lambda x + (1 - \lambda)y$. Assume that $x > a$ or $y > a$, then the contradiction

$$a = \lambda x + (1 - \lambda)y > \lambda a + (1 - \lambda)a = a$$

follows. This shows that $a \in \mathbf{ext}([a, b])$. Similarly, $b \in \mathbf{ext}([a, b])$. All points in (a, b) are clearly not extreme. Hence,

$$\mathbf{co}(\mathbf{ext}([a, b])) = \mathbf{co}\{a, b\} = [a, b].$$

Inductive step $n - 1 \to n$, $n \geq 2$:
a) case "$\dim C = \dim \mathrm{aff}(C) < n$":
In this case we can use the inductive assumption again, since this lower-dimensional set in $\mathbb{R}^n$ is equivalent to a full-dimensional set in $\mathbb{R}^\mu$ with $1 \leq \mu < n$.
b) case "$\dim C = \dim \mathrm{aff}(C) = n$":
From Proposition 3.67 follows that $C = \mathbf{co}(\partial C)$. For $x \in \partial C$ there exists a hyperplane $H(\eta, \alpha)$ with $x \in H(\eta, \alpha)$ and $C \subset \overline{H}_-(\eta, \alpha)$ by Lemma 3.52.
Since $x \in F := C \cap H(\eta, \alpha)$, F is convex, compact and

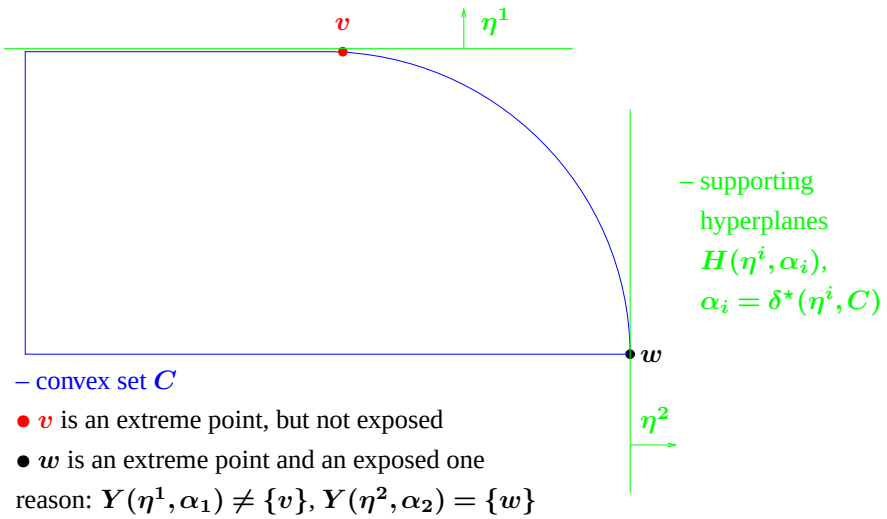
$$\dim(F) = \dim(\mathrm{aff}(C \cap H(\eta, \alpha))) = \dim(H(\eta, \alpha)) = n - 1,$$

we can use part a) to show that $x \in \mathbf{co}(\mathbf{ext}(F))$, i.e. $x = \sum_{j=1}^{k} \lambda_j z^j$ with $z^j \in \mathbf{ext}(F)$.
Since F is an exposed face, $z^j \in \mathbf{ext}(C)$ follows by Proposition 3.60. Hence, x could be written as convex combination of extremal points of C. □

Theorem 3.69 (Straszewicz). *Let $C \subset \mathbb{R}^n$ be compact and convex. Then, $\mathbf{exp}(C) \subset \mathbf{ext}(C) \subset \mathbf{cl}(\mathbf{exp}(C))$, i.e. $C = \overline{\mathbf{co}}(\mathbf{exp}(C))$.*

Proof. cf. [Roc72, Theorem 18.6] □

Not all Extreme Points are Exposed



Corollary 3.70. *Let $C \subset \mathbb{R}^n$ be compact and convex. Then,*

$$C = \overline{\mathbf{co}} \bigcup_{\substack{l \in S_{n-1} \\ Y(l,C) = \{y(l,C)\}}} \{y(l, C)\}.$$

Since $\boldsymbol{Y(l,C) \subset C}$*, we also have*

$$\boldsymbol{C = \mathrm{co} \bigcup_{l \in S_{n-1}} Y(l,C).}$$

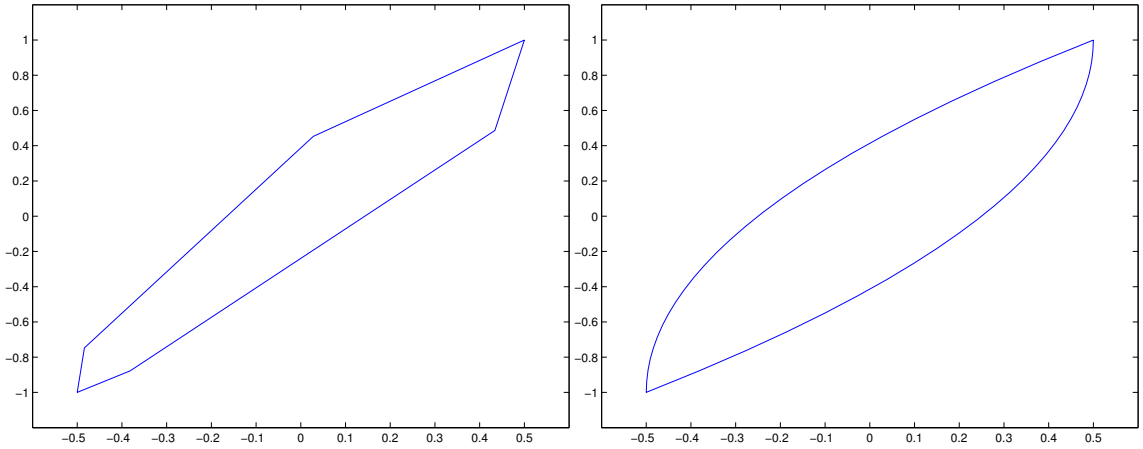
Proof. follows directly from Theorem 3.69 and the definition of

$$\boldsymbol{\exp(C) = \{y(l,C) \,|\, l \in S_{n-1} \text{ with } Y(l,C) = \{y(l,C)\}\}}$$

resp. from Proposition 3.67 and Lemma 3.52 □

representation of a convex sets via supporting points

convex hull based on 12 and 200 outer normals



3.2 Arithmetic Set Operations

Basic Facts
Why do we deal with arithmetic operations on sets?

- approximation methods of reachable sets use arithmetic operations on sets (sum, scalar multiple, linear transformation)
- arithmetic set operations are generalizations of vector arithmetic
- properties of arithmetic operands (=sets) remain valid for its result (compactness, boundedness)
- the operations of convex hull and set arithmetic commute
- the operations of support function and set arithmetic commute (sum, multiple by non-negative scalar)
- formula for support function of linear transformed set uses only support function of original set
- same properties for supporting faces

Basic Facts (continued)

- order of support functions charcterizes inclusion of their sets
- Hausdorff resp. Demyanov metric for convex, compact sets can be defined with support functions resp. supporting points
- the space of compact, nonempty subsets of $\mathbb{R}^n$ and the the space of convex, compact, nonempty subsets of $\mathbb{R}^n$ are complete w.r.t. Hausdorff distance

Basic Facts (continued)
Let $\boldsymbol{U}, \boldsymbol{V} \subset \mathbb{R}^n$, $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{\mu} \in \mathbb{R}$.
For convex hull:

$$\begin{aligned} \mathbf{co}(\boldsymbol{U}+\boldsymbol{V}) &= \mathbf{co}(\boldsymbol{U})+\mathbf{co}(\boldsymbol{V}), \\ \mathbf{co}(\boldsymbol{\mu}\cdot\boldsymbol{U}) &= \boldsymbol{\mu}\cdot\mathbf{co}(\boldsymbol{U}), \\ \mathbf{co}(\boldsymbol{A}\cdot\boldsymbol{U}) &= \boldsymbol{A}\cdot\mathbf{co}(\boldsymbol{U}) \end{aligned}$$

Basic Facts (continued)
Let $\boldsymbol{C}, \boldsymbol{D}$ convex, nonempty sets, $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{\lambda} \geq \mathbf{0}$.
For support functions ($\boldsymbol{\eta} \in \mathbb{R}^n$):

$$\begin{aligned} \delta^\star(\eta, \boldsymbol{C}+\boldsymbol{D}) &= \delta^\star(\eta, \boldsymbol{C})+\delta^\star(\eta, \boldsymbol{D}), \\ \delta^\star(\eta, \boldsymbol{\lambda}\cdot\boldsymbol{C}) &= \boldsymbol{\lambda}\cdot\delta^\star(\eta, \boldsymbol{C}), \\ \delta^\star(\eta, \boldsymbol{A}\cdot\boldsymbol{C}) &= \delta^\star(\boldsymbol{A}^t\cdot\eta, \boldsymbol{C}), \\ \delta^\star(\eta, \boldsymbol{C}) \leq \delta^\star(\eta, \boldsymbol{D}) \text{ for all } \eta \in \boldsymbol{S}_{n-1} &\iff \boldsymbol{C} \subset \boldsymbol{D} \text{ (if } \boldsymbol{C}, \boldsymbol{D} \in \boldsymbol{\mathcal{C}}(\mathbb{R}^n)) \end{aligned}$$

For supporting points ($\boldsymbol{\eta} \in \mathbb{R}^n$):

$$\begin{aligned} \boldsymbol{Y}(\eta, \boldsymbol{C}+\boldsymbol{D}) &= \boldsymbol{Y}(\eta, \boldsymbol{C})+\boldsymbol{Y}(\eta, \boldsymbol{D}), \\ \boldsymbol{Y}(\eta, \boldsymbol{\lambda}\cdot\boldsymbol{C}) &= \boldsymbol{\lambda}\cdot\boldsymbol{Y}(\eta, \boldsymbol{C}) \\ \boldsymbol{Y}(\eta, \boldsymbol{A}\cdot\boldsymbol{C}) &= \boldsymbol{A}\cdot\boldsymbol{Y}(\boldsymbol{A}^t\cdot\eta, \boldsymbol{C}) \end{aligned}$$

Attention:

- space of convex, compact, nonempty sets with Minkowski sum is not a group (only semi-group)!

- space of convex, compact, nonempty sets with Minkowski sum and scalar multiplication is not a vector space!
- in general, no inverse of Minkowsi sum is available
  $(-\mathbf{1}) \cdot \boldsymbol{C}$ is not the inverse of $\boldsymbol{C}$ in general
- second distributive law is not valid for non-convex sets
- second distributive law for non-negative scalars only is valid for convex sets

Important Tools for (Later) Proofs

- valid laws of a vector space (exceptions: inverse, second distributive law)
- second distributive law for non-negative scalars and convex sets
- commuting of sum and multiplication of non-negative scalar and support function/supporting faces
- inclusion of sets corresponds to order of their support functions
- Minkowski theorem for Hausdorff distance
- completeness of set spaces w.r.t. Hausdorff distance
- Theorem of Shapley-Folkman

3.2.1 Definitions and First Properties

Definition 3.71. Let $U, V \subset \mathbb{R}^n$, $\mu \in \mathbb{R}$ and $A \in \mathbb{R}^{m\times n}$. Then,

$$\begin{aligned} \mu \cdot U &:= \{\mu \cdot u \,|\, u \in U\}, & A \cdot U &:= \{A \cdot u \,|\, u \in U\}, \\ U + V &:= \{u + v \,|\, u \in U, v \in V\} \end{aligned}$$

defines the *scalar multiplication*, the *image* of U under the linear map $x \mapsto Ax$ and the *Minkowski sum*.

Example 3.72. Let $v, w \in \mathbb{R}^n$, $A \in \mathbb{R}^{m\times n}$ and $U \subset \mathbb{R}^n$. Then, $U + \{v\}$ coincides with the translation $U + v$. Furthermore, $\mu \cdot \{v\} = \{\mu \cdot v\}$, $A \cdot \{v\} = \{Av\}$ and $\{v\} + \{w\} = \{v + w\}$, i.e. all known vector operations are special cases of the arithmetic set operations.

Proof. $U + \{v\} = \{u + v \,|\, u \in U\} = \{u \,|\, u \in U\} + \{v\} = U + v$ □

Definition 3.73. Let $U, V \subset \mathbb{R}^n$. Then,

$$\begin{aligned} -U &:= \{-u \,|\, u \in U\}, \\ U - V &:= \{u - v \,|\, u \in U, v \in V\} \end{aligned}$$

defines the *pointwise "inverse"* and the *algebraic (pointwise) difference*.

Remark 3.74. *Let $U, V \subset \mathbb{R}^n$. Then,*

$$\begin{aligned} -U &= (-1)\cdot U, \\ U - V &= U + ((-1)\cdot V) \end{aligned}$$

The operations $-U$ and $U - V$ generalize the vector inverse and subtraction, but in general $-U$ does not give the inverse w.r.t. Minkowski sum and $U - U$ is normally bigger than $\{0_{\mathbb{R}^n}\}$!

Remark 3.75. *Let $U, V \subset \mathbb{R}^n$ and V being (point) symmetric to the origin. Then, $-V = V$ and $U - V = U + V$.*

Proposition 3.76. *Let $U, V \subset \mathbb{R}^n$. Then,*

(i) $(U - V) + V \supset U$

(ii) $(U + V) - V \supset U$

(iii) $U - U \supset \{0_{\mathbb{R}^n}\}$

(iv) $U - U = \{0_{\mathbb{R}^n}\}$ *if and only if* $U = \{u\}$ *with some* $u \in \mathbb{R}^n$.

(v) Strict inclusions in (i)–(iii) appear, if e.g. $U = V = B_1(0)$.

Proof. only (iii)–(v) will be proven, (i)–(ii) follow directly from the definition
(iii) follows from $0_{\mathbb{R}^n} = u - u$ for $u \in U$, i.e. $0_{\mathbb{R}^n} \in U - U$
(iv) Clearly, $U \neq \emptyset$, if $U - U \neq \emptyset$ or $u \in U$ exists.
"Only if" follows from Example 3.72.
"If:" For all $v \in U$ follows that $u - v \in U - U$ and $u - v = 0_{\mathbb{R}^n}$. Hence, $u = v$ and $U = \{u\}$.
(v) Then,

$$\begin{aligned} U - U = B_1(0) - B_1(0) &= B_1(0) + \underbrace{((-1)\cdot B_1(0))}_{=B_1(0)} \\ &= 2B_1(0) = B_2(0), \\ (U + V) - V &= (B_1(0) + B_1(0)) - B_1(0) \\ &= 2B_1(0) + ((-1)\cdot B_1(0)) = 3B_1(0) = B_3(0), \\ (U - V) + V &= (B_1(0) - B_1(0)) + B_1(0) \\ &= 2B_1(0) + B_1(0) = 3B_1(0) = B_3(0) \end{aligned}$$

□

Example 3.77 *(reliable computing).* If $x, y \in \mathbb{R}$ are real numbers which should be stored as floating-point numbers in the computer, for which the floating-point accuracy is ε (typically: $\varepsilon \approx 10^{-15}$).

Then, the result of the storage resp. the arithmetical operations yield

$$\begin{aligned}
x &\in [x-\varepsilon, x+\varepsilon],\\
y &\in [y-\varepsilon, y+\varepsilon],\\
x+y &\in [x-\varepsilon, x+\varepsilon]+[y-\varepsilon, y+\varepsilon] = [x+y-2\varepsilon,\ x+y+2\varepsilon],\\
x-y &\in [x-\varepsilon, x+\varepsilon]-[y-\varepsilon, y+\varepsilon] = [x-y-2\varepsilon,\ x-y+2\varepsilon],\\
\mu\cdot x &\in \mu\cdot[x-\varepsilon, x+\varepsilon] = [\mu\cdot x-\mu\cdot\varepsilon, \mu\cdot x+\mu\cdot\varepsilon], \quad (\text{for } \mu \geq 0),\\
\mu\cdot x &\in \mu\cdot[x-\varepsilon, x+\varepsilon] = [\mu\cdot x-|\mu|\cdot\varepsilon, \mu\cdot x+|\mu|\cdot\varepsilon], (\text{for } \mu < 0),
\end{aligned}$$

i.e. the operations in the computer lie definitely in intervals calculated by the Minkowski sum, algebraic difference resp. scalar multiplication.

Remark 3.78. *some literature on interval analysis and reliable computing:*
books: [JKDW01, AH83, AH74, Moo66]
articles: [Kau80, Kul77, Nic78, Mar79, Mar80, Mar95, Mar98]
directed/extended intervals:
[Ort69, Kau77a, Kau77b, Kau80, Mar80, Mar95, Mar98, Dim80, Rat80, BF01a, BF01b]

Remark 3.79. *Let $U, V \subset \mathbb{R}^n$. Then, the Minkowski sum*

$$U+V = \bigcup_{v\in V} (U+v) = \bigcup_{u\in U} (V+u)$$

is the union of all U translated by an element of V and

$$U-V = \bigcup_{v\in V} (U-v) = \bigcup_{u\in U} ((-V)+u).$$

Example 3.80. Let $r, s \geq 0$, $p, q \in \mathbb{R}^n$. Then, $B_r(p)+B_s(q) = B_{r+s}(p+q)$.

Proposition 3.81. *Let $U, V \subset \mathbb{R}^n$. Then,*

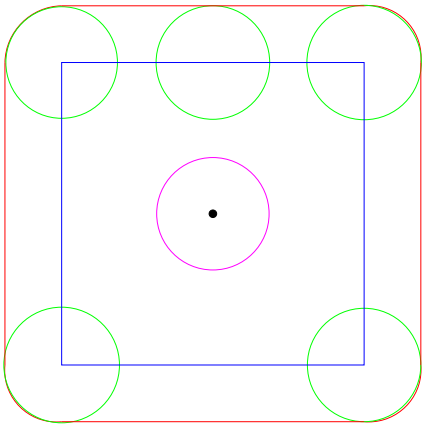
(i) $U+V = V+U$ *(commutative)*

(ii) $U+\{0_{\mathbb{R}^n}\} = U$ *($\{0_{\mathbb{R}^n}\}$ is neutral element)*

(iii) $U+(V+W) = (U+V)+W$ *(associative)*

Proof. Everything is trivial:
(i) $U+V = \{v+u \,|\, u\in U, v\in V\} = V+U$
(ii) $U+\{0_{\mathbb{R}^n}\} = \{u+0_{\mathbb{R}^n} \,|\, u\in U\} = \{u \,|\, u\in U\} = U$
(iii) The left-hand side consists of elements $u+z$ with $u\in U$, $z\in V+W$. But $z = v+w$ for some $v\in V$, $w\in W$. Associativity in $\mathbb{R}^n$ yields $u+(v+w) = (u+v)+w$ which is an element of the right-hand side. Reversing the arguments gives the other inclusion. □

Minkowski Sum of a Big Square and a Small Ball

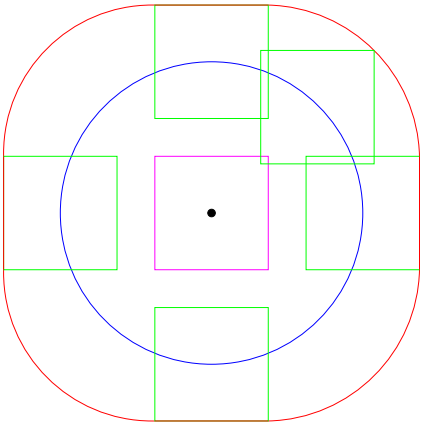


– 1st summand $C = [-1,1]^2$
– 2nd summand $D = B_\varepsilon(0)$
● origin

– translated set $c + D$, $c \in \partial C$

– Minkowski sum $C + D$

Minkowski Sum of a Big Ball and a Small Square

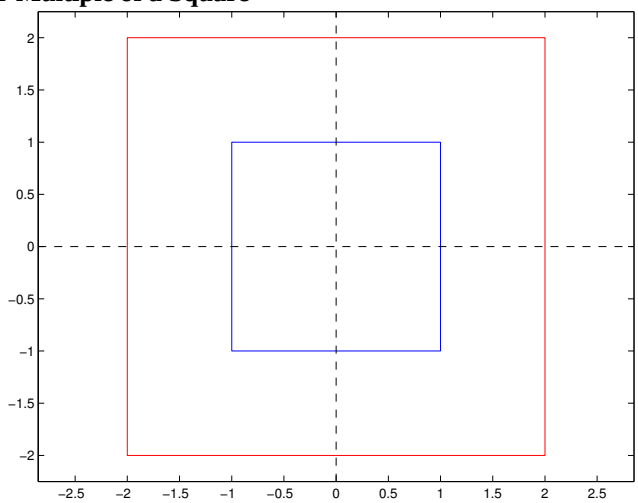


– 1st summand $C = B_1(0)$
– 2nd summand $D = [-\varepsilon,\varepsilon]^2$
● origin

– translated set $c + D$, $c \in \partial C$

– Minkowski sum $C + D$

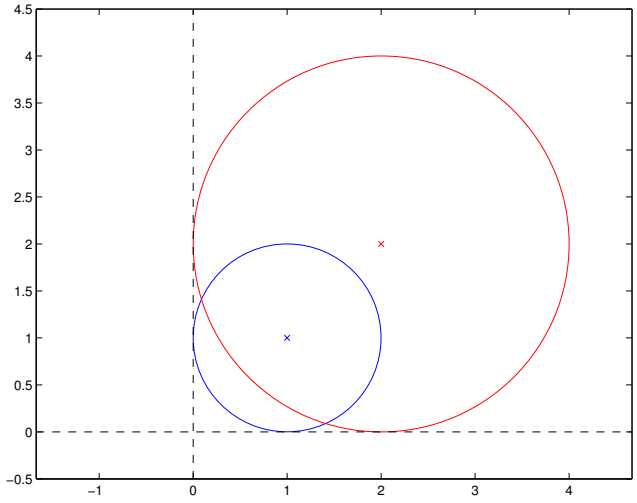
Scalar Multiple of a Square



– $C = [-1,1]^2$

– $D = 2 \cdot C$,
$D = [-2,2]^2$

Scalar Multiple of a Ball



– $C = B_1(\binom{1}{1})$

– $D = 2 \cdot C$,
$D = B_2(\binom{2}{2})$

Lemma 3.83. *Let $U, V, \widetilde{U}, \widetilde{V} \subset \mathbb{R}^n$ with $U \subset \widetilde{U}$, $V \subset \widetilde{V}$. Then, $U + V \subset \widetilde{U} + \widetilde{V}$.*

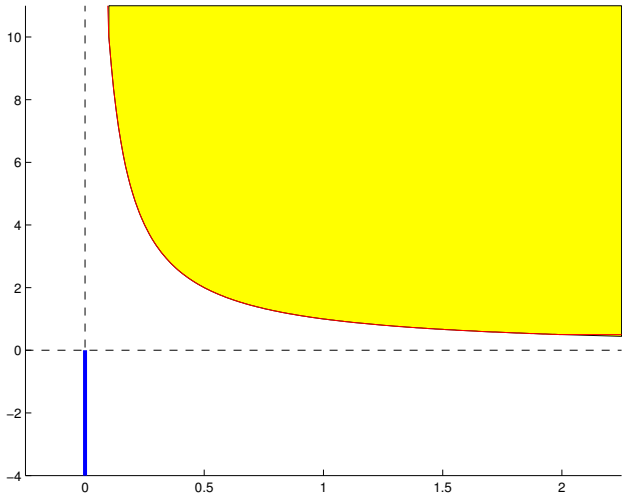
Proof. straight from the definition □

Lemma 3.84. *Let $U, V \subset \mathbb{R}^n$ be closed and one of the sets be bounded. Then, $U + V$ is closed.*

Proof. Let $(w^m)_m \subset U + V$ be a convergent sequence with $w^m \xrightarrow[m\to\infty]{} w$. Each w^m can be represented with $u^m + v^m$ with $u^m \in U$, $v^m \in V$, $m \in \mathbb{N}$. W.l.o.g. let V be bounded.

Proof. Let $(w^m)_m \subset U + V$ be a convergent sequence with $w^m \xrightarrow[m\to\infty]{} w$. Each w^m can be represented with $u^m + v^m$ with $u^m \in U$, $v^m \in V$, $m \in \mathbb{N}$. W.l.o.g. let V be bounded. Then, $(v^m)_m$ contains a convergent subsequence $(v^{m_k})_k \subset (v^m)_m$ with $v^{m_k} \xrightarrow[k\to\infty]{} v$. Since V is closed, $v \in V$. Clearly, $(u^{m_k})_k$ converges to $w - v$. Since U is also closed, $u := w - v \in U$ so that $w = u + v \in U + V$. □

Non-Closed Minkowski Sum of 2 Closed Sets

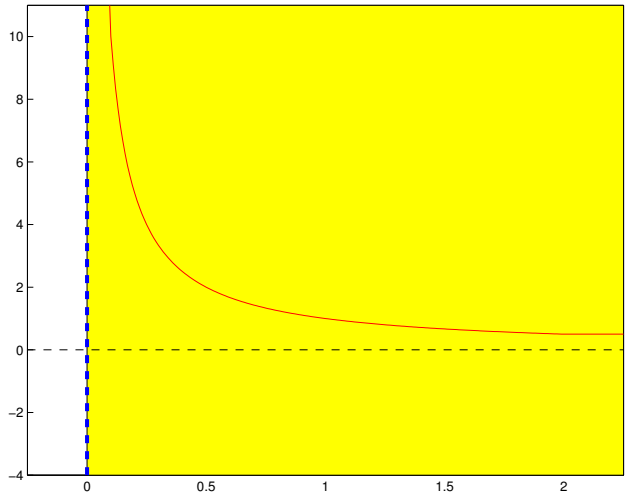


– boundary of C,
$C = \mathbf{epi}(f)$,
$f(x) = \frac{1}{x}$

– interior of C

– convex set
$D = \{\binom{0}{y} \mid y \leq 0\}$

Non-Closed Minkowski Sum of 2 Closed Sets



– boundary of $C = \mathbf{epi}(f)$

– Minkowski sum $C + D$:
$C \subset C + D$, since $0_{\mathbb{R}^2} \in D$,
$C + \binom{0}{y} \subset C + D$ for $y < 0$

--- not part of $C + D$,
belongs to $\partial(\overline{C + D})$

Proposition 3.86. *Let $U, V \subset \mathbb{R}^n$.*

(i) If $U, V \subset \mathbb{R}^n$ is nonempty, then $U + V$ is also nonempty.

(ii) If $U, V \subset \mathbb{R}^n$ is bounded, then $U + V$ is also bounded.

(iii) If $U, V \subset \mathbb{R}^n$ is compact, then $U + V$ is also compact.

(iv) If $U, V \subset \mathbb{R}^n$ is convex, then $U + V$ is also convex.

Proof. (i) If $u \in U$, $v \in V$, then $u + v \in U + V$.
(ii) If $U \subset B_r(0)$, $V \subset B_s(0)$ with $r, s \geq 0$, then $U + V \subset B_r(0) + B_s(0) = B_{r+s}(0)$.
(iii) The boundedness follows from (ii), the closedness from Lemma 3.84.
(iv) Let $w^i \in U + V$, $i = 1, 2$, and $\lambda \in [0, 1]$. Hence, there exists $u^i \in U$, $v^i \in V$ with $w^i = u^i + v^i$, $i = 1, 2$. Since the convex combinations of u^i and v^i lie in U resp. V, we have

$$\begin{aligned} \lambda w^1 + (1-\lambda) w^2 &= \lambda(u^1 + v^1) + (1-\lambda)(u^2 + v^2) \\ &= \underbrace{(\lambda u^1 + (1-\lambda) u^2)}_{\in U} + \underbrace{(\lambda v^1 + (1-\lambda) v^2)}_{\in V} \in U + V \end{aligned}$$

□

Proposition 3.88. *Let $U, V \subset \mathbb{R}^n$ and $A, \widetilde{A} \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{p \times m}$. Then,*

(i) $I_n \cdot U = U$ ($I_n \in \mathbb{R}^{n \times n}$ identity matrix) (I_n is neutral element)

(ii) $B \cdot (A \cdot U) = (B \cdot A) \cdot U$ (associative)

(iii) $A \cdot (U + V) = A \cdot U + A \cdot V$ (1. distributive law)

(iv) $(A + \widetilde{A}) \cdot U \subset A \cdot U + \widetilde{A} \cdot U$ (not the 2. distributive law)

Remark 3.89. *Scalar multiplication of sets is a special case of linear transformation, since with $A := \lambda I_n$ with I_n being the $n \times n$-identity matrix; we have*

$$A \cdot U = \{Au \,|\, u \in U\} = \{(\lambda I_n) u \,|\, u \in U\} = \{\lambda u \,|\, u \in U\} = \lambda \cdot U$$

for all $U \subset \mathbb{R}^n$.

Proposition 3.91. *Let $U, V \subset \mathbb{R}^n$ and $\lambda, \mu \in \mathbb{R}$. Then,*

(i) $1 \cdot U = U$ (1 is neutral element)

(ii) $\lambda \cdot (\mu \cdot U) = (\lambda \cdot \mu) \cdot U$ *(associative)*

(iii) $\lambda \cdot (U + V) = \lambda \cdot U + \lambda \cdot V$ *(1. distributive law)*

(iv) $(\lambda + \mu) \cdot U \subset \lambda \cdot U + \mu \cdot U$ *(not the 2. distributive law)*

Example 3.92. Let $\mu \in \mathbb{R}$, $r \geq 0$ and $p \in \mathbb{R}^n$. Then, $\mu \cdot B_r(p) = B_{|\mu| \cdot r}(\mu \cdot p)$.

Proposition 3.93 ("almost 2nd" distributive law for convex set). *Let $C \subset \mathbb{R}^n$ be convex and $\lambda, \mu \in \mathbb{R}$ with $\lambda \cdot \mu \geq 0$. Then,*

$$(\lambda + \mu) \cdot C = \lambda \cdot C + \mu \cdot C.$$

Proof. Clearly, "$\subset$" is fulfilled by Proposition 3.91.
case "$\lambda, \mu \geq 0$":
$\lambda = \mu = 0$ is trivial, since $\lambda \cdot C = \mu \cdot C = (\lambda + \mu) \cdot C = \{0_{\mathbb{R}^n}\}$.
If $\lambda + \mu > 0$, then consider $c, \widetilde{c} \in C$ and

$$\frac{1}{\lambda + \mu}(\lambda c + \mu \widetilde{c}) = \underbrace{\frac{\lambda}{\lambda + \mu} c + \frac{\mu}{\lambda + \mu} \widetilde{c}}_{\in C}.$$

This is a convex combination of elements in C, hence

$$\lambda c + \mu \widetilde{c} \in (\lambda + \mu) C.$$

case "$\lambda, \mu < 0$":
Since

$$\lambda \cdot C = ((-1) \cdot |\lambda|) \cdot C = (|\lambda| \cdot (-1)) \cdot C = |\lambda| \cdot ((-1) \cdot C),$$

consider $\widetilde{C} = (-1) \cdot C$ which is also convex by Proposition 3.106(v) and $|\lambda|$ resp. $|\mu|$ instead of λ and μ. The first case shows

$$|\lambda| \cdot \widetilde{C} + |\mu| \cdot \widetilde{C} = (|\lambda| + |\mu|) \cdot \widetilde{C} = |\lambda + \mu| \cdot \widetilde{C}.$$

Replacing $\widetilde{C}$ again by $(-1) \cdot C$ and using Proposition 3.91(ii) finishes the proof. □

Example 3.94 *(negative scalar).* Let $C = B_1(0) \subset \mathbb{R}^n$ and $\lambda = 1$, $\mu = -1$. Then, $(\lambda + \mu) C \neq \lambda C + \mu C$.

Example 3.95 *(nonconvex set).* Let $U = \{\pm e^1\} \subset \mathbb{R}^n$ and $\lambda = \mu = \frac{1}{2}$. Then, $(\lambda + \mu) U \neq \lambda U + \mu U$.

Proof.

$$(\lambda + \mu) U = (\frac{1}{2} + \frac{1}{2}) \cdot U = 1 \cdot U = U,$$
$$\lambda U + \mu U = \frac{1}{2}\{\pm e^1\} + \frac{1}{2}\{\pm e^1\} = \{e^1, -e^1, 0_{\mathbb{R}^n}\} \neq U.$$ □

Example 3.96 *(repeated Minkowski sum).* Let $U = \{0_{\mathbb{R}^n}, e^1\} \subset \mathbb{R}^n$ and $N \in \mathbb{N}$. Then, $\sum_{i=1}^{N} \frac{1}{N} U \neq U$, since

$$\sum_{k=1}^{N} \frac{1}{N} U = \{\frac{k}{N} \cdot e^1 \,|\, k = 1, \ldots, N\}.$$

The limit for $N \to \infty$ of these Minkowski sums tend to $\mathbf{co}(U) = \mathbf{co}\{0_{\mathbb{R}^n}, e^1\}$, cf. Lemma 4.28.

Remark 3.97. *Clearly, Example 3.94 and Proposition 3.76 show that $(\mathcal{C}(\mathbb{R}^n), +)$ is not a group (only a semigroup and a convex cone, cf. [Råd52]), since no inverse element is available in general.*
$-U$ is in general not the inverse element of U w.r.t. Minkowski sum.
$(\mathcal{C}(\mathbb{R}^n), +, \cdot)$ is also not a vector space, since the second distributive law is not valid.

Remark 3.98. *To avoid difficulties not having a vector space, often embeddings*

$$J : \mathcal{C}(\mathbb{R}^n) \to \mathcal{V}$$

into a vector space $\mathcal{V}$ with $J(C + D) = J(C) + J(D)$ and $J(\lambda \cdot C) = \lambda \cdot J(C)$ for $\lambda \geq 0$ are used.
Known examples:

(i) $\mathcal{V}_0 = \{(C,D)\,|\,C,D \in \mathcal{C}(\mathbb{R}^n)\}$ *with*

$$\begin{aligned}(C,D)+(\widetilde{C},\widetilde{D}) &:= (C+\widetilde{C}, D+\widetilde{D}),\\ \lambda\cdot(C,D) &:= (\lambda\cdot C, \lambda\cdot D) && (\lambda \geq 0),\\ \lambda\cdot(C,D) &:= (|\lambda|\cdot D, |\lambda|\cdot C) && (\lambda < 0)\end{aligned}$$

and set $\mathcal{V} := \mathcal{V}_0/\sim$ *as quotient space with equivalence relation*

$$(C,D) \sim (\widetilde{C},\widetilde{D}) \iff C+\widetilde{D} = D+\widetilde{C})$$

with $J(C) = (C,\{0_{\mathbb{R}^n}\}) \in \mathcal{V}$*, cf. Rådström in [Råd52] resp. [PU02]*

Remark 3.98 (continued).

(ii) $\mathcal{V} = C(S_{n-1})$*, i.e. the space of continuous, positively homogeneous functions with real values* $J(C) = \delta^\star(\cdot, C)$ *for* $C \in \mathcal{C}(\mathbb{R}^n)$
cf. Hörmander in [Hör54]

(iii) $\mathcal{V}$ *being a q-linear space, an algebraic extension of a quasi-linear space*
see [Mar98, Mar00]

(iv) a huge literature exists on minimal pairs of convex sets (cf. (i)) with additional minimality conditions
for an overview see the book [PU02]

(v) $\mathcal{V}$ *being a space of directed intervals resp. sets*
cf. [Kau80, Mar79, Mar95, BF01a, BF01b]

Remark 3.99. *Many attempts to define a better difference of convex sets are made:*

(i) embedding from the semi-group into a group

$$C \ominus D \mathrel{\widehat{=}} (C,\{0_{\mathbb{R}^n}\}) - (D,\{0_{\mathbb{R}^n}\}) = (C,D)$$

cf. Rådström in [Råd52]

(ii) embedding from the semi-group into a group

$$C \ominus D \mathrel{\widehat{=}} \delta^\star(\cdot, C) - \delta^\star(\cdot, D) \in C(\mathbb{R}^n)$$

cf. Hörmander in [Hör54]

(iii) geometric difference (star-difference, Minkowski difference)

$$\begin{aligned}C \stackrel{*}{-} D &:= \bigcap_{\eta \in S_{n-1}} \{x \in \mathbb{R}^n \,|\, \langle \eta, x\rangle \leq \delta^\star(\eta, C) - \delta^\star(\eta, D)\}\\ &= \{x \in \mathbb{R}^n \,|\, x + D \subset C\} = \bigcap_{d \in D} (C - d)\end{aligned}$$

cf. [Had50, Pon67]

Remark 3.100 (continued).

(iv) Demyanov difference

$$C \dot{-} D := \overline{\text{co}} \bigcup_{\eta \in T_C \cap T_D} \{y(\eta, C) - y(\eta, D)\}$$

cf. Definition 3.165 and [DR95-IAO, RA92-IAO, DKRV97]

(v) difference motivated by q-linear spaces

$$C -^{-} D = (C \stackrel{*}{-} D) \cup (-(D \stackrel{*}{-} C))$$

cf. [Mar98, Mar00]

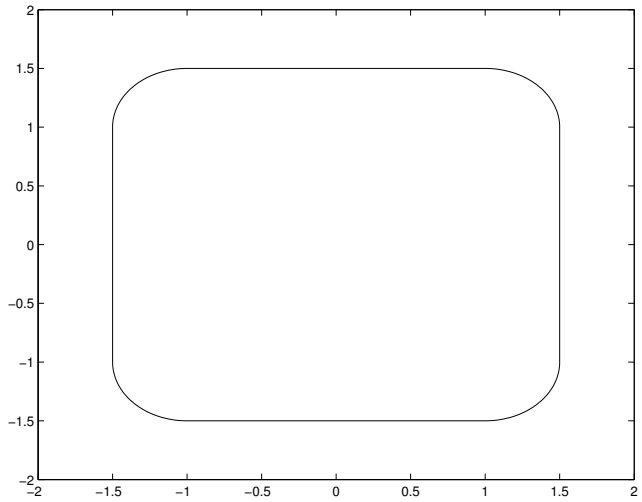
(vi) non-convex, nonempty visualizable difference of directed sets,

$$C \ominus D \mathrel{\widehat{=}} J_n(C) - J_n(D) \in \mathcal{D}(\mathbb{R}^n)$$

$J_n(\cdot)$ *an embedding of* $\mathcal{C}(\mathbb{R}^n)$ *in the space of directed sets, cf. [BF01a, BF01b]*

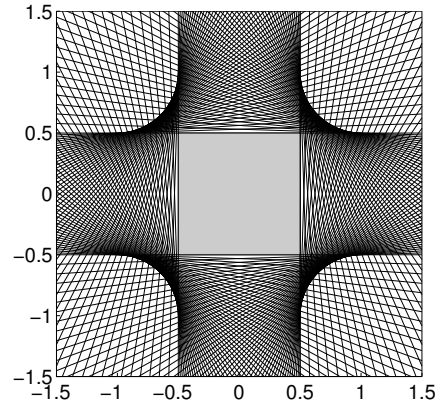
(vii) a unified treatment of algebraic and geometric difference, cf. [Pic03]

Algebraic Difference of $C = [-1,1]^2$ and $D = B_r(0)$ with $r = \frac{1}{2}$



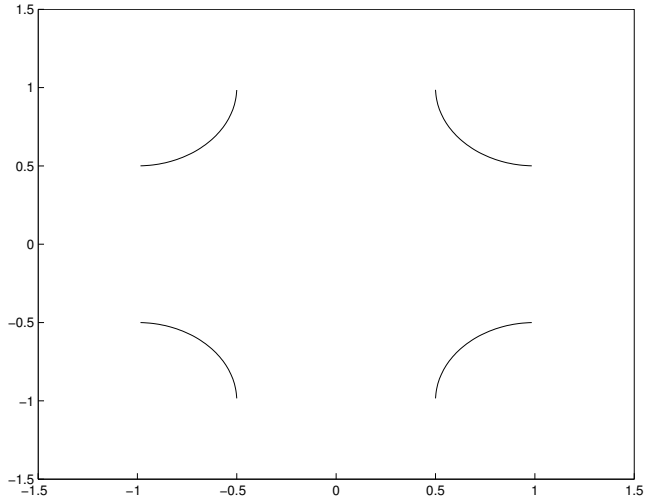
algebr. difference $C - D$
equals Minkowski sum $C + D$
(cf. Remark 3.75)

Geometric Difference of $C = [-1,1]^2$ and $D = B_r(0)$ with $r = \frac{1}{2}$



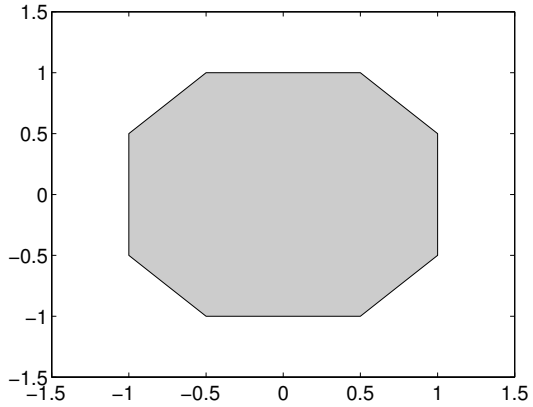
$C \stackrel{*}{-} D$ is the grey square
which is created by
(non-)supporting hyperplanes
$\delta^\star(\eta, C) - \delta^\star(\eta, C)$
is not convex here (cf. Proposition 3.64)

Non-Convexified Part of Demyanov Difference for $C = [-1,1]^2$ and $D = B_r(0)$ with $r = \frac{1}{2}$

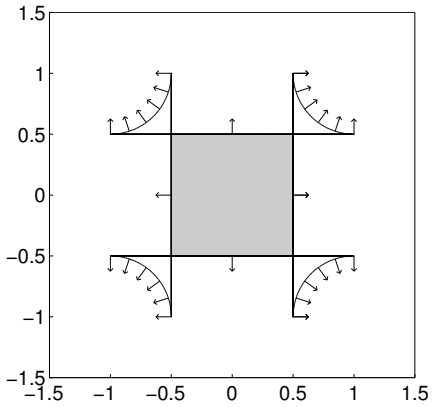


difference of unique supporting points $y(\eta, C) - y(\eta, D)$ for $\eta \in T_C \cap T_D$

Demyanov Difference of $C = [-1, 1]^2$ and $D = B_r(0)$ with $r = \frac{1}{2}$



Directed Difference of $C = [-1, 1]^2$ and $D = B_r(0)$ with $r = \frac{1}{2}$



$J_n(C) - J_n(D)$ has a
non-convex visualization
incl. the grey square $C \stackrel{*}{-} D$
and e.g. four non-convex arcs.
The arrows indicate the orientation
(details in [BF01b]).

Lemma 3.101. *Let $\mu \in \mathbb{R}$, $A \in \mathbb{R}^{m\times n}$ and $U, \widetilde{U} \subset \mathbb{R}^n$ with $U \subset \widetilde{U}$. Then,*

$$A \cdot U \subset A \cdot \widetilde{U} \quad \text{and} \quad \mu \cdot U \subset \mu \cdot \widetilde{U}.$$

Proof. straight from the definition □

Proposition 3.102. *Let $U \subset \mathbb{R}^n$ and $A \in \mathbb{R}^{m\times n}$.*

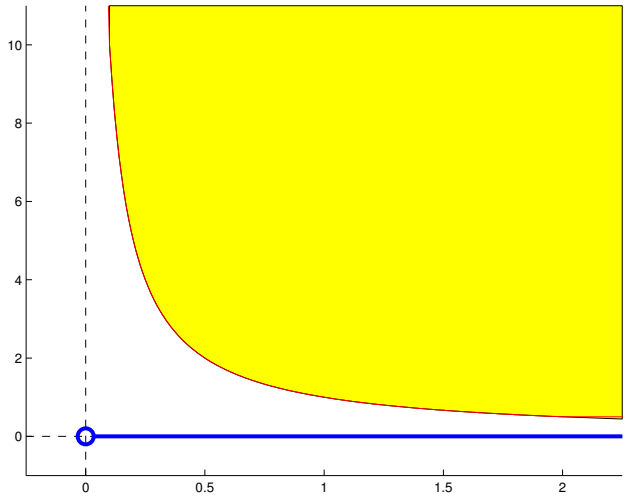
(i) If $U \subset \mathbb{R}^n$ is nonempty, then $A \cdot U$ is also nonempty.

(ii) If $U \subset \mathbb{R}^n$ is bounded, then $A \cdot U$ is also bounded.

(iii) If $U \subset \mathbb{R}^n$ is compact, then $A \cdot U$ is also compact.

(iv) If $U \subset \mathbb{R}^n$ is convex, then $A \cdot U$ is also convex.

Non-Closed Linear Transform of a Closed Set



– boundary of C

$C = \mathbf{epi}(f)$,

$f(x) = \frac{1}{x}, x > 0$

– $D = A \cdot C$,

$A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$,

$D = (0, \infty) \times \{0\}$

is open

and not closed

Remark 3.104. *If $U \subset \mathbb{R}^n$ is closed and $A \in \mathbb{R}^{n\times n}$ is invertible, then $A \cdot U$ is also closed.*

Proof. Consider a converging sequence $(v^m)_m \subset A \cdot U$ with $v^m = Au^m$, $u^m \in U$ for $m \in \mathbb{N}$ approaching $v \in \mathbb{R}^m$. Then, $(A^{-1}v^m)_m = (u^m)_m$ is a converging sequence in U approaching $u := A^{-1}v$. Since U is closed, $u \in U$ and $v = Au$. □

Proposition 3.106. *Let $U \subset \mathbb{R}^n$ and $\mu \in \mathbb{R}$.*

(i) If $U \subset \mathbb{R}^n$ is nonempty, then $\mu \cdot U$ is also nonempty.

(ii) If $U \subset \mathbb{R}^n$ is bounded, then $\mu \cdot U$ is also bounded.

(iii) If $U \subset \mathbb{R}^n$ is closed, then $\mu \cdot U$ is also closed.

(iv) If $U \subset \mathbb{R}^n$ is compact, then $\mu \cdot U$ is also compact.

(v) If $U \subset \mathbb{R}^n$ is convex, then $\mu \cdot U$ is also convex.

Proposition 3.107. *Let $U, V \subset \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$ and $\mu \in \mathbb{R}$. Then,*

$$\begin{aligned} \operatorname{co}(U+V) &= \operatorname{co}(U)+\operatorname{co}(V), \\ \operatorname{co}(A \cdot U) &= A \cdot \operatorname{co}(U) \quad \text{and} \quad \operatorname{co}(\mu \cdot U) = \mu \cdot \operatorname{co}(U). \end{aligned}$$

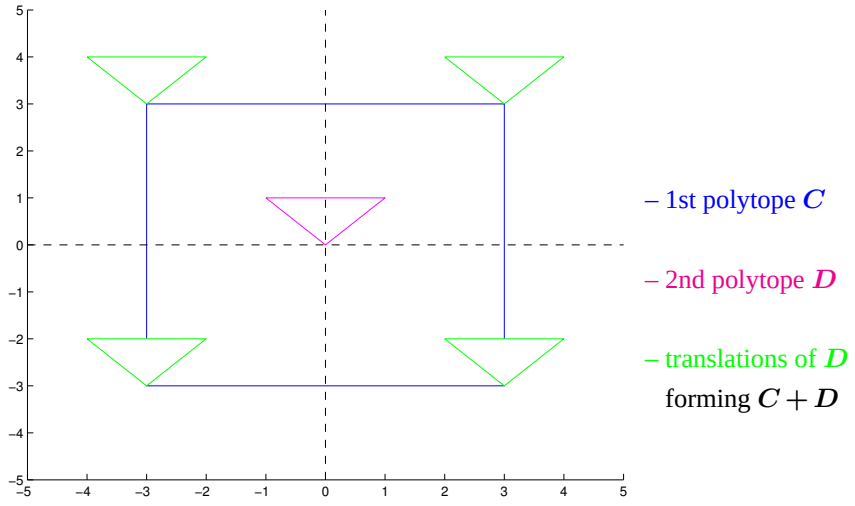
Corollary 3.108. *Let $P = \operatorname{co}\{p^i \mid i = 1, \ldots, r\}$ and $Q = \operatorname{co}\{q^j \mid j = 1, \ldots, s\}$ be two polytopes, $A \in \mathbb{R}^{m \times n}$ and $\lambda \in \mathbb{R}$. Then,*

$$\begin{aligned} P+Q &= \operatorname{co}\{p^i+q^j \mid i = 1, \ldots, r,\ j = 1, \ldots, s\}, \\ A \cdot P &= \operatorname{co}\{A \cdot p^i \mid i = 1, \ldots, r\}, \\ \lambda \cdot P &= \operatorname{co}\{\lambda \cdot p^i \mid i = 1, \ldots, r\} \end{aligned}$$

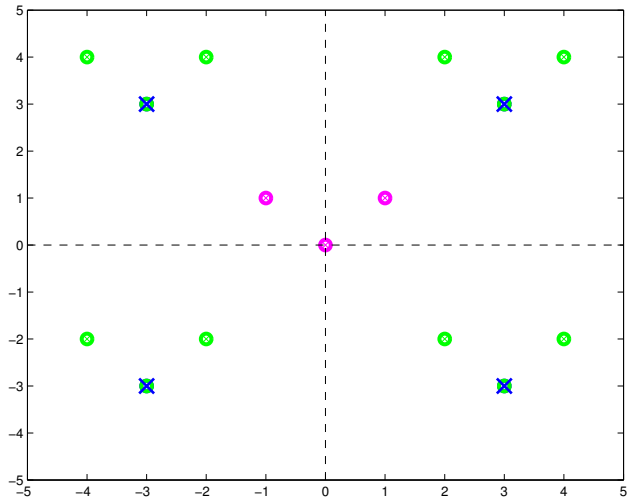
are again polytopes with maximal $r+s$ resp. r vertices.

Proof. Apply Proposition 3.107 to $U := \{p^i \mid i = 1, \ldots, r\}$ and $V := \{q^j \mid j = 1, \ldots, s\}$. □

Minkowski Sum of 2 Polytopes



Minkowski Sum of 2 Polytopes

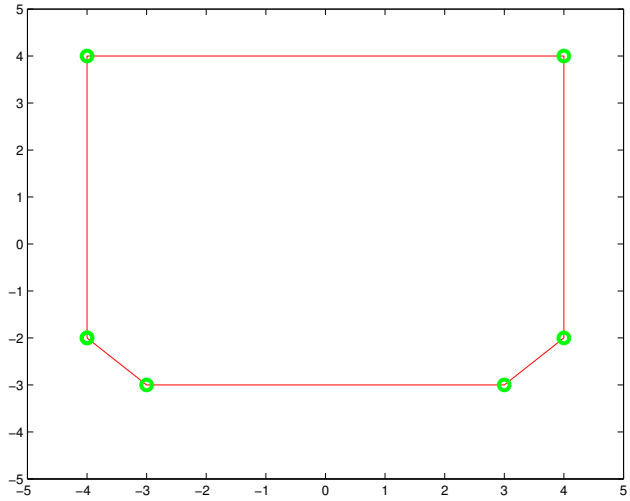


x vertices of C

o vertices of D

o all possible sums of vertices forming $C + D$

Minkowski Sum of 2 Polytopes



– Minkowski sum $C + D$

o vertices of $C + D$

Remark 3.109. *A more intelligent algorithm for computing the Minkowski sum of two polygons in* $\mathbb{R}^2$*, replacing the calculation of all sums of* $r \cdot s$ *pairs of vertices in Corollary 3.108, can be found e.g. in [dvOS97, Algorithm before Theorem 13.10]. Its complexity is* $\mathcal{O}(r + s)$ *for the computation of* $\mathrm{co}(U) + \mathrm{co}(V)$*, but* $\mathcal{O}(r \cdot s)$ *for* $\mathrm{co}(U) + V$ *and* $U + \mathrm{co}(V)$ *(only one convex set) as well as* $\mathcal{O}(r^2 \cdot s^2)$ *in the true non-convex case for the computation of* $U + V$*.*

For Further Reading on Set Arithmetics

References

[HUL93-IAO] J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms I*, volume 305 of *Grundlehren der mathematischen Wissenschaften*. Springer-Verlag, Berlin–Heidelberg–New York–London–Paris–Tokyo–Hong Kong–Barcelona–Budapest, 1993.

[Mar77-IAO] J. T. Marti. *Konvexe Analysis*, volume 54 of *Lehrbücher und Monographien aus dem Gebiet der Exakten Wissenschaften, Mathematische Reihe*. Birkhäuser, Basel–Stuttgart, 1977.

[Roc70-IAO] R. T. Rockafellar. *Convex Analysis*, volume 28 of *Princeton Mathematical Series*. Princeton University Press, Princeton, NJ, (2nd edn) edition, 1972.

[Sch93-IAO] R. Schneider. *Convex Bodies: The Brunn-Minkowski Theory*, volume 44 of *Encyclopedia of Mathematics and Applications*. Cambridge University Press, Cambridge, 1993.

[DR95-IAO] V. F. Demyanov and A. M. Rubinov. *Constructive nonsmooth analysis*, volume 7 of *Approximation and Optimization*. Peter Lang, Frankfurt am Main–Berlin–Bern–New York–Paris–Wien, 1995.

[RA92-IAO] A. M. Rubinov and I. S. Akhundov. Difference of compact sets in the sense of Demyanov and its application to non-smooth analysis. *Optimization*, 23(3):179–188, 1992.

[Bai95-IAO] R. Baier. *Mengenwertige Integration und die diskrete Approximation erreichbarer Mengen*, volume 50 of *Bayreuth. Math. Schr.* Mathematisches Institut der Universität Bayreuth, 1995.

3.2.2 Properties of Support Functions

Remark 3.111. *We use the conventions*

$$\begin{aligned} x+\infty &= \infty \quad \text{for all } x \in \mathbb{R}, \\ \infty+\infty &= \infty, \\ \lambda \cdot \infty &= \infty \quad \text{for all } \lambda > 0, \\ 0 \cdot \infty &= 0. \end{aligned}$$

Proposition 3.112. *Let $U \subset \mathbb{R}^n$ be nonempty and $l \in \mathbb{R}^n$.*

(i) $\delta^\star(\lambda \cdot l, U) = \lambda \cdot \delta^\star(l,U) \quad (\lambda \geq 0)$ *(positively homogeneous)*

(ii) $\delta^\star(l+\eta, U) \leq \delta^\star(l,U) + \delta^\star(\eta,U) \quad (\eta \in \mathbb{R}^n)$ *(subadditive)*

(iii) $\delta^\star(\cdot, U)$ *is convex*

Corollary 3.113. *If $U \subset \mathbb{R}^n$ is included in $B_r(m)$ with $m \in \mathbb{R}^n$, $r \geq 0$. Then,*

$$\begin{aligned} |\delta^\star(l,U) - \langle l, m\rangle| &\leq r \cdot \|l\|_2 \quad \text{for all } l \in \mathbb{R}^n, \\ |\delta^\star(l,U)| &\leq r \cdot \|l\|_2 \quad \text{for all } l \in \mathbb{R}^n, \text{ if } m = 0_{\mathbb{R}^n}. \end{aligned}$$

Proposition 3.114. *Let $U \subset \mathbb{R}^n$ be included in $B_r(0)$ with $r \geq 0$. Then, the function $\delta^\star(\cdot, U)$ is Lipschitz continuous with Lipschitz constant r.*

Proof. Consider $u \in U \subset B_r(0)$ and $l, \eta \in \mathbb{R}^n$. Then,

$$\begin{aligned} \langle l, u\rangle - \delta^\star(\eta, U) &\leq \langle l, u\rangle - \langle \eta, u\rangle = \langle l-\eta, u\rangle \\ &\leq \|l-\eta\|_2 \cdot \|u\|_2 \leq r \cdot \|l-\eta\|_2. \end{aligned}$$

Approaching with $\langle l, u\rangle$ the supremum $\delta^\star(l,U)$ gives

$$\delta^\star(l,U) - \delta^\star(\eta,U) \leq r \cdot \|l-\eta\|_2.$$

The same arguments with interchanged l and η give $\delta^\star(\eta,U) - \delta^\star(l,U) \leq r \cdot \|l-\eta\|_2$. Hence,

$$|\delta^\star(l,U) - \delta^\star(\eta,U)| \leq r \cdot \|l-\eta\|_2.$$

□

Proposition 3.115. *Let $U, V \subset \mathbb{R}^n$ be nonempty, $\lambda \geq 0$ and $l \in \mathbb{R}^n$. Then,*

$$\begin{aligned} \delta^\star(l, U+V) &= \delta^\star(l,U) + \delta^\star(l,V), \\ \delta^\star(l, \lambda \cdot U) &= \lambda \cdot \delta^\star(l,U). \end{aligned}$$

Proof. (i) Let $z \in U+V$, i.e. $z = u+v$ with $u \in U$, $v \in V$. Then,

$$\langle l, z\rangle = \langle l, u\rangle + \langle l, v\rangle \leq \delta^\star(l,U) + \delta^\star(l,V)$$

which shows "$\leq$".
If $u \in U$, $v \in V$, then

$$\langle l, u\rangle + \langle l, v\rangle = \langle l, u+v\rangle \leq \delta^\star(l, U+V)$$

which shows "$\geq$", if both terms on the left-hand side converge to the corresponding suprema $\delta^\star(l,U)$ resp. $\delta^\star(l,V)$.

(ii) If $\lambda = 0$, then $\lambda \cdot U = \{0_{\mathbb{R}^n}\}$ and

$$\delta^\star(l, \lambda \cdot U) = \sup_{x \in \{0_{\mathbb{R}^n}\}} \langle l, x\rangle = \langle l, 0_{\mathbb{R}^n}\rangle = 0$$

which shows equality, even if $\delta^\star(l,U) = \infty$.
Now, let $\lambda > 0$. Let $z \in \lambda \cdot U$, i.e. $z = \lambda \cdot u$ with $u \in U$. Then,

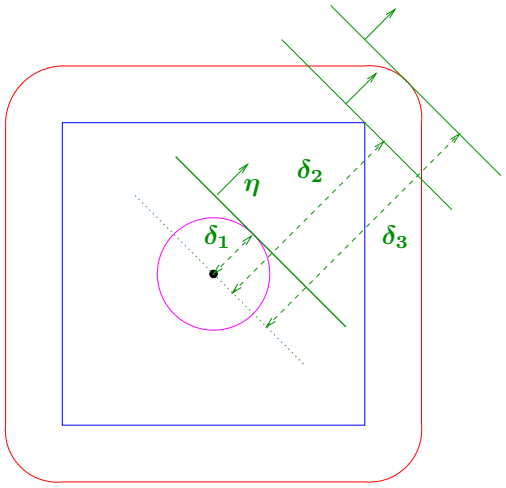
$$\langle l, z\rangle = \lambda \cdot \langle l, u\rangle \leq \lambda \cdot \delta^\star(l,U)$$

(even for $\delta^\star(l,U)=\infty$) which shows "$\leq$", if $\langle l,z\rangle$ approaches the supremum $\delta^\star(l,\lambda\cdot U)$.
"$\geq$": If $u\in U$, then

$$\langle l,u\rangle=\frac{1}{\lambda}\cdot\langle l,\lambda\cdot u\rangle\leq\frac{1}{\lambda}\delta^\star(l,\lambda\cdot U).$$

Approaching with the left-hand side the supremum $\delta^\star(l,U)$ and multiplying with $\lambda>0$ shows $\lambda\cdot\delta^\star(l,U)\leq\delta^\star(l,\lambda\cdot U)$. □

Addition of Support Functions by Minkowski Sums



– convex set $C=[-1,1]^2$

– convex set $D=B_\varepsilon(0)$

– Minkowski sum $C+D$

– corresponding hyperplanes of $C,D,C+D$ in common direction η

--- value of support functions of $C,D,C+D$ in direction η, $\delta_1+\delta_2=\delta_3=\delta^\star(\eta,C+D)$

Proposition 3.116. *Let $U\subset\mathbb{R}^n$ be nonempty, $A\in\mathbb{R}^{m\times n}$ and $l\in\mathbb{R}^n$. Then,*

$$\delta^\star(l,A\cdot U)=\delta^\star(A^\top\cdot l,U).$$

Example 3.117. Let $n\in\mathbb{N}$ and $l=(l_1,\ldots,l_n)^\top\in\mathbb{R}^n$.

(i) point-set $U=\{u\}$ with $u\in\mathbb{R}^n$:

$$\delta^\star(l,U)=\langle l,u\rangle$$

(ii) finite intervall $\mathcal{I}=[a,b]$ with $a\leq b$ and $n=1$:

$$\delta^\star(l,\mathcal{I})=l\cdot b \text{ for } l\geq 0,$$
$$\delta^\star(l,\mathcal{I})=l\cdot a \text{ for } l<0$$

(iii) unit ball $B_1(0)\subset\mathbb{R}^n$:

$$\delta^\star(l,B_1(0))=\|l\|_2$$

(iv) ball $B_r(m)\subset\mathbb{R}^n$ with $m\in\mathbb{R}^n$, $r\geq 0$:

$$\delta^\star(l,B_r(m))=r\cdot\|l\|_2+\langle l,m\rangle$$

Example 3.117 *(continued).*

(v) unit square $[-1,1]^n\subset\mathbb{R}^n$ (unit ball w.r.t. $\|\cdot\|_\infty$):

$$\delta^\star(l,[-1,1]^n)=\|l\|_1=\sum_{i=1}^n|l_i|$$

(vi) unit ball $U\subset\mathbb{R}^n$ w.r.t. $\|\cdot\|_1$, i.e. $U=\operatorname{co}\{\pm e^k\,|\,k=1,\ldots,n\}$:

$$\delta^\star(l,U)=\|l\|_\infty=\max_{k=1,\ldots,n}|l_k|$$

(vii) unit simplex $U = \mathbf{co}\{0_{\mathbb{R}^n}, e^1, \ldots, e^n\} \subset \mathbb{R}^n$:

$$\delta^\star(l, U) = \max(0, \max_{i=1,\ldots,n} l_i)$$

Hence, e.g.

$$\delta^\star(l, P) = \begin{cases} 0, & \text{if } l_i < 0 \text{ for all } i = 1, \ldots, n, \\ l_i, & \text{if } l_j < l_i \text{ for all } j = 1, \ldots, n. \end{cases}$$

(viii) (convex) polytope $P = \mathbf{co}\{p^i : i = 1, \ldots, M\} \subset \mathbb{R}^n$:

$$\delta^\star(l, P) = \max_{i=1,\ldots,M} \langle l, p^i \rangle$$

(ix) ellipsoid $\mathcal{E}(a, Q) := \{x \in \mathbb{R}^n \mid (x-a)^\top Q^{-1}(x-a) \leq 1\} \subset \mathbb{R}^n$ with center $a \in \mathbb{R}^n$ and configuration matrix $Q \in \mathbb{R}^{n \times n}$ which is symmetric and positive definite:

$$\delta^\star(l, \mathcal{E}(a, Q)) = \langle l, a \rangle + \sqrt{\langle l, Ql \rangle}$$

Proposition 3.118. *Let $C_i \in \mathcal{C}(\mathbb{R}^{n_i})$ with $n_i \in \mathbb{N}$, $i = 1, 2$. Then, the support function of $C := C_1 \times C_2 \subset \mathbb{R}^n$ with $n := n_1 + n_2$ in direction $l = \begin{pmatrix} l^1 \\ l^2 \end{pmatrix} \in \mathbb{R}^n$ with $l^i \in \mathbb{R}^{n_i}$, $i = 1, 2$, fulfills:*

$$\delta^\star(l, C) = \delta^\star(l^1, C_1) + \delta^\star(l^2, C_2).$$

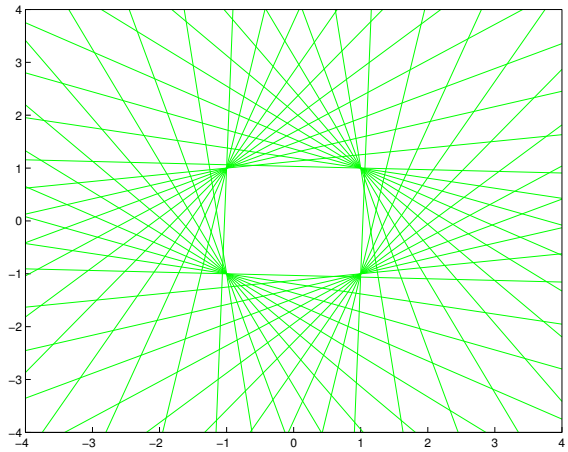
Proof. Proposition 3.7 shows the convexity of C, the equation for the support function follows directly. □

Proposition 3.119. *Let $U, V \subset \mathbb{R}^n$ be nonempty with $U \subset V$. Then,*

$$\delta^\star(l, U) \leq \delta^\star(l, V) \quad \text{for all } l \in \mathbb{R}^n.$$

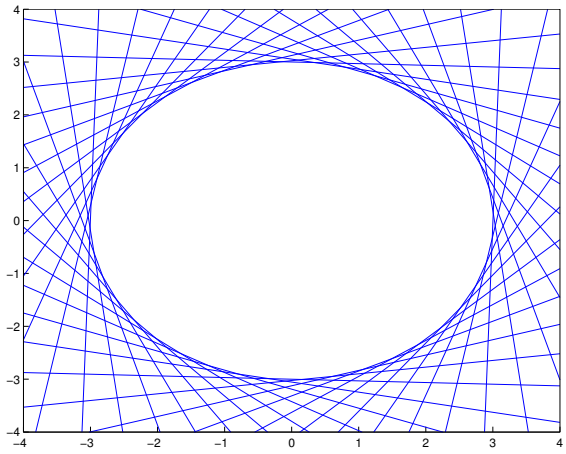
Proposition 3.120. *Let $C, D \in \mathcal{C}(\mathbb{R}^n)$. Then, $C \subset D$ if and only if $\delta^\star(l, C) \leq \delta^\star(l, D)$ for all $l \in S_{n-1}$.*

Ordering of Support Functions by Set Inclusion



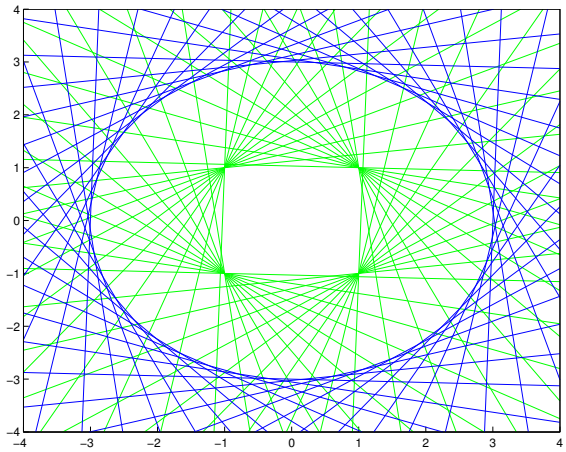
– convex set $C = [-1, 1]^2$
plotted with hyperplanes

Ordering of Support Functions by Set Inclusion



– convex set $D = B_3(0)$
plotted with hyperplanes

Ordering of Support Functions by Set Inclusion



– convex set $C = [-1,1]^2$

– convex set $D = B_3(0)$

In every direction η:

$\delta^\star(\eta, C) \leq \delta^\star(\eta, D)$

which is equivalent to

$C \subset D$

3.2.3 Properties of Supporting Faces

Lemma 3.122. *Let $U \subset \mathbb{R}^n$ be nonempty and $l \in \mathbb{R}^n$. Then,*

$$Y(l,U) \subset Y(l,\operatorname{co}(U)) \quad \textit{and} \quad \operatorname{co}(Y(l,U)) = Y(l,\operatorname{co}(U)).$$

If U is additionally convex, then $Y(l,U)$ is convex.

Lemma 3.123. *Let $U \subset \mathbb{R}^n$ be nonempty and $l \in \mathbb{R}^n$. Then,*

$$Y(l,U) \subset Y(l,\overline{U}).$$

If U is additionally closed, then $Y(l,U)$ is closed, but does not coincide with $\overline{Y(l,U)}$ in general.

Example 3.125. Let $U = \operatorname{co}\{0_{\mathbb{R}^n}, -e^1\} \cup \operatorname{co}\{0_{\mathbb{R}^n}, e^1\} \subset \mathbb{R}^2$ which is a non-convex set and $\eta = \binom{0}{1}$. Then,

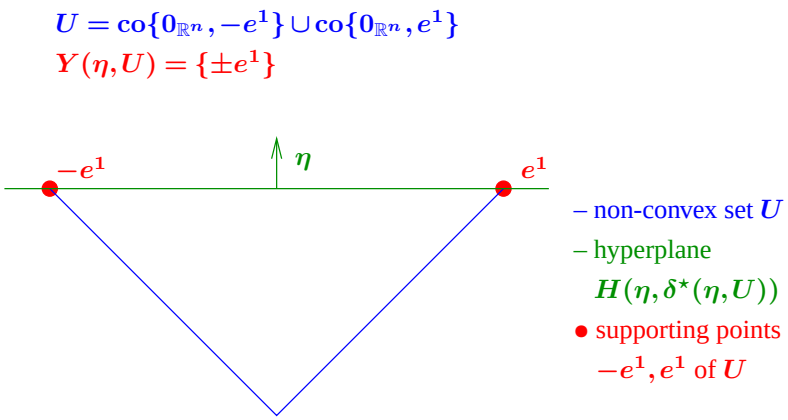
$$\begin{aligned} Y(\eta,U) &= \{-e^1, e^1\}, \\ \operatorname{co}(U) &= \operatorname{co}\{0_{\mathbb{R}^n}, -e^1, e^1\}, \\ Y(\eta,\operatorname{co}(U)) &= \operatorname{co}\{-e^1, e^1\} = \operatorname{co}(Y(\eta,U)). \end{aligned}$$

Let $V = \operatorname{co}\{0_{\mathbb{R}^n}, -e^1\} \cup \{\mu \cdot e^1 \mid \mu \in [0,1)\} \subset \mathbb{R}^2$ which is a non-closed, non-convex set and $\eta = \binom{0}{1}$. Then,

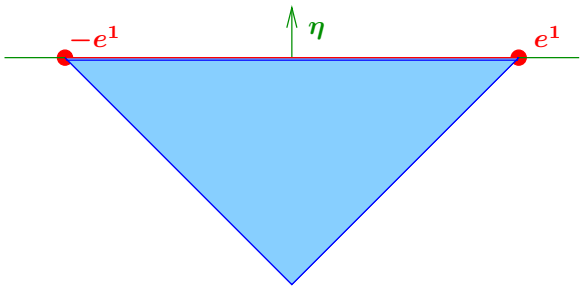
$$\begin{aligned} Y(\eta,V) &= \{-e^1\}, \\ \overline{V} &= U, \\ Y(\eta,\overline{V}) &= \{-e^1, e^1\} \supsetneq \{-e^1\} = \overline{Y(\eta,V)}. \end{aligned}$$

For $W = \operatorname{int} B_1(0)$ follows that $Y(\eta,W) = \emptyset$ for all $\eta \in \mathbb{R}^n$, $\eta \neq 0_{\mathbb{R}^n}$,
but $Y(\eta,\overline{W}) = Y(\eta,B_1(0)) = \{\eta\}$.

Visualization of Example 3.125

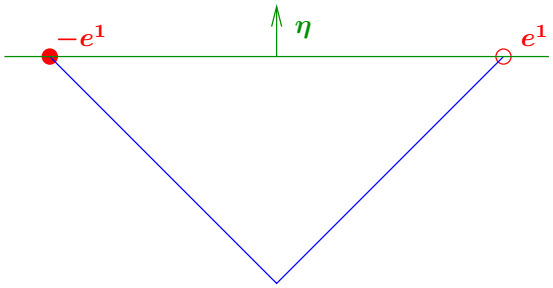


$\mathrm{co}(U) = \mathrm{co}\{0_{\mathbb{R}^n}, -e^1, e^1\}$
$Y(\eta, \mathrm{co}(U)) = \mathrm{co}\{\pm e^1\} = \mathrm{co}(Y(\eta, U))$



– convex set $\mathrm{co}(U)$
– hyperplane
$H(\eta, \delta^\star(\eta, \mathrm{co}(U)))$
– supporting set
$\mathrm{co}\{-e^1, e^1\}$ of $\mathrm{co}(U)$

$V = \mathrm{co}\{0_{\mathbb{R}^n}, -e^1\} \cup \{\mu \cdot e^1 \,|\, \mu \in [0,1)\}$
$Y(\eta, V) = \{-e^1\}, Y(\eta, \overline{V}) = \{\pm e^1\}$
$\overline{Y(\eta, V)} \neq Y(\eta, \overline{V})$



– non-closed set V
– hyperplane
$H(\eta, \delta^\star(\eta, V))$
• supporting point
$-e^1$ of V

Proposition 3.126. *Let $U \subset \mathbb{R}^n$ be nonempty and $l \in \mathbb{R}^n$.*

(i) $Y(\lambda \cdot l, U) = Y(l, U) \quad (\lambda > 0)$ *(positively invariance)*

(ii) $Y(\lambda \cdot l, U) = U \quad (\lambda = 0)$

(iii) $Y(\cdot, U)$ *is convex, if U is additionally convex.*

(iv) The set-valued map $\eta \mapsto Y(\eta, U)$ is u.s.c. (cf. Definition 4.1).

Proposition 3.127. *Let $U, V \subset \mathbb{R}^n$ be nonempty, $\lambda \geq 0$ and $l \in \mathbb{R}^n$. Then,*

$$Y(l, U+V) = Y(l, U) + Y(l, V),$$
$$Y(l, \lambda \cdot U) = \lambda \cdot Y(l, U).$$

Proof. (i) "$\supset$": Let $z \in Y(l,U) + Y(l,V)$, i.e. $z = u + v$ with $u \in Y(l,U)$, $v \in Y(l,V)$. Then,

$$\langle l, z \rangle = \langle l, u \rangle + \langle l, v \rangle = \delta^\star(l, U) + \delta^\star(l, V).$$

"$\subset$": Consider $z = u + v \in U + V$ with $u \in U$, $v \in V$ and $\langle l, z \rangle = \delta^\star(l, U+V)$. Assume that $u \notin Y(l,U)$ or $v \notin Y(l,V)$. Then,

$$\langle l, z \rangle = \langle l, u \rangle + \langle l, v \rangle < \delta^\star(l, U) + \delta^\star(l, V) = \delta^\star(l, U+V)$$

which is a contradiction to $z \in Y(l, U+V)$. Hence, $u \in Y(l,U)$ and $v \in Y(l,V)$ such that z is an element of the right-hand side.
(ii) "$\subset$": The case "$\lambda = 0$" is clear, since

$$Y(l, 0 \cdot U) = Y(l, \{0_{\mathbb{R}^n}\}) = \{0_{\mathbb{R}^n}\} = 0 \cdot Y(l,U).$$

Now, consider $\lambda > 0$.
Let $z \in Y(l, \lambda \cdot U) \subset \lambda \cdot U$, i.e. $z = \lambda \cdot u$ with $u \in U$. Then,

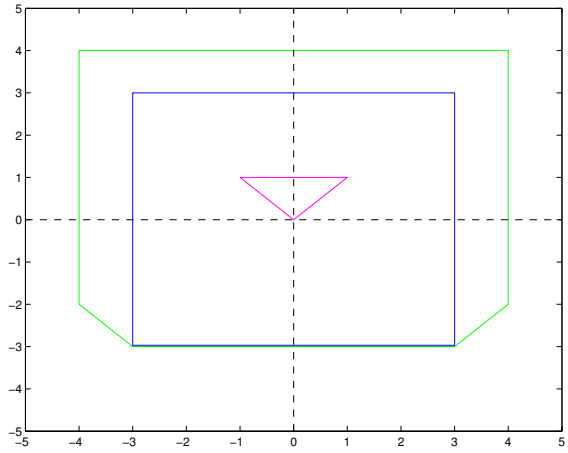
$$\delta^\star(l, \lambda \cdot U) = \langle l, z \rangle = \lambda \cdot \langle l, u \rangle.$$

Since Proposition 3.115 holds, we have $\delta^\star(l,U) = \langle l, u \rangle$, i.e. $u \in Y(l,U)$ and hence $z \in \lambda \cdot Y(l,U)$.
"$\supset$": If $z \in \lambda \cdot Y(l,U)$, then there exists $u \in Y(l,U)$ with $z = \lambda \cdot u$. Then,

$$\langle l, z \rangle = \lambda \cdot \langle l, u \rangle = \lambda \cdot \delta^\star(l,U) = \delta^\star(l, \lambda \cdot U)$$

by Proposition 3.115. Hence, $z \in Y(l, \lambda \cdot U)$. □

Addition of Supporting Faces by Minkowski Sums

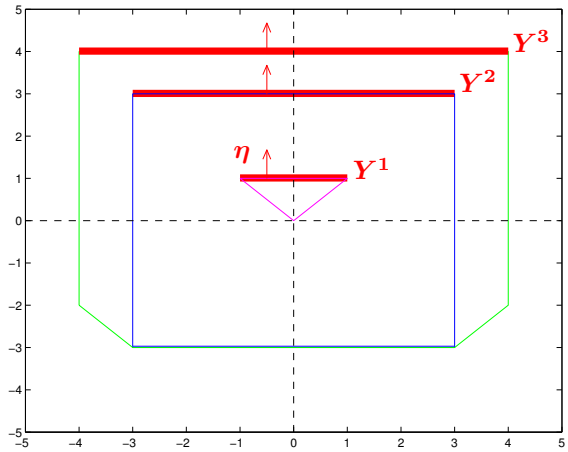


– convex set $C = [-3,3]^2$
plotted with
supporting faces

– convex set D,
$D = \mathrm{co}\{\binom{-1}{1}, \binom{1}{1}, \binom{0}{0}\}$
plotted with
supporting faces

– Minkowski sum $C + D$

Addition of Supporting Faces by Minkowski Sums



– square $C = [-3,3]^2$
– triangle D
– Minkowski sum $C + D$

– supporting faces Y^i in
common direction η:
$Y^1 + Y^2 = Y^3$,
$Y^3 = Y(\eta, C + D)$

Proposition 3.128. *Let $U \subset \mathbb{R}^n$ be nonempty, $A \in \mathbb{R}^{m\times n}$ and $l \in \mathbb{R}^n$. Then,*

$$Y(l, A\cdot U) = A\cdot Y(A^\top \cdot l, U).$$

If you know the supporting face of U in every direction $l \in S_{n-1}$, then you know it also for the set $A \cdot U$.

Example 3.129. Let $n \in \mathbb{N}$ and $l = (l_1, \dots, l_n)^\top \in \mathbb{R}^n$.

(i) point-set $U = \{u\}$ with $u \in \mathbb{R}^n$:

$$Y(l, U) = \{u\}$$

(ii) finite intervall $\mathcal{I} = [a, b]$ with $a \le b$ and $n = 1$:

$$\begin{aligned} Y(l, \mathcal{I}) &= \{b\} \text{ for } l > 0,\\ Y(l, \mathcal{I}) &= \{a\} \text{ for } l < 0,\\ Y(l, \mathcal{I}) &= [a, b] \text{ for } l = 0 \end{aligned}$$

(iii) unit ball $B_1(0) \subset \mathbb{R}^n$, $l \neq 0_{\mathbb{R}^n}$:

$$Y(l, B_1(0)) = \{\frac{1}{\|l\|} \cdot l\}$$

(iv) ball $B_r(m) \subset \mathbb{R}^n$ with $m \in \mathbb{R}^n$, $r \ge 0$ and $l \neq 0_{\mathbb{R}^n}$:

$$Y(l, B_r(m)) = \{r \cdot \frac{1}{\|l\|} \cdot l + m\}$$

Example 3.129 *(continued).*

(v) unit square $[-1, 1]^2 \subset \mathbb{R}^2$ (unit ball w.r.t. $\|\cdot\|_\infty$) and $n = 2$:

$$Y(l, [-1,1]^2) = \begin{cases} \operatorname{co}\{\binom{1}{-1}, \binom{1}{1}\} & \text{for } \varphi = 0,\\ \{\binom{1}{1}\} & \text{for } \varphi \in (0, \frac{\pi}{2}),\\ \operatorname{co}\{\binom{-1}{1}, \binom{1}{1}\} & \text{for } \varphi = \frac{\pi}{2},\\ \{\binom{-1}{1}\} & \text{for } \varphi \in (\frac{\pi}{2}, \pi),\\ \operatorname{co}\{\binom{-1}{-1}, \binom{-1}{1}\} & \text{for } \varphi = \pi,\\ \{\binom{-1}{-1}\} & \text{for } \varphi \in (\pi, \frac{3\pi}{2}),\\ \operatorname{co}\{\binom{-1}{-1}, \binom{1}{-1}\} & \text{for } \varphi = \frac{3\pi}{2},\\ \{\binom{1}{-1}\} & \text{for } \varphi \in (\frac{3\pi}{2}, 2\pi), \end{cases}$$

where $l = \binom{\cos(\varphi)}{\sin(\varphi)}$, $\varphi \in [0, 2\pi)$.

(v') unit square $P = [-1, 1]^n \subset \mathbb{R}^n$ (unit ball w.r.t. $\|\cdot\|_\infty$):
no easy formula available for general n, cf. (viii)

(vi) unit ball $P \subset \mathbb{R}^n$ w.r.t. $\|\cdot\|_1$, i.e. $P = \operatorname{co}\{\pm e^k \mid k = 1, \dots, n\}$:
no easy formula available for general n, cf. (viii)

(vii) unit simplex $P = \operatorname{co}\{0_{\mathbb{R}^n}, e^1, \dots, e^n\} \subset \mathbb{R}^n$:
no easy formula available for general n, cf. (viii), but for special choices of l we have

$$Y(l, P) = \begin{cases} \{0_{\mathbb{R}^n}\}, & \text{if } l_i < 0 \text{ for all } i = 1, \dots, n,\\ \{e^k\}, & \text{if } l_j < l_k \text{ for all } j = 1, \dots, n. \end{cases}$$

(viii) (convex) polytope $P = \operatorname{co}\{p^i : i = 1, \dots, m\} \subset \mathbb{R}^n$, $l \neq 0_{\mathbb{R}^n}$:

$$Y(l, P) = \operatorname{co}\{p^j \mid j = 1, \dots, m \text{ with } \langle \frac{1}{\|l\|} \cdot l, p^j \rangle = \delta^\star(\frac{1}{\|l\|} \cdot l, P)\}$$

(ix) ellipsoid $\mathcal{E}(a,Q) := \{x \in \mathbb{R}^n \mid (x-a)^\top Q^{-1}(x-a) \leq 1\} \subset \mathbb{R}^n$ with center $a \in \mathbb{R}^n$ and configuration matrix $Q \in \mathbb{R}^{n\times n}$ which is symmetric and positive definite:

$$Y(l,\mathcal{E}(a,Q)) = \{\frac{1}{\|Q^{\frac{1}{2}}l\|} \cdot Ql + a\},$$

where $Q^{\frac{1}{2}}$ is the (uniquely defined) square root of the matrix Q which is itself symmetric and positive definite (cf. [Bel70, section 6.5]).

Proposition 3.130. *Let $C_i \in \mathcal{C}(\mathbb{R}^{n_i})$ with $n_i \in \mathbb{N}$, $i = 1,2$. Then, the supporting face of $C := C_1 \times C_2 \subset \mathbb{R}^n$ with $n := n_1 + n_2$ in direction $l = \begin{pmatrix} l^1 \\ l^2 \end{pmatrix} \in \mathbb{R}^n$ with $l^i \in \mathbb{R}^{n_i}$, $i = 1,2$, fulfills:*

$$Y(l,C) = Y(l^1,C_1) \times Y(l^2,C_2).$$

Proof. This follows from Propositions 3.7 and 3.118 by direct calculations. □

Proposition 3.131. *Let $U, V \subset \mathbb{R}^n$ be nonempty with $U \subset V$ and $l \in \mathbb{R}^n$ with $\delta^\star(l,U) = \delta^\star(l,V)$. Then,*

$$Y(l,U) \subset Y(l,V).$$

Corollary 3.132. *If $U \subset \mathbb{R}^n$ is included in $B_r(m)$ with $m \in \mathbb{R}^n$, $r \geq 0$. Then,*

$$\|Y(l,U) - m\| \leq r.$$

3.2.4 Metrics for Sets

Definition 3.134. Let $V \subset \mathbb{R}^n$ be nonempty and $x \in \mathbb{R}^n$. Then,

$$\mathbf{dist}(x, V) := \inf_{v \in V} \|x - v\| \in [0, \infty)$$

is called the *distance* of x to V.
Let additionally $U \subset \mathbb{R}^n$ be nonempty. Then,

$$\mathbf{d}(U, V) := \sup_{u \in U} \mathbf{dist}(u, V) \in [0, \infty) \cup \{+\infty\}$$

denotes the *one-sided distance* of U to V.
The value

$$\mathbf{d_H}(U, V) := \max\{\mathbf{d}(U, V), \mathbf{d}(V, U)\} \in [0, \infty) \cup \{+\infty\}$$

is called the *Hausdorff distance* of U and V.

Lemma 3.135. *Let $u \in \mathbb{R}^n$ and $U, V \subset \mathbb{R}^n$ be nonempty. Then,*

$$\begin{aligned}
\mathbf{dist}(u, V) &= \inf\{\varepsilon > 0 \,|\, u \in V + \varepsilon B_1(0)\},\\
\mathbf{d}(U, V) &= \inf\{\varepsilon > 0 \,|\, U \subset V + \varepsilon B_1(0)\},\\
\mathbf{d_H}(U, V) &= \inf\{\varepsilon > 0 \,|\, U \subset V + \varepsilon B_1(0), V \subset U + \varepsilon B_1(0)\}.
\end{aligned}$$

Proof. The first formula is a special case of the second formula. Set

$$\begin{aligned}
d_1 &:= \sup_{u \in U} \mathbf{dist}(u, V),\\
d_2 &:= \inf\{\varepsilon > 0 \,|\, U \subset V + \varepsilon B_1(0)\}.
\end{aligned}$$

1. case: $d_1 = \infty$
Then, there exists $(u^m)_m \subset U$ with $\mathbf{dist}(u^m, V) > m$. Clearly, $\|u^m - v\| \geq \mathbf{dist}(u^m, V) > m$ for all $v \in V$.

The following reformulations are valid for all $v \in V$:

$$\begin{aligned}
u^m - v &\notin B_m(0),\\
u^m &\notin v + B_m(0),\\
u^m &\notin V + m \cdot B_1(0)
\end{aligned}$$

Since some element (namely u^m) is not included in the right-hand side, $U \not\subset V + m \cdot B_1(0)$ and $d_2 \geq m$. This shows that $d_2 = \infty$.
2. case: $d_2 = \infty$
Then, there exists a sequence $(\varepsilon_m)_m$ tending to infinity with $\varepsilon_m > 0$ and $U \not\subset V + \varepsilon_m B_1(0)$. Hence, there exists $(u^m)_m \subset U$ with

$$u^m \notin V + \varepsilon_m \cdot B_1(0).$$

Hence, for all $v \in V$

$$\begin{aligned}
u^m &\notin v + \varepsilon_m \cdot B_1(0),\\
u^m - v &\notin \varepsilon_m B_1(0),\\
\|u^m - v\| &> \varepsilon_m
\end{aligned}$$

and $\mathbf{dist}(u^m, V) \geq \varepsilon_m$. This shows that $d_1 = \infty$.
3. case $d_1, d_2 < \infty$:
For $u \in U$ we have $\mathbf{dist}(u, V) \leq \sup_{u \in U} \mathbf{dist}(u, V) = d_1$. Hence, there exists $(v^m)_m \subset V$ with

$$\mathbf{dist}(u, V) \leq \|u - v^m\| \leq \mathbf{dist}(u, V) + \frac{1}{m} \leq d_1 + \frac{1}{m}.$$

Hence, the following reformulations are valid:

$$\begin{aligned} u - v^m &\in (d_1 + \frac{1}{m})B_1(0), \\ u &\in v^m + (d_1 + \frac{1}{m})B_1(0) \subset V + (d_1 + \frac{1}{m})B_1(0), \\ u &\in V + \bigcap_{m\in\mathbb{N}} (d_1 + \frac{1}{m})B_1(0) \end{aligned}$$

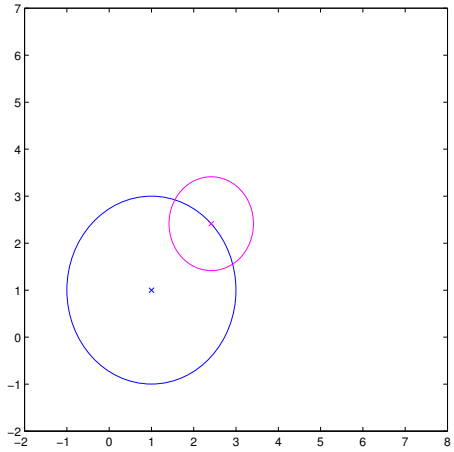
Since $(d_1 + \frac{1}{m})_m$ converges monotone to d_1, the intersection gives simply $d_1 B_1(0)$, so that $u \in V + d_1 B_1(0)$ and hence, $d_2 \leq d_1$.

Let $(\varepsilon_m)_m$ a monotone decreasing sequence converging to d_2 with $U \subset V + \varepsilon_m B_1(0)$. Hence, for all $u \in U$ we have $u \in V + \varepsilon_m B_1(0)$ so that there exists $v^m \in V$ with $u \in v_m + \varepsilon_m B_1(0)$. The following reformulations are valid:

$$\begin{aligned} u - v_m &\in \varepsilon_m B_1(0), \\ \|u - v_m\| &\leq \varepsilon_m, \\ \mathbf{dist}(u, V) &\leq \|u - v_m\| \leq \varepsilon_m \end{aligned}$$

This shows $\mathbf{dist}(u, V) \leq d_2$ and hence, $d_1 \leq d_2$.
In all three cases we have shown that $d_1 = d_2$.
The third equality follows immediately from $\mathbf{d_H}(U, V) = \mathbf{max}\{\mathrm{d}(U, V), \mathrm{d}(V, U)\}$ and the second equality. □

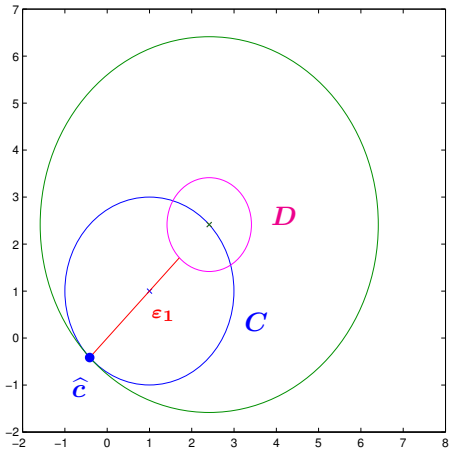
Visualization of Hausdorff Distance



– convex set $C = B_2(\binom{1}{1})$

– convex set $D = B_1(\binom{1+\sqrt{2}}{1+\sqrt{2}})$

Visualization of Hausdorff Distance



– convex set $C = B_2(\binom{1}{1})$

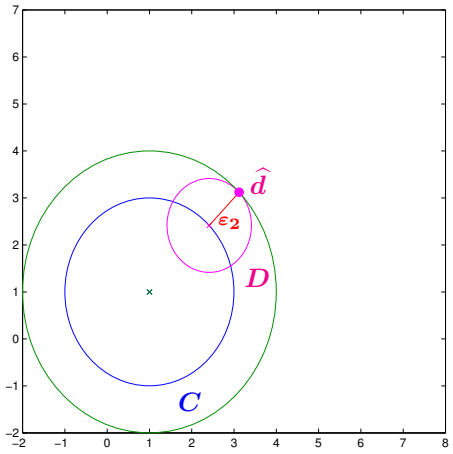
– convex set $D = B_1(\binom{1+\sqrt{2}}{1+\sqrt{2}})$

• farthest point $\widehat{c}$ from D:
$\mathbf{dist}(c, D) = \mathrm{d}(C, D) = \varepsilon_1$

– (one-sided) Hausdorff distance
$\varepsilon_1 = \mathrm{d}(C, D) = \mathrm{d_H}(C, D)$

– suitable neighbourhood V of D:
$C \subset D + \varepsilon_1 B_1(0) = V$

Visualization of Hausdorff Distance



– convex set $C = B_2(\binom{1}{1})$

– convex set $D = B_1(\binom{1+\sqrt{2}}{1+\sqrt{2}})$

• farthest point $\widehat{d}$ from C:
$\mathbf{dist}(c, D) = \mathrm{d}(C, D) = \varepsilon_2$

– one-sided Hausdorff distance
$\varepsilon_2 = \mathrm{d}(D, C)$

– suitable neighbourhood U of C:
$D \subset C + \varepsilon_2 B_1(0) = U$

Definition 3.136. Let $U \subset \mathbb{R}^n$ be nonempty. Then,

$$\|U\| := \sup_{u \in U} \|u\| \in [0, \infty) \cup \{+\infty\}$$

defines the *norm* of U. The *diameter* of U is defined as

$$\mathbf{diam}(U) := \sup_{u,v \in U} \|u - v\| \in [0, \infty) \cup \{+\infty\}$$

The distances and the norm are generalization of the norm distance resp. norm of vectors.

Lemma 3.137. *Let $x, y \in \mathbb{R}^n$ and $U \subset \mathbb{R}^n$ be nonempty. Then,*

(i) $\mathrm{d_H}(\{x\}, \{y\}) = \mathrm{d}(\{x\}, \{y\}) = \mathbf{dist}(x, \{y\}) = \|x - y\|$

(ii) $\mathrm{d}(\{x\}, U) = \mathbf{dist}(x, U)$

(iii) $\|\{x\}\| = \|x\|$

Lemma 3.138. *Let $U, V \subset \mathbb{R}^n$ be nonempty, $x \in \mathbb{R}^n$. Then,*

(i) $\mathbf{dist}(0_{\mathbb{R}^n}, U) = \inf_{u\in U} \|u\|$

(ii) $\mathbf{dist}(x,U) \leq \|x-u\| \leq \|x\| + \|u\| < \infty$ *for all* $u \in U$
Furthermore, $\mathbf{dist}(x,U) \leq \|x\| + \mathbf{dist}(0_{\mathbb{R}^n}, U) = \|x\| + \inf_{u\in U} \|u\|$.

(iii) $\mathbf{dist}(x,U) \leq \|x\| + \|U\| = \|x\| + \sup_{u\in U} \|u\|$, *but* $\mathbf{dist}(x,U) < \infty$, *even if* U *is unbounded*

(iv) $\mathbf{d}(U,V) \leq \mathbf{d_H}(U,V) \leq \|U\| + \|V\|$

(v) *If* $U \subset V$ *with* $V \subset \mathbb{R}^n$, *then* $\|U\| \leq \|V\|$.

(vi) *If* $U \subset B_r(0)$, $V \subset B_s(0)$ *with* $r, s \geq 0$, *then*

$$\mathbf{d}(U,V) \leq \mathbf{d_H}(U,V) \leq r + s < \infty.$$

The next lemma lists basic properties of the distance, when considering set inclusion, closure or convex hull.

Lemma 3.139. *Let* $V, W \subset \mathbb{R}^n$ *be nonempty with* $W \supset V$ *and* $u \in \mathbb{R}^n$. *Then,*

(i) $\mathbf{dist}(u,W) \leq \mathbf{dist}(u,V)$

(ii) $\mathbf{dist}(u,\overline{V}) = \mathbf{dist}(u,V)$

(iii) $\mathbf{dist}(u, \mathbf{co}(V)) \leq \mathbf{dist}(u,V)$

Proof. (i) Clearly, the infimum decreases:

$$\mathbf{dist}(u,W) = \inf_{w\in W} \|u-w\| \leq \inf_{w\in V} \|u-w\| = \mathbf{dist}(u,V)$$

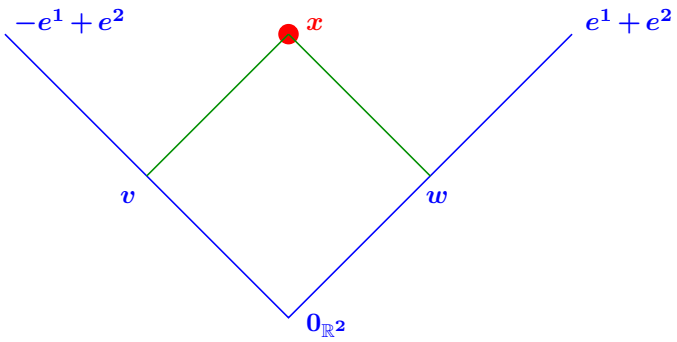
(ii) Since $V \subset \overline{V}$, the inequality "$\leq$" follows by (i).
Let $(w^m)_m \subset \overline{V}$ with $\mathbf{dist}(u,\overline{V}) \leq \|u-w^m\| \leq \mathbf{dist}(u,\overline{V}) + \frac{1}{2m}$ for $m \in \mathbb{N}$. Choose $(v^m)_m \subset V$ with $\|v^m - w^m\| \leq \frac{1}{2m}$ (V is dense in $\overline{V}$).
Then,

$$\begin{aligned}\mathbf{dist}(u,V) &\leq \|u-v^m\| \leq \|u-w^m\| + \|w^m - v^m\| \\ &\leq (\mathbf{dist}(u,\overline{V}) + \frac{1}{2m}) + \frac{1}{2m} = \mathbf{dist}(u,\overline{V}) + \frac{1}{m}\end{aligned}$$

for all $m \in \mathbb{N}$ showing the wanted result.
(iii) Since $V \subset \mathbf{co}(V)$, the inequality "$\leq$" follows by (i). □

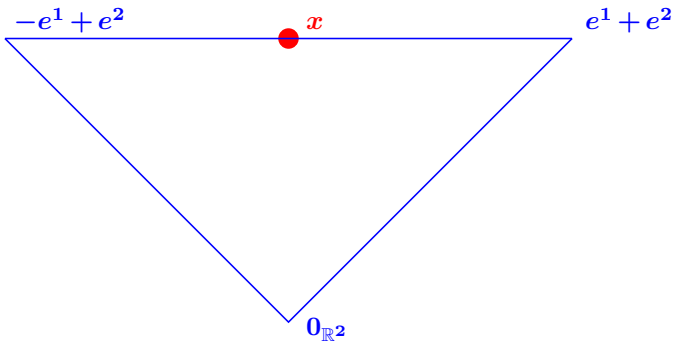
Example 3.140. There are examples for which strict inequality appear in Lemma 3.139(i) and (iii).

Distance of a Point to a Set Decreases by Convexifying



– non-convex set $U = \mathbf{co}\{-e^1+e^2, 0_{\mathbb{R}^2}\} \cup \mathbf{co}\{e^1+e^2, 0_{\mathbb{R}^2}\}$
• point x
– nearest distance from x to U: $\mathbf{dist}(x,U) = \frac{\sqrt{2}}{2}$
nearest point of x in U is non-unique (here: $v, w \in U$)

Distance of a Point to a Set Decreases by Convexifying



– convexified set $\mathbf{co}(U) = \mathbf{co}\{-e^1+e^2, 0_{\mathbb{R}^2}, e^1+e^2\}$

• point x

– nearest distance from x to $\mathbf{co}(U)$: $\mathbf{dist}(x, \mathbf{co}(U)) = 0$

nearest point of x in $\mathbf{co}(U)$ is x itself, since $x \in \mathbf{co}(U)$

The next two lemmas state that the distance of a point to a set behaves like a normal distance. Furthermore, they show estimations of the distance, when applied to set arithmetic.

Lemma 3.142. *Let $U, V, W, Z \subset \mathbb{R}^n$ be nonempty, $u \in U$, $v \in V$ and $\mu \in \mathbb{R}$. Then,*

(i) *If* $\mathbf{dist}(u, V) = 0$, *then* $u \in \overline{V}$ *and vice versa.*

(ii) $\mathbf{dist}(u, W) \leq \|u - v\| + \mathbf{dist}(v, W)$ *for all* $v \in V$
$\mathbf{dist}(u, W) \leq \mathbf{dist}(u, V) + \mathbf{d}(V, W)$

(iii) $\mathbf{dist}(u+v, W+Z) \leq \mathbf{dist}(u, W) + \mathbf{dist}(v, Z)$

We prove only (ii). The other ones are (also) rather trivial.
(ii) Let $v \in V$, $w \in W$. Then,

$$\mathbf{dist}(u, W) \leq \|u - w\| \leq \|u - v\| + \|v - w\|.$$

Choose $(w^m)_m \subset W$ with $\|v - w^m\| \underset{m\to\infty}{\searrow} \mathbf{dist}(v, W)$, then

$$\mathbf{dist}(u, W) \leq \|u - v\| + \mathbf{dist}(v, W).$$

For the second estimation, observe that

$$\mathbf{dist}(u, W) \leq \|u - v\| + \sup_{v \in V} \mathbf{dist}(v, W) = \|u - v\| + \mathbf{d}(V, W).$$

Choose $(v^m)_m \subset V$ with $\|u - v^m\| \underset{m\to\infty}{\searrow} \mathbf{dist}(u, V)$, then

$$\mathbf{dist}(u, W) \leq \mathbf{dist}(u, V) + \mathbf{d}(V, W).$$

□

Lemma 3.143. *Let $V \subset \mathbb{R}^n$ be nonempty, $u \in \mathbb{R}^n$, $A \in \mathbb{R}^{m\times n}$ and $\mu \in \mathbb{R}$. Then,*

(i) $\mathbf{dist}(\mu \cdot u, \mu \cdot V) = |\mu| \cdot \mathbf{dist}(u, V)$

(ii) $\mathbf{dist}(A \cdot u, A \cdot V) \leq \|A\| \cdot \mathbf{dist}(u, V)$, *where $\|\cdot\|$ is a matrix norm compatible with the Euclidean vector norm $\|\cdot\|$ (see Appendix, e.g. Example A.4).*

Proposition 3.145. *Let $A \subset \mathbb{R}^n$ be closed and nonempty. Then, $x \mapsto \mathbf{dist}(x, A)$ is Lipschitz continuous on $\mathbb{R}^n$ with constant $L \leq 1$.*

Proof. Let $x^1, x^2 \in \mathbb{R}^n$ and $\widehat{a}^1, \widehat{a}^2 \in A$ one of the existing best approximations of x^1 resp. x^2 in A (they exist by the assumptions on A, but are in general non-unique). Hence, $\mathbf{dist}(x^1, A) = \|x^1 - \widehat{a}^1\|$ and $\mathbf{dist}(x^2, A) = \|x^2 - \widehat{a}^2\|$. Now, the optimality of $\widehat{a}^1$ for x^1 shows

$$\begin{aligned} \|x^1 - \widehat{a}^1\| \leq \|x^1 - \widehat{a}^2\| &\leq \|x^1 - x^2\| + \|x^2 - \widehat{a}^2\|, \\ \mathbf{dist}(x^1, A) - \mathbf{dist}(x^2, A) &\leq \|x^1 - x^2\|. \end{aligned}$$

In the same spirit, the converse estimation

$$\mathbf{dist}(x^2, A) - \mathbf{dist}(x^1, A) \leq \|x^1 - x^2\|$$

could be proven. □

The next lemma shows the connections of norm and diameter to the distances.

Lemma 3.146. *Let $x \in \mathbb{R}^n$ and $U \subset \mathbb{R}^n$ be nonempty. Then,*

(i) $\|U\| = \mathrm{d}(U, \{0_{\mathbb{R}^n}\}) = \mathbf{d_H}(U, \{0_{\mathbb{R}^n}\})$

(ii) $\mathrm{d}(U, \{x\}) = \|U - x\|$

(iii) $\mathbf{dist}(0_{\mathbb{R}^n}, U) = \mathrm{d}(\{0_{\mathbb{R}^n}\}, U) \leq \mathrm{d}(U, \{0_{\mathbb{R}^n}\}) = \|U\|$

(iv) $\mathbf{diam}(U) = \|U - U\|$

The next lemma is only a technical one and is used below only once.

Lemma 3.147. *Let $U \in \mathcal{K}(\mathbb{R}^n)$ (cf. Definition 3.2), $(u^m)_m \subset \mathbb{R}^n$ and set $U_m := \{u^m\} \in \mathcal{C}(\mathbb{R}^n)$. If $(U_m)_m$ is a sequence converging to U, then the limit of $(u^m)_m$ exists in $\mathbb{R}^n$ and $U = \{u\}$.*

Assume there exists $w \in U \setminus \{u\}$. For $\widetilde{\varepsilon} = \frac{\|w-u\|}{4}$ choose $k \geq \max\{M(\widetilde{\varepsilon}), \widetilde{M}(\widetilde{\varepsilon})\}$. From (**??**) follows

$$\begin{aligned} w \in \{u^{m_k}\} + \widetilde{\varepsilon} B_1(0) &= \{u\} + \{u^{m_k} - u\} + \widetilde{\varepsilon} B_1(0) \\ &\subset \{u\} + B_{\widetilde{\varepsilon}}(0) + B_{\widetilde{\varepsilon}}(0) = \{u\} + 2\widetilde{\varepsilon} B_1(0) \end{aligned}$$

and $\|w - u\| \leq 2\widetilde{\varepsilon} \leq \frac{\|w-u\|}{2}$. This leads to the contradiction $w = u$. Therefore, $U = \{u\}$ and from (**??**) follows the convergence of the complete sequence $(u^m)_m$, since

$$\{u\} \subset \{u^m\} + \varepsilon B_1(0) \quad \text{and} \quad \|u - u^m\| \leq \varepsilon.$$

□

We generalize the former lemmas to the one- and two-sided Hausdorff-distance.

Proposition 3.149. *Let $U, V, W, Z \subset \mathbb{R}^n$ be nonempty, $A \in \mathbb{R}^{m \times n}$ and $\mu \in \mathbb{R}$. Then,*

(i) If additionally $V \subset W$, then $\mathrm{d}(U, W) \leq \mathrm{d}(U, V)$ and $\mathrm{d}(V, U) \leq \mathrm{d}(W, U)$.

(ii) $\mathrm{d}(\overline{U}, V) = \mathrm{d}(U, \overline{V}) = \mathrm{d}(U, V)$

(iii) $\mathrm{d}(U, \mathrm{co}(V)) \leq \mathrm{d}(U, V) \leq \mathrm{d}(\mathrm{co}(U), V)$
$\mathrm{d}(\mathrm{co}(U), \mathrm{co}(V)) \leq \mathrm{d}(\mathrm{co}(U), V)$

(iv) If $\mathrm{d}(U, V) = 0$, then $U \subset \overline{V}$ and vice versa.

(v) $\mathrm{d}(\mu \cdot U, \mu \cdot V) = |\mu| \cdot \mathrm{d}(U, V)$

(vi) $\mathrm{d}(A \cdot U, A \cdot V) \leq \|A\| \cdot \mathrm{d}(U, V)$

(vii) $\mathrm{d}(U, W) \leq \mathrm{d}(U, V) + \mathrm{d}(V, W)$

(viii) $\mathbf{d}(U+V, W+Z) \leq \mathbf{d}(U,W) + \mathbf{d}(V,Z)$

The Hausdorff distance is a metric on the set of all nonempty, bounded, closed subsets of $\mathbb{R}^n$.

Proposition 3.150. *Let $U, V, W, Z \subset \mathbb{R}^n$ be nonempty, $A \in \mathbb{R}^{m\times n}$ and $\mu \in \mathbb{R}$. Then,*

(i) $\mathbf{d_H}(U,V) = \mathbf{d_H}(V,U)$

(ii) $\mathbf{d_H}(\overline{U},V) = \mathbf{d_H}(U,\overline{V}) = \mathbf{d_H}(U,V)$

(iii) $\mathbf{d_H}(\mathbf{co}(U),\mathbf{co}(V)) \leq \mathbf{d_H}(U,V)$
$\mathbf{d_H}(\mathbf{co}(U),\mathbf{co}(V)) \leq \max\{\mathbf{d}(\mathbf{co}(U),V), \mathbf{d}(\mathbf{co}(V),U)\}$

(iv) *If $\mathbf{d_H}(U,V) = 0$, then $\overline{U} = \overline{V}$ and vice versa.*

(v) $\mathbf{d_H}(\mu\cdot U, \mu\cdot V) = |\mu|\cdot \mathbf{d_H}(U,V)$

(vi) $\mathbf{d_H}(A\cdot U, A\cdot V) \leq \|A\|\cdot \mathbf{d_H}(U,V)$

(vii) $\mathbf{d_H}(U,W) \leq \mathbf{d_H}(U,V) + \mathbf{d_H}(V,W)$

(viii) $\mathbf{d_H}(U+V, W+Z) \leq \mathbf{d_H}(U,W) + \mathbf{d_H}(V,Z)$

Corollary 3.151. *Let $C, D \subset \mathbb{R}^n$ be convex, compact. Then,*

$$\begin{aligned}
&\mathbf{d}(\partial C, D) \leq \mathbf{d}(\partial C, \partial D) \leq \mathbf{d}(C, \partial D),\\
&\mathbf{d_H}(C,D) \leq \mathbf{d_H}(\partial C, \partial D),\\
&\mathbf{d_H}(C,D) \leq \max\{\mathbf{d}(C,\partial D), \mathbf{d}(D,\partial C)\}.
\end{aligned}$$

Proof. follows from Propositions 3.149(iii), 3.150(iii) and 3.67 □

The convexification might also change the one-sided and two-sided Hausdorff-distance.

Example 3.152. Let $U := \mathbf{co}\{-e^1+e^2, 0_{\mathbb{R}^n}\} \cup \mathbf{co}\{0_{\mathbb{R}^n}, e^1+e^2\}$ and $V := \mathbf{co}\{-e^1+2e^2, 3e^2\} \cup \mathbf{co}\{3e^2, e^1+2e^2\}$ be subsets of $\mathbb{R}^n$ with e^k the k-th unit vector in $\mathbb{R}^n$.
Then,

$$\begin{aligned}
\mathbf{d}(U,V) &= \mathbf{dist}(0_{\mathbb{R}^n}, V) = \|0_{\mathbb{R}^n} - (e^1+2e^2)\| = \sqrt{5},\\
\mathbf{d}(U,\mathbf{co}(V)) &= \mathbf{d}(\mathbf{co}(U),\mathbf{co}(V)) = \mathbf{dist}(0_{\mathbb{R}^n}, \mathbf{co}(V)) = \|0_{\mathbb{R}^n} - \underbrace{2e^2}_{\in \mathbf{co}(V)}\| = 2,\\
\mathbf{d}(V,U) &= \mathbf{dist}(3e^2, U) = \|3e^2 - (e^1+e^2)\| = \sqrt{5},\\
\mathbf{d}(V,\mathbf{co}(U)) &= \mathbf{d}(\mathbf{co}(V),\mathbf{co}(U)) = \mathbf{dist}(3e^2, \mathbf{co}(U)) = \|3e^2 - \underbrace{e^2}_{\in \mathbf{co}(U)}\| = 2,\\
\mathbf{d_H}(U,V) &= \sqrt{5},\\
\mathbf{d_H}(\mathbf{co}(U),\mathbf{co}(V)) &= 2.
\end{aligned}$$

Hence, $\mathbf{d}(U,\mathbf{co}(V)) < \mathbf{d}(U,V)$ and $\mathbf{d_H}(\mathbf{co}(U),\mathbf{co}(V)) < \mathbf{d_H}(U,V)$.
Another example in $\mathbb{R}^2$ would be $U = S_1$ and $V = \partial([-1,1]^2)$. Clearly, $\mathbf{co}(U) = B_1(0)$, $\mathbf{co}(V) = [-1,1]^2$ and

$$\begin{aligned}
\mathbf{d}(U,V) &= \mathbf{dist}(\frac{1}{\sqrt{2}}\cdot\begin{pmatrix}1\\1\end{pmatrix}, V) = \|\frac{1}{\sqrt{2}}\cdot\begin{pmatrix}1\\1\end{pmatrix} - \begin{pmatrix}\frac{1}{\sqrt{2}}\\1\end{pmatrix}\| = 1 - \frac{1}{\sqrt{2}},\\
\mathbf{d}(\mathbf{co}(U),V) &= \mathbf{dist}(0_{\mathbb{R}^2}, V) = 1,\\
\mathbf{d}(U,\mathbf{co}(V)) &= \mathbf{d}(\mathbf{co}(U),\mathbf{co}(V)) = 0,
\end{aligned}$$

since $U \subset \mathbf{co}(V)$.
Hence, $\mathbf{d}(U,\mathbf{co}(V)) < \mathbf{d}(U,V) < \mathbf{d}(\mathbf{co}(U),V)$.

Proposition 3.153. *Let $U \subset \mathbb{R}^n$ be nonempty and $A, B \in \mathbb{R}^{m\times n}$. Then,*

(i) $\mathbf{d}(A\cdot U, B\cdot U) \leq \|A-B\|\cdot\|U\|$

(ii) $\mathbf{d_H}(A\cdot U, B\cdot U) \leq \|A-B\|\cdot\|U\|$

(iii) $\mathbf{d}(A\cdot U+B\cdot U, (A+B)U) \leq \|A-B\|\cdot\|U\|$, *if* U *is additionally convex.*

(iv) $\mathbf{d_H}(A\cdot U+B\cdot U, (A+B)U) \leq \|A-B\|\cdot\|U\|$, *if* U *is additionally convex.*

Especially, for $\lambda, \mu \in \mathbb{R}$ *we have the estimations*

(v) $\mathbf{d}(\lambda\cdot U, \mu\cdot U) \leq |\lambda-\mu|\cdot\|U\|$

(vi) $\mathbf{d_H}(\lambda\cdot U, \mu\cdot U) \leq |\lambda-\mu|\cdot\|U\|$

(vii) $\mathbf{d}(\lambda\cdot U+\mu\cdot U, (\lambda+\mu)\cdot U) \leq |\lambda-\mu|\cdot\|U\|$, *if* U *is additionally convex.*

(viii) $\mathbf{d_H}(\lambda\cdot U+\mu\cdot U, (\lambda+\mu)\cdot U) \leq |\lambda-\mu|\cdot\|U\|$, *if* U *is additionally convex.*

Proof. (i)–(ii): Let $z \in A\cdot U$. Then, there exists $u \in U$ with $z = Au$. An easy estimation is given by $Bu \in B\cdot U$ with

$$\begin{aligned}\mathbf{dist}(z, B\cdot U) &\leq \|Au - Bu\| = \|(A-B)u\| \\ &\leq \|A-B\|\cdot\|u\| \leq \|A-B\|\cdot\|U\|.\end{aligned}$$

Taking the supremum over all z and interchanging A and B gives the two estimates.
(iii)–(iv): From Proposition 3.88(iv) follows $(A+B)U \subset AU+BU$ so that $\mathbf{d}((A+B)U, AU+BU) = 0$.

Let $z \in A\cdot U + B\cdot U$. Then, there exists $u, v \in U$ with $z = Au + Bv$. z can be rewritten in the form

$$\begin{aligned}Au + Bv &= A\cdot\frac{u+v}{2} + B\cdot\frac{u+v}{2} + A\cdot\frac{u-v}{2} + B\cdot\frac{v-u}{2} \\ &= (A+B)\cdot\frac{u+v}{2} + (A-B)\cdot\frac{u-v}{2}\end{aligned}$$

and a simple estimation with the midpoint $\frac{u+v}{2} \in U$ yields

$$\begin{aligned}\mathbf{dist}(z, (A+B)U) &\leq \|(Au+Bv) - (A+B)\cdot\frac{u+v}{2}\| \\ &= \|(A-B)\cdot\frac{u-v}{2}\| \leq \frac{\|A-B\|}{2}\cdot\|u-v\| \\ &\leq \frac{\|A-B\|}{2}\cdot\mathbf{diam}(U) \leq \|A-B\|\cdot\|U\|.\end{aligned}$$

Hereby, Corollary 3.155(viii) was used. Taking the supremum over all z gives the estimate.
(v)–(viii) follow from Remark 3.89. □

$\|\cdot\|$ is a norm for the set of all nonempty, bounded subsets of $\mathbb{R}^n$.

Corollary 3.154. *Let* $U, V \subset \mathbb{R}^n$ *be nonempty,* $A \in \mathbb{R}^{m\times n}$ *and* $\mu \in \mathbb{R}$. *Then,*

(i) $\|U\| \in [0,\infty)\cup\{+\infty\}$
If $U \subset B_r(m)$ *with* $r \geq 0$, $m \in \mathbb{R}^n$, *then* $\|U-m\| \leq r$ *and* $\|U\| \leq r + \|m\| < \infty$.
If $\|U\| < \infty$, *then* U *is bounded by* $B_r(0_{\mathbb{R}^n})$ *with* $r = \|U\|$.

(ii) $\|U\| = 0$, *if and only if* $U = \{0_{\mathbb{R}^n}\}$

(iii) $\|\mu\cdot U\| = |\mu|\cdot\|U\|$

(iv) $\|A\cdot U\| \leq \|A\|\cdot\|U\|$

(v) $\|U+V\| \leq \|U\| + \|V\|$

(vi) If $U \subset V$, *then* $\|U\| \leq \|V\|$.

(vii) $\|U\| = \|\overline{U}\|$

The diameter of a set has a lot of properties of a norm.

Corollary 3.155. *Let $U, V \subset \mathbb{R}^n$ be nonempty, $A \in \mathbb{R}^{m\times n}$ and $\mu \in \mathbb{R}$. Then,*

(i) $\mathbf{diam}(U) \in [0,\infty) \cup \{+\infty\}$
If $U \subset B_r(m)$ with $r \geq 0$, $m \in \mathbb{R}^n$, then $\mathbf{diam}(U) \leq 2r < \infty$.

(ii) $\mathbf{diam}(U) = 0$, *if and only if* $U = \{u\}$

(iii) $\mathbf{diam}(\mu \cdot U) = |\mu| \cdot \mathbf{diam}(U)$

(iv) $\mathbf{diam}(A \cdot U) \leq \|A\| \cdot \mathbf{diam}(U)$

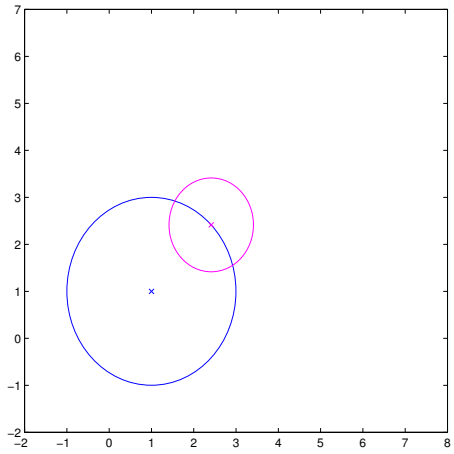
(v) $\mathbf{diam}(U + V) \leq \mathbf{diam}(U) + \mathbf{diam}(V)$

(vi) If $U \subset V$, then $\mathbf{diam}(U) \leq \mathbf{diam}(V)$.

(vii) $\mathbf{diam}(U) = \mathbf{diam}(\overline{U})$

(viii) $\mathbf{diam}(U) \leq 2 \cdot \|U\|$

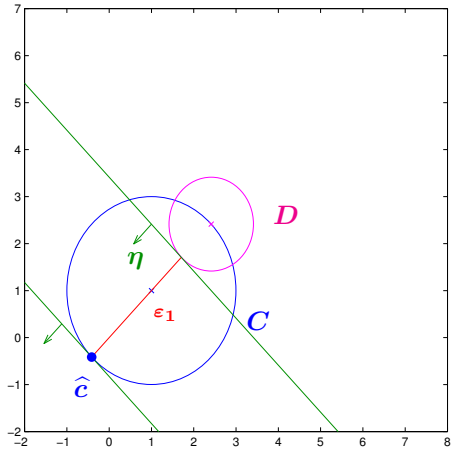
Visualization of Hausdorff Distance



– convex set $C = B_2(\binom{1}{1})$

– convex set $D = B_1(\binom{1+\sqrt{2}}{1+\sqrt{2}})$

Visualization of Hausdorff Distance



– convex set $C = B_2(\binom{1}{1})$

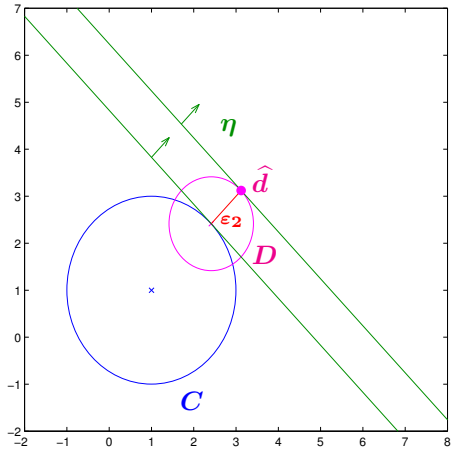
– convex set $D = B_1(\binom{1+\sqrt{2}}{1+\sqrt{2}})$

• farthest point $\hat{c}$ from D:
$\hat{c} = y(\eta, C)$

– (one-sided) Hausdorff distance
$\varepsilon_1 = \mathrm{d}(C, D) = \mathrm{d_H}(C, D)$

– supporting hyperplanes of C, D
in direction η
$\varepsilon_1 = \delta^\star(\eta, C) - \delta^\star(\eta, D)$

Visualization of Hausdorff Distance



– convex set $C = B_2(\binom{1}{1})$

– convex set $D = B_1(\binom{1+\sqrt{2}}{1+\sqrt{2}})$

• farthest point $\widehat{d}$ from C:
$\widehat{d} = y(\eta, D)$

– (one-sided) Hausdorff distance
$\varepsilon_2 = \mathrm{d}(D, C)$

– supporting hyperplanes of C, D
in direction η
$\varepsilon_2 = \delta^\star(\eta, D) - \delta^\star(\eta, C)$

Proposition 3.156 (Minkowski's duality). *Let $U, V \in \mathcal{C}(\mathbb{R}^n)$. Then,*

$$\mathrm{d}(U,V) = \sup_{l \in B_1(0)} (\delta^\star(l,U) - \delta^\star(l,V)),$$
$$\mathrm{d_H}(U,V) = \sup_{l \in B_1(0)} |\delta^\star(l,U) - \delta^\star(l,V)|.$$

Proof. Consider $r > \mathrm{d}(U,V)$. Then, $U \subset V + rB_1(0)$. Hence, Propositions 3.119 and 3.115 yield for $l \in B_1(0)$

$$\begin{aligned}\delta^\star(l,U) &\le \delta^\star(l, V + rB_1(0)) = \delta^\star(l,V) + \delta^\star(l, rB_1(0)) \\ &= \delta^\star(l,V) + r\|l\| \le \delta^\star(l,V) + r, \\ \delta^\star(l,U) - \delta^\star(l,V) &\le r, \\ s := \sup_{l \in B_1(0)} (\delta^\star(l,U) - \delta^\star(l,V)) &\le r\end{aligned}$$

Taking $(r_m)_m$ as a sequence converging monotonously decreasing to $\mathrm{d}(U,V)$, we have proven "$\le$". Since for every $l \in S_{n-1}$

$$\begin{aligned}\delta^\star(l,U) - \delta^\star(l,V) &\le \sup_{l \in B_1(0)} (\delta^\star(l,U) - \delta^\star(l,V)) = s, \\ \delta^\star(l,U) &\le \delta^\star(l,V) + s = \delta^\star(l,V) + s \cdot \|l\| \\ &= \delta^\star(l,V) + \delta^\star(l, B_s(0)) = \delta^\star(l, V + B_s(0)) \\ &= \delta^\star(l, V + s \cdot B_1(0)),\end{aligned}$$

we have by Proposition 3.120 that $U \subset V + s \cdot B_1(0)$, i.e. $\mathrm{d}(U,V) \le s$.

For the equation on the Hausdorff distance, use the first one and show both inequalities. □

Corollary 3.157. *Let $U, V \in \mathcal{C}(\mathbb{R}^n)$. Then, the index set $B_1(0)$ for the directions l in Proposition 3.156 could be replaced by S_{n-1} for the Hausdorff distance $\mathrm{d_H}(U,V)$. For the one-sided Hausdorff distance we have*

$$\mathrm{d}(U,V) = \max\{\sup_{\eta \in S_{n-1}} (\delta^\star(\eta,V) - \delta^\star(\eta,U)), 0\}.$$

In all cases, the supremum is always attained and could be replaced by a maximum.

Proposition 3.159 (cancellation law). *Let $U, V, W \in \mathcal{C}(\mathbb{R}^n)$. Then,*

$$\mathrm{d}(U+W, V+W) = \mathrm{d}(U,V) \quad \textit{and} \quad \mathrm{d_H}(U+W, V+W) = \mathrm{d_H}(U,V).$$

Proof. Proposition 3.156 shows that

$$\mathbf{d}(U+W,V+W) = \sup_{l\in B_1(0)} (\delta^\star(l,U+W)-\delta^\star(l,V+W)),$$
$$\mathbf{d}(U,V) = \sup_{l\in B_1(0)} (\delta^\star(l,U)-\delta^\star(l,V)).$$

Equality follows, since for $l \in S_{n-1}$

$$\begin{aligned}&\delta^\star(l,U+W)-\delta^\star(l,V+W)\\ &= (\delta^\star(l,U)+\delta^\star(l,W)) - (\delta^\star(l,V)+\delta^\star(l,W)) = \delta^\star(l,U)-\delta^\star(l,V).\end{aligned}$$

The same reasoning is true for the Hausdorff distance. □

For a proof of both theorems see e.g. [Sch93, Theorem 3.1.2 and Theorem 3.1.6].

Theorem 3.160 (Shapley-Folkman). *Let $k \in \mathbb{N}$, $(U_i)_{i=1,\dots,k} \subset \mathbb{R}^n$ and $z \in \mathbf{co}(\sum_{i=1}^k U_i)$. Then, there exists an index set $\mathcal{I} \subset \{1,\dots,k\}$ with $\mathbf{card}(\mathcal{I}) \leq n$ such that*

$$z \in \sum_{i\in\mathcal{I}} \mathbf{co}(U_i) + \sum_{i\notin\mathcal{I}} U_i.$$

Theorem 3.161 (Shapley-Folkman-Starr). *Let $k \in \mathbb{N}$, $(U_i)_{i=1,\dots,k} \subset \mathcal{K}(\mathbb{R}^n)$. Then,*

$$\mathbf{d_H}(\sum_{i=1}^k U_i, \sum_{i=1}^k \mathbf{co}(U_i)) \leq \frac{\sqrt{n}}{2} \cdot \max_{i=1,\dots,k} \mathbf{diam}(U_i).$$

Theorem 3.162. *Remember that $\mathcal{K}(\mathbb{R}^n)$ is the set of all compact, nonempty sets of $\mathbb{R}^n$. Then, $(\mathcal{K}(\mathbb{R}^n), \mathbf{d_H})$ and $(\mathcal{C}(\mathbb{R}^n), \mathbf{d_H})$ with the Hausdorff metric are metric spaces.*

Proof. Let $U, V \in \mathcal{K}(\mathbb{R}^n)$. Then, U and V are bounded such that $\mathbf{d_H}(U,V) < \infty$ by Lemma 3.138. Hence, $\mathbf{d_H}(U,V) \in [0,\infty)$.
$\mathbf{d_H}(U,V) = 0$, if and only if $\overline{U} = \overline{V}$ by Proposition 3.150(iv). Since U and V are closed, they coincide with $\overline{U}$ resp. $\overline{V}$.
Proposition 3.150(v),(vii) show the rest of the missing properties for a metric space. □

Theorem 3.163. *Let $\mathcal{K}(\mathbb{R}^n)$ be the set of all compact, nonempty sets of $\mathbb{R}^n$. Then, $(\mathcal{K}(\mathbb{R}^n), \mathbf{d_H})$ and $(\mathcal{C}(\mathbb{R}^n), \mathbf{d_H})$ are complete spaces.*

Proof. see [Sch93, Theorem 1.8.3 and 1.8.5] □

The following theorem is a generalization of the result of Bolzano-Weierstraß about the existence of a convergent subsequence of a bounded sequence in $\mathbb{R}^n$.

Theorem 3.164 (Blaschke's selection theorem). *Let $(C_m)_m \subset \mathcal{C}(\mathbb{R}^n)$ be a bounded sequence of convex, compact sets. Then, there exists a convergent subsequence (w.r.t. Hausdorff metric).*

Proof. see [Sch93, Theorem 1.8.4] □

Definition 3.165. Let $C, D \in \mathcal{C}(\mathbb{R}^n)$. Then,

$$\mathbf{d_D}(C,D) := \sup_{l\in T_C\cap T_D} \|y(l,C) - y(l,D)\|$$

is called the *Demyanov distance* of C, D.
Hereby, T_C and T_D denotes subsets of S_{n-1} of full measure such that $Y(l,C)$ resp. $Y(l,D)$ are singletons.

Remark 3.166. *The definition above does not depend on the actual choice of T_C and T_D.*
Furthermore, $Y(l,C)$ is a singleton for almost every $l \in S_{n-1}$, since it is the derivative of $\delta^\star(l,C)$ which is convex and its derivative coincides almost everywhere with $y(l,C)$.

Proof. The subdifferential of the support function is the supporting face (cf. [BF87a, Theorem 6]) and the support function is convex, finite and subdifferentiable by [BF87a, Corollary of Theorem 3].
The subdifferential consists only of the gradient, if the support function is differentiable (cf. [BF87a, Theorem 5]). But the support function is convex, finite and therefore differentiable a.e. in $\mathbb{R}^n$ by [HUL93, IV, Theorem 4.2.3]. □

Proposition 3.167. *Let $C, D \in \mathcal{C}(\mathbb{R}^n)$. Then,*

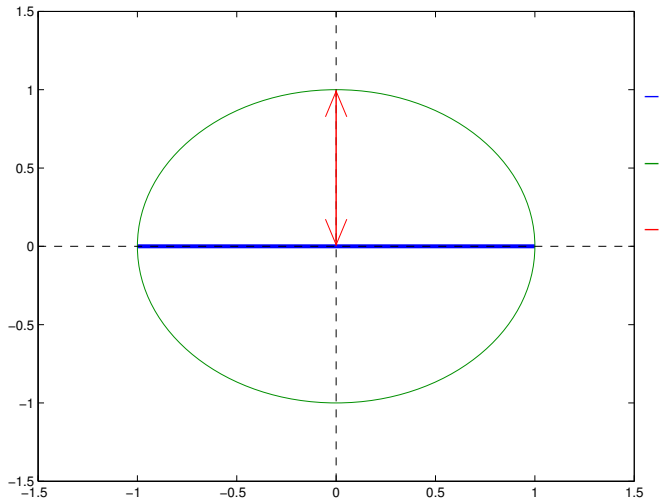
$$\mathbf{d_H}(C,D) \leq \mathbf{d_D}(C,D).$$

Proof. For all $l \in T_C \cap T_D$

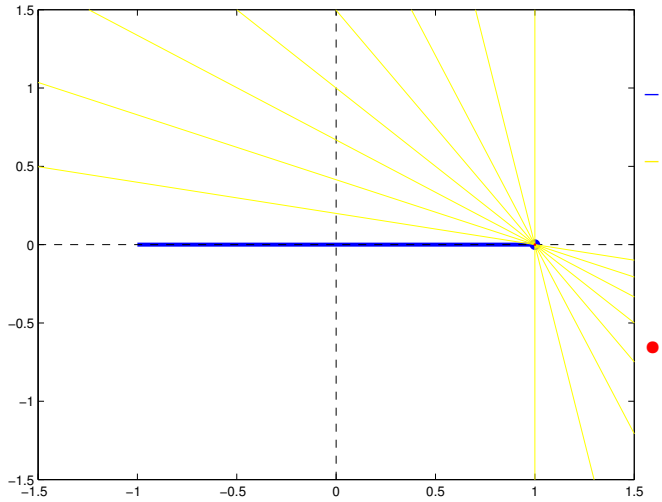
$$\begin{aligned} |\delta^\star(l,C) - \delta^\star(l,D)| &= |\langle l, y(l,C)\rangle - \langle l, y(l,D)\rangle| \\ &= |\langle l, y(l,C) - y(l,D)\rangle| \\ &\leq \underbrace{\|l\|}_{=1} \cdot \|y(l,C) - y(l,D)\| \leq \mathbf{d_D}(C,D). \end{aligned}$$

The Lipschitz continuity of $\delta^\star(\cdot, C)$ and $\delta^\star(\cdot, D)$ by Proposition 3.114 and the density of $T_C \cap T_D$ shows the estimation even for all $l \in S_{n-1}$. Taking the supremum on every $l \in S_{n-1}$ yields $\mathbf{d_H}(C,D) \leq \mathbf{d_D}(C,D)$ by using Proposition 3.156. □

Comparison of Hausdorff and Demyanov Distance



Comparison of Hausdorff and Demyanov Distance

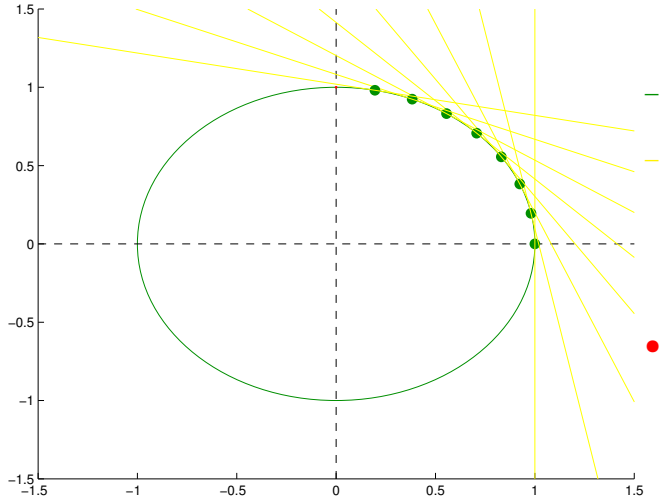


– $C = [-1,1] \times \{0\}$

– supporting hyperplanes for $\eta^m = \begin{pmatrix} \cos(\phi_m) \\ \sin(\phi_m) \end{pmatrix}$ $\phi_m \underset{m\to\infty}{\nearrow} \frac{\pi}{2}$

● supporting point $y(\eta^m, C) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$

Comparison of Hausdorff and Demyanov Distance

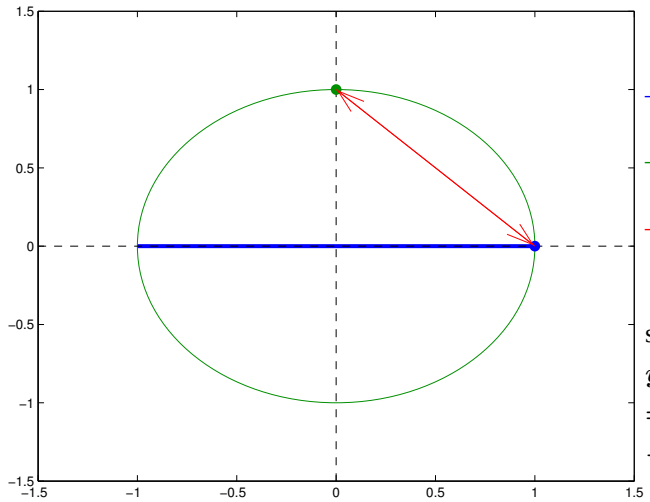


– $D = B_1(0)$

– supporting hyperplanes for $\eta^m = \begin{pmatrix} \cos(\phi_m) \\ \sin(\phi_m) \end{pmatrix}$ $\phi_m \underset{m\to\infty}{\nearrow} \frac{\pi}{2}$

● supporting point $y(\eta^m, C) = \eta^m$ converges for $m \to \infty$ to $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$

Comparison of Hausdorff and Demyanov Distance



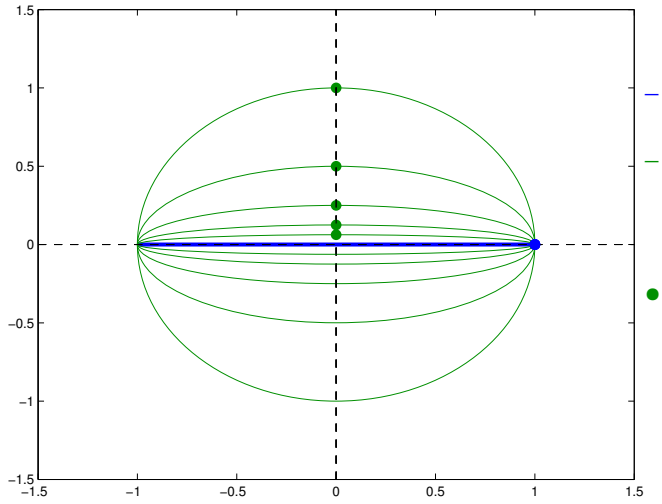
– $C = [-1,1] \times \{0\}$

– $D = B_1(0)$

– Demyanov distance $d_D(C, D) = \sqrt{2}$,

since $y(\eta^m, C) - y(\eta^m, D) = \begin{pmatrix} 1 \\ 0 \end{pmatrix} - \eta^m \xrightarrow[m\to\infty]{} \begin{pmatrix} 1 \\ 0 \end{pmatrix} - \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$

Non-Convergenging Sequence w.r.t. Demyanov Distance

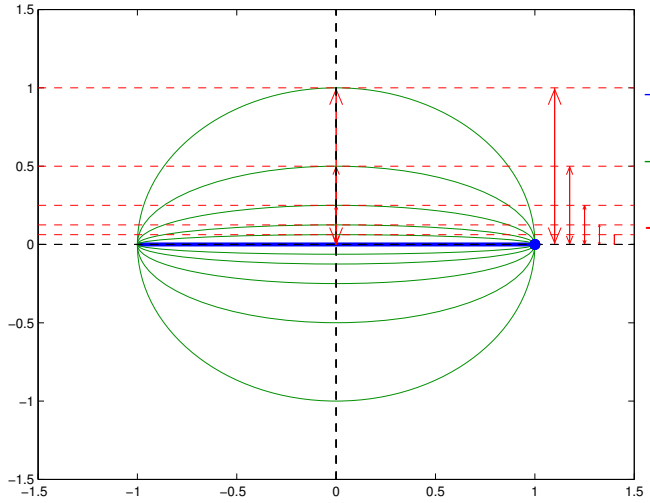


– $C = [-1,1] \times \{0\}$

– $D_k = A_k B_1(0)$,
$A_k = \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{k} \end{pmatrix}$

● farthest point from D_k to C

Non-Convergenging Sequence w.r.t. Demyanov Distance



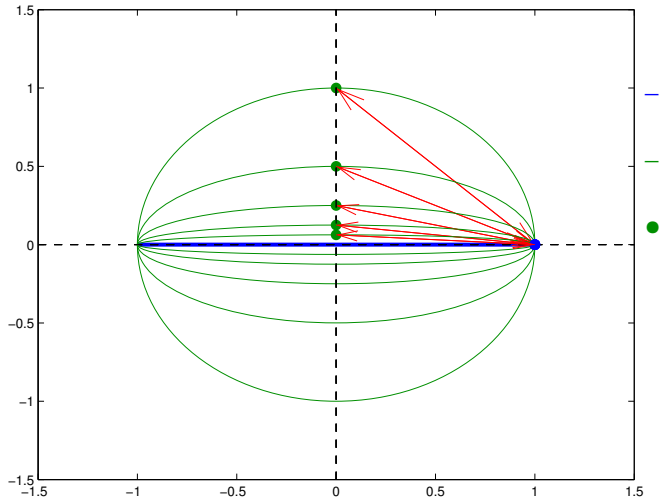
– $C = [-1,1] \times \{0\}$

– $D_k = A_k B_1(0)$,

--- supporting hyperplanes in direction $l = \binom{0}{1}$:

$\mathbf{d_H}(C, D_k) = \frac{1}{k} \xrightarrow[k \to \infty]{} 0$

Non-Convergenging Sequence w.r.t. Demyanov Distance



– $C = [-1,1] \times \{0\}$

– $D_k = A_k B_1(0)$,

● supporting points in direction $l = \binom{0}{1}$:

$\mathrm{d_D}(C, D_k) = \frac{1}{k} \cdot \sqrt{1+k^2} \xrightarrow[k\to\infty]{} 1$

Example 3.169. Set $D := B_1(0)$, $A_k := \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{k} \end{pmatrix}$, $D_k := A_k \cdot D$ for $k \in \mathbb{N}$ and consider $C = \mathrm{co}\{\binom{-1}{0}, \binom{1}{0}\}$. Then, for $l = (l_1, l_2) \in S_1$

$$\begin{aligned} \delta^\star(l, D_k) = \delta^\star(A_k^\top l, D) &= \|A_k^\top l\|_2 = \|\begin{pmatrix} l_1 \\ \frac{1}{k} l_2 \end{pmatrix}\|_2 \\ &= \sqrt{l_1^2 + \frac{1}{k^2} l_2^2} \xrightarrow[k\to\infty]{} |l_1|, \\ \delta^\star(l, C) &= |l_1|, \\ \mathrm{d_H}(C, D_k) &= \sup_{l \in S_1} |\sqrt{l_1^2 + \frac{1}{k^2} l_2^2} - |l_1|\,| \xrightarrow[k\to\infty]{} 0. \end{aligned}$$

4 Set-Valued Integration

Basic Facts
Why do we deal with set-valued maps (*svm*s)?

- generalization of pointwise functions
- reachable sets depend on the end time t and form a svm
- model uncertainty and disturbances as sets, i.e.
  replace $f : \mathcal{I} \to \mathbb{R}^n$ by $F : \mathcal{I} \Rightarrow \mathbb{R}^n$ with $F(t) = B_\varepsilon(f(t))$
- examples:
  - reliable computing (guaranteed end values in computer programs in presence of floating point inaccuracy)
  - geometric modelling (reconstruction of a 3D object $K \subset \mathbb{R}^3$ from its parallel 2D cross-sections), i.e.

    $$F : \mathcal{I} \Rightarrow \mathbb{R}^2 \quad \text{with} \quad K = \bigcup_{t \in \mathcal{I}} (t, F(t))$$

    consider a solution funnel set of a 2D system as a 3D body

Important Tools for (Later) Proofs

- continuity definitions for svms
- measurability + integrably boundedness $\Rightarrow$ set-valued integral
- selections and Castaing representation of svms
- characterization of measurability by Castaing representation or support functions
- "smoothness" definitions for svms = uniform "smooth" support functions

Integral Notions for SVMs

- Riemann integral for svms
  Riemann integrability is minimal "practical" assumption
- properties of Riemann integral:
  - convex set, even if svm has nonconvex images
  - same integral for convexified svm
  - generalization of pointwise Riemann integral
  - similar properties as in pointwise case
- characterization of Riemann integrability via a.e. continuity resp. support functions

Integral Notions for SVMs (continued)

- Aumann integral for svms
  Aumann integrability is minimal "theoretical" assumption
- properties of Aumann integral:
  - shares many properties of the Riemann integral
  - generalization of pointwise Lebesgue integral
  - coincides with Riemann integral for Riemann integrable svm
- support function of Aumann integral is Lebesgue integral of support function of svm
- reachable sets of linear control problems/linear differential inclusions are special Aumann integrals

4.1 Set-Valued Maps

The following definitions generalize known pointwise concepts like measurability, bounded variation, ... (cf. Appendix A.2).

Definition 4.1. Let $\mathcal{I} = [t_0, t_f]$ and $F : \mathcal{I} \Rightarrow \mathbb{R}^n$ with nonempty values, i.e. $F(t) \subset \mathbb{R}^n$ is nonempty for all $t \in \mathcal{I}$.
$F(\cdot)$ is *continuous* in $t \in \mathcal{I}$, if for every $\varepsilon > 0$ there exists $\delta > 0$ such that

$$\mathrm{d_H}(F(\tau), F(t)) \leq \varepsilon$$

for all $\tau \in \mathcal{I}$ with $|t - \tau| \leq \delta$.
$F(\cdot)$ is *continuous*, if it is continuous in every $t \in \mathcal{I}$.
$F(\cdot)$ is *upper semi-continuous* (*u.s.c.*) in $t \in \mathcal{I}$, if for every $\varepsilon > 0$ there exists $\delta > 0$ such that

$$\mathrm{d}(F(\tau), F(t)) \leq \varepsilon$$

for all $\tau \in \mathcal{I}$ with $|t - \tau| \leq \delta$.
$F(\cdot)$ is *lower semi-continuous* (*l.s.c.*) in $t \in \mathcal{I}$, if for every $\varepsilon > 0$ there exists $\delta > 0$ such that

$$\mathrm{d}(F(t), F(\tau)) \leq \varepsilon$$

for all $\tau \in \mathcal{I}$ with $|t - \tau| \leq \delta$.

Example 4.2. Consider $F, G : \mathbb{R} \Rightarrow \mathbb{R}$ with images in $\mathcal{C}(\mathbb{R})$ and

$$F(t) = \begin{cases} \{+1\} & \text{for } t > 0, \\ \{-1\} & \text{for } t < 0, \\ [-1, +1] & \text{for } t = 0, \end{cases}$$

$$G(t) = \begin{cases} [-1, +1] & \text{for } t \neq 0, \\ \{0\} & \text{for } t = 0. \end{cases}$$

Then, $F(\cdot)$ is u.s.c., $G(\cdot)$ is l.s.c. and both are not continuous.

Remark 4.3. *Let $\mathcal{I} = [t_0, t_f]$ and $F : \mathcal{I} \Rightarrow \mathbb{R}^n$ with nonempty values.*
$F(\cdot)$ is continuous in $t \in \mathcal{I}$, if and only if it is u.s.c. and l.s.c. in t.

Definition 4.4. Let $\mathcal{I} = [t_0, t_f]$ and $F : \mathcal{I} \Rightarrow \mathbb{R}^n$.
$F(\cdot)$ is *measurable*, if for every open set $U \subset \mathbb{R}^n$ the inverse image

$$F^{-1}(U) := \{t \in \mathcal{I} : F(t) \cap U \neq \emptyset\}$$

is a measurable set, i.e. it is an element of the underlying σ-algebra. From now on, we will consider the Borel σ-algebra on $\mathbb{R}^n$.

Definition 4.5. Let $\mathcal{I} = [t_0, t_f]$ and $F : \mathcal{I} \Rightarrow \mathbb{R}^n$.
$F(\cdot)$ is *simple*, if there exists a finite partition $(\mathcal{I}_i)_{i=1,\ldots,k}$ of $\mathcal{I}$ and nonempty subsets $F_i \subset \mathbb{R}^n$, $i = 1, \ldots, k$, with

$$F(t) = \sum_{i=1}^{k} \chi_{\mathcal{I}_i}(t) F_i \quad (t \in \mathcal{I}).$$

Hereby, $\chi_{\mathcal{I}_i}(t) = 1$, if $t \in \mathcal{I}_i$ and otherwise equals zero.

Remark 4.6. *Let $\mathcal{I} = [t_0, t_f]$ and $F : \mathcal{I} \Rightarrow \mathbb{R}^n$ be a simple map.*
Then, $F(\cdot)$ is measurable, if each set $\mathcal{I}_i$ in Definition 4.5 is measurable and F_i is closed for all $i = 1, \ldots, k$.

Definition 4.7. Let $\mathcal{I} = [t_0, t_f]$ and $F : \mathcal{I} \Rightarrow \mathbb{R}^n$ with nonempty values.
$F(\cdot)$ is *bounded*, if there exists a constant $C \geq 0$ existiert mit

$$\|F(t)\| \leq C \qquad (t \in I).$$

$F(\cdot)$ is *integrably bounded*, if there exists a function $k(\cdot) \in \mathrm{L}_1(\mathcal{I})$ with

$$\|F(t)\| = \sup_{f(t) \in F(t)} \|f(t)\| \leq k(t)$$

for almost all $t \in I$.

Definition 4.8. Let $\mathcal{I} = [t_0, t_f]$ and $F : \mathcal{I} \Rightarrow \mathbb{R}^n$ with nonempty values.
$F(\cdot)$ is *Riemann-integrable*, if there exists a nonempty set $U \subset \mathbb{R}^n$ such that for every $\varepsilon > 0$ there exists a $\delta > 0$ such that for every partition

$$t_0 = \tau_0 < \tau_1 < \ldots < \tau_{N-1} < \tau_N = t_f \qquad N \in \mathbb{N},$$

with fineness

$$\max_{i=0,\ldots,N-1} (\tau_{i+1} - \tau_i) \leq \delta$$

and intermediate points

$$\xi_i \in [\tau_i, \tau_{i+1}] \quad \text{for } i = 0, \ldots, N-1$$

follows that

$$\mathbf{d_H}(\sum_{i=0}^{N-1} (\tau_{i+1} - \tau_i) F(\xi_i), U) \leq \varepsilon.$$

Definition 4.8 (continued). This unique limit set U is denoted as (R)-$\int_{\mathcal{I}} F(t)dt$, the *Riemann-integral* of $F(\cdot)$.

Remark 4.9. *If $F(\cdot)$ has images in $\mathcal{K}(\mathbb{R}^n)$ or $\mathcal{C}(\mathbb{R}^n)$, then U has to be also an element of $\mathcal{K}(\mathbb{R}^n)$ resp. $\mathcal{C}(\mathbb{R}^n)$ by Theorem 3.163, since it is a element of sets in $\mathcal{K}(\mathbb{R}^n)$ resp. $\mathcal{C}(\mathbb{R}^n)$ (the Minkowski sum of scaled compact resp. convex sets is again compact resp. convex by Proposition 3.86 and 3.106).*

Definition 4.10. Let $\mathcal{I} = [t_0, t_f]$ and $F : \mathcal{I} \Rightarrow \mathbb{R}^n$ with nonempty values.
$F(\cdot)$ has *bounded variation* on $\mathcal{I}$, if for all partitions

$$t_0 < t_1 < \ldots < t_{N-1} < t_N = t_f, \qquad N \in \mathbb{N},$$

the sum

$$\sum_{i=0}^{N-1} \mathbf{d_H}(F(t_{i+1}), F(t_i)) \leq C$$

is bounded by a constant C which is independent from the partition. The infimum of such constants is called *variation* of $F(\cdot)$, namely $\bigvee_{t_0}^{t_f} \mathbf{F}(\cdot)$ or $\bigvee_{\mathcal{I}} \mathbf{F}(\cdot)$.

Definition 4.11. Let $\mathcal{I} = [t_0, t_f]$ and $F : \mathcal{I} \Rightarrow \mathbb{R}^n$ with nonempty values.
$F(\cdot)$ is *Lipschitz continuous* on $\mathcal{I}$, if there exists a constant $L \geq 0$ such that for all $t, \tau \in \mathcal{I}$

$$\mathbf{d_H}(F(t), F(\tau)) \leq L|t - \tau|.$$

As for point-wise function (cf. Lemma 6.23), a Lipschitz svm with constant L has bounded variation $\bigvee_{\mathcal{I}} \mathbf{F}(\cdot) \leq L \cdot (t_f - t_0)$.

Definition 4.12. Let $\mathcal{I} = [t_0, t_f]$ and $F : \mathcal{I} \Rightarrow \mathbb{R}^n$ with nonempty values.
Then, $f : \mathcal{I} \longrightarrow \mathbb{R}^n$ with

$$f(t) \in F(t) \quad \text{(for a.e. } t \in \mathcal{I})$$

is called *selection* of $F(\cdot)$.

Remark 4.13. *Instead of imposing smoothness on the set-valued map $F(\cdot)$, one could demand the corresponding smoothness on the real-valued function $\delta^\star(\eta, F(\cdot))$ (notions are well-known, easier to prove).*
But the smoothness of $\delta^\star(\eta, F(\cdot))$ must be uniform in $\eta \in S_{n-1}$.

Proposition 4.14. *Let $\mathcal{I} = [t_0, t_f]$ and $F : \mathcal{I} \Rightarrow \mathbb{R}^n$ has nonempty values.*
Then, $F(\cdot)$ is bounded by a constant C, if and only if

$$\sup_{\eta \in S_{n-1}} |\delta^\star(\eta, F(t))| \leq C \qquad (t \in I).$$

Proof. clear from $\|F(t)\| = \mathbf{d}(F(t), \{0_{\mathbb{R}^n}\}) = \mathbf{d_H}(F(t), \{0_{\mathbb{R}^n}\})$ and Corollary 3.157 □

Proposition 4.15. *Let $\mathcal{I} = [t_0, t_f]$ and $F : \mathcal{I} \Rightarrow \mathbb{R}^n$ has nonempty values.*
Then, $F(\cdot)$ has bounded variation less or equal V, if

$$\sup_{\eta \in S_{n-1}} \bigvee_{t_0}^{t_f} \delta^\star(\eta, F(\cdot)) \leq V.$$

Proof. The converse is not true in general, since for this direction a bounded joint variation of $(\delta^\star(\eta, F(\cdot)))_{\eta \in S_{n-1}}$ has to be demanded (see [Bai95-Con, 1.6.6 Satz]). □

Proposition 4.16. *Let $\mathcal{I} = [t_0, t_f]$ and $F : \mathcal{I} \Rightarrow \mathbb{R}^n$ with nonempty values.*
Then, $F(\cdot)$ is Lipschitz continuous with constant L, if and only if

$$\sup_{\eta \in S_{n-1}} |\delta^\star(\eta, F(t)) - \delta^\star(\eta, F(\tau))| \leq L|t - \tau|$$

for all $t, \tau \in \mathcal{I}$.

Remark 4.17. *Further appropriate smoothness notions could be found e.g. in [DF90, Bai95-Con].*
No natural differential notion for svms is available, since one could not define a differential quotients due to the missing difference for sets:
Which difference should we use in

$$\frac{F(t+h) - F(t)}{h} \; ?$$

Remark 4.17 (continued). *In the following some publication in this field are listed.*
attempts based on Demyanov difference: [DR95]
attempts based on pairs of sets: [BJ70, DR95, DKRV97]
approximation of svms with "simple" maps: [Art89, Art95, DLZ86, LZ91, Gau90, Sil97]
literature on selections: [AC84-Con, AF90-Con, BF87-Con, Mic56, Her71, Roc76, Dom87, GM92, KBV90, Den98, Den00, Den01, KML93, CR02]

For Further Reading

References

[AC84-Con] J.-P. Aubin and A. Cellina. *Differential Inclusions*, volume 264 of *Grundlehren der mathematischen Wissenschaften*. Springer-Verlag, Berlin–Heidelberg–New York–Tokyo, 1984.

[AF90-Con] J.-P. Aubin and H. Frankowska. *Set-Valued Analysis*, volume 2 of *Systems & Control: Foundations and Applications*. Birkhäuser, Boston–Basel–Berlin, 1990.

[Dei92-Con] K. Deimling. *Multivalued Differential Equations*, volume 1 of *de Gruyter Series in Nonlinear Analysis and Applications*. Walter de Gruyter, Berlin–New York, 1992.

[BF87-Con] V. I. Blagodatskikh and A. F. Filippov. Differential inclusions and optimal control. In *Topology, Ordinary Differential Equations, Dynamical Systems*, volume 1986, issue 4 of *Proceedings of the Steklov Institute of Mathematics*, pages 199–259. AMS, Providence–Rhode Island, 1987.

[Kis91-Con] M. Kisielewicz. *Differential Inclusions and Optimal Control*, volume 44 of *Mathematics and Its Applications*. PWN - Polish Scientific Publishers, Warszawa–Dordrecht–Boston–London, 1991.

[Fil88-Con] A. F. Filippov. *Differential Equations with Discontinuous Righthand Sides*. Mathematics and Its Applications (Soviet Series). Kluwer Academic Publishers, Dordrecht–Boston–London, 1988.

[Bai95-Con] R. Baier. *Mengenwertige Integration und die diskrete Approximation erreichbarer Mengen*, volume 50 of *Bayreuth. Math. Schr.* Mathematisches Institut der Universität Bayreuth, 1995.

4.2 Properties of Measurable Set-Valued Maps

Proposition 4.18. *Let $\mathcal{I} = [t_0, t_f]$, $f : \mathcal{I} \to \mathbb{R}^n$ and set $F : \mathcal{I} \rightrightarrows \mathbb{R}^n$ with $F(t) := \{f(t)\}$. Then, $F(\cdot)$ is measurable, if and only if $f(\cdot)$ is measurable.*

Proof. $f(\cdot)$ is measurable per definition, if its inverse image is measurable. Everything follows by the equality of the inverse images for closed sets $S \subset \mathbb{R}^n$:

$$\begin{aligned} F^{-1}(S) &= \{t \in \mathcal{I} \,|\, F(t) \cap S \neq \emptyset\} = \{t \in \mathcal{I} \,|\, \{f(t)\} \cap S \neq \emptyset\} \\ &= \{t \in \mathcal{I} \,|\, f(t) \in S\} = f^{-1}(S) \end{aligned}$$

□

Proposition 4.19. *Let $\mathcal{I} = [t_0, t_f]$ and $F : \mathcal{I} \rightrightarrows \mathbb{R}^n$ with nonempty values.*
$F(\cdot)$ is measurable, if and only if there exists a sequence of simple, measurable maps $(F_m(\cdot))_m$ with $F_m : \mathcal{I} \rightrightarrows \mathbb{R}^n$ with nonempty, closed values and a set $\mathcal{N}$ of measure 0 such that for every $t \in I \setminus \mathcal{N}$:

$$F_m(t) \xrightarrow[m \to \infty]{} F(t)$$

Proof. see [Jac68-Mea, Corollary 2.5] □

Theorem 4.20. *Let $\mathcal{I} = [t_0, t_f]$ and $F : \mathcal{I} \rightrightarrows \mathbb{R}^n$ be measurable with nonempty, closed values.*
Then, there exists a measurable selection $f(\cdot)$ of $F(\cdot)$.

Proof. see [AF90-Mea, Theorem 8.1.3] for a constructive proof □

Theorem 4.21 (characterization theorem). *Let $\mathcal{I} = [t_0, t_f]$ and $F : \mathcal{I} \rightrightarrows \mathbb{R}^n$ with nonempty, closed values.*
$F(\cdot)$ is measurable, if and only if one of the following conditions hold:

(i) For all $x \in \mathbb{R}^n$ the function $t \mapsto \mathbf{dist}(x, F(t))$ is measurable.

(ii) There exists a Castaing representation of $F(\cdot)$ with measurable selections $(f_m(\cdot))_m$, i.e.

$$F(t) = \overline{\bigcup_{m \in \mathbb{N}} f_m(t)} \quad \text{for all } t \in \mathcal{I}.$$

Proof. see [AF90-Mea, Theorem 8.1.4] for a constructive proof □

Corollary 4.22. *Let $\mathcal{I} = [t_0, t_f]$ and $F : \mathcal{I} \rightrightarrows \mathbb{R}^n$ be measurable and integrably bounded with nonempty, closed values.*
Then, $\|F(\cdot)\|$ is Lebesgue-integrable on $\mathcal{I}$.

Proof. follows from Theorem 4.21(ii), Proposition A.13 and from the definition of integrably boundedness □

Proposition 4.23. *Let $\mathcal{I} = [t_0, t_f]$, $A \in \mathbb{R}^{m \times n}$, $\mu \in \mathbb{R}$ and $F, G : \mathcal{I} \rightrightarrows \mathbb{R}^n$ be measurable with nonempty, closed values.*
Then, the set-valued maps $A \cdot F(\cdot)$ $\mu \cdot F(\cdot)$ and $F(\cdot) + G(\cdot)$ are measurable provided that all images of $A \cdot F(\cdot)$ resp. $F(\cdot) + G(\cdot)$ are closed.

Proof. Let us start to prove the measurability for $A \cdot F(\cdot)$ and the special case $A = 0_{m \times n}$.
A constant map $F(t) = \{v\}$, $v \in \mathbb{R}^n$, is measurable, since $F(t) = \chi_{\mathcal{I}}(t)\{v\}$ is simple, has nonempty, closed images and $\mathcal{I}$ is measurable (see Remark 4.6). Hence, the case $A = 0_{m \times n}$ is trivial.
All other proofs are based on Proposition 4.19.
Let us consider $A \cdot F(\cdot)$ with $A \neq 0_{m \times n}$, i.e. $\|A\| > 0$. For $t \in \mathcal{I} \setminus \mathcal{N}$ choose $\varepsilon > 0$ and $m \in \mathbb{N}$ such that $\mathbf{d_H}(F(t), F_m(t)) \leq \frac{\varepsilon}{\|A\|}$ with a representation $F_m(t) = \sum_{i=1}^{N_m} \chi_{\mathcal{I}_{i,m}}(t) F_{i,m}$.

Propositions 3.150(vi) and 3.88(iii) ensure that

$$\mathbf{d_H}(A \cdot F(t), A \cdot F_m(t)) \leq \|A\| \cdot \mathbf{d_H}(F(t), F_m(t)) \leq \|A\| \cdot \frac{\varepsilon}{\|A\|} = \varepsilon,$$

$$A \cdot F_m(t) = A \cdot \Big(\sum_{i=1}^{N_m} \chi_{\mathcal{I}_{i,m}}(t) F_{i,m}\Big) = \sum_{i=1}^{N_m} \chi_{\mathcal{I}_{i,m}}(t)(A \cdot F_{i,m}),$$

which is also a simple, measurable map with nonempty, closed images approaching $\boldsymbol{A}\cdot\boldsymbol{F}(\cdot)$.
The result for the scalar multiplication follows by setting $\boldsymbol{A}:=\mu\boldsymbol{I}$, where $\boldsymbol{I}$ is the $\boldsymbol{n}\times\boldsymbol{n}$-identity matrix.
Finally, let us study $\boldsymbol{F}(\cdot)+\boldsymbol{G}(\cdot)$.

Now, choose $(\boldsymbol{F}_m)_m$ to be a similar sequence of simple, measurable maps with $\mathbf{d_H}(\boldsymbol{F}(t),\boldsymbol{F}_m(t))\leq\frac{\varepsilon}{2}$. Let $(\boldsymbol{G}_m)_m$ with the representation $\boldsymbol{G}_m(t)=\sum_{j=1}^{\widetilde{N}_m}\chi_{\widetilde{\mathcal{I}}_{j,m}}(t)\boldsymbol{G}_{j,m}$ be the corresponding sequence for $\boldsymbol{G}(\cdot)$. Denote by $\widehat{\mathcal{I}}_{k,m}$, $k=1,\dots,\widehat{N}_m$, the refined partition of $(\mathcal{I}_{i,m})_{i=1,\dots,N_m}$ and $(\widetilde{\mathcal{I}}_{j,m})_{j=1,\dots,\widetilde{N}_m}$. It consists of intersections of measurable sets $\mathcal{I}_{i,m}$ and $\widetilde{\mathcal{I}}_{j,m}$, so all sets from the new partition are also measurable. Define the index functions

$$\begin{aligned}i(k)\in&\{1,\dots,N_m\}\text{ with }\widehat{\mathcal{I}}_{k,m}\subset\mathcal{I}_{i(k),m},\\ j(k)\in&\{1,\dots,\widetilde{N}_m\}\text{ with }\widehat{\mathcal{I}}_{k,m}\subset\widetilde{\mathcal{I}}_{j(k),m}.\end{aligned}$$

for $k=1,\dots,\widehat{N}_m$. Then, we have the new representations

$$\boldsymbol{F}_m(t)=\sum_{k=1}^{\widehat{N}_m}\chi_{\widehat{N}_{k,m}}(t)\boldsymbol{F}_{i(k),m},\quad \boldsymbol{G}_m(t)=\sum_{k=1}^{\widehat{N}_m}\chi_{\widehat{I}_{k,m}}(t)\boldsymbol{G}_{j(k),m},$$

and

$$\boldsymbol{F}_m(t)+\boldsymbol{G}_m(t)=\sum_{k=1}^{\widehat{N}_m}\chi_{\widehat{\mathcal{I}}_{k,m}}(t)(\boldsymbol{F}_{i(k),m}+\boldsymbol{G}_{j(k),m}),$$

where

$$\begin{aligned}&\mathbf{d_H}(\boldsymbol{F}_m(t)+\boldsymbol{G}_m(t),\boldsymbol{F}(t)+\boldsymbol{G}(t))=\mathbf{d_H}(\boldsymbol{F}_{i(k),m}+\boldsymbol{G}_{j(k),m},\boldsymbol{F}(t)+\boldsymbol{G}(t))\\ \leq&\mathbf{d_H}(\boldsymbol{F}_{i(k),m},\boldsymbol{F}(t))+\mathbf{d_H}(\boldsymbol{G}_{j(k),m},\boldsymbol{G}(t))\\ =&\mathbf{d_H}(\boldsymbol{F}_m(t),\boldsymbol{F}(t))+\mathbf{d_H}(\boldsymbol{G}_m(t),\boldsymbol{G}(t))\leq\frac{\varepsilon}{2}+\frac{\varepsilon}{2}=\varepsilon\end{aligned}$$

for $t\in\widehat{\mathcal{I}}_{k,m}$. □

Proposition 4.24. *Let $\mathcal{I}=[t_0,t_f]$ and $\boldsymbol{F}:\mathcal{I}\rightrightarrows\mathbb{R}^n$ with nonempty, closed values. Then:*

(i) If $\boldsymbol{F}(\cdot)$ is measurable, then the support function $t\mapsto\delta^\star(l,\boldsymbol{F}(t))$ is also measurable for all $l\in\mathbb{R}^n$.

(ii) If for all $l\in\mathbb{R}^n$ the support function $\delta^\star(l,\boldsymbol{F}(\cdot))$ is measurable and $\boldsymbol{F}(\cdot)$ has additionally convex and bounded images, then $\boldsymbol{F}(\cdot)$ is measurable.

Proof. see [AF90-Mea, Theorem 8.2.14] □

Definition 4.25. Let $\varphi:\mathcal{I}\times\mathbb{R}^n\to\mathbb{R}^m$ is a (single-valued) function. $\varphi(\cdot,\cdot)$ is a *Carathéodory map*, if

(i) $\varphi(\cdot,x)$ is measurable for every $x\in\mathbb{R}^n$,

(ii) $\varphi(t,\cdot)$ is continuous for every $t\in\mathcal{I}$.

Proposition 4.26. *Let $\boldsymbol{F}:\mathcal{I}\rightrightarrows\mathbb{R}^n$ be measurable with nonempty, closed images and $\varphi:\mathcal{I}\times\mathbb{R}^n\to\mathbb{R}^m$ be a Carathéodory map. Then, the set-valued mapping $\boldsymbol{G}:\mathcal{I}\rightrightarrows\mathbb{R}^n$ with $\boldsymbol{G}(t)=\overline{\varphi(t,\boldsymbol{F}(t))}$ is measurable with nonempty, closed images in $\mathbb{R}^m$.*

Proof. see [AF90-Mea, Theorem 8.2.8] □

Proposition 4.27. *Let $\mathcal{I}=[t_0,t_f]$ and $\boldsymbol{F}:\mathcal{I}\rightrightarrows\mathbb{R}^n$ has convex and bounded images.*
Then $\boldsymbol{F}(\cdot)$ is measurable and integrably bounded by a L_1-function $k(\cdot)$, if and only if all support functions $\delta^\star(\eta,\boldsymbol{F}(\cdot))$ are measurable for $\eta\in\boldsymbol{S}_{n-1}$ and

$$\sup_{\eta\in\boldsymbol{S}_{n-1}}|\delta^\star(\eta,\boldsymbol{F}(t))|\leq k(t)$$

for almost all $t\in\boldsymbol{I}$, i.e. $\delta^\star(\eta,\boldsymbol{F}(\cdot))\in L_1(\mathcal{I})$ with uniform L_1-norm.

Proof. follows from Proposition 4.24 with Lebesgue's dominated convergence theorem (see A.14) and the same arguments as in Proposition 4.14. □

For Further Reading

References

[AC84-Mea] J.-P. Aubin and A. Cellina. *Differential Inclusions*, volume 264 of *Grundlehren der mathematischen Wissenschaften*. Springer-Verlag, Berlin–Heidelberg–New York–Tokyo, 1984.

[AF90-Mea] J.-P. Aubin and H. Frankowska. *Set-Valued Analysis*, volume 2 of *Systems & Control: Foundations and Applications*. Birkhäuser, Boston–Basel–Berlin, 1990.

[Dei92-Mea] K. Deimling. *Multivalued Differential Equations*, volume 1 of *de Gruyter Series in Nonlinear Analysis and Applications*. Walter de Gruyter, Berlin–New York, 1992.

[Kis91-Mea] M. Kisielewicz. *Differential Inclusions and Optimal Control*, volume 44 of *Mathematics and Its Applications*. PWN - Polish Scientific Publishers, Warszawa–Dordrecht–Boston–London, 1991.

[CV77-Mea] C. Castaing and M. Valadier. *Convex Analysis and Measurable Multifunctions*, volume 580 of *Lecture Notes in Math.* Springer-Verlag, Berlin–Heidelberg–New York, 1977.

[BF87-Mea] V. I. Blagodatskikh and A. F. Filippov. Differential inclusions and optimal control. In *Topology, Ordinary Differential Equations, Dynamical Systems*, volume 1986, issue 4 of *Proceedings of the Steklov Institute of Mathematics*, pages 199–259. AMS, Providence–Rhode Island, 1987.

[Jac68-Mea] M. Q. Jacobs. Measurable multivalued mappings and Lusin's theorem. *Trans. Amer. Math. Soc.*, 134:471–481, 1968.

[HU77-Mea] F. Hiai and H. Umegaki. Integrals, Conditional Expectations, and Martingales of Multivalued Functions. *J. Multivariate Anal.*, 7(1):149–182, 1977.

[Fil88-Mea] A. F. Filippov. *Differential Equations with Discontinuous Righthand Sides*. Mathematics and Its Applications (Soviet Series). Kluwer Academic Publishers, Dordrecht–Boston–London, 1988.

[Coh80-Mea] D. L. Cohn. *Measure Theory*. Birkhäuser, Boston–Basel–Stuttgart, 1980.

Literature Recommended by Participants of the Intensive Course

References

[dB03-Mea] G. de Barra. *Measure Theory and Integration*. Ellis Horwood Series in Mathematics and Its Applications. Ellis Horwood Limited, Publisher, Chichester, 2003. 2nd updated ed.

[Mir90-Mea] B. Mirković. *Theory of Measures and Integrals*. Naučna knjiga, Beograd, 1990. (Serbian).

[Pap02a-Mea] E. Pap, editor. *Handbook of Measure Theory. Volume I*. North-Holland, Amsterdam, 2002.

[Pap02b-Mea] E. Pap, editor. *Handbook of Measure Theory. Volume II*. North-Holland, Amsterdam, 2002.

[Roy86-Mea] H. L. Royden. *Real Analysis*. Macmillan Publishing Company/Collier Macmillan Publisher, New York–London, 1986. 3rd ed.

4.3 Set-Valued Integrals

4.3.1 Riemann-Integral

Lemma 4.28. *Let $U \subset \mathbb{R}^n$ be nonempty, bounded and $N \in \mathbb{N}$. Then,*

$$\mathbf{d_H}(\frac{1}{N}\sum_{i=1}^{N} U, \mathbf{co}(U)) \leq \frac{2n}{N} \cdot \|U\|.$$

Proof. We will use Lemma 3.83 and Propositions 3.91(iii) and 3.93.

$$\frac{1}{N}\sum_{i=1}^{N} U \subset \frac{1}{N}\sum_{i=1}^{N} \mathbf{co}(U) = \sum_{i=1}^{N} \frac{1}{N}\mathbf{co}(U) = (\sum_{i=1}^{N}\frac{1}{N})\,\mathbf{co}(U) = \mathbf{co}(U),$$

hence

$$\mathbf{d}(\frac{1}{N}\sum_{i=1}^{N} U, \mathbf{co}(U)) = 0.$$

Let $z = \sum_{j=1}^{n+1} \lambda_j u^j \in \mathbf{co}(U)$ with $u^j \in U$ for $j = 1, \ldots, n+1$. Set $t_\nu := \frac{\nu}{N}$ for $\nu = 0, \ldots, N$ and choose numbers $m(j) \in \{0, \ldots, N-1\}$ with

$$\lambda_j \in \begin{cases} [t_{m(j)}, t_{m(j)+1}) & \text{for } m(j) \leq N-2, \\ [t_{N-1}, 1] & \text{if } m(j) = N-1. \end{cases}$$

for $j = 1, \ldots, n$. Set $m(n+1) := N - \sum_{j=1}^{n} m(j)$. For abbreviation, set $\widetilde{\lambda}_j := t_{m(j)} = \frac{m(j)}{N}$, $j = 1, \ldots, n+1$.
Let us first study $m(n+1)$.

$$\sum_{j=1}^{n} \frac{m(j)}{N} \leq \sum_{j=1}^{n} \lambda_j = 1 - \lambda_{n+1} < 1,$$

$$\sum_{j=1}^{n} m(j) < N \quad \text{and} \quad \sum_{j=1}^{n+1} m(j) = N,$$

$$\widetilde{\lambda}_{n+1} = \frac{m(n+1)}{N} = \frac{1}{N} \cdot (N - \sum_{j=1}^{n} m(j))$$

$$\widetilde{\lambda}_{n+1} = 1 - \sum_{j=1}^{n} \frac{m(j)}{N} = 1 - \sum_{j=1}^{n} \widetilde{\lambda}_j,$$

$$|\lambda_{n+1} - \widetilde{\lambda}_{n+1}| = |(1 - \sum_{j=1}^{n} \lambda_j) - (1 - \sum_{j=1}^{n} \widetilde{\lambda}_j)| = |\sum_{j=1}^{n} (\lambda_j - \widetilde{\lambda}_j)|$$

$$\leq \sum_{j=1}^{n} \underbrace{|\lambda_j - \widetilde{\lambda}_j|}_{\leq \frac{1}{N}} \leq \frac{n}{N}.$$

Estimate

$$\|z - \sum_{j=1}^{n+1} \widetilde{\lambda}_j u^j\| = \|\sum_{j=1}^{n+1} (\lambda_j - \widetilde{\lambda}_j) u^j\| \leq \sum_{j=1}^{n+1} |\lambda_j - \widetilde{\lambda}_j| \cdot \|u^j\|,$$

$$\|z - \sum_{j=1}^{n+1} \widetilde{\lambda}_j u^j\| \leq \sum_{j=1}^{n} \underbrace{|\lambda_j - \widetilde{\lambda}_j|}_{\leq \frac{1}{N}} \cdot \|u^j\| + \underbrace{|\lambda_{n+1} - \widetilde{\lambda}_{n+1}|}_{\leq \frac{n}{N}} \cdot \|u^{n+1}\|$$

$$\leq (\sum_{j=1}^{n} \frac{1}{N} + \frac{n}{N}) \cdot \|U\| = \frac{2n}{N} \cdot \|U\|$$

Hence, with $W := \frac{2n}{N} \cdot \|U\| \cdot B_1(0)$:

$$\begin{aligned} z = \sum_{j=1}^{n+1} \lambda_j u^j \in \sum_{j=1}^{n+1} \widetilde{\lambda}_j u^j + W = \sum_{j=1}^{n+1} \frac{m(j)}{N} u^j + W \\ = \frac{1}{N} \sum_{j=1}^{n+1} \underbrace{\sum_{k=1}^{m(j)} u^j}_{=m(j)u^j} + W \subset \frac{1}{N} \sum_{j=1}^{n+1} \sum_{k=1}^{m(j)} U + W = \frac{1}{N} \sum_{i=1}^{N} U + W \end{aligned}$$

We can prove now with Lemma 3.142(ii) and Proposition 3.149(viii)

$$\begin{aligned} \mathbf{dist}(z, \frac{1}{N} \sum_{i=1}^{N} U) \leq \underbrace{\mathbf{dist}(z, \frac{1}{N} \sum_{i=1}^{N} U + W)}_{=0} + \mathrm{d}(\frac{1}{N} \sum_{i=1}^{N} U + W, \frac{1}{N} \sum_{i=1}^{N} U) \\ \leq \underbrace{\mathrm{d}(\frac{1}{N} \sum_{i=1}^{N} U, \frac{1}{N} \sum_{i=1}^{N} U)}_{=0} + \mathrm{d}(W, \{0_{\mathbb{R}^n}\}) = \|W\|. \end{aligned}$$

Altogether, we can prove

$$\mathrm{d}(\mathbf{co}(U), \frac{1}{N} \sum_{i=1}^{N} U) = \sup_{z \in \mathbf{co}(U)} \mathbf{dist}(z, \frac{1}{N} \sum_{i=1}^{N} U) \leq \|W\| = \frac{2n}{N} \cdot \|U\|.$$

□

Lemma 4.29. *Let $U \subset \mathbb{R}^n$ be nonempty, bounded and $N \in \mathbb{N}$. Then,*

$$\mathrm{d_H}(\frac{1}{N} \sum_{i=1}^{N} U, \mathbf{co}(U)) \leq \frac{\sqrt{n}}{2N} \cdot \mathbf{diam}(U) \leq \frac{\sqrt{n}}{N} \cdot \|U\|.$$

Proof. First, from Propositions 3.91(iii) and 3.93:

$$\frac{1}{N} \sum_{i=1}^{N} \mathbf{co}(U) = \sum_{i=1}^{N} \frac{1}{N} \mathbf{co}(U) = (\sum_{i=1}^{N} \frac{1}{N}) \mathbf{co}(U) = \mathbf{co}(U)$$

Then, we will essentially use Theorem 3.161 and once again, Proposition 3.91(iii).

$$\begin{aligned} \mathrm{d_H}(\frac{1}{N} \sum_{i=1}^{N} U, \mathbf{co}(U)) &= \mathrm{d_H}(\frac{1}{N} \sum_{i=1}^{N} U, \frac{1}{N} \sum_{i=1}^{N} \mathbf{co}(U)) \\ &= \mathrm{d_H}(\sum_{i=1}^{N} \frac{1}{N} U, \sum_{i=1}^{N} \frac{1}{N} \mathbf{co}(U)) \\ &\leq \frac{\sqrt{n}}{2} \cdot \mathbf{diam}(\frac{1}{N} \cdot U) = \frac{\sqrt{n}}{2} \cdot \frac{1}{N} \cdot \underbrace{\mathbf{diam}(U)}_{\leq 2 \cdot \|U\|} \\ &\leq \frac{\sqrt{n}}{N} \cdot \|U\|. \end{aligned}$$

□

Proposition 4.30. *Let $\mathcal{I} = [t_0, t_f]$ and $F : \mathcal{I} \rightrightarrows \mathbb{R}^n$ has nonempty values.*
(i) If $F(\cdot)$ is Riemann-integrable, then $\mathbf{co}\, F(\cdot)$ is Riemann-integrable with

$$\text{(R)-}\int_{\mathcal{I}} \mathbf{co}\, F(t) dt = \mathbf{co}(\text{(R)-}\int_{\mathcal{I}} F(t) dt). \tag{18}$$

(ii) If $\boldsymbol{F}(\cdot)$ has values in $\mathcal{K}(\mathbb{R}^n)$, is bounded and $\mathbf{co}\,\boldsymbol{F}(\cdot)$ is Riemann-integrable, then $\boldsymbol{F}(\cdot)$ is Riemann-integrable with

$$\text{(R)-}\int_{\mathcal{I}} \mathbf{co}\,\boldsymbol{F}(t)dt = \text{(R)-}\int_{\mathcal{I}} \boldsymbol{F}(t)dt. \tag{19}$$

Proof. (i) Let $\varepsilon > \mathbf{0}$ and choose $\delta > \mathbf{0}$ as for the function $\boldsymbol{F}(\cdot)$ in Definition 4.8. Proposition 3.107 yields

$$\sum_{i=0}^{N-1} (\tau_{i+1} - \tau_i)\,\mathbf{co}\,\boldsymbol{F}(\xi_i) = \mathbf{co}(\sum_{i=0}^{N-1} (\tau_{i+1} - \tau_i)\boldsymbol{F}(\xi_i)).$$

From Proposition 3.150(iii) follows that

$$\begin{aligned}
&\mathbf{d_H}(\sum_{i=0}^{N-1} (\tau_{i+1} - \tau_i)\,\mathbf{co}\,\boldsymbol{F}(\xi_i), \mathbf{co}((\text{R})\text{-}\int_{\mathcal{I}} \boldsymbol{F}(t)dt)) \\
&= \mathbf{d_H}(\mathbf{co}(\sum_{i=0}^{N-1} (\tau_{i+1} - \tau_i)\boldsymbol{F}(\xi_i), \mathbf{co}((\text{R})\text{-}\int_{\mathcal{I}} \boldsymbol{F}(t)dt)) \\
&\leq \mathbf{d_H}(\sum_{i=0}^{N-1} (\tau_{i+1} - \tau_i)\boldsymbol{F}(\xi_i), (\text{R})\text{-}\int_{\mathcal{I}} \boldsymbol{F}(t)dt) \leq \varepsilon.
\end{aligned}$$

(ii) Let $\varepsilon > \mathbf{0}$ and choose $\delta = \delta(\frac{\varepsilon}{2}) > \mathbf{0}$ as for the function $\mathbf{co}\,\boldsymbol{F}(\cdot)$ in Definition 4.8, but additionally with $\delta \leq \frac{\varepsilon}{2\sqrt{n}\cdot R}$, where $\boldsymbol{F}(t) \subset \boldsymbol{B_R}(\mathbf{0})$ for all $t \in \mathcal{I}$.

Then, Theorem 3.161 yields

$$\begin{aligned}
&\mathbf{d_H}(\sum_{i=0}^{N-1} (\tau_{i+1} - \tau_i)\boldsymbol{F}(\xi_i), (\text{R})\text{-}\int_{\mathcal{I}} \mathbf{co}\,\boldsymbol{F}(t)dt) \\
\leq &\mathbf{d_H}(\sum_{i=0}^{N-1} (\tau_{i+1} - \tau_i)\boldsymbol{F}(\xi_i), \sum_{i=0}^{N-1} (\tau_{i+1} - \tau_i)\,\mathbf{co}\,\boldsymbol{F}(\xi_i)) \\
&+ \mathbf{d_H}(\sum_{i=0}^{N-1} (\tau_{i+1} - \tau_i)\,\mathbf{co}\,\boldsymbol{F}(\xi_i), (\text{R})\text{-}\int_{\mathcal{I}} \mathbf{co}\,\boldsymbol{F}(t)dt) \\
\leq &\frac{\sqrt{n}}{\mathbf{2}} \cdot \max_{i=0,\ldots,N-1} \mathbf{diam}((\tau_{i+1} - \tau_i)\boldsymbol{F}(\xi_i)) + \frac{\varepsilon}{\mathbf{2}}. \\
\leq &\frac{\sqrt{n}}{\mathbf{2}} \cdot \max_{i=0,\ldots,N-1} (\tau_{i+1} - \tau_i) \cdot \mathbf{diam}(\boldsymbol{F}(\xi_i)) + \frac{\varepsilon}{\mathbf{2}}.
\end{aligned}$$

Since $\mathbf{diam}(\boldsymbol{F}(\xi_i)) \leq \mathbf{diam}(\boldsymbol{B_R}(\mathbf{0})) = \mathbf{2}\boldsymbol{R}$,

$$\begin{aligned}
&\mathbf{d_H}(\sum_{i=0}^{N-1} (\tau_{i+1} - \tau_i)\boldsymbol{F}(\xi_i), (\text{R})\text{-}\int_{\mathcal{I}} \mathbf{co}\,\boldsymbol{F}(t)dt) \\
\leq &\sqrt{n} \cdot \boldsymbol{R} \cdot \boldsymbol{\delta} + \frac{\varepsilon}{\mathbf{2}} \leq \frac{\varepsilon}{\mathbf{2}} + \frac{\varepsilon}{\mathbf{2}} = \varepsilon.
\end{aligned}$$

This shows that the Riemann integral of $\boldsymbol{F}(\cdot)$ and $\mathbf{co}\,\boldsymbol{F}(\cdot)$ coincide. □

Corollary 4.31 (convexity of Riemann-integral). *Let $\mathcal{I} = [t_0, t_f]$ and $\boldsymbol{F} : \mathcal{I} \Rightarrow \mathbb{R}^n$ be Riemann-integrable (or $\boldsymbol{F}(\cdot)$ be bounded and $\mathbf{co}\,\boldsymbol{F}(\cdot)$ be Riemann-integrable) with values in $\mathcal{K}(\mathbb{R}^n)$.*
Then, the Riemann-integral of $\boldsymbol{F}(\cdot)$ and $\mathbf{co}\,\boldsymbol{F}(\cdot)$ coincide and are both elements of $\mathcal{C}(\mathbb{R}^n)$.

Proof. follows from Remark 4.9 and from (18) and (19) in Proposition 4.30:
If $\boldsymbol{F}(\cdot)$ is Riemann-integrable, it is bounded by a constant $C \geq \mathbf{0}$ (assume the contrary and bring this to a contradiction). Furthermore, $\mathbf{co}\,\boldsymbol{F}(\cdot)$ is Riemann-integrable and its Riemann integral is compact, convex, nonempty by

Remark 4.9. Now, let $\varepsilon > 0$ and take a partition $(\mathcal{I}_i)_{i=0,\dots,N-1}$ of $\mathcal{I}$ of fineness $\delta \leq \varepsilon$ such that

$$\mathbf{d_H}(\sum_{i=0}^{N-1}(\tau_{i+1}-\tau_i)\,\mathbf{co}\,F(\xi_i), \text{(R)-}\int_{\mathcal{I}}\mathbf{co}\,F(t)dt) \leq \varepsilon.$$

Then, Theorem 3.161 with properties of the diameter (cf. Corollary 3.155) guarantees that

$$\begin{aligned}
&\mathbf{d_H}(\sum_{i=0}^{N-1}(\tau_{i+1}-\tau_i)\,\mathbf{co}\,F(\xi_i), \sum_{i=0}^{N-1}(\tau_{i+1}-\tau_i)F(\xi_i))\\
\leq&\frac{\sqrt{n}}{2}\cdot\max_{i=1,\dots,k}\mathbf{diam}((\tau_{i+1}-\tau_i)F(\xi_i))\\
\leq&\frac{\sqrt{n}}{2}\cdot\max_{i=1,\dots,k}(\tau_{i+1}-\tau_i)\cdot\mathbf{diam}(F(\xi_i))\\
\leq&\frac{\sqrt{n}}{2}\cdot\max_{i=1,\dots,k}(\tau_{i+1}-\tau_i)\cdot 2\cdot\underbrace{\|F(\xi_i)\|}_{\leq C}\\
\leq&\sqrt{n}\cdot\max_{i=1,\dots,k}(\tau_{i+1}-\tau_i)\cdot C \leq \sqrt{n}\cdot C\cdot\delta \leq \sqrt{n}\cdot C\cdot\varepsilon.
\end{aligned}$$

Therefore, both Riemann integrals of $F(\cdot)$ and $\mathbf{co}\,F(\cdot)$ coincide so that equation (18) shows the convexity of the integral. □

Example 4.32 *(constant set-valued map).* Let $\mathcal{I} = [t_0, T]$, $U \in \mathcal{K}(\mathbb{R}^n)$ and $F : \mathcal{I} \Rightarrow \mathbb{R}^n$ with $F(t) = U$ for all $t \in \mathcal{I}$. Then, $F : \mathcal{I} \Rightarrow \mathbb{R}^n$ is Riemann-integrable with

$$\text{(R)-}\int_{\mathcal{I}} F(t)dt = (T - t_0)\,\mathbf{co}(U).$$

Proof. Let us study first $\mathbf{co}\,F(\cdot)$, since for a partition $([\tau_i, \tau_{i+1}])_{i=0,\dots,N}$ with $\xi_i \in [\tau_i, \tau_{i+1}]$ for $i = 0, \dots, N$, we have

$$\sum_{i=0}^{N-1}(\tau_{i+1}-\tau_i)\,\mathbf{co}\,F(\xi_i) = \sum_{i=0}^{N-1}(\tau_{i+1}-\tau_i)\,\mathbf{co}\,U = \underbrace{(\sum_{i=0}^{N-1}(\tau_{i+1}-\tau_i))}_{=(T-t_0)}\mathbf{co}\,U$$

which shows that $\mathbf{co}(U)$ is the Riemann-integral of $\mathbf{co}\,F(\cdot)$. Corollary 4.31 shows that $F(\cdot)$ is Riemann-integrable with the same integral.

If $U = \{0, 1\}$, then one sees the convexifying impact of the Minkowski sum, since for $N \in \mathbb{N}$

$$\begin{aligned}
\frac{1}{N}\sum_{i=0}^{N-1}U &= \{\frac{k}{N}\,|\,k = 0, \dots, N\} \supsetneq U,\\
\text{(R)-}\int_0^1 U\,dt &= \mathbf{co}(U) = [0, 1].
\end{aligned}$$

□

The next four lemmas are easy to prove just by applying the definition of the Riemann integral.

Lemma 4.33. *Let $\mathcal{I} = [t_0, t_f]$ and $F, G : \mathcal{I} \Rightarrow \mathbb{R}^n$ be Riemann-integrable with nonempty values and $F(t) \subset G(t)$ for all $t \in \mathcal{I}$. Then,*

$$\textit{(R)-}\int_{\mathcal{I}} F(t)dt \subset \textit{(R)-}\int_{\mathcal{I}} G(t)dt.$$

Lemma 4.34. *Let $\mathcal{I} = [t_0, t_f]$, $\tau \in (t_0, t_f)$ and $\widetilde{F} : \widetilde{\mathcal{I}} \rightrightarrows \mathbb{R}^n$ be Riemann-integrable on $\widetilde{\mathcal{I}} = [t_0, \tau]$ with nonempty values. Then, $F : \mathcal{I} \rightrightarrows \mathbb{R}^n$*

$$F(t) := \begin{cases} \widetilde{F}(t) & \text{for } t \in \widetilde{\mathcal{I}}, \\ \{0_{\mathbb{R}^n}\} & \text{for } t \in \mathcal{I} \setminus \widetilde{\mathcal{I}} \end{cases}$$

is Riemann-integrable with

$$\text{(R)-}\int_{\mathcal{I}} F(t)dt = \text{(R)-}\int_{\widetilde{\mathcal{I}}} \widetilde{F}(t)dt.$$

Lemma 4.35. *Let $\mathcal{I} = [t_0, t_f]$ and $F : \mathcal{I} \rightrightarrows \mathbb{R}^n$ be Riemann-integrable with nonempty values. Then, $F|_{\widetilde{\mathcal{I}}}(\cdot)$ is Riemann-integrable on $\widetilde{\mathcal{I}} = [t_0, \tau]$ for every $\tau \in (t_0, t_f)$.*

Lemma 4.36. *Let $\mathcal{I} = [t_0, t_f]$ and $F, G : \mathcal{I} \rightrightarrows \mathbb{R}^n$ be Riemann-integrable with nonempty values. Then, $F(\cdot) + G(\cdot)$ is Riemann-integrable with*

$$\text{(R)-}\int_{\mathcal{I}} (F(t) + G(t))dt = \text{(R)-}\int_{\mathcal{I}} F(t)dt + \text{(R)-}\int_{\mathcal{I}} G(t)dt.$$

Proposition 4.37. *Let $\mathcal{I} = [t_0, t_f]$, $(U_i)_{i=0,\dots,N-1} \subset \mathcal{K}(\mathbb{R}^n)$ and and $([\tau_i, \tau_{i+1}])_{i=0,\dots,N-1}$ a partition of $\mathcal{I}$. Then, the staircase function $F : \mathcal{I} \rightrightarrows \mathbb{R}^n$ with*

$$F(t) = \sum_{i=0}^{N-2} \chi_{[\tau_i, \tau_{i+1})}(t) U_i + \chi_{[\tau_{N-1}, \tau_N]}(t) U_{N-1}$$

is Riemann-integrable with

$$\text{(R)-}\int_{\mathcal{I}} F(t)dt = \sum_{i=0}^{N-1} (\tau_{i+1} - \tau_i) \operatorname{co}(U_i).$$

Proof. Set $\widetilde{F}_i(t) := \chi_{[\tau_i, \tau_{i+1})}(t) U_i$ for $i = 0, \dots, N-2$ resp. $\widetilde{F}_{N-1}(t) := \chi_{[\tau_{N-1}, \tau_N]}(t) U_{N-1}$, then $\widetilde{F}_i(\cdot)$ is contant on $[\tau_i, \tau_{i+1} - \varepsilon]$ for every $\varepsilon > 0$. Lemma 4.35 shows the Riemann-integrability with

$$\text{(R)-}\int_{\tau_i}^{\tau_{i+1}-\varepsilon} \widetilde{F}_i(t)dt = (\tau_{i+1} - \varepsilon - \tau_i) \operatorname{co}(U_i).$$

Taking the limit $\varepsilon \searrow 0$ shows integrability on $[\tau_i, \tau_{i+1}]$ with value $(\tau_{i+1} - \tau_i) \operatorname{co}(U_i)$ (for $i = N-1$ this is easier to show).
With Lemma 4.34 we know the Riemann-integrability of $\widetilde{F}_i(\cdot)$ with

$$\text{(R)-}\int_{\mathcal{I}} \widetilde{F}_i(t)dt = \text{(R)-}\int_{\tau_i}^{\tau_{i+1}} \widetilde{F}_i(t)dt = (\tau_{i+1} - \tau_i) \operatorname{co}(U_i).$$

Since the sum of all $\widetilde{F}_i(\cdot)$ coincides with $F(\cdot)$ on $\mathcal{I}$, we have by Lemma 4.36 the Riemann-integrability of $F(\cdot)$ with

$$\text{(R)-}\int_{\mathcal{I}} F(t)dt = \sum_{i=0}^{N-1} \text{(R)-}\int_{\mathcal{I}} \widetilde{F}_i(t)dt = \sum_{i=0}^{N-1} (\tau_{i+1} - \tau_i) \operatorname{co}(U_i).$$

□

Proposition 4.38. *Let $\mathcal{I} = [t_0, t_f]$, $f : \mathcal{I} \to \mathbb{R}^n$ and $F : \mathcal{I} \rightrightarrows \mathbb{R}^n$ with $F(t) = \{f(t)\}$. Then, $F(\cdot)$ is Riemann-integrable, if and only if $f(\cdot)$ is Riemann-integrable and (R)-$\int_{\mathcal{I}} F(t)dt = \{$(R)-$\int_{\mathcal{I}} f(t)dt\}$.*

Proof. "⇒": Let $\varepsilon_m = \frac{1}{m}$ and choose a partition with $\tau_i^{(m)} := t_0 + ih_m$, fineness $h_m := \frac{t_f - t_0}{N_m} \leq \delta$ and $\xi_i^{(m)} := \tau_i^{(m)}$. By the Riemann-integrability of $F(\cdot)$ we have

$$\mathrm{d_H}(\sum_{i=0}^{N_m-1}(\tau_{i+1}^{(m)} - \tau_i^{(m)})\{f(\xi_i^{(m)})\}, (\mathrm{R})\text{-}\int_{\mathcal{I}} F(t)dt) \leq \varepsilon.$$

By Lemma 3.147 the Riemann-integral of $F(\cdot)$ consists only of a single element $x \in \mathbb{R}^n$. Now, choose $\varepsilon > 0$ arbitrary, take δ from Definition 4.8 and consider an arbitrary partition a partition $([\tau_i, \tau_{i+1}])_{i=0,\ldots,N-1}$ with fineness not exceeding δ and $\xi_i \in [\tau_i, \tau_{i+1}]$.

By the Riemann-integrability of $F(\cdot)$ we use Lemma 3.137(i) and have

$$\varepsilon \geq \mathrm{d_H}(\sum_{i=0}^{N-1}(\tau_{i+1} - \tau_i)F(\xi_i), (\mathrm{R})\text{-}\int_{\mathcal{I}} F(t)dt) = \|\sum_{i=0}^{N-1}(\tau_{i+1} - \tau_i)f(\xi_i) - x\|,$$

i.e. the Riemann-integrability of $f(\cdot)$ with (R)-$\int_{\mathcal{I}} f(t)dt = x$.
"⇐": If $f(\cdot)$ is Riemann-integrable, we can reverse by Lemma 3.137(i) the arguments showing that $F(\cdot)$ is Riemann-integrable with

$$(\mathrm{R})\text{-}\int_{\mathcal{I}} F(t)dt = \{(\mathrm{R})\text{-}\int_{\mathcal{I}} f(t)dt\}.$$

□

Proposition 4.39. *Let $\mathcal{I} = [t_0, t_f]$ and $F : \mathcal{I} \Rightarrow \mathbb{R}^n$ be bounded with values in $\mathcal{K}(\mathbb{R}^n)$. If $F(\cdot)$ is Riemann-integrable, then for all $\eta \in \mathbb{R}^n$ the support function $\delta^\star(\eta, F(\cdot))$ is Riemann-integrable.*
If for all $\eta \in S_{n-1}$ the support function $\delta^\star(\eta, F(\cdot))$ is uniformly Riemann-integrable (i.e. δ in Definition 4.8 is independent from η), then $F(\cdot)$ is Riemann-integrable.
In both cases,

$$\delta^\star(\eta, (R)\text{-}\int_{\mathcal{I}} F(t)dt) = (R)\text{-}\int_{\mathcal{I}} \delta^\star(\eta, F(t))dt.$$

Proof. From [Pol83-Int, Bai95] follows that

$$\delta^\star(\eta, \sum_{i=0}^{N-1}(\tau_{i+1} - \tau_i)F(\xi_i)) = \sum_{i=0}^{N-1}(\tau_{i+1} - \tau_i)\delta^\star(\eta, F(\xi_i)).$$ □

Proposition 4.40. *Let $\mathcal{I} = [t_0, t_f]$ and $F : \mathcal{I} \Rightarrow \mathbb{R}^n$ be bounded with values in $\mathcal{K}(\mathbb{R}^n)$. Then, $F(\cdot)$ is Riemann-integrable, if and only if it is continuous a.e. on $\mathcal{I}$.*

Proof. see [Pol83-Int, Theorem 8] □

Proposition 4.41. *Let $\mathcal{I} = [t_0, t_f]$ and $F, G : \mathcal{I} \Rightarrow \mathbb{R}^n$ be Riemann-integrable with values in $\mathcal{K}(\mathbb{R}^n)$. Then, $t \mapsto \mathrm{d_H}(F(t), G(t))$ is Riemann-integrable and*

$$\mathrm{d_H}((R)\text{-}\int_{\mathcal{I}} F(t)dt, (R)\text{-}\int_{\mathcal{I}} G(t)dt) \leq (R)\text{-}\int_{\mathcal{I}} \mathrm{d_H}(F(t), G(t))dt.$$

Proof. Consider the map $\varphi(t) := \mathrm{d_H}(F(t), G(t))$. From the assumptions, it is bounded, since for all $t \in \mathcal{I}$

$$\mathrm{d_H}(F(t), G(t)) \leq \|F(t)\| + \|G(t)\| \leq C_1 + C_2 < \infty.$$

From Proposition 4.40 follows that there exists subsets $\mathcal{N}_\mu \subset \mathcal{I}$ of measure zero such that for all $\varepsilon > 0$ and all $t \in \mathcal{I} \setminus \mathcal{N}_\mu$ there exists $\delta_\mu > 0$ such that for all $\tau \in \mathcal{I} \setminus \mathcal{N}_\mu$ with $|t - \tau| \leq \delta$ follows that

$$\mathrm{d_H}(F_\mu(t), F_\mu(\tau)) \leq \frac{\varepsilon}{2}$$

with $\mu = 1, 2$ and $F_1(\cdot) := F(\cdot)$, $F_2(\cdot) := G(\cdot)$. For the set $\mathcal{N} := \mathcal{N}_1 \cup \mathcal{N}_2$ of measure zero and given $\varepsilon > 0$ we set $\delta := \min\{\delta_1, \delta_2\} > 0$ and consider $t \in \mathcal{I} \setminus \mathcal{N}$. Then,

$$\begin{aligned} \mathrm{d_H}(F(t), G(t)) &\le \mathrm{d_H}(F(t), F(\tau)) + \mathrm{d_H}(F(\tau), G(\tau)) + \mathrm{d_H}(G(\tau), G(t)), \\ \mathrm{d_H}(F(\tau), G(\tau)) &\le \mathrm{d_H}(F(\tau), F(t)) + \mathrm{d_H}(F(t), G(t)) + \mathrm{d_H}(G(t), G(\tau)) \end{aligned}$$

for all $\tau \in \mathcal{I} \setminus \mathcal{N} \subset \mathcal{I} \setminus \mathcal{N}_\mu$, $\mu = 1, 2$. Now,

$$\begin{aligned} |\,\mathrm{d_H}(F(t), G(t)) - \mathrm{d_H}(G(\tau), G(\tau))| &\le \mathrm{d_H}(F(t), F(\tau)) + \mathrm{d_H}(G(t), G(\tau)) \\ &\le \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon. \end{aligned}$$

Hence, $\varphi(\cdot)$ is continuous a.e. and by Propositions 4.40 and 4.38 also Riemann-integrable.

Now, let $\varepsilon_m = \frac{1}{3m}$ and choose a partition with $\tau_i^{(m)} := t_0 + i h_m$, fineness $h_m := \frac{t_f - t_0}{N_m} \le \delta$ and $\xi_i^{(m)} := \tau_i^{(m)}$, where δ is the minimum of $(\delta_\mu)_{\mu=1,2,3}$. $\delta_\mu = \delta_\mu(\varepsilon_m)$ are the numbers in Definition 4.8 for the functions $F(\cdot)$, $G(\cdot)$ resp. $\varphi(\cdot)$.
Then,

$$\begin{aligned} &\mathrm{d_H}((\mathrm{R})\text{-}\int_{\mathcal{I}} F(t)dt, (\mathrm{R})\text{-}\int_{\mathcal{I}} G(t)dt) \\ \le\, &\mathrm{d_H}((\mathrm{R})\text{-}\int_{\mathcal{I}} F(t)dt, \sum_{i=0}^{N_m-1} (\tau_{i+1}^{(m)} - \tau_i^{(m)}) F(\xi_i^{(m)})) \end{aligned}$$

$$\begin{aligned} &+ \mathrm{d_H}(\sum_{i=0}^{N_m-1} (\tau_{i+1}^{(m)} - \tau_i^{(m)}) F(\xi_i^{(m)}), \sum_{i=0}^{N_m-1} (\tau_{i+1}^{(m)} - \tau_i^{(m)}) G(\xi_i^{(m)})) \\ &+ \mathrm{d_H}(\sum_{i=0}^{N_m-1} (\tau_{i+1}^{(m)} - \tau_i^{(m)}) G(\xi_i^{(m)}), (\mathrm{R})\text{-}\int_{\mathcal{I}} G(t)dt) \\ \le\, &\frac{1}{3m} + \sum_{i=0}^{N_m-1} (\tau_{i+1}^{(m)} - \tau_i^{(m)}) \cdot \mathrm{d_H}(F(\xi_i^{(m)}), G(\xi_i^{(m)})) + \frac{1}{3m} \end{aligned}$$

Now, by letting m tend to infinity, the second term on the right-hand side approaches the Riemann integral, the other terms vanish and we have

$$\mathrm{d_H}((\mathrm{R})\text{-}\int_{\mathcal{I}} F(t)dt, (\mathrm{R})\text{-}\int_{\mathcal{I}} G(t)dt) \le (\mathrm{R})\text{-}\int_{\mathcal{I}} \mathrm{d_H}(F(t), G(t))dt.$$

□

4.3.2 Aumann's Integral

Definition 4.42. Let $\mathcal{I} = [t_0, t_f]$ and $F : \mathcal{I} \Rightarrow \mathbb{R}^n$ be a set-valued mapping. Then,

$$\begin{aligned} \int_{\mathcal{I}} F(\tau)d\tau = \big\{ z \in \mathbb{R}^n : &\text{ there exists an integrable selection} \\ &f(\cdot) \text{ of } F(\cdot) \text{ on } I \text{ with } z = \int_{\mathcal{I}} f(\tau)d\tau \big\} \end{aligned}$$

is called *Aumann's integral* of $F(\cdot)$.

Proposition 4.43. *Let $\mathcal{I} = [t_0, t_f]$, $f : \mathcal{I} \to \mathbb{R}^n$ be (Lebesgue-)integrable and $F : \mathcal{I} \Rightarrow \mathbb{R}^n$ with $F(t) = \{f(t)\}$. Then,*

$$\int_{\mathcal{I}} F(t)dt = \{\int_{\mathcal{I}} f(t)dt\}.$$

Theorem 4.44. *Let $\mathcal{I} = [t_0, t_f]$ and $\boldsymbol{F} : \mathcal{I} \Rightarrow \mathbb{R}^n$ be a measurable set-valued mapping with nonempty and closed images. Then,*

$$\int_{\mathcal{I}} \boldsymbol{F}(\tau) d\tau = \int_{\mathcal{I}} \overline{\mathrm{co}}\, \boldsymbol{F}(\tau) d\tau$$

is convex.
If, moreover, $\boldsymbol{F}(\cdot)$ is integrably bounded, then

$$\int_{\mathcal{I}} \boldsymbol{F}(\tau) d\tau = \int_{\mathcal{I}} \mathrm{co}(\boldsymbol{F}(\tau)) d\tau$$

is nonempty, compact, and convex.

Proof. see [Aum65-Int] □

Proposition 4.45. *Let $\mathcal{I} = [t_0, t_f]$ and $\boldsymbol{F} : \mathcal{I} \Rightarrow \mathbb{R}^n$ be a measurable, integrably bounded set-valued mapping with nonempty and closed images. Then,*

$$\delta^\star(l, \int_{\mathcal{I}} \boldsymbol{F}(t) dt) = \int_{\mathcal{I}} \delta^\star(l, \boldsymbol{F}(t)) dt \quad (l \in \mathbb{R}^n).$$

Proof. see [AF90-Int, Proposition 8.6.2, 3.] □

Example 4.46 *(constant set-valued map).* Let $\mathcal{I} = [t_0, t_f]$, $\boldsymbol{U} \in \mathcal{K}(\mathbb{R}^n)$ and $\boldsymbol{F} : \mathcal{I} \Rightarrow \mathbb{R}^n$ with $\boldsymbol{F}(t) = \boldsymbol{U}$ for all $t \in \mathcal{I}$. Then, the Aumann-integral gives

$$\int_{\mathcal{I}} \boldsymbol{F}(t) dt = (t_f - t_0)\, \mathrm{co}(\boldsymbol{U}).$$

Proof. $\boldsymbol{F}(\cdot) = \chi_{\mathcal{I}}(\cdot)\boldsymbol{U}$ is a simple map with measurable $\mathcal{I}$, hence measurable by Remark 4.6. Since

$$\int_{\mathcal{I}} \|\boldsymbol{F}(t)\| dt = \int_{\mathcal{I}} \|\boldsymbol{U}\| dt = (t_f - t_0)\|\boldsymbol{U}\| < \infty,$$

the map is also integrably bounded.

Now, apply Theorem 4.44 and Proposition 4.45:

$$\begin{aligned} \delta^\star(l, \int_{\mathcal{I}} \boldsymbol{F}(t) dt) &= \int_{\mathcal{I}} \delta^\star(l, \underbrace{\boldsymbol{F}(t)}_{=\boldsymbol{U}}) dt \\ &= (t_f - t_0)\delta^\star(l, \boldsymbol{U}) = (t_f - t_0)\delta^\star(l, \mathrm{co}\,\boldsymbol{U}) \end{aligned}$$

□

Corollary 4.47. *Let $\mathcal{I} = [t_0, t_f]$, $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, $\mu \in \mathbb{R}$ and $\boldsymbol{F}, \boldsymbol{G} : \mathcal{I} \Rightarrow \mathbb{R}^n$ be measurable, integrably bounded set-valued mappings with nonempty and closed images. Then,*

$$\begin{aligned} &\textit{(i)} && \int_{\mathcal{I}} (\boldsymbol{F}(\tau) + \boldsymbol{G}(\tau)) d\tau = \int_{\mathcal{I}} \boldsymbol{F}(\tau) d\tau + \int_{\mathcal{I}} \boldsymbol{G}(\tau) d\tau, \\ &\textit{(ii)} && \int_{\mathcal{I}} (\mu \cdot \boldsymbol{F}(\tau)) d\tau = \mu \cdot \int_{\mathcal{I}} \boldsymbol{F}(\tau) d\tau, \\ &\textit{(iii)} && \int_{\mathcal{I}} (\boldsymbol{A} \cdot \boldsymbol{F}(\tau)) d\tau = \boldsymbol{A} \cdot \int_{\mathcal{I}} \boldsymbol{F}(\tau) d\tau, \end{aligned}$$

if additionally in (i) and (iii) the images of $\boldsymbol{A} \cdot \boldsymbol{F}(\cdot)$ resp. $\boldsymbol{F}(\cdot) + \boldsymbol{G}(\cdot)$ are all closed.

Proof. We will prove all claims by showing that there support functions are equal and use Proposition 3.61 and Proposition 4.45. This is justified by remembering Theorem 4.44.
(ii) Let $\mu \geq \mathbf{0}$. Then, Proposition 3.115 shows

$$\begin{aligned}\delta^\star(l, \int_{\mathcal{I}} \mu \cdot F(\tau) d\tau) &= \int_{\mathcal{I}} \delta^\star(l, \mu \cdot F(\tau)) d\tau \\ = \int_{\mathcal{I}} \mu \cdot \delta^\star(l, F(\tau)) d\tau &= \mu \cdot \int_{\mathcal{I}} \delta^\star(l, F(\tau)) d\tau \\ = \mu \cdot \delta^\star(l, \int_{\mathcal{I}} F(\tau) d\tau) &= \delta^\star(l, \mu \cdot \int_{\mathcal{I}} F(\tau) d\tau)\end{aligned}$$

and hence equality in (ii).

Now, let $\mu < \mathbf{0}$. We will use the first part for non-negative scalars for $|\mu|$ instead of μ:

$$\begin{aligned}\delta^\star(l, \int_{\mathcal{I}} \mu \cdot F(\tau) d\tau) &= \int_{\mathcal{I}} \delta^\star(l, \mu \cdot F(\tau)) d\tau \\ = \int_{\mathcal{I}} \delta^\star(-l, |\mu| \cdot F(\tau)) d\tau &= \delta^\star(-l, |\mu| \cdot \int_{\mathcal{I}} F(\tau) d\tau) \\ = \delta^\star(l, \mu \cdot \int_{\mathcal{I}} F(\tau) d\tau)&\end{aligned}$$

The other proofs are similar and use again Proposition 3.115 resp. 3.116. □

Corollary 4.48. *Let $\mathcal{I} = [t_0, t_f]$ and $F, G : \mathcal{I} \Rightarrow \mathbb{R}^n$ be measurable, integrably bounded set-valued mappings with nonempty and closed images and $F(t) \subset G(t)$ for almost every $t \in \mathcal{I}$. Then,*

$$\int_{\mathcal{I}} F(\tau) d\tau \subset \int_{\mathcal{I}} G(\tau) d\tau.$$

Proof. We will use Proposition 3.120 as well as Theorem 4.44 and Proposition 4.45:

$$\begin{aligned}\delta^\star(l, \int_{\mathcal{I}} F(\tau) d\tau) &= \int_{\mathcal{I}} \delta^\star(l, F(\tau)) d\tau \\ \leq \int_{\mathcal{I}} \delta^\star(l, G(\tau)) d\tau &= \delta^\star(l, \int_{\mathcal{I}} G(\tau) d\tau) \quad (l \in S_{n-1})\end{aligned}$$

□

The next result is the set-valued version of Lebesgue's dominated convergence theorem (cf. Theorem A.14).

Proposition 4.49. *Let $\mathcal{I} = [t_0, t_f]$ and $F_m : \mathcal{I} \Rightarrow \mathbb{R}^n$ be measurable and integrably bounded set-valued mappings with nonempty images with the same function $k(\cdot) \in L_1(\mathcal{I})$, i.e. there exists $\mathcal{N} \subset \mathcal{I}$ of measure zero such that for every $m \in \mathbb{N}$*

$$\|F_m(t)\| \leq k(t) \quad \text{for } t \in \mathcal{I} \setminus \mathcal{N}.$$

Let $F : \mathcal{I} \Rightarrow \mathbb{R}^n$ be defined as the limit of $(F_m(\cdot))_{m \in \mathbb{N}}$, i.e.

$$\mathbf{d_H}(F_m(t), F(t)) \xrightarrow[m \to \infty]{} \mathbf{0} \quad \text{for all } t \in \mathcal{I}.$$

Then,

$$\mathbf{d_H}(\int_{\mathcal{I}} F_m(\tau) d\tau, \int_{\mathcal{I}} F_m(\tau) d\tau) \xrightarrow[m \to \infty]{} \mathbf{0}.$$

Proof. cf. [Aum65, Theorem 5] □

Proposition 4.50. *Let $\mathcal{I} = [t_0, t_f]$ and $F : \mathcal{I} \Rightarrow \mathbb{R}^n$ be Riemann integrable with values in $\mathcal{C}(\mathbb{R}^n)$. Then, the Aumann integral and the Riemann integral coincide.*

Proof. see [Pol75-Int, Theorem 4] □

Corollary 4.51. *Let $\mathcal{I} = [t_0, t_f]$ and $\boldsymbol{F} : \mathcal{I} \Rightarrow \mathbb{R}^n$ be measurable, integrably bounded set-valued mappings with nonempty and closed images. Then, for a partition $(\mathcal{I}_i)_{i=0,\dots,N-1}$ of $\mathcal{I}$ we have*

$$\int_{\mathcal{I}} \boldsymbol{F}(\tau)d\tau = \sum_{i=0}^{N-1} \int_{\mathcal{I}_i} \boldsymbol{F}(\tau)d\tau.$$

Proof. Consider the Carathéodory maps $\varphi_i(t,x) = \chi_{\mathcal{I}_i}(t)x$ and the set-valued mappings $\boldsymbol{F}_i(t) := \chi_{\mathcal{I}_i}(t)\boldsymbol{F}(t) = \varphi_i(t, \boldsymbol{F}(t)) = \mathbf{cl}(\varphi_i(t, \boldsymbol{F}(t)))$ which are measurable by Proposition 4.26. Clearly, they are also integrably bounded and have closed images by Proposition 3.106(iii). Proposition 4.23 shows that the sum

$$\sum_{i=0}^{N-1} \boldsymbol{F}_i(t) = \boldsymbol{F}(t) \tag{20}$$

is measurable and has closed images by Lemma 3.84.

Then, Propositions 4.27, 4.45 and Theorem 4.44 and allows us to prove (21) by looking at the support functions.

$$\begin{aligned}
\delta^\star(l, \int_{\mathcal{I}} \boldsymbol{F}_i(\tau)d\tau) &= \int_{\mathcal{I}} \delta^\star(l, \boldsymbol{F}_i(\tau))d\tau = \int_{\mathcal{I}} \delta^\star(l, \chi_{\mathcal{I}_i}(\tau)\boldsymbol{F}(\tau))d\tau \\
&= \int_{\mathcal{I}} \chi_{\mathcal{I}_i}(\tau)\delta^\star(l, \boldsymbol{F}(\tau))d\tau = \int_{\mathcal{I}_i} \delta^\star(l, \boldsymbol{F}(\tau))d\tau \\
&= \delta^\star(l, \int_{\mathcal{I}_i} \boldsymbol{F}(\tau)d\tau),
\end{aligned}$$

i.e.

$$\int_{\mathcal{I}} \boldsymbol{F}_i(\tau)d\tau = \int_{\mathcal{I}_i} \boldsymbol{F}(\tau)d\tau. \tag{21}$$

Corollary 4.47(i) and (20) justifies to sum the equations which yields

$$\begin{aligned}
\int_{\mathcal{I}} \boldsymbol{F}(\tau)d\tau &= \int_{\mathcal{I}} \sum_{i=0}^{N-1} \boldsymbol{F}_i(\tau)d\tau \\
&= \sum_{i=0}^{N-1} \int_{\mathcal{I}} \boldsymbol{F}_i(\tau)d\tau = \sum_{i=0}^{N-1} \int_{\mathcal{I}_i} \boldsymbol{F}(\tau)d\tau
\end{aligned}$$

□

Proposition 4.52. *Let $\mathcal{I} = [t_0, t_f]$ and $\boldsymbol{F}, \boldsymbol{G} : \mathcal{I} \Rightarrow \mathbb{R}^n$ be measurable, integrably bounded with values in $\mathcal{K}(\mathbb{R}^n)$. Then, $t \mapsto \mathbf{d_H}(\boldsymbol{F}(t), \boldsymbol{G}(t))$ is integrable and*

$$\mathbf{d_H}(\int_{\mathcal{I}} \boldsymbol{F}(t)dt, \int_{\mathcal{I}} \boldsymbol{G}(t)dt) \leq \int_{\mathcal{I}} \mathbf{d_H}(\boldsymbol{F}(t), \boldsymbol{G}(t))dt.$$

Proof. see [Pol83-Int, Theorem 2] □

For Further Reading

References

[AC84-Int] J.-P. Aubin and A. Cellina. *Differential Inclusions*, volume 264 of *Grundlehren der mathematischen Wissenschaften*. Springer-Verlag, Berlin–Heidelberg–New York–Tokyo, 1984.

[AF90-Int] J.-P. Aubin and H. Frankowska. *Set-Valued Analysis*, volume 2 of *Systems & Control: Foundations and Applications*. Birkhäuser, Boston–Basel–Berlin, 1990.

[Kis91-Int] M. Kisielewicz. *Differential Inclusions and Optimal Control*, volume 44 of *Mathematics and Its Applications*. PWN - Polish Scientific Publishers, Warszawa–Dordrecht–Boston–London, 1991.

[BF87-Int] V. I. Blagodatskikh and A. F. Filippov. Differential inclusions and optimal control. In *Topology, Ordinary Differential Equations, Dynamical Systems*, volume 1986, issue 4 of *Proceedings of the Steklov Institute of Mathematics*, pages 199–259. AMS, Providence–Rhode Island, 1987.

[CV77-Int] C. Castaing and M. Valadier. *Convex Analysis and Measurable Multifunctions*, volume 580 of *Lecture Notes in Math.* Springer-Verlag, Berlin–Heidelberg–New York, 1977.

[Pol75-Int] E. S. Polovinkin. *Riemannian Integral of Set-Valued Function*, pages 405–410. Lecture Notes in Computer Science 27, Optimization Techniques, IFIP Technical Conference, Novosibirsk, July 1–7, 1974. Springer-Verlag, Berlin–Heidelberg–New York, 1975.

[Pol83-Int] E. S. Polovinkin. On integration of multivalued mappings. *Dokl. Akad. Nauk SSSR*, 28(1):223–228, 1983.

[Aum65-Int] R. J. Aumann. Integrals of Set-Valued Functions. *J. Math. Anal. Appl.*, 12(1):1–12, 1965.

[HU77-Int] F. Hiai and H. Umegaki. Integrals, Conditional Expectations, and Martingales of Multivalued Functions. *J. Multivariate Anal.*, 7(1):149–182, 1977.

[Deb67-Int] G. Debreu. Integration of correspondences. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Held at the Statistical Laboratory, University of California, June 21–July 18, 1965 and December 27, 1965–January 7, 1966*, Volume II: Contributions to Probability Theory, Part 1, pages 351–372, Berkeley–Los Angeles, 1967. University of California Press.

[Bri70-Int] T. F. Bridgland, Jr. Trajectory integrals of set valued functions. *Pac. J. Math.*, 33(1):43–68, 1970.

[Ole65-Int] C. Olech. A note concerning set-valued measurable functions. *Bull. Polish Acad. Sci. Math.*, 13:317–1965, 1965.

[Pap85a-Int] N. S. Papageorgiou. On the Theory of Banach Space Valued Multifunctions. 1. Integration and Conditional Expectation. *J. Multivariate Anal.*, 17(2):185–206, 1985.

[Sam99-Int] A. R. Sambucini. Remarks on set valued integrals of multifunctions with non empty bounded closed and convex values. *Ann. Soc. Math. Pol., Ser. I, Commentat. Math.*, 39:153–165, 1999.

[Mik78-Int] J. Mikusiński. *The Bochner Integral*, volume 55 of *Lehrbücher und Monographien aus dem Gebiete der exakten Wissenschaften: Math. Reihe*. Birkhäuser Verlag, Basel–Stuttgart, 1978.

5 Numerical Solution of Initial Value Problems

Initial Value Problems

Problem 5.1 (Initial Value Problem). *Let $f : [t_0, t_f] \times \mathbb{R}^{n_x} \to \mathbb{R}^{n_x}$ and $x_0 \in \mathbb{R}^{n_x}$ be given. Find x with*

$$\begin{aligned} x(t_0) &= x_0, \\ \dot{x}(t) &= f(t, x(t)) \end{aligned}$$

in the interval $[t_0, t_f]$.

5.1 Existence and Uniqueness

Existence

Theorem 5.2 (Peano, cf. Walter [Wal90], Paragraphs 7,10).

(a) Local Existence: Let $D \subseteq \mathbb{R} \times \mathbb{R}^{n_x}$ be open and $(t_0, x_0) \in D$. Furthermore, let f be continuous in D. Then, the initial value problem 5.1 possesses at least one locally defined solution around t_0. The solution can be continued to the boundary of D.

(b) Global Existence: Let f be continuous and bounded in $[t_0, t_f] \times \mathbb{R}^{n_x}$. Then, there exists at least one differentiable solution of the initial value problem 5.1 in $[t_0, t_f]$.

Existence and Uniqueness I

Theorem 5.3 (Picard-Lindelöf, cf. Walter [Wal90], Paragraph 10). *Local Existence and Uniqueness: Let $D \subseteq \mathbb{R} \times \mathbb{R}^{n_x}$ open and $(t_0, x_0) \in D$. Furthermore, let f be continuous in D and locally lipschitz-continuous w.r.t. x, i.e. for every $(\hat{t}, \hat{x}) \in D$ there exists a neighborhood $U_\varepsilon(\hat{t}, \hat{x})$ and a constant $L = L(\hat{t}, \hat{x})$ with*

$$\|f(t, y) - f(t, z)\| \leq L\|y - z\| \qquad \forall (t, y), (t, z) \in D \cap U_\varepsilon(\hat{t}, \hat{x}).$$

Then, the initial value problem 5.1 possesses a locally unique solution around t_0. The solution can be continued to the boundary of D.

Existence and Uniqueness II

Theorem 5.4 (Picard-Lindelöf, cf. Walter [Wal90], Paragraph 10). *Global Existence and Uniqueness: Let f be continuous in $[t_0, t_f] \times \mathbb{R}^{n_x}$ and lipschitz-continuous w.r.t. x, i.e. there exists a constant L with*

$$\|f(t, y) - f(t, z)\| \leq L\|y - z\| \qquad \forall (t, y), (t, z) \in [t_0, t_f] \times \mathbb{R}^{n_x}.$$

Then, there exists a unique solution of the initial value problem 5.1 in $[t_0, t_f]$.

Idea of Proof

Proof. The proof of the theorems of Picard-Lindelöf is based on the *fixed point iteration*

$$y_{i+1}(t) = (Ty_i)(t), \qquad i = 0, 1, 2, \ldots$$

for the operator

$$(Ty)(t) := y(t_0) + \int_{t_0}^{t} f(\tau, y(\tau)) d\tau$$

and the application of *Banach's fixed-point theorem*. □

Examples I

Example 5.5 *(Multiple Solutions).* Consider

$$\dot{x}(t) = -2\sqrt{1 - x(t)} =: f(x(t)), \qquad x(0) = 1.$$

Verify that

$$\begin{aligned} x(t) &= 1, \\ x(t) &= 1 - t^2 \end{aligned}$$

both are solutions. *f does not satisfy a lipschitz-condition at $x = 1$.*

Examples II

Example 5.6 *(Locally Unique Solution).* Consider

$$\dot{x}(t) = x(t)^2 =: f(x(t)), \quad x(0) = x_0 \in \mathbb{R}.$$

$f(x) = x^2$ is *not lipschitz-continuous on* $\mathbb{R}$ but only locally lipschitz-continuous!

Locally unique solution:

$$x(t) = -\frac{x_0}{x_0 t - 1}.$$

If $t = 1/x_0$, $x_0 \neq 0$, then x is unbounded.

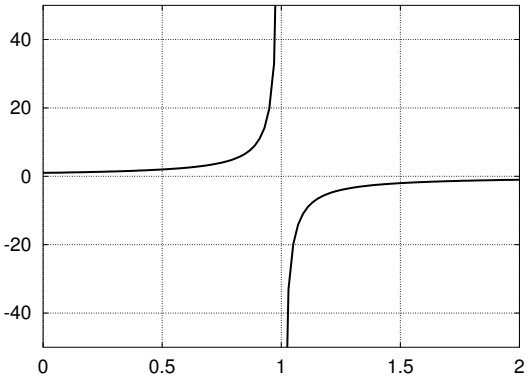
If $x_0 > 0$, the solution is defined in $(-\infty, 1/x_0)$.

If $x_0 < 0$, the solution is defined in $(1/x_0, \infty)$.

If $x_0 = 0$, then $x(t) \equiv 0$ is defined on $\mathbb{R}$.

Examples III

Solution for $x(0) = x_0 = 1$:



Examples IV

Example 5.7 *(Globally Unique Solution).* Consider

$$\dot{x}(t) = \lambda x(t) =: f(x(t)), \qquad x(0) = x_0 \in \mathbb{R}$$

with $\lambda \in \mathbb{R}$. f is globally lipschitz-continuous. The globally unique solution is given by

$$x(t) = x_0 \cdot \exp(\lambda t).$$

5.2 One-Step Methods

One-Step Methods

Grid

$$\mathbb{G}_h := \{t_0 < t_1 < \ldots < t_N = t_f\}$$

Stepsize

$$h_i = t_{i+1} - t_i, \quad i = 0, 1, \ldots, N-1, \quad h = \max_{i=0,\ldots,N-1} h_i.$$

Increment function

$$\Phi(t, x, h)$$

General One-Step-Method:

Construct grid function $x_h : \mathbb{G}_h \to \mathbb{R}^{n_x}$ by

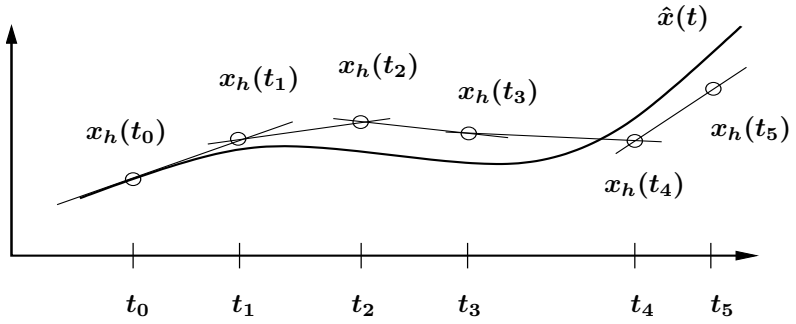
$$\begin{aligned} x_h(t_0) &= x_0, \\ x_h(t_{i+1}) &= x_h(t_i) + h_i \Phi(t_i, x_h(t_i), h_i), \quad i = 0, \ldots, N-1. \end{aligned}$$

Explicit Euler Method

Example 5.8 *(Explicit Euler's Method).* Special choice $\Phi(t,x,h) := f(t,x)$ yields

$$\begin{aligned} x_h(t_0) &= x_0, \\ x_h(t_{i+1}) &= x_h(t_i) + h_i f(t_i, x_h(t_i)), \quad i = 0, \ldots, N-1. \end{aligned}$$

Idea: local linearization



Connection: Explicit Euler Method and Riemann Sums

Special case: f does not depend on x, i.e. $f = f(t)$!
Solution of $\dot{x}(t) = f(t)$, $x(t_0) = 0$:

$$x(t) = \int_{t_0}^{t} f(\tau) d\tau$$

Euler's method: equidistant grid, step size $h = (t_f - t_0)/N$

$$x_h(t_i) = \sum_{j=0}^{i-1} h f(t_j) \approx \int_{t_0}^{t_i} f(\tau) d\tau = x(t_i)$$

Explicit Euler's method corresponds to a Riemann sum!

Implicit Euler Method

Example 5.9 *(Implicit Euler's Method).*

$$\begin{aligned} x_h(t_0) &= x_0, \\ x_h(t_{i+1}) &= x_h(t_i) + h_i f(t_{i+1}, x_h(t_{i+1})), \quad i = 0, \ldots, N-1. \end{aligned}$$

Increment function is defined implicitly:

$$\Phi(t,x,h) = f(t+h, x+h\Phi(t,x,h)).$$

Heun's Method

Example 5.10 *(Heun's Method).*

$$\begin{aligned} x_h(t_0) &= x_0, \\ k_1 &= f(t_i, x_h(t_i)), \\ k_2 &= f(t_i + h_i, x_h(t_i) + h_i k_1), \\ x_h(t_{i+1}) &= x_h(t_i) + \frac{h_i}{2}(k_1 + k_2), \quad i = 0, \ldots, N-1. \end{aligned}$$

Increment function:

$$\Phi(t,x,h) = \frac{1}{2}\left(f(t,x) + f(t+h, x+hf(t,x))\right).$$

Connection: Heun's method and Trapezoidal Rule

Special case: f does not depend on x, i.e. $f = f(t)$!
Solution of $\dot{x}(t) = f(t)$, $x(t_0) = 0$:

$$x(t) = \int_{t_0}^{t} f(\tau) d\tau$$

Heun's method: equidistant grid, step size $h = (t_f - t_0)/N$

$$x_h(t_i) = \frac{h}{2} \sum_{j=0}^{i-1} (f(t_j) + f(t_{j+1})) = \underbrace{h \left(\frac{f(t_0)}{2} + \sum_{j=1}^{i-1} f(t_j) + \frac{f(t_i)}{2} \right)}_{\approx \int_{t_0}^{t_i} f(\tau) d\tau = x(t_i)}$$

Heun's method corresponds to the iterated trapezoidal rule!

Modified Euler's Method

Example 5.11 *(Modified Euler's Method).*

$$\begin{aligned} x_h(t_0) &= x_0, \\ k_1 &= f(t_i, x_h(t_i)), \\ k_2 &= f(t_i + h_i/2, x_h(t_i) + h_i/2 \cdot k_1), \\ x_h(t_{i+1}) &= x_h(t_i) + h_i k_2, \quad i = 0, \ldots, N-1. \end{aligned}$$

Increment function:

$$\Phi(t, x, h) = f(t + h/2, x + h/2 \cdot f(t, x)).$$

Runge-Kutta Methods I

Example 5.12 *(Runge-Kutta-Methods).* For $s \in \mathbb{N}$ and $b_j, c_j \in [0, 1], a_{ij}$, $i, j = 1, \ldots, s$ the s-stage Runge-Kutta-Method is defined by

$$\begin{aligned} x_h(t_{i+1}) &= x_h(t_i) + h_i \sum_{j=1}^{s} b_j k_j(t_i, x_h(t_i); h_i) \\ k_j(t_i, x_h(t_i); h_i) &= f(t_i + c_j h_i, x_h(t_i) + h_i \sum_{l=1}^{s} a_{jl} k_l(t_i, x_h(t_i); h_i)) \end{aligned}$$

Increment function:

$$\Phi(t, x, h) = \sum_{j=1}^{s} b_j k_j(t, x; h)$$

Runge-Kutta Methods II

Butcher-Table:

$$\begin{array}{c|cccc} c_1 & a_{11} & a_{12} & \cdots & a_{1s} \\ c_2 & a_{21} & a_{22} & \cdots & a_{2s} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_s & a_{s1} & a_{s2} & \cdots & a_{ss} \\ \hline & b_1 & b_2 & \cdots & b_s \end{array} \quad \Leftrightarrow \quad \begin{array}{c|c} c & A \\ \hline & b^\top \end{array}$$

Explicit methods: $a_{ij} = 0$ for $j \geq i$.

Classic Runge-Kutta Method

Example 5.13 *(Classic Runge-Kutta-Method of order 4).*

$$\begin{array}{c|cccc} 0 & & & & \\ 1/2 & 1/2 & & & \\ 1/2 & 0 & 1/2 & & \\ 1 & 0 & 0 & 1 & \\ \hline & 1/6 & 1/3 & 1/3 & 1/6 \end{array}$$

$$\begin{aligned} x_h(t_{i+1}) &= x_h(t_i) + h_i\left(k_1/6 + k_2/3 + k_3/3 + k_4/6\right) \\ k_1 &= f(t_i, x_h(t_i)), \\ k_2 &= f\left(t_i + h_i/2, x_h(t_i) + h_i/2 \cdot k_1\right), \\ k_3 &= f\left(t_i + h_i/2, x_h(t_i) + h_i/2 \cdot k_2\right), \\ k_4 &= f\left(t_i + h_i, x_h(t_i) + h_i k_3\right). \end{aligned}$$

Radau-IIA Method

Example 5.14 *(Radau-IIA Method).* implicit, 2-stage, order 3

$$\begin{array}{c|cc} 1/3 & 5/12 & -1/12 \\ 1 & 3/4 & 1/4 \\ \hline & 3/4 & 1/4 \end{array}$$

$$\begin{aligned} x_h(t_{i+1}) &= x_h(t_i) + h_i\left(\frac{3}{4}k_1 + \frac{1}{4}k_2\right) \\ k_1 &= f\left(t_i + \frac{h_i}{3}, x_h(t_i) + h_i\left(\frac{5}{12}k_1 - \frac{1}{12}k_2\right)\right), \\ k_2 &= f\left(t_i + h_i, x_h(t_i) + h_i\left(\frac{3}{4}k_1 + \frac{1}{4}k_2\right)\right). \end{aligned}$$

5.3 Convergence of One-Step Methods

Notations

One-Step-Method

$$\begin{aligned} x_h(t_0) &= x_0, \\ x_h(t_{i+1}) &= x_h(t_i) + h\Phi(t_i, x_h(t_i), h), \quad i = 0, \ldots, N-1. \end{aligned}$$

Exact Solution: $\hat{x}$

Restriction Operator on $\mathbb{G}_h$:

$$\Delta_h : \{x : [t_0, t_f] \to \mathbb{R}^{n_x}\} \to \{x_h : \mathbb{G}_h \to \mathbb{R}^{n_x}\}, \qquad x \mapsto \Delta_h(x)$$

Norm for grid functions:

$$\|x_h\|_\infty = \max_{t_i \in \mathbb{G}_h} \|x_h(t_i)\|$$

Convergence

Definition 5.15 (global error, convergence). The global error $e_h : \mathbb{G}_h \to \mathbb{R}^{n_x}$ is defined by

$$e_h := x_h - \Delta_h(\hat{x}).$$

The one-step method is called convergent if

$$\lim_{h \to 0} \|e_h\|_\infty = 0.$$

The one-step method is convergent of order p if

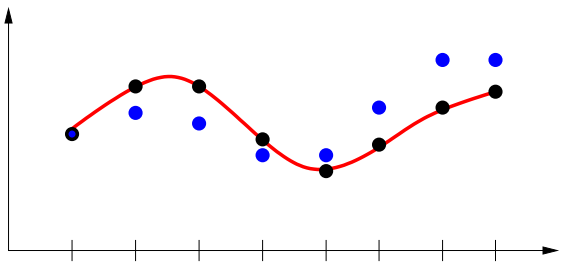
$$\|e_h(\cdot)\|_\infty = \mathcal{O}(h^p) \qquad \text{for } h \to 0.$$

Convergence

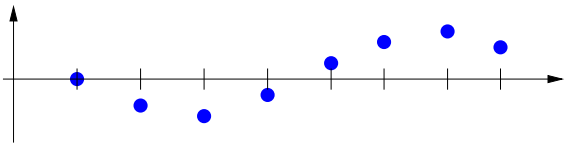
$\hat{x}(\cdot)$

$\Delta_h(\hat{x})(\cdot)$

$x_h(\cdot)$



$e_h(\cdot)$



Consistency: Preliminaries

In the sequel:

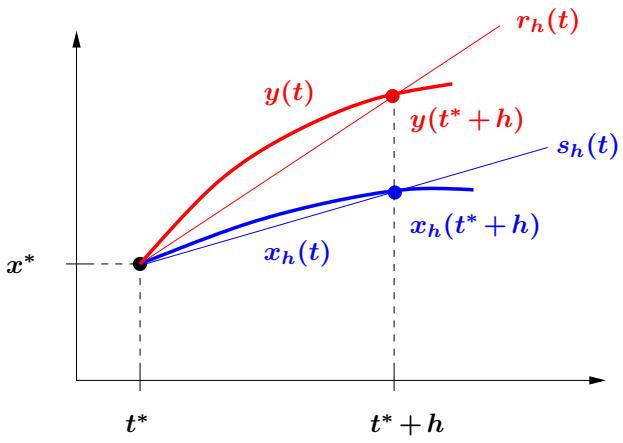
Let y denote the *local solution* of the initial value problem

$$\dot{y}(t) = f(t, y(t)), \qquad y(t^*) = x^*$$

for arbitrary

$$x^* \in \mathbb{R}^{n_x}, \qquad t^* \in [t_0, t_f].$$

Consistency: Motivation



Consistency: Motivation

'Linearizations':

$$\begin{aligned} r_h(t) &= x^* + \frac{t-t^*}{h}\,(y(t^*+h) - x^*) \\ s_h(t) &= x^* + \frac{t-t^*}{h}\,(x_h(t^*+h) - x^*) \end{aligned}$$

Postulation:

$$\lim_{h\to 0}\left(r_h'(t)-s_h'(t)\right)=0$$

i.e.

$$\begin{aligned}0 &= \lim_{h\to 0}\left(r_h'(t)-s_h'(t)\right)\\ &= \lim_{h\to 0}\left(\frac{y(t^*+h)-x^*}{h}-\frac{x_h(t^*+h)-x^*}{h}\right)\\ &= \lim_{h\to 0}\left(\frac{y(t^*+h)-x^*}{h}-\Phi(t^*,x^*,h)\right)\end{aligned}$$

Consistency

Definition 5.16 (local discretization error, consistency). The local discretization error at (t^*,x^*) is defined by

$$\ell_h(t^*,x^*):=\frac{y(t^*+h)-x^*}{h}-\Phi(t^*,x^*,h).$$

The one-step method is called consistent if

$$\lim_{h\to 0}\ell_h(t^*,x^*)=0\qquad\forall(t^*,x^*)\in[t_0,t_f]\times\mathbb{R}^{n_x}.$$

The one-step method is consistent of order p if

$$\ell_h(t^*,x^*)=\mathcal{O}(h^p)\qquad\forall(t^*,x^*)\in[t_0,t_f]\times\mathbb{R}^{n_x}.$$

Remarks

Remark 5.17.

- *Since*
$$\ell_h(t^*,x^*)=\frac{y(t^*+h)-(x^*+h\Phi(t^*,x^*,h))}{h}=\frac{y(t^*+h)-x_h(t^*+h)}{h}$$
the local discretization error often is called local error per unit step. Notice, that the local error is of order $p+1$, if the method is consistent of order p.
- *It is sufficient to postulate consistency only in a neighborhood of the exact solution $\hat{x}$.*
- *The maximum achievable order of implicit Runge-Kutta methods is $2s$, that of explicit Runge-Kutta methods is s, where s denotes the number of stages.*

Consistency of Euler's Method

Example 5.18 *(Explicit Euler's Method).*

$$x_h(t^*+h)=x^*+hf(t^*,x^*).$$

Taylor expansion of local solution:

$$y(t^*+h)=x^*+\dot{y}(t^*)h+\mathcal{O}(h^2)=x^*+f(t^*,x^*)h+\mathcal{O}(h^2)$$

Taylor expansion of local discretization error:

$$\ell_h(t^*,x^*)=\frac{y(t^*+h)-x^*}{h}-f(t^*,x^*)=\mathcal{O}(h)$$

Explicit Euler's method is *consistent of order* 1 (if f is smooth enough).

Consistency of Runge-Kutta Methods

Order conditions for Runge-Kutta Methods: (f sufficiently smooth)

$$
\begin{aligned}
p = 1 &: \sum_{i=1}^{s} b_i = 1, \\
p = 2 &: \sum_{i=1}^{s} b_i c_i = \frac{1}{2}, \\
p = 3 &: \sum_{i=1}^{s} b_i c_i^2 = \frac{1}{3}, \sum_{i,j=1}^{s} b_i a_{ij} c_j = \frac{1}{6}, \\
p = 4 &: \sum_{i=1}^{s} b_i c_i^3 = \frac{1}{4}, \sum_{i,j=1}^{s} b_i c_i a_{ij} c_j = \frac{1}{8}, \sum_{i,j=1}^{s} b_i a_{ij} c_j^2 = \frac{1}{12}, \sum_{i,j,k=1}^{s} b_i a_{ij} a_{jk} c_k = \frac{1}{24}
\end{aligned}
$$

In addition, the node conditions must hold:

$$c_i = \sum_{j=1}^{s} a_{ij}, \qquad i = 1, \ldots, s.$$

Node Conditions

The node conditions

$$c_i = \sum_{j=1}^{s} a_{ij}, \qquad i = 1, \ldots, s.$$

ensure that the Runge-Kutta method applied to the non-autonomous equation

$$x'(t) = f(t, x(t))$$

and the equivalent autonomous equation

$$\begin{pmatrix} x'(\tau) \\ t'(\tau) \end{pmatrix} = \begin{pmatrix} f(t(\tau), x(\tau)) \\ 1 \end{pmatrix}$$

yields the *same solution*.

Stability: Motivation

Consistency alone is not enough for convergence. In addition, we need stability!

Stability of a function $F : \mathbb{R} \to \mathbb{R}$ in $\hat{x}$:

There exist $R > 0$ and $S > 0$ such that

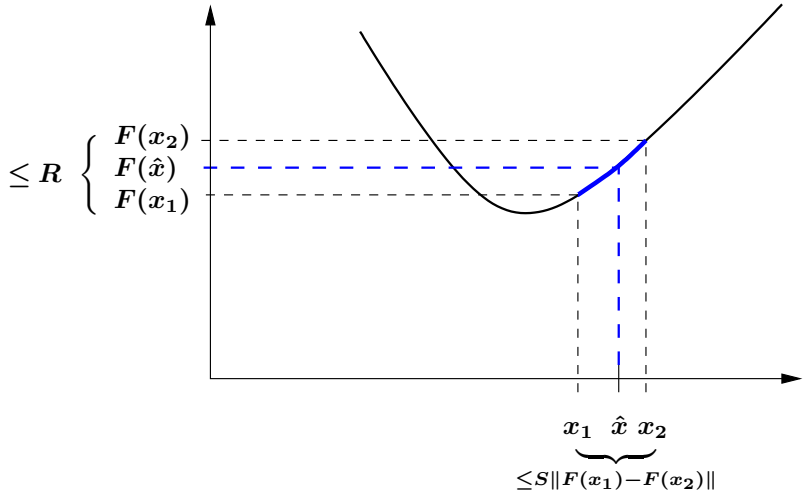
$$\|F(x_i) - F(\hat{x})\| < R,\ i = 1, 2 \quad \Rightarrow \quad \|x_1 - x_2\| \leq S \|F(x_1) - F(x_2)\|$$

i.e.

small deviation in function values $\Rightarrow$ *small deviation in arguments!*

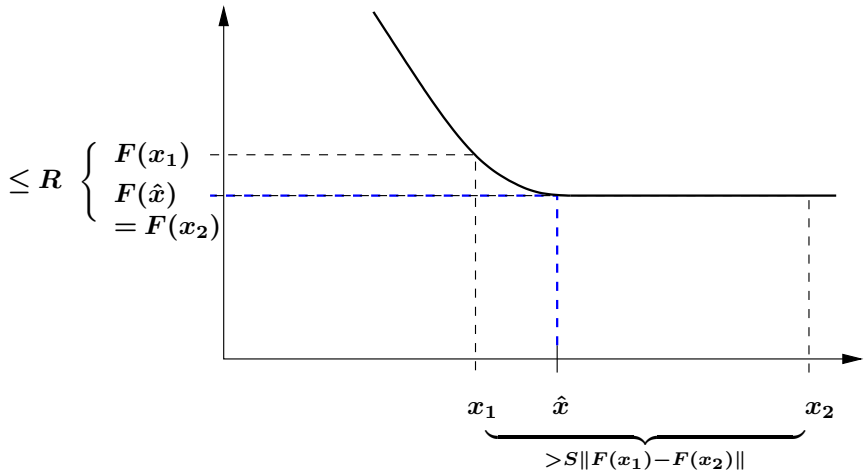
Stability: Motivation

The stable case:



Stability: Motivation

The unstable case:



Stability: Preliminaries

For an arbitrary grid function $y_h : \mathbb{G}_h \to \mathbb{R}^{n_x}$ let

$$\begin{aligned} \delta_h(t_0) &:= y_h(t_0) - x_0, \\ \delta_h(t_i) &:= \frac{y_h(t_i) - y_h(t_{i-1})}{h} - \Phi(t_{i-1}, y_h(t_{i-1}), h), \qquad i = 1, \ldots, N. \end{aligned}$$

The function $\delta_h : \mathbb{G}_h \to \mathbb{R}^{n_x}$ is called defect of y_h.

Stability

Definition 5.19 (Stability). Let $\{x_h\}_h$, $h = (t_f - t_0)/N$, $N \in \mathbb{N}$ be a sequence of solutions of the one-step method. Furthermore, let $\{y_h\}_h$ be a sequence of grid functions $y_h : \mathbb{G}_h \to \mathbb{R}^{n_x}$ with defect δ_h.

The one-step method is called stable, if there exist constants $S, R \geq 0$ independent of h (and N), such that for almost all $h = (t_f - t_0)/N$, $N \in \mathbb{N}$ it holds:

$$\|\delta_h\|_\infty < R \qquad \Rightarrow \qquad \|y_h - x_h\|_\infty \leq S\|\delta_h\|_\infty$$

R is called stability threshold and S stability bound.

Convergence Theorem

Consistency and stability ensure convergence.

Theorem 5.20 (Convergence). *Let the one-step method be consistent and stable. Then it is convergent.*
In addition, if the one-step method is consistent of order p, then it is also convergent of order p.

Convergence: Proof

Proof. Exact solution $\hat{x}$ yields a defect when introduced into difference scheme:

$$\begin{aligned} \delta_h(t_0) &= \hat{x}(t_0) - x_0 = 0, \\ \delta_h(t_i) &= \frac{\hat{x}(t_i) - \hat{x}(t_{i-1})}{h} - \Phi(t_{i-1}, \hat{x}(t_{i-1}), h), \qquad i = 1, \dots, N. \end{aligned}$$

It holds $\|\delta_h\|_\infty < R$ for sufficiently small h since the method is consistent, i.e.

$$0 = \lim_{h \to 0} \ell_h(t_i, \hat{x}(t_i)) = \lim_{h \to 0} \delta_h(t_{i+1}) \qquad \forall i = 0, \dots, N-1.$$

Convergence: Proof

Stability of the method yields

$$\|e_h\|_\infty = \|x_h - \Delta_h(\hat{x})\|_\infty \leq S \|\delta_h\|_\infty.$$

With $\delta_h(t_0) = 0$ and $\delta_h(t_i) = \ell_h(t_{i-1}, \hat{x}(t_{i-1}))$, $i = 1, \dots, N$ consistency yields

$$\lim_{h \to 0} \|\delta_h\|_\infty = 0$$

resp.

$$\|\delta_h\|_\infty = \mathcal{O}(h^p) \qquad \text{for } h \to 0$$

if the order of consistency is p. □

Stability Condition

Sufficient condition for stability:

Proposition 5.21. *Let the increment function Φ of the one-step method be lipschitz-continuous w.r.t. x for all sufficiently small step sizes h and all $t \in [t_0, t_f]$.*
Then, the one-step method is stable.

In particular: the increment function of Runge-Kutta methods is lipschitz-continuous, if f is lipschitz-continuous.

Stability Condition: Proof

Proof. Lipschitz continuity of $\Phi \Rightarrow \exists h_0 > 0,\ L > 0$:

$$\|\Phi(t, y, h) - \Phi(t, z, h)\| \leq L \|y - z\| \qquad \forall y, z,\ 0 < h \leq h_0,\ t \in [t_0, t_f].$$

Let $\mathbb{G}_h$ be a grid with $0 < h \leq h_0$. Let y_h be a grid function with defect δ_h and

$$\|\delta_h\|_\infty < R.$$

Then,

$$\|y_h(t_0) - x_h(t_0)\| = \|x_0 + \delta_h(t_0) - x_0\| = \|\delta_h(t_0)\|$$

and

Stability Condition: Proof

for $j = 1, \ldots, N$:

$$\begin{aligned}
\|y_h(t_j) - x_h(t_j)\| &= \| \quad y_h(t_{j-1}) + h\Phi(t_{j-1}, y_h(t_{j-1}), h) + h\delta_h(t_j) \\
&\quad - x_h(t_{j-1}) - h\Phi(t_{j-1}, x_h(t_{j-1}), h)\| \\
&\leq \|y_h(t_{j-1}) - x_h(t_{j-1})\| \\
&\quad + h\|\Phi(t_{j-1}, y_h(t_{j-1}), h) - \Phi(t_{j-1}, x_h(t_{j-1}), h)\| \\
&\quad + h\|\delta_h(t_j)\| \\
&\leq (1 + hL)\|y_h(t_{j-1}) - x_h(t_{j-1})\| + h\|\delta_h(t_j)\|.
\end{aligned}$$

Stability Condition: Proof

Recursive evaluation yields

$$\begin{aligned}
\|y_h(t_j) - x_h(t_j)\| &\leq (1 + hL)^j \|y_h(t_0) - x_h(t_0)\| + h \sum_{k=1}^{j} (1 + hL)^{j-k} \|\delta_h(t_k)\| \\
&\leq \exp((t_j - t_0)L)\|\delta_h(t_0)\| \\
&\quad + \max_{k=1,\ldots,j} \|\delta_h(t_k)\| \exp((t_j - t_0)L)(t_j - t_0) \\
&\leq C \exp((t_j - t_0)L) \max_{k=0,\ldots,j} \|\delta_h(t_k)\|,
\end{aligned}$$

where $C := \max\{t_f - t_0, 1\}$, $(1 + hL) \leq \exp(hL)$, $jh = t_j - t_0$.With $S := C\exp((t_f - t_0)L)$ we finally obtain

$$\|y_h - x_h\|_\infty \leq S\|\delta_h\|_\infty.$$

□

Remarks

Remark 5.22.

- *The assumption can be weakened: It suffices, that* Φ *is locally lipschitz-continuous at the exact solution* $\hat{x}$
- *During the definition of stability we implicitely assumed that the numerical solution* x_h *satisfies the equations of the one-step method exactly. In praxis, this is usually not the case since roundoff errors occur. However, the above definitions can be extended to this situation. Moreover, Stetter [Ste73] develops a general convergence theory, which is applicable to many other problem classes.*
- *There are different stability definitions being very important for the numerical solution of initial value problems. For instance, stiff differential equations require A-stable resp. A(*α*)-stable methods, which necessarily leads to implicit methods in connection with Runge-Kutta-Methods.*

5.4 Step-Size Control

Step-size Control

Efficiency: Need for algorithm for automatic step-size selection

Idea:

- numerical estimation of local (or global) error
- choose step-size such that error is within given error bounds

Ansatz: employ two Runge-Kutta methods with

- neighboring order of convergence, i.e. p and $p + 1$
- increment functions Φ and $\bar{\Phi}$
- approximations η and $\bar{\eta}$

Local Errors

Starting at t_{j-1} with $\eta(t_{j-1}) = \bar{\eta}(t_{j-1}) = x_{j-1}$ one step with stepsize h is performed for both Runge-Kutta methods.

local discretization error of first method:

$$\ell_h(t_{j-1}, x_{j-1}) = \frac{y(t_{j-1}+h) - x_{j-1}}{h} - \Phi(t_{j-1}, x_{j-1}, h) = C(t_{j-1})h^p + \mathcal{O}(h^{p+1})$$

local discretization error of second method:

$$\bar{\ell}_h(t_{j-1}, x_{j-1}) = \frac{y(t_{j-1}+h) - x_{j-1}}{h} - \bar{\Phi}(t_{j-1}, x_{j-1}, h) = \bar{C}(t_{j-1})h^{p+1} + \mathcal{O}(h^{p+2})$$

Estimation of Principal Error Term

local errors:

$$\begin{aligned} \eta(t_{j-1}+h) - y(t_{j-1}+h) &= -C(t_{j-1})h^{p+1} + \mathcal{O}(h^{p+2}) \\ \bar{\eta}(t_{j-1}+h) - y(t_{j-1}+h) &= -\bar{C}(t_{j-1})h^{p+2} + \mathcal{O}(h^{p+3}) \end{aligned}$$

Estimation of $C(t_{j-1})$:

$$-C(t_{j-1}) = \frac{1}{h^{p+1}}\left(\eta(t_{j-1}+h) - \bar{\eta}(t_{j-1}+h)\right) + \mathcal{O}(h)$$

New Step Size

New step-size: Estimation of local error using first method with new step-size h_{new}:

$$\begin{aligned} &\eta(t_{j-1}+h_{new}) - y(t_{j-1}+h_{new}) \\ &\quad = \left(\eta(t_{j-1}+h) - \bar{\eta}(t_{j-1}+h)\right)\left(\frac{h_{new}}{h}\right)^{p+1} + \mathcal{O}(h \cdot h_{new}^{p+1}) + \mathcal{O}(h_{new}^{p+2}) \end{aligned}$$

Postulation:

$$\|\eta(t_{j-1}+h_{new}) - y(t_{j-1}+h_{new})\| \leq \text{tol}$$

Result:

$$h_{new} \leq \left(\frac{\text{tol}}{\text{err}}\right)^{p+1} \cdot h, \qquad \text{err} := \|\eta(t_{j-1}+h) - \bar{\eta}(t_{j-1}+h)\|$$

Imbedded Runge-Kutta Methods

Efficiency: imbedded Runge-Kutta methods

c_1	a_{11}	a_{12}	$\cdots$	a_{1s}
c_2	a_{21}	a_{22}	$\cdots$	a_{2s}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
c_s	a_{s1}	a_{s2}	$\cdots$	a_{ss}
RK_1	b_1	b_2	$\cdots$	b_s
RK_2	$\tilde{b}_1$	$\tilde{b}_2$	$\cdots$	$\tilde{b}_s$

Imbedded Runge-Kutta Methods

Example 5.23. Runge-Kutta-Fehlberg of order 2(3)

$\mathbf{0}$			
$\mathbf{1}$	$\mathbf{1}$		
$\mathbf{1/2}$	$\mathbf{1/4}$	$\mathbf{1/4}$	
$\boldsymbol{RK_1}$	$\mathbf{1/2}$	$\mathbf{1/2}$	$\mathbf{0}$
$\boldsymbol{RK_2}$	$\mathbf{1/6}$	$\mathbf{1/6}$	$\mathbf{4/6}$

Check: $\boldsymbol{RK_1}$ has order $\mathbf{2}$, $\boldsymbol{RK_2}$ has order $\mathbf{3}$!

Algorithm for Step-size Control I

Algorithm for step-size control:

(0) Init: $t = t_0$, $x = x_0$. Choose initial step-size h.

(1) If $t + h > t_f$, set $h = t_f - t$.

(2) Compute approximations η and $\bar{\eta}$ at $t + h$ with $\boldsymbol{RK_1}$ resp. $\boldsymbol{RK_2}$ starting at x.

(3) Compute err and h_{new} according to

$$err = \max_{i=1,\ldots,n_x} \left(\frac{|\eta_i - \bar{\eta}_i|}{sk_i} \right)$$

scaling factors $sk_i = atol + \max(|\eta_i|, |x_i|) \cdot rtol$, absolute error tolerance $atol$, relative error tolerance $rtol$,

Algorithm for Step-size Control II

$$h_{new} = \min(\alpha_{max}, \max(\alpha_{min}, \alpha \cdot (1/err)^{1/(1+p)})) \cdot h$$

with $\alpha_{max} = 1.5$, $\alpha_{min} = 0.2$ and $\alpha = 0.8$.

(4) If $h_{new} < h_{min} := 10^{-8}$, *stop with error message.*

(5) If $err \leq 1$ (accept step):

(i) Set $x = \eta$, $t = t + h$.

(ii) If $|t - t_f| < 10^{-8}$, *stop with success.*

(iii) Set $h = h_{new}$ and go to (1).

if $err > 1$ (reject step): Set $h = h_{new}$ and go to (1).

Numerical Example

Differential equation:

$$\begin{aligned} \ddot{x}(t) &= x(t) + 2\dot{y}(t) - \bar{\mu}\tfrac{x(t)+\mu}{D_1} - \mu\tfrac{x(t)-\bar{\mu}}{D_2}, \\ \ddot{y}(t) &= y(t) - 2\dot{x}(t) - \bar{\mu}\tfrac{y(t)}{D_1} - \mu\tfrac{y(t)}{D_2}, \end{aligned}$$

where $\mu = 0.012277471$, $\bar{\mu} = 1 - \mu$, and

$$D_1 = \sqrt{((x(t)+\mu)^2 + y(t)^2)^3}, \qquad D_2 = \sqrt{((x(t)-\bar{\mu})^2 + y(t)^2)^3}.$$

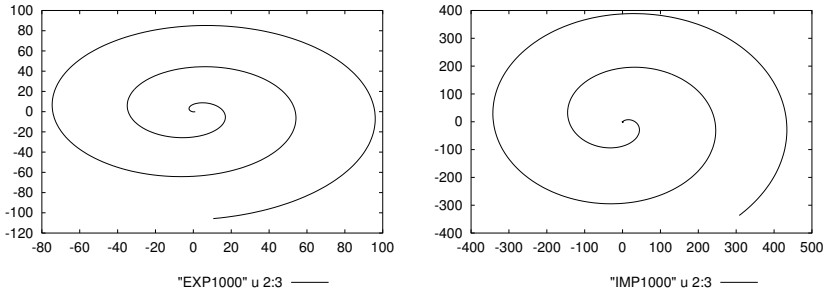
Initial values:

$$\begin{aligned} x(0) &= 0.994, & y(0) &= 0, \\ \dot{x}(0) &= 0, & \dot{y}(0) &= -2.001585106379. \end{aligned}$$

Numerical Example

Solution: explicit/implicit Euler, $(x(t), y(t))$, $[a, b] = [0, 17.065216560158]$, $h = (b-a)/N$.

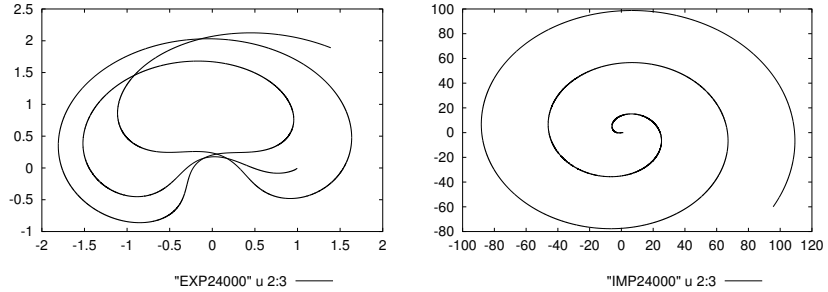
N=1000:



Numerical Example

Solution: explicit/implicit Euler, $(x(t), y(t))$, $[a, b] = [0, 17.065216560158]$, $h = (b-a)/N$.

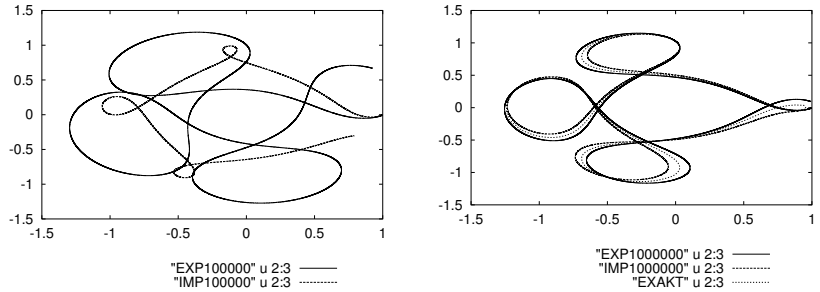
N=24000:



Numerical Example

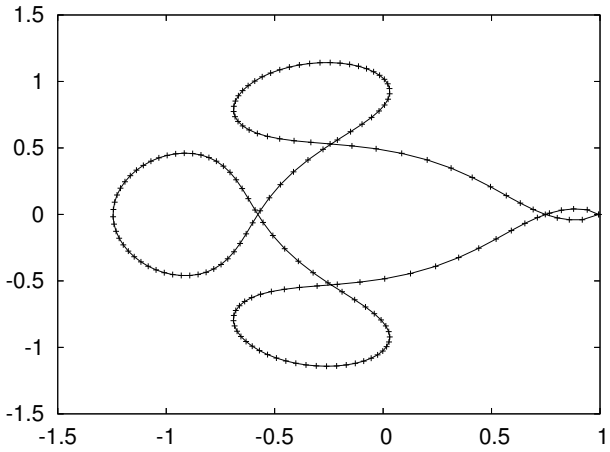
Solution: explicit/implicit Euler, $(x(t), y(t))$, $[a, b] = [0, 17.065216560158]$, $h = (b-a)/N$.

N=100000: *N=1000000:*



Numerical Example

Solution: Runge-Kutta method RKF2(3)



5.5 Sensitivity Analysis

Sensitivity Analysis

Often, initial value problems depend on parameters, e.g. air density, damping constants, reaction constants. Hence, we consider

Problem 5.24 (Parametric Initial Value Problem). *For given functions* $f : [t_0, t_f] \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_p} \to \mathbb{R}^{n_x}$ *and* $x_0 : \mathbb{R}^{n_p} \to \mathbb{R}^{n_x}$ *and given parameter* $p \in \mathbb{R}^{n_p}$ *solve the initial value problem*

$$\dot{x}(t) = f(t, x(t), p), \qquad x(t_0) = x_0(p). \tag{22}$$

Since the solution depends on p, it is denoted by $x(t; p)$.

Questions

Questions:

- Under which conditions is the solution of the initial value problem a continuous or even continuously differentiable function w.r.t. the parameter p?
- How can the sensitivities
$$S(t) := \frac{\partial x(t; p)}{\partial p}$$
be computed?

Gronwall Lemma

We need an auxiliary result.

Lemma 5.25 (Gronwall). *Let* $u, z : [t_0, t_f] \to \mathbb{R}$ *be continuous functions with*

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

for all $t \in [t_0, t_f]$ *and some constants* $c, L \geq 0$.

Then:

$$u(t) \leq \left(c + \max_{t_0 \leq \tau \leq t} |z(\tau)| \right) \exp(L(t - t_0)) \quad \forall t \in [t_0, t_f]$$

Gronwall Lemma: Proof

Proof. Define

$$v(t) := \int_{t_0}^{t} u(\tau)d\tau.$$

Then, $v(t_0) = 0$ and

$$v'(t) = u(t) \leq c + L\int_{t_0}^{t} u(\tau)d\tau + z(t) = c + Lv(t) + z(t).$$

Furthermore, it holds

$$\begin{aligned}\frac{d}{dt}\left(v(t)\exp(-L(t-t_0))\right) &= (v'(t) - Lv(t))\exp(-L(t-t_0)) \\ &\leq (c + z(t))\exp(-L(t-t_0)).\end{aligned}$$

Gronwall Lemma: Proof

Integration leads to

$$\begin{aligned}v(t)\exp(-L(t-t_0)) &\leq \int_{t_0}^{t}(c + z(\tau))\exp(-L(\tau - t_0))d\tau \\ &\leq \left(c + \max_{t_0\leq\tau\leq t}|z(\tau)|\right)\frac{1-\exp(-L(t-t_0))}{L}.\end{aligned}$$

We get

$$v(t) \leq \left(c + \max_{t_0\leq\tau\leq t}|z(\tau)|\right)\frac{\exp(L(t-t_0))-1}{L}.$$

Gronwall Lemma: Proof

With this estimate we finally obtain

$$\begin{aligned}u(t) &\leq c + Lv(t) + z(t) \\ &\leq c + \max_{t_0\leq\tau\leq t}|z(\tau)| + \left(c + \max_{t_0\leq\tau\leq t}|z(\tau)|\right)(\exp(L(t-t_0)) - 1) \\ &= \left(c + \max_{t_0\leq\tau\leq t}|z(\tau)|\right)\exp(L(t-t_0)).\end{aligned}$$

□

Dependence of Parameters

Theorem 5.26 (Continuous Dependence of Parameters). *Assumptions:*

- *Lipschitz condition: For all $t \in [t_0, t_f], x_1, x_2 \in \mathbb{R}^{n_x}, p_1, p_2 \in \mathbb{R}^{n_p}$:*

$$\|f(t, x_1, p_1) - f(t, x_2, p_2)\| \leq L\left(\|x_1 - x_2\| + \|p_1 - p_2\|\right) \tag{23}$$

- *Let $x_0 : \mathbb{R}^{n_p} \to \mathbb{R}^{n_x}$ be continuous.*

Dependence of Parameters

Theorem 5.26 (continued). *Assertions:*

- *$x(t;p)$ is a continuous function of p for every $t \in [t_0, t_f]$, i.e.*

$$\lim_{p\to\hat{p}} x(t;p) = x(t;\hat{p}) \qquad \forall t \in [t_0, t_f],\ \hat{p} \in \mathbb{R}^{n_p}.$$

- *If x_0 is lipschitz-continuous, then there exists a constant S with*

$$\|x(t;p_1) - x(t;p_2)\| \leq S\|p_1 - p_2\| \qquad \forall t \in [t_0, t_f],\ p_1, p_2 \in \mathbb{R}^{n_p}.$$

Dependence of Parameters: Proof

Proof. Solution x for parameter p:

$$x(t;p) = x_0(p) + \int_{t_0}^{t} f(t, x(t), p)dt.$$

The lipschitz condition guarantees the global existence on $[t_0, t_f]$ for every $p \in \mathbb{R}^{n_p}$. For given parameters p_1 and p_2 and the corresponding solutions $x(t;p_1)$ and $x(t;p_2)$ Gronwall's Lemma for every $t \in [t_0, t_f]$ yields

$$\begin{aligned} \|x(t;p_1) - x(t;p_2)\| &\leq \|x_0(p_1) - x_0(p_2)\| \\ &\quad + \int_{t_0}^{t} \|f(\tau, x(\tau;p_1), p_1) - f(\tau, x(\tau;p_2), p_2)\| d\tau \end{aligned}$$

Dependence of Parameters: Proof

$$\begin{aligned} \dots &\leq \|x_0(p_1) - x_0(p_2)\| \\ &\quad + L \int_{t_0}^{t} \|x(\tau;p_1) - x(\tau;p_2)\| d\tau \\ &\quad + L(t - t_0)\|p_1 - p_2\| \\ &\leq \Big(\|x_0(p_1) - x_0(p_2)\| \\ &\qquad + L(t - t_0)\|p_1 - p_2\| \Big) \exp(L(t - t_0)). \end{aligned}$$

□

Special Case

Special case: Dependence on initial values

With

$$x_0(p) = p, \qquad f = f(t, x)$$

it follows

$$\|x(t;p_1) - x(t;p_2)\| \leq \|p_1 - p_2\| \exp(L(t - t_0)).$$

Computation of Sensitivities I

Observation: IVP is identity in p!

Total differentiation w.r.t. p:

$$\begin{aligned} \frac{\partial x(t_0;p)}{\partial p} &= x_0'(p), \\ \frac{\partial}{\partial p}\left(\frac{d}{dt} x(t;p)\right) &= f_x'(t, x(t;p), p) \cdot \frac{\partial x(t;p)}{\partial p} + f_p'(t, x(t;p), p). \end{aligned}$$

Assumption:

$$\frac{\partial}{\partial p}\left(\frac{d}{dt} x(t;p)\right) = \frac{d}{dt}\left(\frac{\partial}{\partial p} x(t;p)\right)$$

Computation of Sensitivities II

Sensitivity Differential Equation

$$\begin{aligned} S(t_0) &= x_0'(p), \\ \dot{S}(t) &= f_x'(t, x(t;p), p) \cdot S(t) + f_p'(t, x(t;p), p). \end{aligned}$$

Special case: Dependence on initial values

- $x_0(p) = p \Rightarrow x_0'(p) = I$
- $f = f(t, x) \Rightarrow f_p' \equiv 0$

Notice: In this case S is a fundamental system!

For Further Reading

References

[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.

[Wal90] Walter, W. *Gewöhnliche Differentialgleichungen*. Springer, Berlin-Heidelberg-New York, 4th edition, 1990.

[Dem91] Demailly, J.-P. *Gewöhnliche Differentialgleichungen*. Vieweg, Braunschweig, 1991.

[HNW93] Hairer, E., Norsett, S. P., and Wanner, G. *Solving Ordinary Differential Equations I: Nonstiff Problems*. volumes 8 of *Springer Series in Computational Mathematics*. Springer-Verlag Berlin Heidelberg New York, 1993.

[SW95] Strehmel, K. and Weiner, R. *Numerik gewöhnlicher Differentialgleichungen*. Teubner, Stuttgart, 1995.

[MM02] Mattheij, R. M. M. and Molenaar, J. *Ordinary Differential Equations in Theory and Practice*, volume 43 of *Classics In Applied Mathematics*. SIAM, Philadelphia, 2002.

6 Discrete Approximation of Reachable Sets

Basic Ideas for Set-Valued Quadrature Methods

- numerical approximation of Aumann's integral uses that this integral is a convex, compact set under weak assumptions
- integration of the support function of the svm gives the support function of Aumann's integral
- pointwise quadrature methods with non-negative weights applied to support functions define set-valued quadrature methods
- appropriate smoothness for set-valued quadrature methods means the smoothness of the support function w.r.t. time t uniformly in directions $l \in S_{n-1}$
- for appropriate smoothness the same order of convergence could be reached for set-valued quadrature methods

Basic Ideas for Smoothness Conditions

- the averaged modulus of smoothness is an appropriate tool to overcome the difficulty in proofs for convergence order that the support function is often not in C^2
- "smoothness" of $F(\cdot) = A(\cdot)U$ up to "order 2" is given, if $A(\cdot)$ is smooth
- examples of "smooth" svms could be given

6.1 Set-Valued Quadrature Methods

Basic Ideas
Let $\mathcal{I} = [t_0, t_f]$ and $F : \mathcal{I} \Rightarrow \mathbb{R}^n$ be measurable, integrably bounded with values in $\mathcal{K}(\mathbb{R}^n)$.
By Theorem 4.44, the (Aumann) integrals of $F(\cdot)$ and $\mathbf{co}\, F(\cdot)$ coincide, i.e.

$$\int_{\mathcal{I}} F(t)dt = \int_{\mathcal{I}} \mathbf{co}\, F(t)dt.$$

Therefore, we can assume from now on that $F(\cdot)$ has values in $\mathcal{C}(\mathbb{R}^n)$.
Please notice that convex sets could be far easier represented in the computer than only compact sets.
Let us remark that $\int_{\mathcal{I}} F(t)dt$ is under these assumptions itself convex, compact and nonempty by Theorem 4.44.

Basic Ideas
To define set-valued quadrature methods, we first look on the support function of the integral which coincides with the integral of the support function (cf. Proposition 4.45):

$$\int_{\mathcal{I}} F(t)dt = \bigcap_{\eta \in S_{n-1}} \{x \in \mathbb{R}^n \,|\, \langle \eta, x\rangle \leq \int_{\mathcal{I}} \delta^\star(\eta, F(t))dt\}$$

The representing function $\delta^\star(\eta, F(\cdot))$ is a single-valued function for which quadrature methods are well known.
If $Q(f)$ is a quadrature formula for a function $f(\cdot)$ we will study set-valued analogues

$$Q(F) := \bigcap_{\eta \in S_{n-1}} \{x \in \mathbb{R}^n \,|\, \langle \eta, x\rangle \leq Q(\delta^\star(\eta, F(\cdot)))\}.$$

Attention: $Q(\delta^\star(\eta, F(\cdot)))$ is only a support function (espec. convex), if all weights are non-negative (cf. Proposition 3.64). Otherwise, $Q(F)$ could be empty!

Notation 6.1. *Let $\mathcal{I} = [0,1]$ and $f : \mathcal{I} \to \mathbb{R}^n$. Then,*

$$Q(f;\mathcal{I}) := Q(f) := \sum_{\mu=1}^{m} b_\mu f(t_\mu)$$

defines a quadrature method with weights b_μ and m different nodes $t_\mu \in \mathcal{I}$, $\mu = 1, \ldots, m$.
The remainder term describing the error is denoted by

$$R(f;\mathcal{I}) := R(f) := \int_{\mathcal{I}} f(\tau)d\tau - Q(f).$$

Point-Wise Quadrature Methods
Apply the quadrature method for a common intervall $\mathcal{I} = [t_0, t_f]$ and $f : \mathcal{I} \to \mathbb{R}^n$ and use that for the function $\widetilde{f}(t) := f(t_0 + t \cdot (t_f - t_0))$, $t \in [0,1]$:

$$Q(\widetilde{f}; [0,1]) = \sum_{\mu=1}^{m} b_\mu \widetilde{f}(t_\mu) = \sum_{\mu=1}^{m} b_\mu f(t_0 + t_\mu(t_f - t_0)),$$
$$\int_{t_0}^{t_f} f(t)dt = (t_f - t_0) \cdot \int_0^1 f(t_0 + t \cdot (t_f - t_0))dt = (t_f - t_0) \cdot \int_0^1 \widetilde{f}(t)dt$$

This motivates to set

$$Q(f;\mathcal{I}) := (t_f - t_0)Q(\widetilde{f}; [0,1]) = (t_f - t_0) \sum_{\mu=1}^{m} b_\mu f(t_0 + t_\mu \cdot (t_f - t_0)).$$

The corresponding quadrature method on $[t_0, t_f]$ uses the weights $(t_f - t_0)b_\mu$ and the nodes $t_0 + t_\mu \cdot (t_f - t_0)$, $\mu = 1, \ldots, m$, with remainder term $R(f;\mathcal{I}) = (t_f - t_0)R(\widetilde{f}; [0,1])$.

Iterated Quadrature Methods

A common technique to decrease the error measured by the remainder term is the partition of the integral: Let $(t_j)_{j=0,\dots,N}$ be an equidistant partition of $\mathcal{I} = [t_0, t_f]$ with step-size $h := \frac{t_f - t_0}{N}$, i.e.

$$t_0 < t_1 < \ldots < t_{N-1} < t_N = t_f.$$

Then, motivated by

$$\int_{t_0}^{t_f} f(t)dt = \sum_{j=0}^{N-1} \int_{t_j}^{t_{j+1}} f(t)dt$$

we define

$$Q_N(f;\mathcal{I}) := Q_N(f) := \sum_{j=0}^{N-1} Q(f;[t_j, t_{j+1}]),$$
$$R_N(f;\mathcal{I}) := R_N(f) := \int_{t_0}^{t_f} f(t)dt - Q_N(f;\mathcal{I})$$

as the *iterated quadrature method* with corresponding remainder term.

Example 6.2. Easy quadrature methods are:

(i) (iterated) special Riemann sum (staircase sum):

$$Q(f) = f(0),$$
$$Q_N(f;\mathcal{I}) = h \sum_{j=0}^{N-1} f(t_j)$$

(ii) (iterated) trapezoidal rule:

$$Q(f) = \frac{1}{2}(f(0) + f(1)),$$
$$Q_N(f;\mathcal{I}) = h \sum_{j=0}^{N-1} \frac{1}{2}(f(t_j) + f(t_{j+1})) = \frac{h}{2}(f(t_0) + f(t_f)) + h \sum_{j=1}^{N-1} f(t_j)$$

Example 6.2 *(continued).*

(iii) (iterated) Simpson's rule:

$$Q(f) = \frac{1}{6}(f(0) + 4f(\frac{1}{2}) + f(1)),$$
$$\begin{aligned} Q_N(f;\mathcal{I}) &= h \sum_{j=0}^{N-1} \frac{1}{6}(f(t_j) + 4f(t_j + \frac{h}{2}) + f(t_{j+1})) \\ &= \frac{h}{6}(f(t_0) + f(t_f)) + \frac{h}{3} \sum_{j=1}^{N-1} f(t_j) + \frac{2h}{3} \sum_{j=0}^{N-1} f(t_j + \frac{h}{2}) \end{aligned}$$

(iv) (iterated) midpoint rule:

$$Q(f) = f(\frac{1}{2}),$$
$$Q_N(f;\mathcal{I}) = h \sum_{j=0}^{N-1} f(t_j + \frac{h}{2})$$

Proposition 6.3. *Let $\mathcal{I} = [t_0, t_f]$ and $f : \mathcal{I} \to \mathbb{R}^n$. Then, the iterated quadrature method*

$$Q_N(f;\mathcal{I}) = \sum_{j=0}^{N-1} Q(f;[t_j, t_{j+1}])$$

for a partition $(t_j)_{j=0,\dots,N}$ with step-size h yields

$$R_N(f;\mathcal{I}) = \sum_{j=0}^{N-1} R(f;[t_j, t_{j+1}]).$$

Proposition 6.4. *Let $\mathcal{I} = [t_0, t_f]$ and $f : \mathcal{I} \to \mathbb{R}^n$ with $f \in C^2$. Then, the (iterated) trapezoidal rule fulfills:*

$$R(f;\mathcal{I}) = -\frac{(t_f - t_0)^3}{12} \cdot f^{(2)}(\xi), \quad \|R(f;\mathcal{I})\| \le \frac{(t_f - t_0)^3}{12} \max_{t\in\mathcal{I}} \|f^{(2)}(t)\|,$$

$$R_N(f;\mathcal{I}) = -\frac{h^3}{12} \cdot \sum_{j=0}^{N-1} f^{(2)}(\xi_j), \quad \|R_N(f;\mathcal{I})\| \le \frac{t_f - t_0}{12} \cdot h^2 \cdot \max_{t\in\mathcal{I}} \|f^{(2)}(t)\|,$$

where $\xi \in (t_0, t_f)$, $\xi_j \in (t_j, t_{j+1})$, $j = 0, \dots, N-1$.

Notation 6.5. *For a quadrature formula $Q(\cdot)$ with remainder term $R(\cdot)$ we denote*

$$\begin{aligned} Q(\eta, F) &:= Q(\delta^\star(\eta, F(\cdot))), \\ R(\eta, F) &:= R(\delta^\star(\eta, F(\cdot))), \\ Q_N(\eta, F) &:= Q_N(\delta^\star(\eta, F(\cdot))), \\ R_N(\eta, F) &:= R_N(\delta^\star(\eta, F(\cdot))) \end{aligned}$$

for $\eta \in \mathbb{R}^n$. For a measurable and integrably bounded set-valued mapping $F : \mathcal{I} \rightrightarrows \mathbb{R}^n$ with images in $\mathcal{K}(\mathbb{R}^n)$, we set

$$Q(F) := \bigcap_{\eta\in S_{n-1}} \{x \in \mathbb{R}^n \,|\, \langle \eta, x\rangle \le Q(\eta, F)\}, \tag{24}$$

$$Q_N(F) := \bigcap_{\eta\in S_{n-1}} \{x \in \mathbb{R}^n \,|\, \langle \eta, x\rangle \le Q_N(\eta, F)\}$$

as the (iterated) set-valued quadrature method.

Proposition 6.6. *Let $F : \mathcal{I} \rightrightarrows \mathbb{R}^n$ be a measurable and integrably bounded set-valued mapping with images in $\mathcal{K}(\mathbb{R}^n)$, and let the quadrature formula $Q(\cdot;\mathcal{I})$ have non-negative weights b_μ, nodes $t_\mu \in \mathcal{I}$, $\mu = 1, \dots, m$, and remainder term $R(\cdot;\mathcal{I})$.*
Then, the following error estimate holds

$$\mathrm{d_H}\Big(\int_{\mathcal{I}} F(t)dt, Q(F)\Big) = \sup_{\eta\in S_{n-1}} |R(\eta, F)|,$$

where

$$Q(F) = \sum_{\mu=1}^{m} b_\mu \operatorname{co}(F(t_\mu)).$$

Proof. Let us study $Q(\eta, F)$ first:

$$Q(\eta, F) = \sum_{\mu=1}^{m} b_\mu \underbrace{\delta^\star(\eta, F(t_\mu))}_{=\delta^\star(\eta, \operatorname{co} F(t_\mu))} = \delta^\star(\eta, \sum_{\mu=1}^{m} b_\mu \operatorname{co} F(t_\mu)),$$

since Proposition 3.115 and Lemma 3.51 were applied. Hence, $Q(\eta, F)$ is the support function of $Q(F)$ by Proposition 3.64 so that with the help of Corollary 3.157

$$\begin{aligned} \mathrm{d_H}\Big(\int_{\mathcal{I}} F(t)dt, Q(F)\Big) &= \sup_{l\in S_{n-1}} |\delta^\star(l, \int_{\mathcal{I}} F(t)dt) - \delta^\star(l, Q(F))| \\ &= \sup_{l\in S_{n-1}} |\int_{\mathcal{I}} \delta^\star(l, F(t))dt - Q(l, F)| \\ &= \sup_{l\in S_{n-1}} |R(l, F)|. \end{aligned}$$

□

Example 6.7 *(further quadrature methods).*

(i) Newton-Cotes-formulae of closed type, i.e. the interpolation polynomial $p_d(\cdot)$ of degree $d \in \mathbb{N}_0$ on $\mathcal{I} = [t_0, t_f]$ at equidistant nodes with step-size $h = \frac{t_f - t_0}{d}$ (resp. at t_0 for $d = 0$), including the boundary points t_0, t_f is integrated, i.e.

$$Q(f) := \int_{t_0}^{t_f} p_d(t)dt$$

non-negative weights up to convergence order 10, cf. Table 1

(ii) Newton-Cotes-formulae of open type
as in (i) with the same polynomial degree, but the equidistant nodes with step-size $h = \frac{t_f - t_0}{d+2}$ avoid the boundary points of $\mathcal{I}$
non-negative weights up to convergence order 4, cf. Table 2

Example 6.7 *(further quadrature methods, continued).* (iii) for Newton-Cotes-formulae of closed type with $d = 8$ or $d \geq 10$ resp. for Newton-Cotes-formulae of open type with $d = 2$ or $d \geq 4$ negative weights appear cf. saved convergence order in [Bai95] under additional geometrical condition on the integrand

(iv) Gaussian integration (interpolation polynomial of degree d, for which the nodes are chosen "optimal", i.e. the order of convergence reaches $2d$)
The weights are always non-negative.
cf. [Sto93, Lem98]

(v) Romberg's integration (extrapolation of the iterated trapezoidal rule) with Romberg's step sequence $h_i = \frac{t_f - t_0}{2^i}$, $i \in \mathbb{N}$, generates quadrature formulae of degree $2p$, $p \in \mathbb{N}$, if sufficient smoothness is present. Helping tableaus generate lower and upper bounds for the (theoretical) value of the integral.
cf. [Sto93, Lem98]

Remark 6.8. *For set-valued generalizations cf. [DF90] and [Bai95, BL94b, BL94a] (incl. inner and outer approximations of the Aumann integral with Romberg's method).*

Example 6.9. The following tables 1 and 2 show, up to which polynomial degrees the Newton-Cotes formulae have non-negative weights. The order of convergence in the tables is meant for the iterated quadrature methods under sufficient smoothness.

polynomial degree d	convergence order	name of the method
0	1	iterated staircase sum (special Riemann sum)
1	2	iterated trapezoidal rule
2	4	iterated Simpson's rule
3	4	iterated $3/8$-rule (pulcherrima, Faßregel)
4	6	iterated Milne's rule
5	6	(no name)
6	8	iterated Weddle's rule
7	8	(no name)
9	10	(no name)

Table 1: closed Newton-Cotes formulae with non-negative weights

To calculate $Q(F)$ in the computer we need to replace the (infinite) intersection in (24) by a finite one:

Proposition 6.10. *Let* $C \in \mathcal{C}(\mathbb{R}^n)$ *and* $\mathcal{G}_M = \{\eta^i \mid i = 1, \ldots, M\} \subset S_{n-1}$ *with* $\mathrm{d}(S_{n-1}, \mathcal{G}_M) \leq \varepsilon$. *If* $\varepsilon \leq \frac{1}{2}$ *and*

$$C_M := \bigcap_{i=1,\ldots,M} \left\{ x \in \mathbb{R}^n \mid \langle \eta^i, x \rangle \leq \delta^\star(\eta^i, C) \right\}$$

polynomial degree d	convergence order	name of the method
0	2	iterated midpoint-rule (rectangle rule)
1	2	(no name)
3	4	(no name)

Table 2: open Newton-Cotes formulae with non-negative weights

then C_M *is bounded by* $2 \cdot \|C\|$ *in the norm and*

$$\mathbf{d_H}(C, C_M) \leq 3 \cdot \|C\| \cdot \varepsilon.$$

Proof. It is clear that $C \subset C_M$. Let us estimate $\|C_M\|$ by $\|C\|$. For this purpose, choose an arbitrary $\eta \in S_{n-1}$ and an appropriate $\eta^i \in \mathcal{G}_M$ with $\|\eta - \eta^i\| = \mathbf{dist}(\eta, \mathcal{G}_M) \leq \mathbf{d}(S_{n-1}, \mathcal{G}_M) \leq \varepsilon$.

In the same manner, there exists $\eta^j \in \mathcal{G}_M$ with $\|(-\eta) - \eta^j\| \leq \varepsilon$.

For $x \in C_M$:

$$\begin{aligned}
\langle \eta, x \rangle &= \langle \eta - \eta^i, x \rangle + \langle \eta^i, x \rangle \leq \|\eta - \eta^i\| \cdot \|x\| + \delta^\star(\eta^i, C) \\
&\leq \varepsilon \cdot \|x\| + \max_{\xi \in S_{n-1}} |\delta^\star(\xi, C)| = \varepsilon \cdot \|x\| + \|C\|, \\
\langle -\eta, x \rangle &= \langle (-\eta) - \eta^j, x \rangle + \langle \eta^j, x \rangle \leq \varepsilon \cdot \|x\| + \|C\|, \\
|\langle \eta, x \rangle| &\leq \varepsilon \cdot \|x\| + \|C\|
\end{aligned}$$

for all $\eta \in S_{n-1}$. Now, choose $\eta := \frac{1}{\|x\|} x$, if $x \neq 0_{\mathbb{R}^n}$ (otherwise, the following estimation is trivial):

$$\|x\| \leq \varepsilon \cdot \|x\| + \|C\| \quad \text{and} \quad \|x\| \leq \frac{1}{1-\varepsilon} \cdot \|C\|$$

for all $x \in C_M$. Hence, we arrive at the estimation $\|C_M\| \leq \frac{1}{1-\varepsilon} \cdot \|C\| \leq 2 \cdot \|C\|$.

Let us show that $\delta^\star(\eta^\mu, C_M) = \delta^\star(\eta^\mu, C)$ for $\mu = 1, \ldots, M$.

$$\begin{aligned}
\delta^\star(\eta^\mu, C) &\leq \delta^\star(\eta^\mu, C_M), \quad \text{since } C \subset C_M, \\
\langle \eta^\mu, x \rangle &\leq \delta^\star(\eta^\mu, C) \quad \text{for all } x \in C_M, \\
\delta^\star(\eta^\mu, C_M) &\leq \delta^\star(\eta^\mu, C)
\end{aligned}$$

Now, choose $\eta \in S_{n-1}$ such that $\mathbf{d}(C_M, C) = \delta^\star(\eta, C_M) - \delta^\star(\eta, C)$. Again, take an appropriate $\eta^i \in \mathcal{G}_M$ with $\|\eta - \eta^i\| = \mathbf{dist}(\eta, \mathcal{G}_M)$. Hence, we have for $\varepsilon \leq \frac{1}{2}$

$$\begin{aligned}
\mathbf{d}(C_M, C) &= \big(\delta^\star(\eta, C_M) - \delta^\star(\eta^i, C_M)\big) + \big(\delta^\star(\eta^i, C) - \delta^\star(\eta, C)\big) \\
&\quad + \big(\delta^\star(\eta^i, C_M) - \delta^\star(\eta^i, C)\big) \\
&\leq \|\eta - \eta^i\| \cdot \|C_M\| + \|\eta^i - \eta\| \cdot \|C\| + 0 \leq \varepsilon \cdot (\|C_M\| + \|C\|) \\
&\leq \varepsilon \cdot \left(\frac{1}{1-\varepsilon} \cdot \|C\| + \|C\| \right) \leq 3 \|C\| \cdot \varepsilon.
\end{aligned}$$

□

To calculate $Q(F)$ with supporting points in a finite number of directions, we need the following justification.

Proposition 6.11. *Let* $C \in \mathcal{C}(\mathbb{R}^n)$ *and* $\mathcal{G}_M = \{\eta^i \,|\, i = 1, \ldots, M\} \subset S_{n-1}$ *with* $\mathbf{d}(S_{n-1}, \mathcal{G}_M) \leq \varepsilon$. *If*

$$\widetilde{C}_M := \mathbf{co}\{y(\eta^\mu, C) \,|\, \mu = 1, \ldots, M\},$$

then $\widetilde{C}_M$ *is bounded and*

$$\mathbf{d_H}(C, \widetilde{C}_M) \leq \mathbf{diam}(C) \cdot \varepsilon \leq 2 \cdot \|C\| \cdot \varepsilon.$$

Proof. Clearly, $\widetilde{C}_M \subset C$ which shows the boundedness. Hence, choose $c \in C$ and an appropriate $\widetilde{c}_M \in \widetilde{C}_M$ with

$$\|c - \widetilde{c}_M\| = \mathbf{dist}(c, \widetilde{C}_M) = \mathbf{d}(C, \widetilde{C}_M).$$

The case that $\|c - \widetilde{c}_M\| = 0$ is trivial. Otherwise, the characterization of the best approximation yields for all $w \in \widetilde{C}_M$

$$\begin{aligned} \langle c - \widetilde{c}_M, w - \widetilde{c}_M \rangle &\leq 0, \\ \mathbf{dist}(c, \widetilde{C}_M)^2 &= \langle c - \widetilde{c}_M, c - \widetilde{c}_M \rangle \\ &= \langle c - \widetilde{c}_M, c - w \rangle + \langle c - \widetilde{c}_M, w - \widetilde{c}_M \rangle \\ &\leq \langle c - \widetilde{c}_M, c - w \rangle. \end{aligned} \tag{25}$$

Set $v := \frac{1}{\|c - \widetilde{c}_M\|}(c - \widetilde{c}_M)$, choose $\eta^i \in \mathcal{G}_M$ with $\|v - \eta^i\| = \mathbf{dist}(v, \mathcal{G}_M) \leq \varepsilon$ and set $w = y(\eta^i, C)$.
From (25) follows

$$\begin{aligned} \mathbf{dist}(c, \widetilde{C}_M) &\leq \frac{1}{\mathbf{dist}(c, \widetilde{C}_M)} \cdot \langle c - \widetilde{c}_M, c - y(\eta^i, C) \rangle \\ &= \langle v, c - y(\eta^i, C) \rangle \\ &\leq \underbrace{\langle \eta^i, y(\eta^i, C) \rangle - \langle \eta^i, c \rangle}_{\leq 0} - \langle v, y(\eta^i, C) - c \rangle \\ &= \langle \eta^i - v, y(\eta^i, C) - c \rangle \\ &\leq \|\eta^i - v\| \cdot \|y(\eta^i, C) - c\| \leq \mathbf{diam}(C) \cdot \varepsilon. \end{aligned}$$

□

To approximate the Hausdorff distance numerically by choosing only a finite number of directions, we need the following justification.

Proposition 6.12. *Let $C, D \in \mathcal{C}(\mathbb{R}^n)$ and $\mathcal{G}_M = \{\eta^i \,|\, i = 1, \ldots, M\} \subset S_{n-1}$ with $\mathbf{d}(S_{n-1}, \mathcal{G}_M) \leq \varepsilon$. If C_M, D_M and $\widetilde{C}_M, \widetilde{D}_M$ are defined according to Proposition 6.10 resp. 6.11, then*

$$|\,\mathbf{d_H}(C, D) - \mathbf{d_H}(\widetilde{C}_M, \widetilde{D}_M)| \leq 2 \cdot (\|C\| + \|D\|) \cdot \varepsilon.$$

If additionally $\varepsilon \leq \frac{1}{2}$, then

$$\begin{aligned} |\,\mathbf{d_H}(C, D) - \mathbf{d_H}(C_M, D_M)| &\leq 3 \cdot (\|C\| + \|D\|) \cdot \varepsilon. \\ |\,\mathbf{d_H}(C, D) - \max_{\mu=1,\ldots,M} |\delta^\star(\eta^\mu, C) - \delta^\star(\eta^\mu, D)|\,| &\leq 5 \cdot (\|C\| + \|D\|) \cdot \varepsilon. \end{aligned}$$

Proof.
uses above Propositions 6.10 and 6.11 and estimations of $\max_{\mu=1,\ldots,M} |\delta^\star(\eta^\mu, C) - \delta^\star(\eta^\mu, D)|$ by $\mathbf{d_H}(C_M, D_M)$
□

Algorithm 6.13 (for iterated trapezoidal rule).
(1) Choose $N \in \mathbb{N}$ and set $h := \frac{t_f - t_0}{N}$.
(2) Please notice that the iteration

$$\begin{aligned} Q_j &:= Q_{j-1} + \frac{h}{2}(F(t_{j-1}) + F(t_j)) \quad (j = 1, \ldots, N), \\ Q_0 &:= \{0_{\mathbb{R}^n}\} \end{aligned}$$

resp. slightly more efficient

$$Q_j := Q_{j-1} + hF(t_j) \qquad (j = 1, \ldots, N-1), \tag{26}$$

$$Q_0 := \frac{h}{2}(F(t_0) + F(t_f)) \tag{27}$$

produces (set $Q_N := Q_{N-1}$ in the second case)

$$Q_N = \frac{h}{2}\sum_{j=0}^{N-1}(F(t_j)+F(t_{j+1})) = Q_N(F).$$

Choose one of the following alternatives of the algorithms:

Algorithm 6.13 (for iterated trapezoidal rule based on convex hull).
(3a) realization with convex hull algorithm (cf. Remark 3.11)
Assume the knowledge of the vertices of the polytope $F(t_j)$, i.e.

$$F(t_j) = \mathrm{co}\{p_{j,\mu} \,|\, \mu = 1,\dots,m_j\}$$

with $m_j \in \mathbb{N}$, $p_{j,\mu} \in \mathbb{R}^n$ for $\mu = 1,\dots,m_j$ and $j = 0,\dots,N$.
Set

$$\begin{aligned} S_0 &:= \{\frac{h}{2}(p_{0,\mu}+p_{N,\nu}) \,|\, \mu = 1,\dots,m_0,\ \nu = 1,\dots,m_N\},\\ Q_0 &:= \mathrm{co}(S_0). \end{aligned}$$

Use a convex hull algorithm to drop elements $\frac{h}{2}(p_{0,\mu}+p_{N,\nu})$ which are in $\mathrm{int}\, Q_0$ so that for $j = 0$:

$$Q_j = \mathrm{co}\{q_{j,r} \,|\, r = 1,\dots,k_j\} \tag{28}$$

Algorithm 6.13 (for iterated trapezoidal rule based on convex hull).
Now, iterate for $j = 1,\dots,N-1$ and set

$$\begin{aligned} S_j &:= \{q_{j-1,r} + hp_{j,\mu} \,|\, \mu = 1,\dots,m_j,\ r = 1,\dots,k_{j-1}\},\\ Q_j &:= \mathrm{co}(S_j). \end{aligned}$$

Use again a convex hull algorithm to drop elements $q_{j-1,r} + hp_{j,\mu}$ which are in $\mathrm{int}\, Q_j$ to get the representation (28).
Then, Q_{N-1} coincides with $Q_N(F)$.

disadvantage: *number of vertices could explode*

Algorithm 6.13 (for iterated trapezoidal rule based on affine inequalities).
(3b) realization with affine inequalities
Assume the knowledge of the half-spaces of the polyhedral set $F(t_j)$, i.e.

$$F(t_j) = \{x \in \mathbb{R}^n \,|\, A_j x \le b^j\}$$

with $m_j \in \mathbb{N}$, $b^j \in \mathbb{R}^{m_j}$ and $A_j \in \mathbb{R}^{m_j \times n}$ for $j = 0,\dots,N$.
Set

$$S_0 := \{z \in \mathbb{R}^n \,|\, z = \frac{h}{2}(x+\widetilde{x}),\ A_0 x \le b^0,\ A_N \widetilde{x} \le b^N\}.$$

Insert $\widetilde{x} = \frac{2}{h}z - x$ in second inequality and set

$$\begin{aligned} \widetilde{A}_0 &:= \left(\begin{array}{c|c} A_0 & 0 \\ \hline -A_N & \frac{2}{h}A_N \end{array}\right), \quad \widetilde{b}^0 := \begin{pmatrix} b^0 \\ b^N \end{pmatrix}, \quad \widetilde{x}^0 := \begin{pmatrix} x \\ z \end{pmatrix},\\ \widetilde{m}_0 &:= m_0 + m_N, \quad \widetilde{n}_0 := 2n. \end{aligned}$$

Algorithm 6.13 (trapezoidal rule based on affine inequalities, continued). *With the projection $P_z : \mathbb{R}^{\widetilde{n}^0} \to \mathbb{R}^n$ with $P_z(\widetilde{x}^0) = z$, we set*

$$Q_0 := \{P_z(\widetilde{x}^0) \,|\, \widetilde{A}_0\widetilde{x}^0 \le \widetilde{b}^0,\ \widetilde{x}^0 \in \mathbb{R}^{\widetilde{n}^0}\} \tag{29}$$

Use an algorithm to simplify the representation in S_0 and drop redundant inequalities in Q_0.
Now, iterate for $j = 1, \dots, N-1$ and set

$$S_j := \{z \in \mathbb{R}^n \mid z = P_z(\widetilde{x}^{j-1}) + h\widetilde{x},\ \widetilde{A}_{j-1}\widetilde{x}^{j-1} \le \widetilde{b}^{j-1},\ A_j\widetilde{x} \le b^j\}.$$

Insert $\widetilde{x} = \frac{1}{h}z - P_z(\widetilde{x}^{j-1})$ in second inequality and set

$$\widetilde{A}_j := \left(\begin{array}{c|c} \widetilde{A}_{j-1} & 0 \\ \hline -\frac{1}{h}A_jP_z & \frac{1}{h}A_N \end{array}\right), \quad \widetilde{b}^j := \begin{pmatrix} b^{j-1} \\ b_j \end{pmatrix}, \quad \widetilde{x}^j := \begin{pmatrix} \widetilde{x}^{j-1} \\ z \end{pmatrix},$$
$$\widetilde{m}_j := \widetilde{m}_{j-1} + m_j, \quad \widetilde{n}_j := \widetilde{n}_{j-1} + n.$$

Algorithm 6.13 (trapezoidal rule based on affine inequalities, continued). *With the projection $P_z : \mathbb{R}^{\widetilde{n}^j} \to \mathbb{R}^n$ with $P_z(\widetilde{x}^j) = z$, we set*

$$Q_j := \{P_z(\widetilde{x}^j) \mid \widetilde{A}_j\widetilde{x}^j \le \widetilde{b}^j,\ \widetilde{x}^j \in \mathbb{R}^{\widetilde{n}_j}\} \tag{30}$$

Hereby,

$$\widetilde{A}_j \in \mathbb{R}^{\widetilde{m}_j \times \widetilde{n}_j}, \quad \widetilde{b}^j \in \mathbb{R}^{\widetilde{m}_j}.$$

Use again an algorithm to simplify the representation in Q_j and drop redundant inequalities to simplify the representation (30)*.*
Then, Q_{N-1} coincides with $Q_N(F)$.

disadvantage: *number of inequalities could explode*

Algorithm 6.13 (for iterated trapezoidal rule based on support functions).
(3c) realization with support functions
Choose $\mathcal{G}_M = \{\eta^\mu \mid \mu = 1, \dots, M\} \subset S_{n-1}$ with $\mathrm{d}(S_{n-1}, \mathcal{G}_M) \le \varepsilon$. Assume the knowledge of the support functions of $F(t_j)$, $j = 0, \dots, N$, in the directions η^μ, $\mu = 1, \dots, M$.
For $\mu = 1, \dots, M$ set

$$d_0^\mu := \frac{h}{2}(\delta^\star(\eta^\mu, F(t_0)) + \delta^\star(\eta^\mu, F(t_f))) \in \mathbb{R}$$

and calculate iteratively for $j = 1, \dots, N-1$:

$$d_j^\mu := d_{j-1}^\mu + h\delta^\star(\eta^\mu, F(t_j)) \in \mathbb{R}$$

Algorithm 6.14 (trapezoidal rule, based on support functions). *Set*

$$Q_{N-1} := \bigcap_{\mu=1,\dots,M} \{x \in \mathbb{R}^n \mid \langle \eta^\mu, x\rangle \le d_{N-1}^\mu\}$$

which approximates $Q_N(F)$ within $\mathcal{O}(\varepsilon)$, if $\varepsilon \le \frac{1}{2}$ (cf. Proposition 6.10).
If $F(t) = A(t)U$, then the support function of $F(t_j)$ could be calculated only by the knowledge of the support function of U (cf. Proposition 3.116).

disadvantage: *number of directions should be adapted to convergence order, i.e., if*

$$\mathrm{d_H}(\int_{t_0}^{t_f} F(t)dt, Q_N(F)) = \mathcal{O}(h^p),$$

then choose $M \in \mathbb{N}$ and hence $\varepsilon = h^p$ so that

$$\mathrm{d_H}(Q_N(F), Q_{N,M}(F)) = \mathcal{O}(\varepsilon) = \mathcal{O}(h^p),$$
$$\mathrm{d_H}(\int_{t_0}^{t_f} F(t)dt, Q_{N,M}(F)) = \mathcal{O}(h^p).$$

Hereby, $Q_{N,M}(F)$ is the set C_M for $C := Q_N(F)$ from Proposition 6.10 constructed by directions only from $\mathcal{G}_M$.

Algorithm 6.14 (for iterated trapezoidal rule based on supporting points).
(3d) realization with supporting points
Choose $\mathcal{G}_M = \{\eta^\mu \,|\, \mu = 1,\dots,M\} \subset S_{n-1}$ *with* $\mathrm{d}(S_{n-1}, \mathcal{G}_M) \leq \varepsilon$. *Assume the knowledge of some supporting point of* $F(t_j)$, $j = 0,\dots,N$, *in the directions* η^μ, $\mu = 1,\dots,M$.
For $\mu = 1,\dots,M$ *set*

$$y_0^\mu := \frac{h}{2}(y(\eta^\mu, F(t_0)) + y(\eta^\mu, F(t_f))) \in \mathbb{R}^n$$

and calculate iteratively for $j = 1,\dots,N-1$*:*

$$y_j^\mu := y_{j-1}^\mu + h y(\eta^\mu, F(t_j)) \in \mathbb{R}^n$$

Algorithm 6.15 (trapezoidal rule based on supporting points, continued). *Set*

$$Q_{N-1} := \mathrm{co}\{y_{N-1}^\mu \,|\, \mu = 1,\dots,M\}$$

which approximates $Q_N(F)$ *within* $\mathcal{O}(\varepsilon)$ *(cf. Proposition 6.11).*
If $F(t) = A(t)U$, *then a supporting point of* $F(t_j)$ *could be calculated only by the knowledge of a supporting point of* U *(cf. Proposition 3.128).*

disadvantage: *number of directions should be adapted to convergence order, i.e., if*

$$\mathrm{d_H}(\int_{t_0}^{t_f} F(t)dt, Q_N(F)) = \mathcal{O}(h^p),$$

then choose $M \in \mathbb{N}$ *and hence* $\varepsilon = h^p$ *so that*

$$\mathrm{d_H}(Q_N(F), \widetilde{Q}_{N,M}(F)) = \mathcal{O}(\varepsilon) = \mathcal{O}(h^p),$$
$$\mathrm{d_H}(\int_{t_0}^{t_f} F(t)dt, \widetilde{Q}_{N,M}(F)) = \mathcal{O}(h^p).$$

Hereby, $\widetilde{Q}_{N,M}(F)$ *is the set* $\widetilde{C}_M$ *for* $C := Q_N(F)$ *from Proposition 6.11 constructed by directions only from* $\mathcal{G}_M$.

Algorithm 6.15 (for iterated trapezoidal rule based on other ideas).
(3e) other ideas
Choose anything else to represent sets and to calculate the Minkowski sum and the scalar multiplication in equations (26)–(27)*.*

Very popular are ellipsoidal methods, where the Minkowski sum of two ellipsoids are numerically approximated by another (inner/outer) ellipsoid. If all ellipsoids in (26)–(27) *are described with a center and a positive (semi-)definite matrix, an ordinary differential equation could be derived yielding a final ellipsoidal approximation of* $Q_N(F)$.
cf. [Sch68] and the references in the books [Che94, KV97]

Citations:
for theoretical results on quadrature methods:
[Pol75, Pol83, Bal82, Vel89a, DF90, DV93, BL94b, Bai95]
for numerical implementations of quadrature methods:
[BL94b, KK94, BL94a, Bai95, Che94, KV97]

Example 6.16. Calculate $\int_{t_0}^{t_f} F(t)dt$ with $F(t) = t^p U \subset \mathbb{R}^2$ and

$$U = \begin{cases} [-1,1]^2, \\ \mathrm{co}\{\binom{-1}{-1}, \binom{1}{-1}, \binom{3}{1}, \binom{-2}{1}\} \end{cases}$$

with iterated trapezoidal rule and $N = 10000$ subintervals.

$p = 0$, $[t_0, t_f] = [0, 1] \Rightarrow \int_{t_0}^{t_f} F(t)dt = U$

Please notice that $\int_{-1}^{1} dt = 1$!

$p = 3$, $[t_0, t_f] = [-1, 1] \Rightarrow \int_{t_0}^{t_f} F(t)dt = \frac{1}{4}(U + (-U)) \neq \{0_{\mathbb{R}^n}\}$

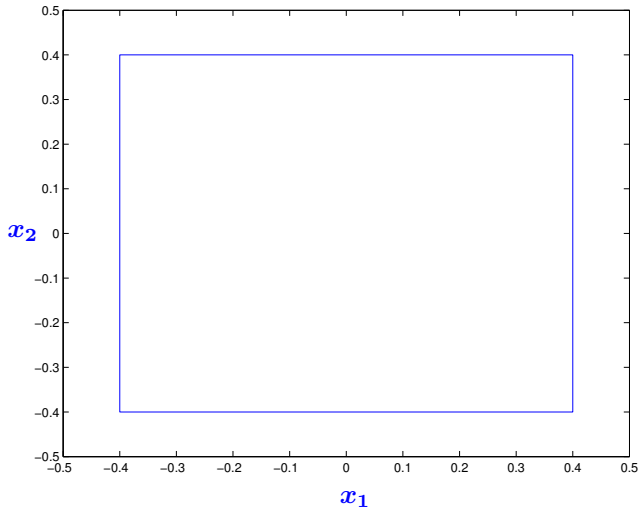
Please notice that $\int_{-1}^{1} t^3 dt = 0$!

$p = 4$, $[t_0, t_f] = [-1, 1] \Rightarrow \int_{t_0}^{t_f} F(t)dt = \frac{2}{5}U$

Please notice that $\int_{-1}^{1} t^4 dt = \frac{2}{5}$!

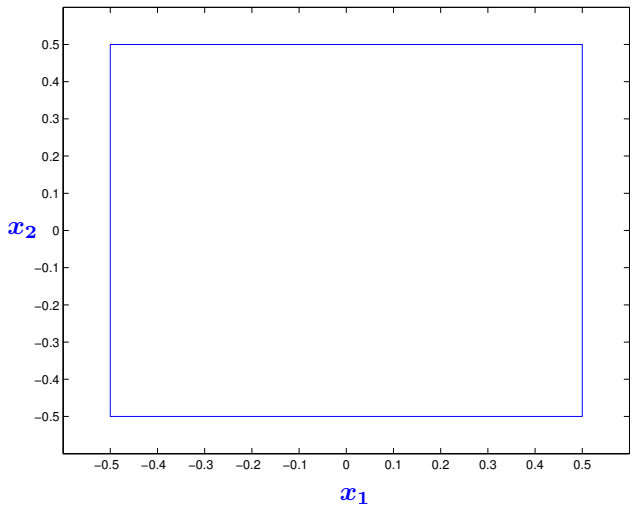
Aumann's Integral

$\int_{-1}^{1} t^4 U dt, U = [-1, 1]^2$



Aumann's integral is scaled set $\frac{2}{5} \cdot U$

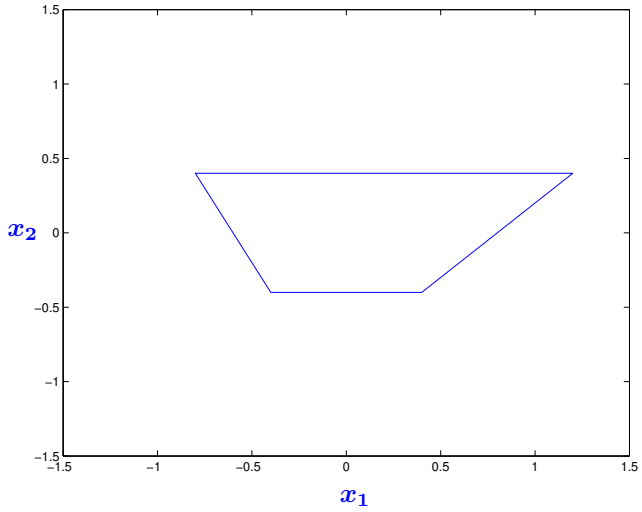
$\int_{-1}^{1} t^3 U dt, U = [-1, 1]^2$



Aumann's integral is not $\{0_{\mathbb{R}^n}\}$, but is the scaled set $\frac{1}{2} \cdot U$, since $U = -U$ is symmetric

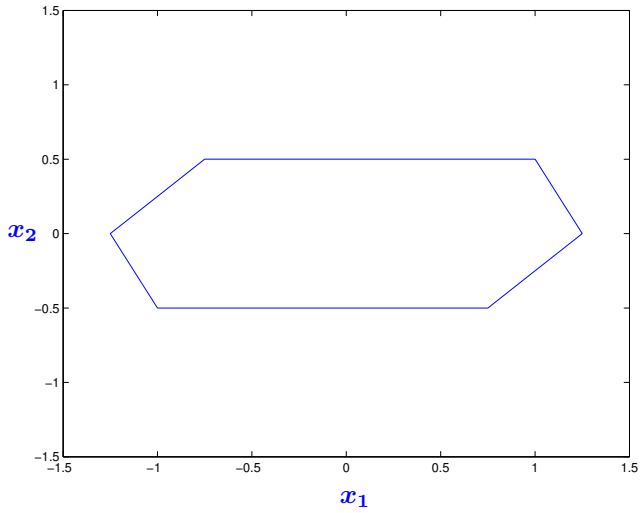
Aumann's Integral

$\int_{-1}^{1} t^4 U dt, U = \mathrm{co}\{\binom{-1}{-1}, \binom{1}{-1}, \binom{3}{1}, \binom{-2}{1}\}$



Aumann's integral
is scaled set $\frac{2}{5} \cdot U$

$\int_{-1}^{1} t^3 U dt, U = \mathrm{co}\{\binom{-1}{-1}, \binom{1}{-1}, \binom{3}{1}, \binom{-2}{1}\}$



Aumann's integral
is not $\{0_{\mathbb{R}^n}\}$,
but is the
set $\frac{1}{4}(U + (-U))$
and has 6 vertices
(U is not symmetric)

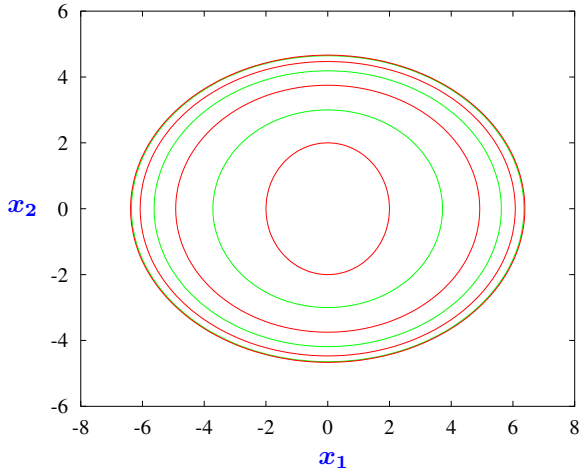
Example 6.17. Calculate $\int_0^2 F(t)dt$ with $F(t) = A(t)U$ and

$$A(t) = \begin{pmatrix} e^t & 0 \\ 0 & t^2+1 \end{pmatrix}, \quad U = B_1(0) \subset \mathbb{R}^2$$

The support function is smooth uniformly in $\eta \in S_1$, since $\delta^\star(\eta, F(t)) = \|A(t)^\top \eta\|_2$ and $A(t)$ is smooth and invertible.

Iterated Quadrature Methods

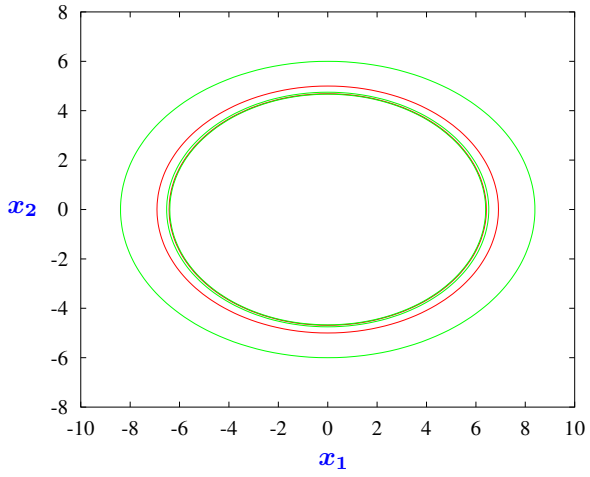
Iterated Staircase Sum



$h = 2$ (smallest set), 1, $0.5, 0.25, 0.1, 0.01$, 0.001 (biggest set)

convergence order 1: Hausdorff distance to biggest set approx. multiplied by $\frac{1}{2}$ each time the step-size halfens

Iterated Trapezoidal Rule



$h = 2$ (biggest set), 1, $0.5, 0.25$, 0.1 (smallest set)

convergence order 2: Hausdorff distance to biggest set approx. multiplied by $\frac{1}{4}$ each time the step-size halfens

Example 6.17 *(continued).* reference set: set-valued Romberg's method at tableau entry $(10, 10)$ with accuracy $\mathcal{O}(h_{10}^{22})$, $h_{10} = \frac{2}{2^{10}}$.

N	Hausdorff distance to the reference set	estimated order of convergence
1	**4.389056**	———
2	**2.670774**	**0.716653**
4	**1.464710**	**0.866643**
8	**0.765390**	**0.936348**
20	**0.314129**	**0.971941**
200	**0.031892**	**0.993426**
2000	**0.003194**	**0.999348**
20000	**0.000319**	**0.999935**

Table 3: order of convergence for the iterated staircase sum

Example 6.17 *(continued).* A method with high order of convergence (iter. Simpson's rule with $\mathcal{O}(h^4)$) is more efficient than one with lower order (here: iter. trapezoidal rule with $\mathcal{O}(h^2)$ and iter. staircase sum with $\mathcal{O}(h)$).

N	Hausdorff distance to the reference set	estimated order of convergence
1	**2.00000000000**	———
2	**0.52375377899**	**1.93303935**
4	**0.13255401055**	**1.98230843**
8	**0.03324172250**	**1.99551328**
20	**0.00532332626**	**1.99904613**
200	**0.00005324205**	**1.99992835**
2000	**0.00000053242**	**1.99999928**
20000	**0.00000000532**	**2.00000089**

Table 4: order of convergence for the iterated trapezoidal rule

N	Hausdorff distance to the reference set	estimated order of convergence
2	**0.0316717053249596**	———
4	**0.0021577697845663**	**3.87558**
8	**0.0001401261042595**	**3.94474**
20	**0.0000036242154211**	**3.98880**
200	**0.0000000003630625**	**3.99923**
2000	**0.0000000000000462**	**3.89548**
20000	**0.0000000000000222**	**0.31806**

Table 5: order of convergence for the iterated Simpson's rule

prescribed accuracy	iterated staircase sum		iterated trapezoidal rule	
	N	factor in CPU-time	N	factor in CPU-time
$\leq$ **1.0**	**7**	**2.0**	**2**	**1.5**
$\leq$ **0.1**	**64**	**5.0**	**5**	**2.0**
$\leq$ **0.01**	**639**	**37.0**	**15**	**4.0**
$\leq$ **0.001**	**6389**	**176.0**	**146**	**8.5**
$\leq$ **0.0001**	**63891**	**1752.25**	**1460**	**78.5**

Table 6: comparison of CPU-times of the 3 methods

prescribed accuracy	iterated trapezoidal rule		iterated Simpson's rule	
	N	factor in CPU-time	N	factor in CPU-time
$\leq$ **1.0**	**2**	**1.5**	**2**	**1.0**
$\leq$ **0.1**	**5**	**2.0**	**2**	**1.0**
$\leq$ **0.01**	**15**	**4.0**	**4**	**1.0**
$\leq$ **0.001**	**146**	**8.5**	**6**	**1.0**
$\leq$ **0.0001**	**1460**	**78.5**	**10**	**1.0**

Table 7: comparison of CPU-times of the 3 methods

6.2 Appropriate Smoothness of Set-Valued Mappings

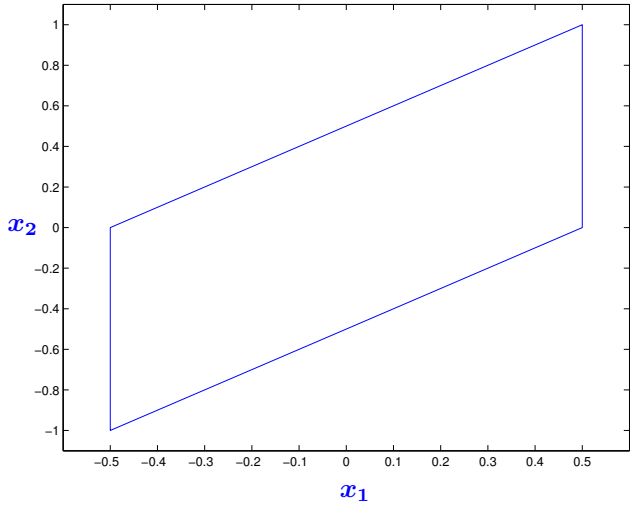
Example 6.18 *(very similar to [Vel92, Example (2.8)]).* Let $\mathcal{I} = [0,1]$ and $F : \mathcal{I} \Rightarrow \mathbb{R}^2$ with $F(t) = A(t)U$ with $A(t) = \binom{1-t}{1}$, $U = [-1,1]$.
Although $A(\cdot)$ is very smooth, the set-valued integration problem is not very smooth, since

$$\begin{aligned}\delta^\star(\eta, F(t)) &= \delta^\star(A(t)^\top \eta, [-1,1]) = \delta^\star(\eta_1(1-t)+\eta_2, [-1,1]) \\ &= |\eta_1(1-t)+\eta_2|\end{aligned}$$

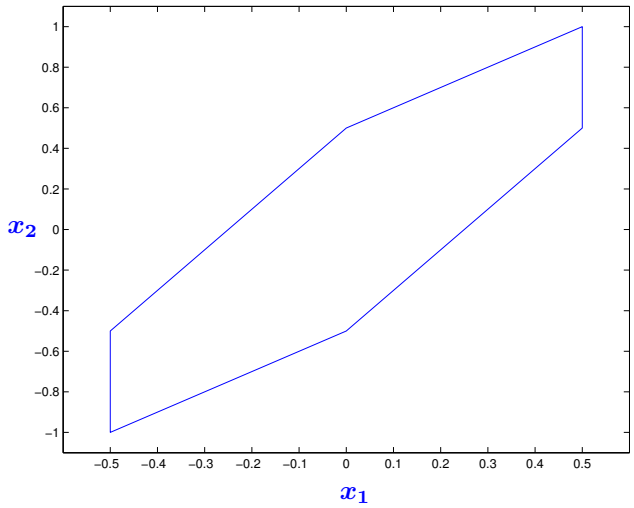
which is not even in C^2 for some $\eta \in S_{n-1}$.
But still this function could be estimated via an averaged modulus of smoothness of order 2 (cf. Example 6.28(iv)).

Iterated Trapezoidal Rule

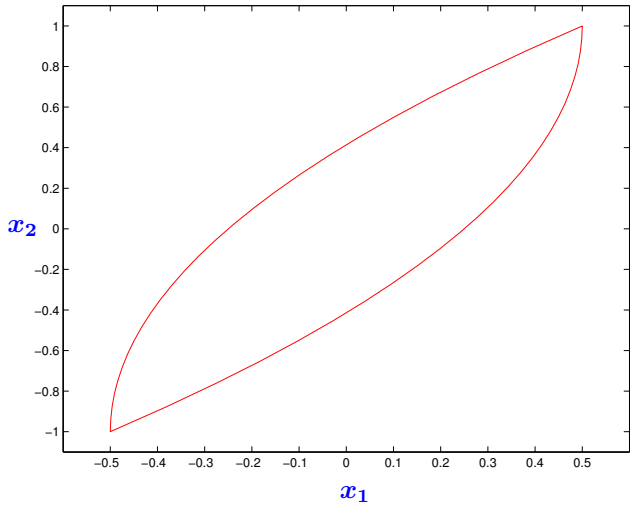
Iterated Trapezoidal Rule, $N = 1$, Scaled Set



Iterated Trapezoidal Rule, $N = 2$

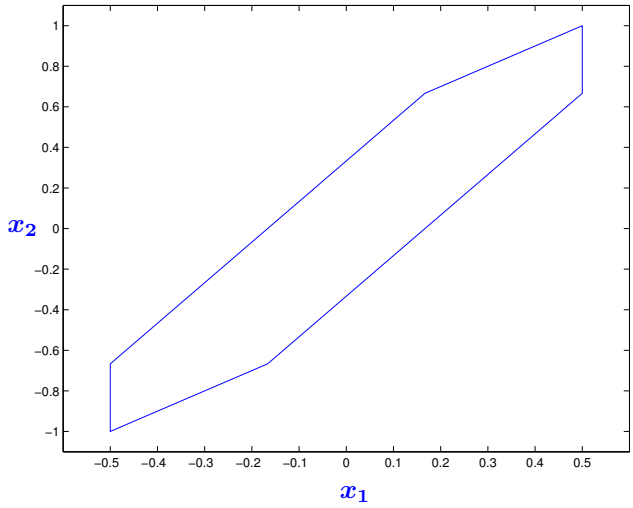


Reference Set: Iterated Trapezoidal Rule, $N = 100000$

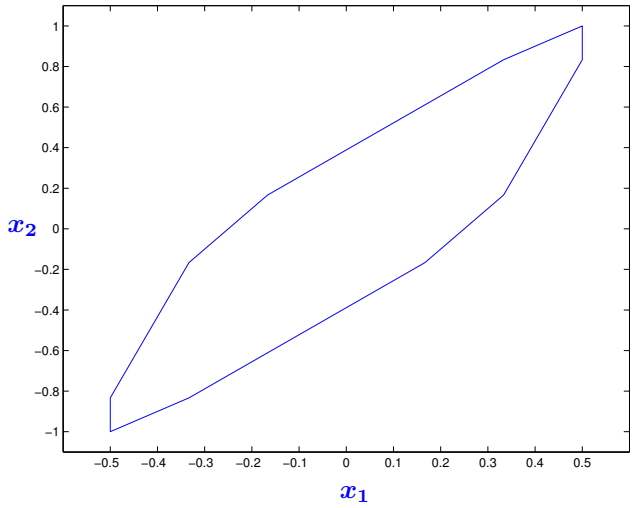


Iterated Simpson's Rule

Iterated Simpson's Rule, $N = 2$



Iterated Simpson's Rule, $N = 4$



Example 6.18 *(continued).* reference set: iterated trapezoidal rule with $N = 100000$

N	Hausdorff distance to the reference set	estimated order of convergence
1	**0.22542277**	———
2	**0.06049223**	**1.89781**
4	**0.01550004**	**1.96448**
8	**0.00389837**	**1.99133**
16	**0.00097605**	**1.99784**
32	**0.00022931**	**2.08964**
64	**0.00005804**	**1.98227**
128	**0.00001458**	**1.99313**

Table 8: order of convergence for the iterated trapezoidal rule

N	Hausdorff distance to the reference set	estimated order of convergence
2	**0.06867323**	———
4	**0.01804166**	**1.92842**
8	**0.00494920**	**1.86606**
16	**0.00126520**	**1.96783**
32	**0.00031805**	**1.99204**
64	**0.00006428**	**2.30686**
128	**0.00001629**	**1.98020**
256	**0.00000210**	**2.95502**

Table 9: order of convergence for the iterated Simpson's rule

Example 6.18 *(continued).* If we compare the results of the iterated trapezoidal rule and the Simpson's rule, we see that both methods show order of convergence 2. This could be expected for the trapezoidal rule, but the Simpson's rule suffers an order-breakdown from 4 to 2, since the support function is not smooth enough to allow order greater than 2.

Definition 6.19 (Sendov/Popov in [SP88]). Let $\mathcal{I} = [t_0, t_f]$ and $f : \mathcal{I} \to \mathbb{R}$ be bounded. Then, the *local*

modulus of smoothness of order $k \in \mathbb{N}$ is defined as

$$\omega_k(f;x,\delta) := \sup\left\{|\Delta_h^k f(t)| : t, t+kh \in [x-\frac{k\delta}{2}, x+\frac{k\delta}{2}] \cap \mathcal{I}\right\},$$

where $\delta \in [0, \frac{t_f - t_0}{k}]$ and $\Delta_h^k f(t)$ is the *k-th difference* of $f(\cdot)$ with step-size h at the point $t \in \mathcal{I}$, i.e.

$$\Delta_h^k f(t) := \sum_{i=0}^{k} (-1)^{k+i} \binom{k}{i} f(t+ih).$$

Definition 6.20 (Sendov/Popov in [SP88]). The *averaged modulus of smoothness of order* k, $k \in \mathbb{N}$, of a measurable and bounded function $f : \mathcal{I} \to \mathbb{R}$, $\mathcal{I} = [t_0, t_f]$, is defined as

$$\tau_k(f;\delta)_p := \|\omega_k(f;\cdot;\delta)\|_{L_p(\mathcal{I})}$$

where $\delta \in [0, \frac{t_f - t_0}{k}]$ and $p \in [1, \infty]$. We use the abbreviation $\tau_k(f;\delta)$ for $\tau_k(f;\delta)_1$.

Lemma 6.21. *Let $\mathcal{I} = [t_0, t_f]$, $k \in \mathbb{N}$, $\delta \in [0, \frac{t_f - t_0}{k}]$ and $f : \mathcal{I} \to \mathbb{R}$ be measurable and bounded. Then,*

(i) $\tau_k(f;\delta) \le 2^{k-1} \tau_1(f; k\delta)$,

(ii) $\tau_k(f;\delta) \le \dfrac{k^\nu}{\prod_{j=0}^{\nu-1}(k-j)} \delta^\nu \cdot \tau_{k-\nu}(f^{(\nu)}; \dfrac{k}{k-\nu}\delta)$ *(if $f^{(\nu)}$ exists, is bounded and measurable and $\nu \in \{0, \dots, k-1\}$),*

(iii) $\tau_k(f; n\delta)_p \le (2n)^{k+1} \tau_k(f;\delta)_p$ *(for all $n \in \mathbb{N}$),*

(iv) $\tau_1(f; n\delta)_p \le n \tau_1(f;\delta)_p$ *(for all $n \in \mathbb{N}$)*

Proof. see [SP88, (1)–(5),(5")] □

Lemma 6.22. *Let $\mathcal{I} = [t_0, t_f]$, $k \in \mathbb{N}$, $\delta \in [0, \frac{t_f - t_0}{k}]$ and $f : \mathcal{I} \to \mathbb{R}$ be measurable and bounded. Then:*

(i) If $f(\cdot)$ is Riemann-integrable, then for all $\varepsilon > 0$ there exists $\delta_0 > 0$ such that for all $0 < \delta \le \delta_0$

$$\tau_1(f;\delta)_p \le \varepsilon.$$

(ii) If $f(\cdot)$ has bounded variation, then

$$\tau_1(f;\delta)_p \le \delta \cdot \bigvee_{t_0}^{t_f} f(\cdot).$$

(iii) If $f(\cdot)$ is absolutely continuous and $f'(\cdot) \in L_p(\mathcal{I})$, then for all $\varepsilon > 0$ there exists $\delta_0 > 0$ such that for all $0 < \delta \le \delta_0$

$$\tau_2(f;\delta)_p \le 16\varepsilon\delta.$$

Lemma 6.22 (continued).

(iv) If $f(\cdot)$ is absolutely continuous, $f'(\cdot)$ has bounded L_1-variation, then

$$\tau_2(f;\delta)_p \le 2\delta^2 \bigvee_{L_1([t_0,t_f])} f'(\cdot).$$

Hereby, the L_1-variation is the infimum over all variations of L_1-representatives of $f'(\cdot)$.

(v) For a partition $(t_i)_{i=0,\dots,N-1}$ of $\mathcal{I}$ one has

$$\sum_{i=0}^{N-1} \tau_k(f|_{[t_i,t_{i+1}]};\delta) \le \tau_k(f;\delta).$$

Proof. (i) see [SP88, Theorem 1.2], for (ii) see [SP88, 1.3. (7)]
(iii) see [SP88, (1.28)] and [Bai95, 0.2.15 Satz (iii)]
(iv) see Lemma 6.21 and [Bai95, 0.2.15 Satz (iv)]
(v) see [Bai95, 0.2.14 Hilfssatz] □

Lemma 6.23. *Let* $\mathcal{I} = [t_0, t_f]$ *and* $f : \mathcal{I} \to \mathbb{R}$.

$$\begin{aligned}
\tau_1(f;\delta) &= o(1), && \text{if } f(\cdot) \text{ is continuous},\\
\tau_1(f;\delta) &= \mathcal{O}(\delta), && \text{if } f(\cdot) \text{ Lipschitz or } f \in C^1,\\
\tau_2(f;\delta) &= o(\delta), && \text{if } f(\cdot) \in C^1,\\
\tau_2(f;\delta) &= \mathcal{O}(\delta^2), && \text{if } f(\cdot) \in C^2
\end{aligned}$$

Proof (based on Lemma 6.22).
If $f(\cdot)$ is continuous, then $f(\cdot)$ is Riemann-integrable.
If $f(\cdot) \in \mathrm{C}^1$, then $f(\cdot)$ is Lipschitz (with constant $L = \max_{x\in\mathcal{I}} |f'(x)|$) and hence has bounded variation $\bigvee_{t_0}^{t_f} \mathrm{f}(\cdot) \le L \cdot (t_f - t_0)$. Then, $\tau_1(f;\delta) \le L \cdot \delta$.
If $f(\cdot) \in \mathrm{C}^1$, then $f(\cdot)$ is absolutely continuous and $\tau_2(f;\delta) = o(\delta)$.
If $f(\cdot) \in \mathrm{C}^2$, then $f(\cdot)$ is absolutely continuous and $f'(\cdot)$ has bounded variation so that $\tau_2(f;\delta) = \mathcal{O}(\delta^2)$. □

We will discuss on one example the advantage of the averaged modulus of smoothness and the difference to classical Taylor expansion for the error representation resp. estimation.

Example 6.24. Let $\mathcal{I} = [t_0, t_f]$ and $f : \mathcal{I} \to \mathbb{R}$ be measurable and bounded. Consider $Q_N(f)$ the (iterated) special Riemannian sum. Then,

$$|\int_{\mathcal{I}} f(t)dt - Q_N(f)| \le \tau_1(f; 2h).$$

If $f(\cdot) \in \mathrm{C}^1$, then

$$|\int_{\mathcal{I}} f(t)dt - Q_N(f)| \le \frac{t_f - t_0}{2} \cdot h \cdot \max_{x\in\mathcal{I}} |f'(x)|.$$

Proof. $Q(f) = (t_f - t_0) f(t_0)$ is the special Riemannian sum.
Let us prove first the estimation for $N = 1, h = t_f - t_0$.

$$\begin{aligned}
|R(f)| &\le \int_{t_0}^{t_f} |f(t) - f(t_0)|dt = \int_{t_0}^{t_f} |\Delta^1_{t-t_0} f(t_0)|dt\\
&\le \int_{t_0}^{t_f} \sup\left\{ |\Delta^1_h f(t_0)| : t_0, t_0 + h \in [t_0, t_f] \right\} dt\\
&= \int_{t_0}^{t_f} \omega_1(f; \frac{t_0 + t_f}{2}, t_f - t_0)dt \le \int_{t_0}^{t_f} \omega_1(f; t, 2(t_f - t_0))dt\\
&= \tau_1(f; 2(t_f - t_0)),
\end{aligned}$$

(cf. Definition 6.19)
For the iterated formula

$$R_N(f) = \int_{t_0}^{t_f} f(t)dt - h \sum_{j=0}^{N-1} f(t_j)$$

with $h = \frac{t_f - t_0}{N}$, $N \in \mathbb{N}$:

$$\begin{aligned}|R_N(f)| &= |\sum_{j=0}^{N-1} \left(\int_{t_j}^{t_{j+1}} f(t)dt - hf(t_j) \right)| \leq \sum_{j=0}^{N-1} \left| \int_{t_j}^{t_{j+1}} f(t)dt - hf(t_j) \right| \\ &= \sum_{j=0}^{N-1} R(f;[t_j,t_{j+1}]) \leq \sum_{j=0}^{N-1} \tau_1(f|_{[t_j,t_{j+1}]};2h) \leq \tau_1(f;2h)\end{aligned}$$

For $Q(f)$ one estimates via Taylor expansion:

$$\begin{aligned}|R(f)| &\leq \int_{t_0}^{t_f} |f(t) - f(t_0)|dt = \int_{t_0}^{t_f} |(t-t_0) \cdot f'(\xi_t)|dt \\ &= \int_{t_0}^{t_f} (t-t_0) \cdot |f'(\xi_t)|dt \leq \max_{x \in \mathcal{I}} |f'(x)| \cdot \int_{t_0}^{t_f} (t-t_0)dt \\ &= \frac{(t_f - t_0)^2}{2} \cdot \max_{x \in \mathcal{I}} |f'(x)|.\end{aligned}$$

For the iterated quadrature formula:

$$\begin{aligned}|R_N(f)| &\leq \sum_{j=0}^{N-1} \left| \int_{t_j}^{t_{j+1}} f(t)dt - hf(t_j) \right| = \sum_{j=0}^{N-1} R(f;[t_j,t_{j+1}]) \\ &\leq \sum_{j=0}^{N-1} \frac{h^2}{2} \cdot \max_{x \in [t_j,t_{j+1}]} |f'(x)| \leq \frac{t_f - t_0}{2} \cdot h \cdot \max_{x \in \mathcal{I}} |f'(x)|.\end{aligned}$$

□

Remark 6.25. *The estimation with the averaged modulus of smoothness is slightly worse for $f(\cdot) \in C^1$, since*

$$\tau_1(f;2h) \leq 2h \cdot \max_{x \in \mathcal{I}} |f'(x)|.$$

But the Taylor expansion demands before the proof that $f(\cdot)$ is continuously differentiable. Such proofs do not help, if this assumption is not fulfilled.
Estimation via the averaged modulus of smoothness doesn't make smoothness assumptions in advance. Hence, one could use them even if $f(\cdot)$ is Riemann-integrable or has bounded variation.

Proposition 6.26. *Let $\mathcal{I} = [t_0, t_f]$ and $F : \mathcal{I} \Rightarrow \mathbb{R}^n$ measurable and integrably bounded with images in $\mathcal{K}(\mathbb{R}^n)$. Then,*

$$\mathrm{d_H}(\int_{\mathcal{I}} F(t)dt, h \sum_{j=0}^{N-1} \mathrm{co}\, F(t_j)) \leq \sup_{\eta \in S_{n-1}} \tau_1(\delta^*(l, F(\cdot));2h).$$

Proposition 6.27. *Let $\mathcal{I} = [t_0, t_f]$ and $f : \mathcal{I} \to \mathbb{R}$ be measurable and bounded. Let $Q(f)$ be a quadrature method of order $k \in \mathbb{N}_0$, i.e. it is exact for polynomials up to degree k.*
Then, the iterated quadrature method fulfills:

$$\begin{aligned}&|\int_{t_0}^{t_f} f(t)dt - h \sum_{j=0}^{N-1} \sum_{\mu=1}^{m} b_\mu f(t_j + c_\mu h)| \\ \leq &(1 + \sum_{\mu=1}^{m} |b_\mu|) \cdot W_{k+1} \cdot \tau_{k+1}(f; \frac{2}{k+1}h),\end{aligned}$$

where $h = \frac{t_f - t_0}{N}$, $N \in \mathbb{N}$ and W_{k+1} is the Whitney constant (which is always less or equal 1, see [SP88, 2.1.1, p. 21–23]).

Proof. see [SP88, Theorem 2.4] □

Example 6.28. Let $\mathcal{I} = [t_0, t_f]$, $A : \mathcal{I} \to \mathbb{R}^{m \times n}$, $U \in \mathcal{C}(\mathbb{R}^n)$ and $F : \mathcal{I} \rightrightarrows \mathbb{R}^n$ with $F(t) = A(t)U$. Then,

(i) If $A(\cdot)$ is Riemann-integrable, then $F(\cdot)$ is Riemann-integrable.

(ii) If $A(\cdot)$ has bounded variation, then $F(\cdot)$ has bounded variation.

(iii) If $A(\cdot)$ is absolutely continuous, then $\delta^\star(\eta, F(\cdot))$ is absolutely continuous uniform in $l \in S_{n-1}$.

(iv) If $A(\cdot)$ is absolutely continuous with a L_1-representative of $A'(\cdot)$ with bounded variation, then $\delta^\star(\eta, F(\cdot))$ is absolutely continuous and suitable L_1-representatives of its derivative have bounded variation uniform in $l \in S_{n-1}$.

Proof. see [Bai95, 1.6.13 Satz] and references therein □

Now, some explicit support functions are listed which are k-times continuously differentiable uniformly in $\eta \in S_{n-1}$.

Proposition 6.29. *Let $\mathcal{I} = [t_0, t_f]$ and $F : \mathcal{I} \rightrightarrows \mathbb{R}^n$ with images in $\mathcal{C}(\mathbb{R}^n)$. Then, the support functions $\delta^\star(\eta, F(\cdot))$ are in C^k with k-th derivative uniform continuous in $\eta \in S_{n-1}$, in one of the following cases:*

(i) $F(t) = B_{r(t),p}(m(t))$ *Euclidean ball with radius $r(t) \geq 0$ and midpoint $m(t) \in \mathbb{R}^n$ for $t \in \mathcal{I}$ and $r(\cdot), m(\cdot) \in C^k$*

(ii) like (i), only for ball w.r.t. $\|\cdot\|_p$, $p \in [1, \infty]$, instead of $\|\cdot\|_2$

(iii) $F(t) = \prod_{i=1}^{n} [a_i(t), b_i(t)]$ *Cartesian product of intervals with $a_i(t) \leq b_i(t)$ in $t \in \mathcal{I}$ and $a_i(\cdot), b_i(\cdot) \in C^k$*

(iv) $F(t) = r(t)U$ *scalar multiple of fixed set with $r(t) \geq 0$ for $t \in \mathcal{I}$, $r(\cdot) \in C^k$*

(v) $F(t) = A(t)B_1(0) + b(t)$ *affine transformation of the Euclidean unit ball with $A(t) \in \mathbb{R}^{m \times n}$, $b(t) \in \mathbb{R}^m$ and $\operatorname{rang} A(t) = m$ for $t \in \mathcal{I}$ and $A(\cdot), b(\cdot) \in C^k$*

6.3 Reachable Sets/Differential Inclusions

Basic Ideas for the Approximation of Reachable Sets

- linear differential inclusions (LDIs) and linear control problems are "equivalent"
- reachable sets of LDIs are special Aumann integrals involving the fundamental solution of the homogeneous matrix differential equation (DE)
- an optimality criterion in linear optimal control problems is easily obtained via previous results
- set-valued quadrature methods could be used to approximate reachable sets, if the fundamental solution is known

Basic Ideas for the Approximation of Reachable Sets (continued)

- set-valued combination methods allow to approximate missing values of the fundamental solution by (pointwise) DE solver
- convergence order p is preserved, if the DE Solver and the set-valued quadrature method have this order and the problem is "smooth" enough
- proof uses global or local disturbances of set-valued quadrature methods
- set-valued Runge-Kutta methods could be defined by fixing a selection strategy
- the convergence order depends on a suitable selection strategy (and on the underlying set-valued quadrature method);
  break-downs in the convergence order are due to the missing second distributive law

Problem 6.31 (linear control problem). $\mathcal{I} = [t_0, t_f]$, $A : \mathcal{I} \to \mathbb{R}^{n\times n}$ *and* $B : \mathcal{I} \to \mathbb{R}^{n\times m}$ *with* $A(\cdot), B(\cdot) \in L_1$, $X_0 \subset \mathbb{R}^n$ *nonempty set of starting points,* $U \subset \mathbb{R}^m$ *nonempty set of controls.*
Given: control function $u : \mathcal{I} \to \mathbb{R}^m$ *with* $u(\cdot) \in L_1(\mathcal{I}, \mathbb{R}^m)$
Find: corresponding solution $x(\cdot)$ *(absolutely continuous) of the linear control problem (LCP) with*

$$x'(t) = A(t)x(t) + B(t)u(t) \quad \text{(a.e. } t \in \mathcal{I}\text{)}, \tag{31a}$$
$$x(t_0) = x_0 \in X_0, \tag{31b}$$
$$u(t) \in U \qquad \text{(a.e. } t \in \mathcal{I}\text{)}. \tag{31c}$$

Remark 6.32. *automatically:* $u(\cdot) \in L_\infty(\mathcal{I}, \mathbb{R}^m)$ *for* $U \in \mathcal{K}(\mathbb{R}^m)$
If $A(\cdot) \in L_\infty$, $B(\cdot) \in L_\infty$,
then $x(\cdot) \in W^{1,\infty}(\mathcal{I}, \mathbb{R}^n)$, *since* $x'(t) = A(t)x(t) + B(t)u(t)$.

In the following, we will deal with linear differential inclusion (cf. Problem 1.3 for the general, non-linear setting).

Problem 6.33 (linear differential inclusion). $\mathcal{I}$, $A(\cdot)$, $B(\cdot)$, X_0, U *as in Problem 6.31.*
$x(\cdot) : \mathcal{I} \to \mathbb{R}^n$ *is solution of the linear differential inclusion (LDI),*
if $x(\cdot)$ *is absolutely continuous and*

$$x'(t) \in A(t)x(t) + B(t)U, \tag{32}$$
$$x(t_0) \in X_0 \tag{33}$$

for almost all $t \in \mathcal{I}$.

Definition 6.34. linear control problem (LCP) in Problem 6.31, $t \in \mathcal{I}$. Then,

$$\mathcal{R}(t, t_0, x_0) := \{y \in \mathbb{R}^n \mid \exists u(\cdot) \text{ control function and } \exists x(\cdot) \text{ corresponding solution of Problem 6.31 with } x(t) = y\}$$

is called the *reachable set at time* t of the corresponding control problem.

Remark 6.35. *If we skip the control function $u(\cdot)$ in Definition 6.34 and replace Problem 6.31 by Problem 6.33, then we could define the reachable set for differential inclusions (cf. Remark 1.4 for the general, non-linear case).*

Remark 6.36. *Sometines, the term attainable set is used instead of reachable set. Especially, to distinguish points which could be attained in forward time (i.e., $t > t_0$) starting from x_0 from points which reach x_0 at time t (attainable set in backward time, i.e. time reversal).*
Since we do not consider time reversal from now on, we still use the term reachable set which is more commonly used in the literature.

Proposition 6.37 (equivalence of (LCP) and (LDI)). *$\mathcal{I}$, $A(\cdot)$, $B(\cdot)$, X_0, U as in Problem 6.31.*
If $x(\cdot)$ is a solution to a given control function $u(\cdot)$ of (LCP) in Problem 6.31, then $x(\cdot)$ is a solution of (LDI) in Problem 6.33.
If $x(\cdot)$ is a solution of (LDI) in Problem 6.33, then there exists a selection $u(\cdot)$ of U by [BF87a, Theorem 27] such that

$$x'(t) - A(t)x(t) = B(t)u(\cdot) \in B(t)U$$

and $(x(\cdot), u(\cdot))$ is a solution of (LCP) in Problem 6.31.
Clearly, the reachable set for both problems coincide (cf. Remark 6.35).

Proposition 6.38. *The reachable set of Problem 6.31 or 6.33 at time $t \in \mathcal{I}$ equals*

$$\mathcal{R}(t_f, t_0, X_0) = \Phi(t_f, t_0)X_0 + \int_{t_0}^{t_f} \Phi(t_f, t)B(t)U\,dt,$$

where $\Phi(t, \tau) \in \mathbb{R}^{n\times n}$ is the fundamental solution of the homogeneous matrix system, i.e.

$$\begin{aligned} \frac{d}{dt}\Phi(t,\tau) &= A(t)\Phi(t,\tau) && \text{(a.e. in } t \in \mathcal{I}\text{)},\\ \Phi(\tau,\tau) &= I_n && \text{where } I_n \text{ is the } n\times n\text{-identity matrix}. \end{aligned}$$

Therefore, $\mathcal{R}(t_f, t_0, X_0) \in \mathcal{C}(\mathbb{R}^n)$, if $X_0 \in \mathcal{C}(\mathbb{R}^n)$ and $U \in \mathcal{K}(\mathbb{R}^n)$.

Proof. cf. e.g. [Sv65, Theorem 1]. A direct proof would use Theorem 4.44 and Propositions 3.86, 3.88 and 6.37. □

Remark 6.39. *From Theorem 4.44 and Proposition 6.38 follows an existence result for solutions of linear differential inclusions:*
If

- *$U \neq \emptyset$, U is closed and $B(t)U$ is closed for a.e. $t \in \mathcal{I}$*
- *$B(\cdot)$ is measurable and $B(\cdot)U$ is integrably bounded*
  or $B(\cdot)$ is Lebesgue-integrable and U is bounded
- *$A(\cdot)$ is Lebesgue-integrable*
- *$X_0 \neq \emptyset$*

then the reachable set is nonempty (the integral part is in $\mathcal{C}(\mathbb{R}^n)$) and hence, solutions exist.
classical assumptions so that the reachable set is in $\mathcal{C}(\mathbb{R}^n)$:
$U \in \mathcal{K}(\mathbb{R}^n)$, $A(\cdot), B(\cdot)$ are Lebesgue-integrable, $X_0 \in \mathcal{C}(\mathbb{R}^n)$

Problem 6.40 (linear optimal control problem). *$\mathcal{I} = [t_0, t_f]$, $\eta \in \mathbb{R}^n$, $A : \mathcal{I} \to \mathbb{R}^{n\times n}$ and $B : \mathcal{I} \to \mathbb{R}^{n\times m}$ with $A(\cdot), B(\cdot) \in L_1(\mathcal{I})$, $X_0 \subset \mathbb{R}^n$ nonempty set of starting points, $U \subset \mathbb{R}^m$ nonempty set of controls.*
Find: optimal control function $\widehat{u} : \mathcal{I} \to \mathbb{R}^m$ with $\widehat{u}(\cdot) \in L_1(\mathcal{I}, \mathbb{R}^m)$ and corresponding optimal solution $\widehat{x}(\cdot)$ (absolutely continuous) of the linear optimal control problem (LOCP) with

$$\begin{aligned} &\min \ \langle \eta, x(t_f)\rangle && && \text{(34a)}\\ &\text{s.t.} && x'(t) = A(t)x(t) + B(t)u(t) \quad \text{(a.e. } t \in \mathcal{I}\text{)}, && \text{(34b)}\\ & && x(t_0) = x_0 \in X_0, && \text{(34c)}\\ & && u(t) \in U \quad \text{(a.e. } t \in \mathcal{I}\text{)}. && \text{(34d)} \end{aligned}$$

Proposition 6.41. *Consider the linear optimal control problem (LOCP) in Problem 6.40. Then, the optimal value of (LOCP) equals* $-\delta^\star(-\eta, \mathcal{R}(t_f, t_0, X_0))$.
If $\widehat{u}(\cdot)$ *satisfies*

$$\langle B(t)^\top \Phi(t_f, t)^\top(-\eta), \widehat{u}(t)\rangle = \delta^\star(B(t)^\top \Phi(t_f, t)^\top(-\eta), U)$$

for a.e. $t \in \mathcal{I}$ *and* $\widehat{x_0}$ *satisfies*

$$\langle \Phi(t_f, t_0)^\top(-\eta), \widehat{x_0}\rangle = \delta^\star(\Phi(t_f, t_0)^\top(-\eta), X_0),$$

then the corresponding solution

$$\widehat{x}(t) := \Phi(t, t_0)\widehat{x_0} + \int_{t_0}^{t} \Phi(t_f, t)B(t)\widehat{u}(t)dt$$

is a minimizer of (LOCP).

Sketch of proof. This essentially follows from

$$\begin{aligned}\langle -\eta, \widehat{x}(t_f)\rangle &= \delta^\star(-\eta, \Phi(t_f, t_0)X_0) + \int_{t_0}^{t_f} \delta^\star(-\eta, \Phi(t_f, t)B(t)\widehat{u}(t))dt \\ &= \delta^\star(-\eta, \mathcal{R}(t_f, t_0, X_0))\end{aligned}$$

and Proposition 4.45. □

Remark 6.42. *Under the assumptions of Proposition 6.41 we have*

$$\begin{aligned}\widehat{u}(t) &\in Y(B(t)^\top \Phi(t_f, t)^\top(-\eta), U) \subset \partial U \quad (\text{a.e. } t \in \mathcal{I}), \\ \widehat{x}(t_f) &\in Y(-\eta, \mathcal{R}(t_f, t_0, X_0)) \subset \partial \mathcal{R}(t_f, t_0, X_0),\end{aligned}$$

i.e. is a supporting point of the reachable set at time t_f *in direction* $-\eta$.

Set-valued quadrature methods could approximate the reachable set at time t_f, if the values of the fundamental solution are known.

Proposition 6.43. $\mathcal{I} = [t_0, t_f]$, $Q(\cdot; [0,1])$ *be a quadrature formula with nonnegative weights* b_μ, *nodes* $c_\mu \in [0,1]$, $\mu = 1, \dots, m$, *and remainder term* $R(\cdot; \mathcal{I})$. *Then, the set-valued quadrature method fulfills:*

$$\begin{aligned}&\mathrm{d_H}(\mathcal{R}(t_f, t_0, X_0), \Phi(t_f, t_0)X_0 + Q(\Phi(t_f, \cdot)B(\cdot)U; \mathcal{I})) \\ &= \sup_{\eta \in S_{n-1}} |R(l, \Phi(t_f, \cdot)B(\cdot)U)|\end{aligned}$$

Hereby, $t_\mu = t_0 + c_\mu(t_f - t_0)$, $\mu = 1, \dots, m$, *and*

$$Q(\Phi(t_f, \cdot)B(\cdot)U; \mathcal{I}) = \sum_{\mu=1}^{m} b_\mu \Phi(t_f, t_\mu)B(t_\mu)\,\mathrm{co}\,U.$$

Proof. cf. [DF90, BL94b] □

Remark 6.44. *Using the semi-group property of the fundamental solution, i.e.*

$$\Phi(t, \tau) = \Phi(t, s) \cdot \Phi(s, \tau) \qquad (t, \tau, s \in \mathcal{I}), \tag{35}$$

we could show the semi-group property of the reachable set

$$\mathcal{R}(t, \tau, X_0) = \mathcal{R}(t, s, \mathcal{R}(s, \tau, X_0)) \quad (t, \tau, s \in \mathcal{I} \text{ with } \tau \le s \le t) \tag{36}$$

with Proposition 6.38.

Proposition 6.45 (iterated quadrature method). *Assumptions as in Proposition 6.43. If the quadrature method has order $p \in \mathbb{N}$ and $\delta^\star(\eta, \Phi(t_f,\cdot)B(\cdot)U)$ has absolutely continuous $(p-1)$-st derivative and if the p-th derivative is of bounded variation uniformly w.r.t. $\eta \in S_{n-1}$, then the iterated quadrature method $\mathcal{R}_N(t_f, t_0, \Phi(t_f,\cdot)B(\cdot)U))$ with partition $(t_j)_{j=0,\dots,N}$ and step-size h fulfills*

$$\begin{aligned} &\mathrm{d_H}(\mathcal{R}(t_f,t_0,X_0), \mathcal{R}_N(t_f,t_0,\Phi(t_f,\cdot)B(\cdot)U)) \\ \leq &\sup_{\eta\in S_{n-1}} |R_N(l, \Phi(t_f,\cdot)B(\cdot)U)| = \mathcal{O}(h^{p+1}). \end{aligned}$$

If the quadrature method has order $p = 0$, then we assume that $\Phi(t_f,\cdot)B(\cdot)U$ has bounded variation to get

$$\begin{aligned} &\mathrm{d_H}(\mathcal{R}(t_f,t_0,X_0), \mathcal{R}_N(t_f,t_0,\Phi(t_f,\cdot)B(\cdot)U)) \\ \leq &\sup_{\eta\in S_{n-1}} |R_N(l, \Phi(t_f,\cdot)B(\cdot)U)| = \mathcal{O}(h). \end{aligned}$$

Proposition 6.45 (continued). *Hereby,*

$$Q_N(\Phi(t_f,\cdot)B(\cdot)U) = h \sum_{j=0}^{N-1} \sum_{\mu=1}^{m} b_\mu \Phi(t_f, t_j + c_\mu h) B(t_j + c_\mu h)\,\mathrm{co}\,U$$

and

$$\mathcal{R}_N(t_f,t_0,\Phi(t_f,\cdot)B(\cdot)U) := \Phi(t_f,t_0)X_0 + Q_N(\Phi(t_f,\cdot)B(\cdot)U)$$

denotes the set-valued iterated quadrature method for the approximation of reachable sets.

Proof. follows from Proposition 6.43, Remarks 6.44 and 6.47 with (35) □

Using the semi-group property (35) of the fundamental solution we could rewrite the iterated quadrature method iteratively.

Algorithm 6.46. *set-valued quadrature method in iterative form:*

$$\begin{aligned} Q^N_{j+1} &= \Phi(t_{j+1},t_j)Q^N_j + Q_N(\Phi(t_{j+1},\cdot)B(\cdot)U; [t_j, t_{j+1}]), \\ Q^N_0 &= X_0 \end{aligned}$$

Then,

$$Q^N_N = \Phi(t_f,t_0)X_0 + Q_N(\Phi(t_f,\cdot)B(\cdot)U)$$

This permits later in combination methods to allow (local) disturbances in the iteration above and motivate these methods.

The following remark shows the similarity of the calculation of the reachable set by iterated quadrature methods.

Remark 6.47.

$$\begin{aligned} &\textit{reachable set} \quad \begin{cases} \mathcal{R}^N_{j+1} &= \Phi(t_{j+1},t_j)\mathcal{R}^N_j + \int_{t_j}^{t_{j+1}} \Phi(t_{j+1},t)B(t)U\,dt, \\ \mathcal{R}^N_0 &= X_0 \\ \Rightarrow\ \mathcal{R}^N_N &= \mathcal{R}(t_f,t_0,X_0). \end{cases} \\ &\textit{quadrature method} \quad \begin{cases} Q^N_{j+1} &= \Phi(t_{j+1},t_j)Q^N_j + Q_N(\Phi(t_{j+1},\cdot)B(\cdot)U; [t_j,t_{j+1}]), \\ Q^N_0 &= X_0 \\ \Rightarrow\ Q^N_N &= \Phi(t_f,t_0)X_0 + Q_N(\Phi(t_f,\cdot)B(\cdot)U) \end{cases} \end{aligned}$$

Reason: semi-group property of fundamental solution and reachable sets in (35) and (36) □

Example 6.48. set-valued staircase sum/special Riemannian sum for $\mathcal{I} = [t_0, t_f]$:

$$Q(F;\mathcal{I}) = (t_f - t_0)F(t_0), \quad Q_N(F;\mathcal{I}) = h\sum_{j=0}^{N-1} F(t_j),$$

$$Q_N(\Phi(t_f,\cdot)B(\cdot)U;\mathcal{I}) = h\sum_{j=0}^{N-1} \Phi(t_f,t_j)B(t_j)U$$

in iterative form:

$$\begin{aligned} Q_{j+1}^N &= \Phi(t_{j+1},t_j)Q_j^N + h\Phi(t_{j+1},t_j)B(t_j)U, \\ Q_0^N &= X_0 \end{aligned}$$

Example 6.49. set-valued trapezoidal rule

$$Q(F;\mathcal{I}) = \frac{t_f - t_0}{2}\big(F(t_0) + F(t_f)\big),$$

$$Q_N(F;\mathcal{I}) = \frac{h}{2}\sum_{j=0}^{N-1} \big(F(t_j) + F(t_{j+1}\big),$$

$$Q_N(\Phi(t_f,\cdot)B(\cdot)U;\mathcal{I}) = \frac{h}{2}\sum_{j=0}^{N-1} \big(\Phi(t_f,t_j)B(t_j)U + \Phi(t_f,t_{j+1})B(t_{j+1})U\big)$$

in iterative form:

$$Q_{j+1}^N = \Phi(t_{j+1},t_j)Q_j^N + \frac{h}{2}\big(\Phi(t_{j+1},t_j)B(t_j)U + \underbrace{\Phi(t_{j+1},t_{j+1})}_{=\mathrm{I}_n}B(t_{j+1})U\big),$$

$$Q_0^N = X_0$$

Example 6.50. Consider the differential inclusion

$$\begin{aligned} x'(t) &\in A(t)x(t) + B(t)U \quad \text{for a.e. } t \in [0,1], \\ x(0) &\in X_0 \end{aligned}$$

with the data

$$A(t) = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad B(t) = \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$
$$U = [-1,1], \qquad X_0 = \{0_{\mathbb{R}^n}\}.$$

Example 6.50 *(continued).* Calculations show that the fundamental solution is

$$\Phi(t,\tau) = \begin{pmatrix} 1 & t-\tau \\ 0 & 1 \end{pmatrix}.$$

From Proposition 6.38 follows the equation

$$\mathcal{R}(1,0,X_0) = \Phi(1,0)\{0_{\mathbb{R}^n}\} + \int_0^1 \Phi(1,t)B(t)U\,dt = \int_0^1 \widetilde{A}(t)[-1,1]dt$$

with

$$\widetilde{A}(t) = \Phi(1,t)B(t) = \begin{pmatrix} 1-t \\ 1 \end{pmatrix}.$$

This coincides with the Aumann integral of Example 6.18.

Remark 6.51. *problems with quadrature methods:*

- *no generalization for nonlinear differential inclusions possible*
- *values of fundamental solutions $\Phi(t_{j+1}, t_j + c_\mu h)$ resp. $\Phi(t_f, t_j + c_\mu h)$ must be known in advance*

The second disadvantage could be solved by considering set-valued combination methods, the first disadvantage could be improved (at least to some extent) by the use of set-valued Runge-Kutta methods.

6.4 Set-Valued Combination Methods

Set-valued combination methods combine set-valued quadrature methods and differential equation solver to skip the demand for theoretical knowledge of the fundamental solution.

Proposition 6.52 (global disturbances). *$\mathcal{I} = [t_0, t_f]$, $Q(\cdot)$ be a quadrature formula with nonnegative weights b_μ, nodes $c_\mu \in [0,1]$, $\mu = 1, \dots, m$. Let $Q(\cdot;\mathcal{I})$ have order $p \in \mathbb{N}_0$ with remainder term $R(\cdot;\mathcal{I})$ and $\widetilde{\Phi}_\mu(t_f, t_\mu) \in \mathbb{R}^{n\times n}$ with $t_\mu = t_0 + c_\mu(t_f - t_0)$, $\mu = 1, \dots, m$, and $\widetilde{\Phi}(t_f, t_0) \in \mathbb{R}^{n\times n}$ be given. Assume for the error term and disturbances*

$$\begin{aligned} \sup_{\eta \in S_{n-1}} |R(l, \Phi(t_f, \cdot)B(\cdot)U)| &\le \widetilde{R}_1, \\ \|\widetilde{\Phi}(t_f, t_0) - \Phi(t_f, t_0)\| &\le \widetilde{R}_2, \\ \max_{\mu=1,\dots,m} \|\widetilde{\Phi}_\mu(t_f, t_\mu) - \Phi(t_f, t_\mu)\| &\le \widetilde{R}_3. \end{aligned}$$

Proposition 6.52 (continued). *Then, the set-valued combination method $\widetilde{R}(t_f, t_0, X_0)$ fulfills:*

$$\begin{aligned} &\mathbf{d_H}(\mathcal{R}(t_f, t_0, X_0), \widetilde{R}(t_f, t_0, X_0)) \\ \le &\widetilde{R}_1 + \|X_0\| \cdot \widetilde{R}_2 + \|U\| \cdot Q(\|B(\cdot)\|) \cdot \widetilde{R}_3 \end{aligned}$$

Hereby,

$$\widetilde{Q}(\widetilde{\Phi}(t_f, \cdot)B(\cdot)\,\mathbf{co}\,U) := \sum_{\mu=1}^{m} b_\mu \widetilde{\Phi}_\mu(t_f, t_\mu) B(t_\mu)\,\mathbf{co}\,U$$

is the disturbed set-valued quadrature method and

$$\widetilde{R}(t_f, t_0, X_0) := \widetilde{\Phi}(t_f, t_0)X_0 + \widetilde{Q}(\widetilde{\Phi}(t_f, \cdot)B(\cdot)\,\mathbf{co}\,U)$$

denotes the set-valued combination method.

Proof. cf. [BL94b, Bai95] □

Proposition 6.53 (iterated quadrature method, global disturbances). *Assumptions as in Proposition 6.52. If the quadrature method has order $p \in \mathbb{N}$ we assume that $\delta^\star(\eta, \Phi(t_f, \cdot)B(\cdot)\,\mathbf{co}\,U)$ has absolutely continuous $(p-1)$-st derivative and that the p-th derivative is of bounded variation uniformly w.r.t. $\eta \in S_{n-1}$.*
If the quadrature method has order $p = 0$, then we assume that $\Phi(t_f, \cdot)B(\cdot)\,\mathbf{co}\,U$ has bounded variation. Then, in both cases

$$\sup_{\eta \in S_{n-1}} |R(l, \Phi(t_f, \cdot)B(\cdot)\,\mathbf{co}\,U)| \le C_1 h^{p+1}.$$

Consider a partition $(t_j)_{j=0,\dots,N}$ with step-size h.

Proposition 6.53 (continued). *Furthermore, we assume the following estimations of the disturbances $\widetilde{\Phi}_{N,\mu}(t_f, t_j + c_\mu h) \in \mathbb{R}^{n\times n}$, $j = 0, \dots, N-1$, $\mu = 1, \dots, m$, and $\widetilde{\Phi}_N(t_f, t_0) \in \mathbb{R}^{n\times n}$:*

$$\begin{aligned} \|\widetilde{\Phi}_N(t_f, t_0) - \Phi(t_f, t_0)\| &\le C_2 h^{p+1}, \\ \max_{\substack{j=0,\dots,N-1 \\ \mu=1,\dots,m}} \|\widetilde{\Phi}_{N,\mu}(t_f, t_j + c_\mu h) - \Phi(t_f, t_j + c_\mu h)\| &\le C_3 h^{p+1}, \\ Q_N(\|B(\cdot)\|) &\le C_4 \end{aligned}$$

Proposition 6.53 (continued). *Then, the iterated combination method $\widetilde{\mathcal{R}}_N(t_f, t_0, X_0)$ for this partition fulfills*

$$\begin{aligned} &\mathbf{d_H}(\mathcal{R}(t_f, t_0, X_0), \widetilde{\mathcal{R}}_N(t_f, t_0, X_0)) \\ \le &(C_1 + C_2 \cdot \|X_0\| + C_3 \cdot C_4 \cdot \|U\|) h^{p+1} = \mathcal{O}(h^{p+1}). \end{aligned}$$

Hereby,

$$\widetilde{Q}_N(\widetilde{\Phi}_N(t_f, \cdot)B(\cdot)\,\mathbf{co}\,U) = h \sum_{j=0}^{N-1} \sum_{\mu=1}^{m} b_\mu \widetilde{\Phi}_{N,\mu}(t_f, t_j + c_\mu h) B(t_j + c_\mu h)\,\mathbf{co}\,U$$

and

$$\widetilde{\mathcal{R}}_N(t_f,t_0,X_0) := \widetilde{\Phi}_N(t_f,t_0)X_0 + \widetilde{Q}_N(\widetilde{\Phi}_N(t_f,\cdot)B(\cdot)\operatorname{co}U)$$

denotes the set-valued iterated combination method.

Proof. cf. [BL94b, Bai95] □

Proposition 6.54 (iterated quadrature method, local disturbances). *Assumptions as in Proposition 6.52. If the quadrature method has order $p \in \mathbb{N}$ we assume that $\delta^\star(\eta,\Phi(t_f,\cdot)B(\cdot)\operatorname{co}U)$ has absolutely continuous $(p-1)$-st derivative and that the p-th derivative is of bounded variation uniformly w.r.t. $\eta \in S_{n-1}$.*
If the quadrature method has order $p=0$, then we assume that $\Phi(t_f,\cdot)B(\cdot)\operatorname{co}U$ has bounded variation. Then, in both cases

$$\sup_{\eta\in S_{n-1}} |R(l,\Phi(t_f,\cdot)B(\cdot)\operatorname{co}U)| \le C_1 h^{p+1}.$$

Consider a partition $(t_j)_{j=0,\dots,N}$ with step-size h.

Proposition 6.55 (quadrature method, local disturbances; continued). *Furthermore, we assume the following estimations of the local disturbances $\widetilde{U}_{N,\mu}(t_{j+1},t_j+c_\mu h) \in \mathcal{C}(\mathbb{R}^n)$, $j=0,\dots,N-1$, $\mu=1,\dots,m$, and $\widetilde{\Phi}_N(t_{j+1},t_j) \in \mathbb{R}^{n\times n}$:*

$$\max_{j=0,\dots,N-1} \|\widetilde{\Phi}_N(t_{j+1},t_j) - \Phi(t_{j+1},t_j)\| \le C_2 h^{p+1},$$

$$\max_{\substack{j=0,\dots,N-1\\ \mu=1,\dots,m}} \mathrm{d_H}(\widetilde{U}_{N,\mu}(t_j+c_\mu h), \Phi(t_{j+1},t_j+c_\mu h)B(t_j+c_\mu h)\operatorname{co}U) \le C_3 h^{p+1},$$

$$Q_N(\|B(\cdot)\|) \le C_4$$

Proposition 6.55 (continued). *Then, the iterated combination method X_N^N for this partition defined iteratively by*

$$X_{j+1}^N = \widetilde{\Phi}(t_{j+1},t_j)X_j^N + h\sum_{\mu=1}^m b_\mu \widetilde{U}_{N,\mu}(t_j+c_\mu h) \quad (j=0,\dots,N-1), \tag{37}$$

$$X_0^N \in \mathcal{C}(\mathbb{R}^n) \text{ with } \mathrm{d_H}(X_0,X_0^N) \le C_5 h^p = \mathcal{O}(h^p) \tag{38}$$

fulfills the global estimate

$$\mathrm{d_H}(\mathcal{R}(t_f,t_0,X_0),X_N^N) = \mathcal{O}(h^{p+1}). \tag{39}$$

Proposition 6.55 (continued). *Especially, if approximations $\widetilde{\Phi}_{N,\mu}(t_f,t_j+c_\mu h) \in \mathbb{R}^{n\times n}$ of the values of the fundamental solution with*

$$\widetilde{\Phi}_{N,\mu}(t_{j+1},t_j+c_\mu h) = \Phi(t_{j+1},t_j+c_\mu h) + \mathcal{O}(h^p)$$

for $j=0,\dots,N-1$, $\mu=1,\dots,m$, are given, then the estimation (39) *above also holds for the same indices with the following setting:*

$$\widetilde{U}_{N,\mu}(t_j+c_\mu h) = \widetilde{\Phi}_{N,\mu}(t_{j+1},t_j+c_\mu h)B(t_j+c_\mu h)\operatorname{co}U.$$

Proof. cf. [Bai05] □

Example 6.56. w.l.o.g. $U \in \mathcal{C}(\mathbb{R}^n)$
combination method:
iter. Riemannian sum/Euler for matrix differ. equation

$$\begin{aligned} X'(t) &= A(t)X(t) \quad (t\in[t_j,t_{j+1}]),\\ X(t_j) &= \mathrm{I}_n \end{aligned}$$

combination method:

$$\begin{aligned} X_{j+1}^N &= \widetilde{\Phi}_N(t_{j+1},t_j)X_j^N + h\widetilde{\Phi}_{N,1}(t_{j+1},t_j)B(t_j)U,\ (j=0,\dots,N-1)\\ \widetilde{\Phi}_N(t_{j+1},t_j) &= \widetilde{\Phi}_N(t_j,t_j) + hA(t_j)\widetilde{\Phi}_N(t_j,t_j),\\ \widetilde{\Phi}_{N,1}(t_{j+1},t_j) &= \widetilde{\Phi}_N(t_{j+1},t_j). \end{aligned}$$

Hence,

$$X_{j+1}^N = (\mathrm{I}_n + hA(t_j))X_j^N + h(\mathrm{I}_n + hA(t_j))B(t_j)U\ (j=0,\dots,N-1).$$

Example 6.56 *(continued).* Other possibility for calculation: Euler for adjoint equation

$$\begin{aligned} Y'(t) &= -Y(t)A(t) \quad (t\in[t_0,t_f]),\\ Y(t_f) &= \mathrm{I}_n \end{aligned}$$

with given end-value gives

$$\begin{aligned} X_N^N &= \widetilde{\Phi}_N(t_f,t_j)X_0^N + h\sum_{j=0}^{N-1}\widetilde{\Phi}_{N,1}(t_f,t_j)B(t_j)U,\\ \widetilde{\Phi}_N(t_f,t_j) &= N-j \text{ (backward) steps of Euler for adjoint equation,}\\ \widetilde{\Phi}_{N,1}(t_f,t_j) &= \widetilde{\Phi}_N(t_f,t_j). \end{aligned}$$

Remark 6.57. *usual combination of set-valued quadrature method and pointwise DE solver which provides approximations to the values of the fundamental solution at the quadrature nodes:*

set-valued iter. quadrature method	*solver for diff. equations*	*step-size of DE solver*	*overall order*
Riemannian sum	*Euler*	h	$\mathcal{O}(h)$
trapezoidal rule	*Euler-Cauchy/Heun*	h	$\mathcal{O}(h^2)$
midpoint rule	*modified Euler*	$\frac{h}{2}$	$\mathcal{O}(h^2)$
Simpson's rule	*classical RK(4)*	$\frac{h}{2}$	$\mathcal{O}(h^4)$
Romberg's method	*extrapolation of midpoint rule (with Euler as starting procedure)*	$h_i = \frac{t_f-t_0}{2^i}$	$\mathcal{O}(\prod_{\nu=0}^{j} h_{i-\nu}^2)$

(under suitable smoothness assumptions)

Example 6.58. Consider the differential inclusion

$$\begin{aligned} x'(t) &\in A(t)x(t) + B(t)U \quad \text{for a.e. } t\in[0,2],\\ x(0) &\in X_0 \end{aligned}$$

with the data

$$\begin{aligned} A(t) &= \begin{pmatrix} 0 & 1\\ -2 & -3\end{pmatrix}, \quad B(t) = \mathrm{I}_2,\\ U &= B_1(0)\subset\mathbb{R}^2, \qquad X_0 = \{0_{\mathbb{R}^n}\}. \end{aligned}$$

Example 6.58 *(continued).* Calculations show that the fundamental solution is

$$\Phi(t,\tau) = \begin{pmatrix} 2e^{-(t-\tau)} - e^{-2(t-\tau)} & e^{-(t-\tau)} - e^{-2(t-\tau)}\\ -2e^{-(t-\tau)} + 2e^{-2(t-\tau)} & -e^{-(t-\tau)} + 2e^{-2(t-\tau)}\end{pmatrix}.$$

From Proposition 6.38 follows the equation

$$\mathcal{R}(2,0,X_0) = \int_0^2 \phi(2,t)B_1(0)dt.$$

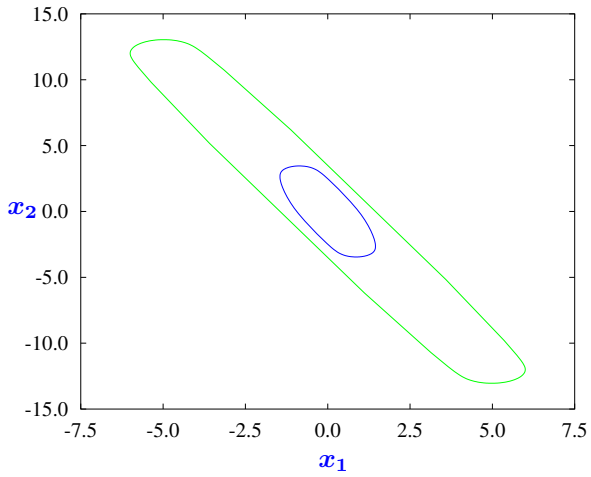
The support function

$$\delta^\star(\eta, \phi(2,\tau)B_1(0)) = \|\phi(2,\tau)^*\eta\|_2$$

is smooth uniformly in $\eta \in S_1$, since $A(\cdot)$ is constant, $\Phi(2,\cdot)$ is invertible and arbitrarily smooth.
reference set: set-valued Romberg's method at tableau entry $(10,10)$ with accuracy $\mathcal{O}(h_{10}^{22})$, $h_{10} = \frac{2}{2^{10}}$.

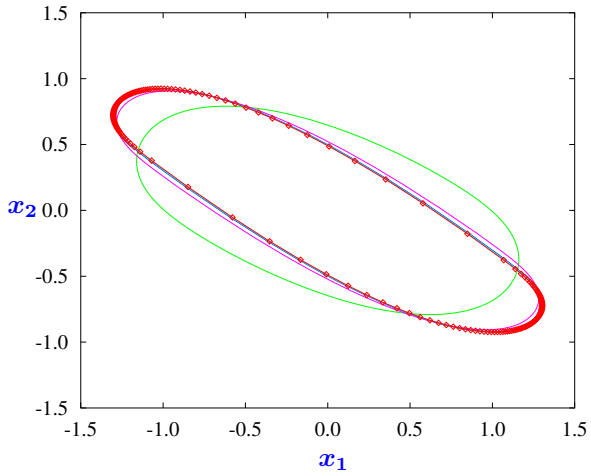
Set-Valued Combination Method

Iterated Trapezoidal Rule with Heun's method, $N = 1$ and $N = 2$



worse approximations
for first big step-sizes
– $N = 1$
– $N = 2$

Iterated Trapezoidal Rule with Heun's method, $N = 4, 8, 16, 20, 200$



good approximations
for smaller step-sizes
– $N = 4$
– $N = 8$
– $N = 16$
– $N = 20, 200$ and
reference set with
◇ supporting points

Example 6.58 *(continued).* reference set: set-valued Romberg's method at tableau entry $(10,10)$

Remark 6.59. *problems with these combination methods:*

- *no generalization for nonlinear differential inclusions possible*
- *values of fundamental solutions* $\Phi(t_{j+1}, t_j), \Phi(t_j + c_\mu h, t_j)$ *resp.* $\Phi(t_f, t_j), \Phi(t_f, t_j + c_\mu h)$ *must be approximated additionally*
- *approximations for* $\Phi(t_j + c_\mu h, t_j)$ *resp.* $\Phi(t_f, t_j + c_\mu h)$ *are calculated too accurately (*$\mathcal{O}(h^{p+1})$ *instead of* $\mathcal{O}(h^p)$*)*

N	Hausdorff distance to the reference set	estimated order of convergence
1	**12.33074198**	———
2	**5.03983763**	**1.29081046**
4	**0.49084218**	**3.36004612**
8	**0.24181874**	**1.02133316**
16	**0.12097455**	**0.99922246**
20	**0.09674847**	**1.00144359**
200	**0.00962392**	**1.00229196**
2000	**0.00096169**	**1.00031491**
20000	**0.00009616**	**1.00003200**
200000	**0.00000962**	**1.00000321**

Table 10: order of convergence for the iterated staircase sum/Euler's method

N	Hausdorff distance to the reference set	estimated order of convergence
1	**12.7776205317754**	———
2	**2.5354954375123**	**2.3332796**
4	**0.2862571241233**	**3.1468842**
8	**0.0413561282627**	**2.7911386**
16	**0.0087482798902**	**2.2410298**
20	**0.0054487041696**	**2.1218596**
200	**0.0000496413276**	**2.0404498**
2000	**0.0000004920011**	**2.0038773**
20000	**0.0000000049156**	**2.0003866**
200000	**0.0000000000492**	**2.0000426**

Table 11: order of convergence for the iterated trapezoidal rule/Heun's method

N	Hausdorff distance to the reference set	estimated order of convergence
2	**0.5738839013430377**	———
4	**0.0130316902051055**	**5.4607**
8	**0.0008327343160452**	**3.9680**
16	**0.0000457277029451**	**4.1867**
20	**0.0000180766413602**	**4.1591**
200	**0.0000000018139237**	**3.9985**
2000	**0.0000000000001839**	**3.9941**

Table 12: order of convergence for the iterated Simpson's rule/RK(4)

6.5 Set-Valued Runge-Kutta Methods

Runge-Kutta methods could be expressed by the Butcher array (cf. [But87]):

c_1	a_{11}	a_{12}	$\dots$	$a_{1,s-2}$	$a_{1,s-1}$	$a_{1,s}$
c_2	a_{21}	a_{22}	$\dots$	$a_{2,s-2}$	$a_{2,s-1}$	$a_{2,s}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
c_{s-1}	$a_{s-1,1}$	$a_{s-1,2}$	$\dots$	$a_{s-1,s-2}$	$a_{s-1,s-1}$	$a_{s-1,s}$
c_s	$a_{s,1}$	$a_{s,2}$	$\dots$	$a_{s,s-2}$	$a_{s,s-1}$	$a_{s,s}$
	b_1	b_2	$\dots$	b_{s-2}	b_{s-1}	b_s

Explicit Runge-Kutta methods satisfy $a_{\mu,\nu} = 0$, if $\mu \leq \nu$, and $c_1 = 0$.

The set-valued Runge-Kutta method for (LDI) is defined as follows:
Choose a starting set $X_0^N \in \mathcal{C}(\mathbb{R}^n)$ and define for $j = 0, \dots, N-1$ and $\mu = 1, \dots, s$:

$$\eta_{j+1}^N = \eta_j^N + h \sum_{\mu=1}^{s} b_\mu \xi_j^{(\mu)}, \tag{40}$$

$$\xi_j^{(\mu)} = A(t_j + c_\mu h)(\eta_j^N + h \sum_{\nu=1}^{\mu-1} a_{\mu,\nu} \xi_j^{(\nu)}) + B(t_j + c_\mu h) u_j^{(\mu)}, \tag{41}$$

$$u_j^{(\mu)} \in U, \tag{42}$$

$$\eta_0^N \in X_0^N, \tag{43}$$

$$X_{j+1}^N = \{\eta_{j+1}^N \mid \eta_{j+1}^N \text{ is defined by (40)–(43)}\}. \tag{44}$$

Remark 6.60. *If nonlinear DIs are considered with* $F(t,x) = \bigcup_{u \in U} \{f(t,x,u)\}$, *equation* (41) *must be replaced by*

$$\xi_j^{(\mu)} = f(t_j + c_\mu h, \eta_j^N + h \sum_{\nu=1}^{\mu-1} a_{\mu,\nu} \xi_j^{(\nu)}, u_j^{(\mu)}).$$

For some selection strategies, some of the selections $u_j^{(\mu)}$ *depend on others (e.g., they could be all equal).*

Remark 6.60 (continued). *If* $f(t,x,u) = f(t,u)$, *i.e.* $F(t,x) = F(t)$, *and* $X_0^N = \{0_{\mathbb{R}^n}\}$, *we arrive at the underlying quadrature method of the Runge-Kutta method.*

$$\eta_{j+1}^N = \eta_j^N + h \sum_{\mu=1}^{s} b_\mu f(t_j + c_\mu h, u_j^{(\mu)}), \quad u_j^{(\mu)} \in U,$$

$$X_{j+1}^N = X_j^N + h \sum_{\mu=1}^{s} b_\mu F(t_j + c_\mu h),$$

$$X_N^N = h \sum_{j=0}^{N-1} \sum_{\mu=1}^{s} b_\mu F(t_j + c_\mu h) = Q_N(F; [t_0, t_f])$$

Remark 6.61. *If* $f(t,x,u) = f(t,x)$, *i.e.* $F(t,x) = \{f(t,x)\}$, *then* $X_j^N = \{\eta_j^N\}$ *coincides with the pointwise Runge-Kutta method.*

Remark 6.62. *Grouping in equation* (40) *by matrices multiplied by* η_j^N *and* $u_j^{(\mu)}$, $\mu = 1, \dots, s$ *we arrive at the form*

$$X_{j+1}^N = \widetilde{\Phi}_N(t_{j+1}, t_j) X_j^N + h \bigcup_{\substack{u_j^{(\mu)} \in U \\ \mu = 1, \dots, s}} \{\sum_{\mu=1}^{s} b_\mu \widetilde{\Psi}_{N,\mu}(t_{j+1}, t_j + c_\mu h) u_j^{(\mu)}\}$$

with suitable matrices $\widetilde{\Phi}_N(t_{j+1}, t_j)$ *(involving matrix values of* $A(\cdot)$*) and* $\widetilde{\Psi}_{N,\mu}(t_{j+1}, t_j + c_\mu h)$ *(involving matrix values of* $A(\cdot)$ *and* $B(\cdot)$*).*
$\widetilde{\Phi}_N(t_{j+1}, t_j)$ *is the same matrix as in the pointwise case for* $f(t,x,u) = A(t)x$, *hence it approximates* $\Phi(t_{j+1}, t_j)$ *with the same order of convergence as in the pointwise case.*

Questions:

- What is the order of the set-valued Runge-Kutta method,
  i.e. $\mathbf{d_H}(\mathcal{R}(t_f, t_0, X_0), X_0^N) = \mathcal{O}(h^p)$?
  Does the order coincide with the single-valued case?
- What selection strategy is preferrable?
- Should the chosen selection strategy depend on the Runge-Kutta method?
- What smoothness assumptions do we need?

Answers in the literature:

set-valued RK-method	iter. quadrature method	global order	distur-bance term for ...	local order of disturbance	overall global order
Euler	Riemannian sum	$\mathcal{O}(h)$	η_j^N $u_j^{(1)}$	$\mathcal{O}(h^2)$ $\mathcal{O}(h)$	$\mathcal{O}(h)$
Heun (constant sel.)	midpoint rule	$\mathcal{O}(h^2)$	η_j^N $u_j^{(1)}$	$\mathcal{O}(h^3)$ $\mathcal{O}(h^2)$	$\mathcal{O}(h^2)$
Heun (2 free sel.)	trapezoidal rule	$\mathcal{O}(h^2)$	η_j^N $u_j^{(1)}$ $u_j^{(2)}$	$\mathcal{O}(h^3)$ $\mathcal{O}(h^2)$ $\mathcal{O}(h^2)$	$\mathcal{O}(h^2)$

Euler's method (see Subsection 6.5.1):
cf. [Nik88], [DF89], [Wol90] for nonlinear DIs,
for extensions see [Art94], [Gra03]
Euler-Cauchy method:
cf. [Vel92] as well as [Vel89b]
for strongly convex nonlinear DIs
modified Euler method (see Subsection 6.5.2)
cf. [Bai05]

6.5.1 Euler's Method

Remark 6.63. *Consider Euler's method, i.e. the Butcher array*

$$\begin{array}{c|c} \mathbf{0} & \mathbf{0} \\ \hline & \mathbf{1} \end{array} \quad .$$

underlying quadrature method = special Riemannian sum:

$$Q_N(F; [t_0, t_f]) = h \sum_{j=0}^{N-1} F(t_j)$$

Grouping by η_j^N and the single selection $u_j^{(1)}$ yields

$$X_{j+1}^N = (\underbrace{I_n + hA(t_j)}_{=\widetilde{\Phi}_N(t_{j+1}, t_j)}) X_j^N + h \underbrace{B(t_j)}_{=\widetilde{\Psi}_{N,1}(t_{j+1}, t_j)} U \quad (j = 0, \ldots, N-1).$$

Proposition 6.64. *Euler's method is a combination method with the following settings:*

$$\begin{aligned} Q_N(F; [t_0, t_f]) &= h \sum_{j=0}^{N-1} F(t_j), \\ \widetilde{\Phi}_N(t_{j+1}, t_j) &= I_n + hA(t_j), \\ \widetilde{\Phi}_{N,1}(t_{j+1}, t_j) &= I_n. \end{aligned}$$

Proposition 6.65. *(cf. [Nik88], [DF89], [Wol90], see also [Art94], [Gra03])*
If

- $A(\cdot)$ *is Lipschitz,*
- $B(\cdot)$ *is bounded,*
- $\tau_1(\delta^\star(l,\Phi(t_f,\cdot)B(\cdot)U),h) \le Ch$ *uniformly in* $l \in S_{n-1}$*, e.g., if* $B(\cdot)$ *is Lipschitz,*
- $\mathrm{d_H}(X_0, X_0^N) = \mathcal{O}(h)$,

then Euler's method converges at least with order $\mathcal{O}(h)$.

Proof. The quadrature method has precision 0.
If $B(\cdot)$ is Lipschitz, then $\Phi(t_f,\cdot)B(\cdot)$ and hence also $\delta^\star(l,\Phi(t_f,\cdot)B(\cdot)U)$ is Lipschitz (uniformly in $l \in S_{n-1}$).
This shows that $\tau_1(\delta^\star(l,\Phi(t_f,\cdot)B(\cdot)U),h) = \mathcal{O}(h)$. The following estimations are valid:

$$\begin{aligned}\|\widetilde{\Phi}_N(t_{j+1},t_j) - \Phi(t_{j+1},t_j)\| &= \|(\mathrm{I}_n + hA(t_j)) - \Phi(t_{j+1},t_j)\| = \mathcal{O}(h^2),\\ \|\widetilde{\Phi}_{N,1}(t_{j+1},t_j) - \Phi(t_{j+1},t_j)\| &= \|\mathrm{I}_n - \Phi(t_{j+1},t_j)\| = \mathcal{O}(h).\end{aligned}$$

Hence, Proposition 6.54 can be applied yielding $\mathcal{O}(h)$. □

For order of convergence 1, it is sufficient that $A(\cdot)$ and $B(\cdot)$ (resp. $\delta^\star(l,\Phi(t_f,\cdot)B(\cdot)U)$, uniformly in $l \in S_{n-1}$) have bounded variation.

6.5.2 Modified Euler Method

Remark 6.66. *Consider modified Euler method, i.e. the Butcher array*

$$\begin{array}{c|cc} 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ \hline & 0 & 1 \end{array} \quad .$$

underlying quadrature method = iterated midpoint rule:

$$Q_N(F;[t_0,t_f]) = h\sum_{j=0}^{N-1} F(t_j + \frac{h}{2})$$

Grouping by η_j^N *and the two selections* $u_j^{(1)}$ *and* $u_j^{(2)}$ *yields*

$$\begin{aligned}X_{j+1}^N = \Big(I_n + hA(t_j+\frac{h}{2}) + \frac{h^2}{2}A(t_j+\frac{h}{2})A(t_j)\Big)X_j^N \\ + h \bigcup_{u_j^{(1)},u_j^{(2)}\in U} \Big(\frac{h}{2}A(t_j+\frac{h}{2})B(t_j)u_j^{(1)} + B(t_j+\frac{h}{2})u_j^{(2)}\Big)\end{aligned}$$

for $j = 0,\dots,N-1$.

Proposition 6.67. *(cf. [Bai05])*
Modified Euler method with constant selection strategy "$u_j^{(1)} = u_j^{(2)}$*" is a combination method with the following settings:*

$$\begin{aligned}Q_N(F;[t_0,t_f]) &= h\sum_{j=0}^{N-1} F(t_j+\frac{h}{2}),\\ \widetilde{\Phi}_N(t_{j+1},t_j) &= I_n + hA(t_j+\frac{h}{2}) + \frac{h^2}{2}A(t_j+\frac{h}{2})A(t_j),\\ \widetilde{U}_{N,1}(t_j+\frac{h}{2}) &:= \Big(B(t_j+\frac{h}{2}) + \frac{h}{2}A(t_j+\frac{h}{2})B(t_j)\Big)U.\end{aligned}$$

constant approximation by the quadrature method (midpoint rule) on $[t_j,t_{j+1}]$
$\Rightarrow$ constant selection in modified Euler is appropriate

Proposition 6.68. *(cf. [Bai05])*
If

- $A'(\cdot)$ *and* $B(\cdot)$ *are Lipschitz,*
- $\tau_2(\delta^\star(l,\Phi(t_f,\cdot)B(\cdot)U),h) \le Ch^2$ *uniformly in* $l \in S_{n-1}$*, e.g., if* $B'(\cdot)$ *is Lipschitz,*
- $\mathrm{d_H}(X_0, X_0^N) = \mathcal{O}(h^2)$*,*

then modified Euler method with constant selection strategy converges at least with order $\mathcal{O}(h^2)$*.*

For order of convergence 2, it is sufficient that $A'(\cdot)$ and $B'(\cdot)$ (resp. $\frac{d}{dt}\delta^\star(l,\Phi(t_f,\cdot)B(\cdot)U)$, uniformly in $l \in S_{n-1}$) have bounded variation.
Proof. The quadrature method has precision 1.
Careful Taylor expansion shows (as in the pointwise case) that

$$\begin{aligned}&\|\widetilde{\Phi}_N(t_{j+1},t_j) - \Phi(t_{j+1},t_j)\|\\ =&\|(\mathrm{I}_n + hA(t_j+\frac{h}{2}) + \frac{h^2}{2}A(t_j+\frac{h}{2})A(t_j)) - \Phi(t_{j+1},t_j)\| = \mathcal{O}(h^3).\end{aligned}$$

The following estimations are valid:

$$\begin{aligned}\mathrm{d_H}(\widetilde{U}_{N,1}(t_j+\frac{h}{2}), (\mathrm{I}_n + \frac{h}{2}A(t_j+\frac{h}{2}))B(t_j+\frac{h}{2})U) = \mathcal{O}(h^2),\\ \mathrm{d_H}((\mathrm{I}_n + \frac{h}{2}A(t_j+\frac{h}{2}))B(t_j+\frac{h}{2})U, \Phi(t_{j+1},t_j+\frac{h}{2})B(t_j+\frac{h}{2})U) = \mathcal{O}(h^2)\end{aligned}$$

Hence,

$$\mathrm{d_H}(\widetilde{U}_{N,1}(t_j+\frac{h}{2}), \Phi(t_{j+1},t_j+\frac{h}{2})B(t_j+\frac{h}{2})U) = \mathcal{O}(h^2)$$

follows with Proposition 3.153(i).
Alltogether, Proposition 6.54 can be applied yielding $\mathcal{O}(h^2)$. □

Proposition 6.69. *(cf. [Bai05])*
Modified Euler method with two free choices $u_j^{(1)}, u_j^{(2)} \in U$ *is a combination method with the following settings:*

$$\begin{aligned}\text{(i)}\quad Q_N(F;[t_0,t_f]) &= h\sum_{j=0}^{N-1} F(t_j+\frac{h}{2}),\\ \widetilde{\Phi}_N(t_{j+1},t_j) &= I_n + hA(t_j+\frac{h}{2}) + \frac{h^2}{2}A(t_j+\frac{h}{2})A(t_j),\\ \widetilde{U}_{N,1}(t_j+\frac{h}{2}) &= B(t_j+\frac{h}{2})U + \frac{h}{2}A(t_j+\frac{h}{2})B(t_j)U\end{aligned}$$

Proposition 6.69 (continued). *resp.*

$$\begin{aligned}\text{(ii)}\quad Q_N(F;[t_0,t_f]) &= \frac{h}{2}\sum_{j=0}^{N-1}\big(F(t_j)+F(t_{j+1})\big),\\ \widetilde{\Phi}_N(t_{j+1},t_j) &= I_n + hA(t_j+\frac{h}{2}) + \frac{h^2}{2}A(t_j+\frac{h}{2})A(t_j),\\ \widetilde{U}_{N,1}(t_j) &= B(t_j+\frac{h}{2})U + hA(t_j+\frac{h}{2})B(t_j)U,\\ \widetilde{U}_{N,2}(t_{j+1}) &= B(t_j+\frac{h}{2})U.\end{aligned}$$

problem in (i):
Minkowski sum of 2 sets in $\widetilde{U}_{N,1}(t_j+\frac{h}{2})$, hence disturbance term $\mathcal{O}(h)$
problem in (ii): Minkowski sum of 2 sets and $B(t_j+\frac{h}{2})$ instead of $B(t_j)$ in $\widetilde{U}_{N,1}(t_j)$
resp. $B(t_j+\frac{h}{2})$ instead of $B(t_{j+1})$ in $\widetilde{U}_{N,2}(t_{j+1})$
The problem with two selections was also observed in the approximation of nonlinear optimal controls (cf. [DHV00a]).

Proposition 6.70. *(cf. [Bai05])*
If

- $A(\cdot)$ *is Lipschitz and* $B(\cdot)$ *is bounded,*
- $\tau_1(\delta^\star(l,\Phi(t_f,\cdot)B(\cdot)U),h)\le Ch$ *uniformly in* $l\in S_{n-1}$*, e.g., if* $B(\cdot)$ *is Lipschitz,*
- $\mathrm{d_H}(X_0,X_0^N)=\mathcal{O}(h)$*,*

then modified Euler method with two free selections converges at least with order $\mathcal{O}(h)$*.*

Proof. The underlying quadrature method has precision 1, hence also 0.
Careful Taylor expansion shows (as in the pointwise case) that

$$\begin{aligned}
&\|\widetilde{\Phi}_N(t_{j+1},t_j)-\Phi(t_{j+1},t_j)\|\\
=&\|\left(\mathrm{I}_n+hA(t_j+\frac{h}{2})+\frac{h^2}{2}A(t_j+\frac{h}{2})A(t_j)\right)-\Phi(t_{j+1},t_j)\|\\
\le&\|\left(\mathrm{I}_n+hA(t_j+\frac{h}{2})\right)-\Phi(t_{j+1},t_j)\|+\frac{h^2}{2}\|A(t_j+\frac{h}{2})\|\cdot\|A(t_j)\|=\mathcal{O}(h^2).
\end{aligned}$$

The following estimations for (i) in Proposition 6.69 are valid:

$$\begin{aligned}
&\mathrm{d_H}(\widetilde{U}_{N,1}(t_j+\frac{h}{2}),\Phi(t_{j+1},t_j+\frac{h}{2})B(t_j+\frac{h}{2})U)\\
=&\mathrm{d_H}(B(t_j+\frac{h}{2})U+\frac{h}{2}A(t_j+\frac{h}{2})B(t_j)U,\Phi(t_{j+1},t_j+\frac{h}{2})B(t_j+\frac{h}{2})U)\\
\le&\mathrm{d_H}(B(t_j+\frac{h}{2})U,\Phi(t_{j+1},t_j+\frac{h}{2})B(t_j+\frac{h}{2})U)+\frac{h}{2}\|A(t_j+\frac{h}{2})B(t_j)U\|\\
\le&\|\mathrm{I}_n-\Phi(t_{j+1},t_j+\frac{h}{2})\|\cdot\|B(t_j+\frac{h}{2})\|\cdot\|U\|+\mathcal{O}(h)=\mathcal{O}(h).
\end{aligned}$$

Hence, Proposition 6.54 can be applied yielding $\mathcal{O}(h)$. □

Remark 6.71. *If we assume that* $A'(\cdot)$ *is Lipschitz, it would be valid that*

$$\begin{aligned}
&\|\widetilde{\Phi}_N(t_{j+1},t_j)-\Phi(t_{j+1},t_j)\|\\
=&\|(I_n+hA(t_j+\frac{h}{2})+\frac{h^2}{2}A(t_j+\frac{h}{2})A(t_j))-\Phi(t_{j+1},t_j)\|=\mathcal{O}(h^3).
\end{aligned}$$

But the disturbances in $\widetilde{U}_{N,1}(t_j+\frac{h}{2})$ *are not of order* $\mathcal{O}(h^2)$*.*
Please notice that in (i)

$$\begin{aligned}
\widetilde{U}_{N,1}(t_j+\frac{h}{2})&=B(t_j+\frac{h}{2})U+\frac{h}{2}A(t_j+\frac{h}{2})B(t_j)U\\
&\neq\left(B(t_j+\frac{h}{2})+\frac{h}{2}A(t_j+\frac{h}{2})B(t_j)\right)U\\
&=\Phi(t_{j+1},t_j+\frac{h}{2})B(t_j+\frac{h}{2})U+\mathcal{O}(h^2)
\end{aligned}\tag{45}$$

Remark 6.71 (continued). *and*

$$\mathbf{d_H}(B(t_j+\frac{h}{2})U+\frac{h}{2}A(t_j+\frac{h}{2})B(t_j)U,$$
$$(B(t_j+\frac{h}{2})+\frac{h}{2}A(t_j+\frac{h}{2})B(t_j))U)=\mathcal{O}(h).$$

constant approximation by the quadrature method (midpoint rule) on $[t_j,t_{j+1}]$
$\Rightarrow$ *two free selections in modified Euler do not fit well, possible order breakdown*

Example 6.72 *(very similar to [Vel92, Example (2.8)], cf. Example 6.18).* Let $n=2$, $m=1$, $I=[0,1]$ and set

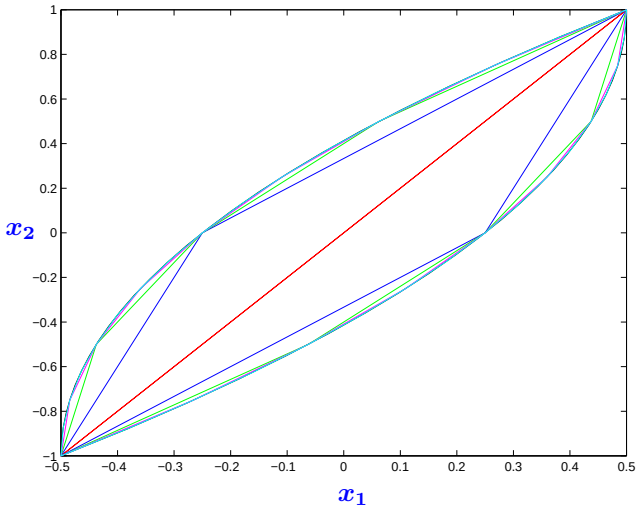
$$A(t)=\begin{pmatrix}0&1\\0&0\end{pmatrix},\quad B(t)=\begin{pmatrix}0\\1\end{pmatrix}\quad\text{and}\quad U=[-1,1].$$

Since (45) is fulfilled here, both selection strategies for modified Euler differ.
data for the reference set:

- combination method "iterated trapezoidal rule/Heun's method"
- $N=10000$ subintervals
- calculated supporting points in $M=200$ directions

Set-Valued Runge-Kutta Methods

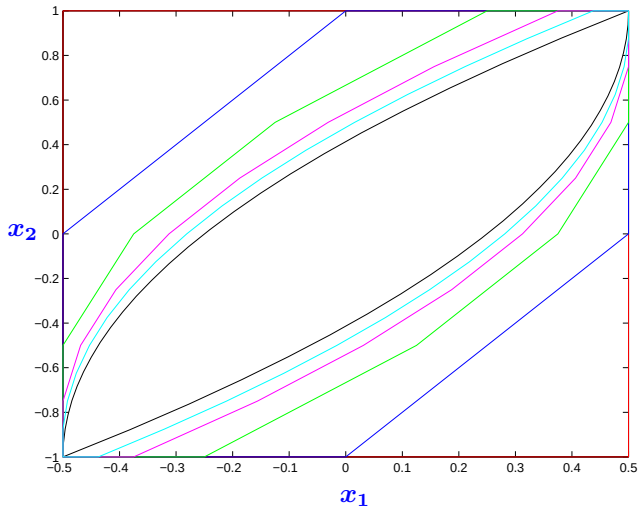
Modified Euler with Constant Selections, $N=1,2,4,8,16$



– $N=1$ (line segment)
– $N=2$
– $N=4$
– $N=8$
– $N=16$ (biggest set)

order of convergence: $\mathcal{O}(h^2)$

Modified Euler with 2 Free Selections, $N=1,2,4,8,16$



– reference set

– $N = 1$ (square)

– $N = 2$

– $N = 4$

– $N = 8$

– $N = 16$ (smallest set)

order of convergence: $\mathcal{O}(h)$

Example 6.72 *(continued).* computed estimations of the order of convergence:

N	Hausdorff distance to reference set	estimated order of convergence	Hausdorff distance to reference set	estimated order of convergence
1	0.21434524	———	0.75039466	———
2	0.05730861	1.90311	0.36454336	1.04156
4	0.01517382	1.91717	0.17953522	1.02182
8	0.00384698	1.97979	0.08841414	1.02192
16	0.00096510	1.99498	0.04419417	1.00042
	(constant selections)		(2 free selections)	

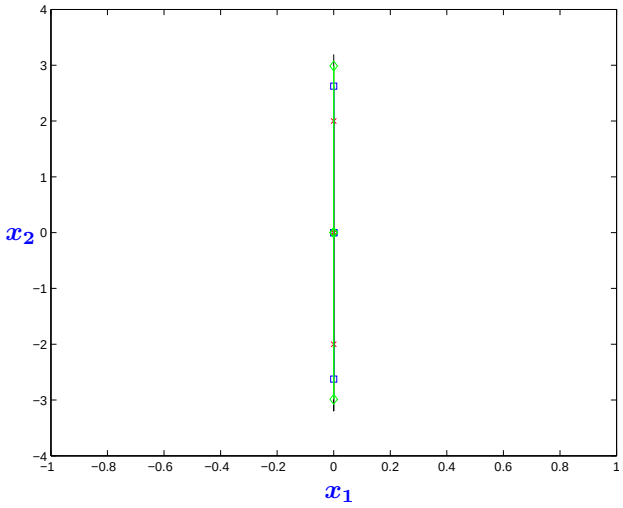
Possible order breakdown to $\mathcal{O}(h)$ in Proposition 6.70 for modified Euler with two free selections can occur!

Example 6.73. data as in Example 6.72, only $A(t) = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$. Both selection strategies for modified Euler coincide, since

$$\begin{aligned}
(B + \frac{h}{2}AB)U &= (\begin{pmatrix} 0 \\ 1 \end{pmatrix} + \frac{h}{2}\begin{pmatrix} 0 \\ 2 \end{pmatrix})\,\mathrm{co}\{-1, 1\} \\
&= \mathrm{co}\{\begin{pmatrix} 0 \\ -1-h \end{pmatrix}, \begin{pmatrix} 0 \\ 1+h \end{pmatrix}\}, \\
BU + \frac{h}{2}ABU &= \begin{pmatrix} 0 \\ 1 \end{pmatrix}\mathrm{co}\{-1, 1\} + \frac{h}{2}\begin{pmatrix} 0 \\ 2 \end{pmatrix}\mathrm{co}\{-1, 1\} \\
&= \mathrm{co}\{\begin{pmatrix} 0 \\ -1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}\} + \mathrm{co}\{\begin{pmatrix} 0 \\ -h \end{pmatrix}, \begin{pmatrix} 0 \\ h \end{pmatrix}\} \\
&= \mathrm{co}\{\begin{pmatrix} 0 \\ -1-h \end{pmatrix}, \begin{pmatrix} 0 \\ 1+h \end{pmatrix}\}.
\end{aligned}$$

Set-Valued Runge-Kutta Methods

Modified Euler with Constant Selections, $N = 1, 2, 4$

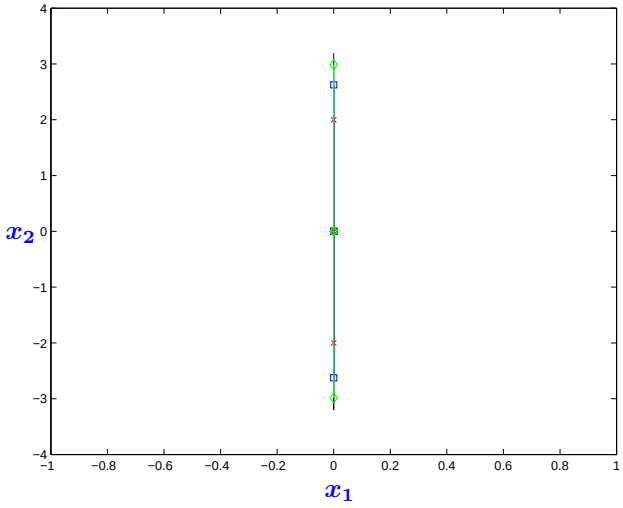


– reference set
(longest line)

marked end points:
x $N = 1$
□ $N = 2$
◇ $N = 4$

order of
convergence: $\mathcal{O}(h^2)$

Modified Euler with 2 Free Selections, $N = 1, 2, 4$



– reference set
(longest line)

marked end points:
x $N = 1$
□ $N = 2$
◇ $N = 4$

order of
convergence: $\mathcal{O}(h^2)$

Example 6.73 *(continued).* computed estimations of the order of convergence:

N	Hausdorff distance to reference set	estimated order of convergence	Hausdorff distance to reference set	estimated order of convergence
1	1.19452805	———	1.19452805	———
2	0.56952805	1.06860	0.56952805	1.06860
4	0.20807785	1.45264	0.20807785	1.45264
8	0.06340445	1.71447	0.06340445	1.71447
16	0.01748660	1.85833	0.01748660	1.85833
32	0.00458787	1.93035	0.00458787	1.93035
64	0.00117462	1.96562	0.00117462	1.96562
	(constant selections)		(2 free selections)	

Possible order breakdown to $\mathcal{O}(h)$ in Proposition 6.70 for modified Euler with two free selections does not occur always!

Proposition 6.74. *Modified Euler method with linear interpolated selections* $u_j^{(1)}, u_j^{(3)} \in U$ *and* $u_j^{(2)} = \frac{1}{2}(u_j^{(1)} +$

$u_j^{(3)}$) *is a combination method with the settings:*

$$\begin{aligned} Q_N(F;[t_0,t_f]) &= \frac{h}{2}\sum_{j=0}^{N-1}\big(F(t_j)+F(t_{j+1})\big),\\ \widetilde{\Phi}_N(t_{j+1},t_j) &= I_n + hA(t_j+\frac{h}{2}) + \frac{h^2}{2}A(t_j+\frac{h}{2})A(t_j),\\ \widetilde{U}_{N,1}(t_j) &= \big(B(t_j+\frac{h}{2}) + hA(t_j+\frac{h}{2})B(t_j)\big)U,\\ \widetilde{U}_{N,2}(t_{j+1}) &= B(t_j+\frac{h}{2})U. \end{aligned}$$

problem: $B(t_j+\frac{h}{2})$ instead of $B(t_j)$ resp. $B(t_j+\frac{h}{2})$ instead of $B(t_{j+1})$
This strategy was used in the approximation of the value function of Hamilton-Jacobi-Bellman equations in [Fer94] and caused two unexpected results in one test example.

Proposition 6.75. *If*

- $A(\cdot)$ *is Lipschitz and* $B(\cdot)$ *is bounded,*
- $\tau_1(\delta^\star(l,\Phi(t_f,\cdot)B(\cdot)U),h) \le Ch$ *uniformly in* $l \in S_{n-1}$*, e.g., if* $B(\cdot)$ *is Lipschitz,*
- $\mathrm{d_H}(X_0, X_0^N) = \mathcal{O}(h)$,

then modified Euler method with linear interpolated selections converges at least with order $\mathcal{O}(h)$.

Proof. The quadrature method has precision 1, hence also 0.
Careful Taylor expansion shows as for two free selections that

$$\|\widetilde{\Phi}_N(t_{j+1},t_j) - \Phi(t_{j+1},t_j)\| = \mathcal{O}(h^2).$$

The following estimations in Proposition 6.74 are valid:

$$\begin{aligned} &\mathrm{d_H}(\widetilde{U}_{N,1}(t_j), \Phi(t_{j+1},t_j)B(t_j)U)\\ =&\mathrm{d_H}(\big(B(t_j+\frac{h}{2}) + hA(t_j+\frac{h}{2})B(t_j)\big)U, \big(B(t_j)+hA(t_j)B(t_j)\big)U)\\ &+\mathrm{d_H}((\mathrm{I}_n + hA(t_j))B(t_j)U, \Phi(t_{j+1},t_j)B(t_j)U)\\ \le&\big(\|B(t_j+\frac{h}{2}) - B(t_j)\| + h\|A(t_j+\frac{h}{2}) - A(t_j)\|\cdot\|B(t_j)\|\big)\cdot\|U\|\\ &+\|(\mathrm{I}_n + hA(t_j)) - \Phi(t_{j+1},t_j)\|\cdot\|B(t_j)\|\cdot\|U\| = \mathcal{O}(h),\\ &\mathrm{d_H}(\widetilde{U}_{N,2}(t_{j+1}), \Phi(t_{j+1},t_{j+1})B(t_{j+1})U)\\ \le&\|B(t_j+\frac{h}{2}) - B(t_{j+1})\|\cdot\|U\| = \mathcal{O}(h) \end{aligned}$$

Hence, Proposition 6.54 can be applied yielding $\mathcal{O}(h)$. □

Remark 6.76. *Assuming more smoothness, we could show that*

$$\|\widetilde{\Phi}_N(t_{j+1},t_j) - \Phi(t_{j+1},t_j)\| = \mathcal{O}(h^3),$$

for time-independent situations it is valid that

$$\begin{aligned} \mathrm{d_H}(\widetilde{U}_{N,1}(t_j), \Phi(t_{j+1},t_j)B(t_j)U) &= \mathcal{O}(h^2),\\ \mathrm{d_H}(\widetilde{U}_{N,2}(t_{j+1}), \Phi(t_{j+1},t_{j+1})B(t_{j+1})U) &= \mathcal{O}(h^2) \end{aligned}$$

and hence global order of convergence $\mathcal{O}(h^2)$.

Example 6.77. (cf. [BL94a]) Let $n = m = 2$, $I = [0,1]$ and set

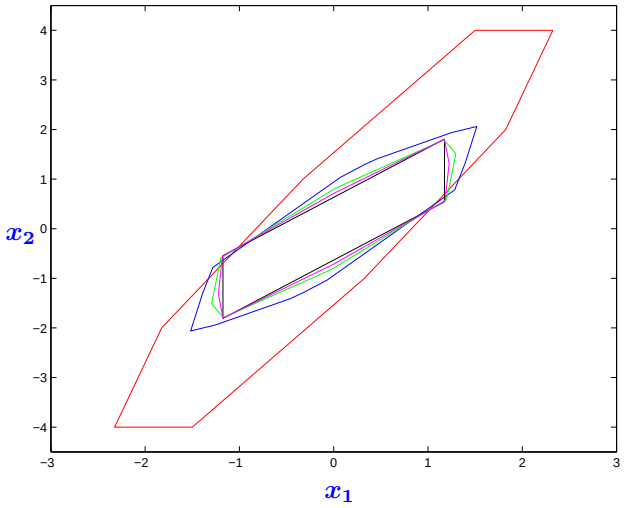
$$A(t) = \begin{pmatrix} 1 & -1 \\ 4 & -3 \end{pmatrix}, \; B(t) = \begin{pmatrix} 1-t & t \cdot e^t \\ 3-2t & (-1+2t) \cdot e^t \end{pmatrix} \text{ and } U = [-1,1]^2.$$

data for the reference set:

- combination method "iterated Simpson's rule/RK(4)"
- $N = 100000$ subintervals
- calculated supporting points in $M = 200$ directions

Set-Valued Runge-Kutta Methods

Modified Euler with Linear Interpolation, $N = 1, 2, 4, 8$



– reference set (smallest set)

– $N = 1$ (biggest set)

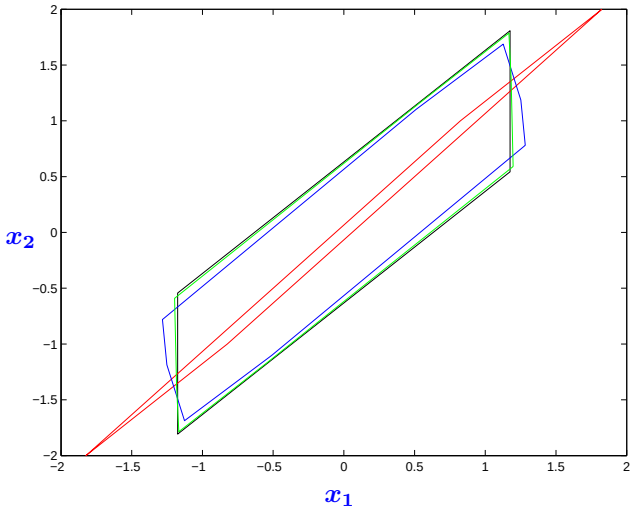
– $N = 2$

– $N = 4$

– $N = 8$

order of convergence: $\mathcal{O}(h)$

Modified Euler with Constant Selections, $N = 1, 2, 4$



– reference set

– $N = 1$ (flattest set)

– $N = 2$

– $N = 4$

order of convergence: $\mathcal{O}(h^2)$

Example 6.77 *(continued).* computed estimations of the order of convergence:

N	Hausdorff distance to reference set	estimated order of convergence	Hausdorff distance to reference set	estimated order of convergence
1	2.47539809	———	0.67713923	———
2	0.42619535	2.53807	0.12998374	2.38112
4	0.12006081	1.82775	0.02271635	2.51653
8	0.05540102	1.11578	0.00498557	2.18790
16	0.02687764	1.04351	0.00119539	2.06027
32	0.01321630	1.02409	0.00029294	2.02881
64	0.00655070	1.01260	0.00007252	2.01407
	(selections by linear interpolation)		(constant selections)	

Possible order breakdown to $\mathcal{O}(h)$ in Proposition 6.75 for modified Euler with linear interpolated selections can occur!

7 Discrete Approximation of Optimal Control

Optimal Control Problem

Problem 7.1 (Optimal Control Problem). *Minimize*

$$\varphi(x(t_0), x(t_f)) + \int_{t_0}^{t_f} f_0(t, x(t), u(t))dt \tag{46}$$

subject to the differential equation

$$\dot{x}(t) = f(t, x(t), u(t)), \qquad t_0 \le t \le t_f, \tag{47}$$

the mixed control-state constraints

$$c(t, x(t), u(t)) \le 0, \qquad t_0 \le t \le t_f, \tag{48}$$

the pure state constraints

$$s(t, x(t)) \le 0, \qquad t_0 \le t \le t_f, \tag{49}$$

Optimal Control Problem

Problem 7.1 (continued). *the boundary conditions*

$$\psi(x(t_0), x(t_f)) = 0, \tag{50}$$

and the set constraints

$$u(t) \in \mathcal{U} \subseteq \mathbb{R}^{n_u}, \qquad t_0 \le t \le t_f. \tag{51}$$

Terminology

Depending on the structure of the objective function we call the problem

- Bolza-Problem, if $\varphi \not\equiv 0$ and $f_0 \not\equiv 0$
- Mayer-Problem, if $\varphi \not\equiv 0$ and $f_0 \equiv 0$
- Lagrange-Problem , if $\varphi \equiv 0$ and $f_0 \not\equiv 0$

The problem is called autonomous, if the functions f_0, f, c, s do not depend explicitly of the time t.

Transformation: Free Initial/Final Time to Fixed Time

Let t_0 and/or t_f be *free*.
New time τ by time transformation

$$t(\tau) := t_0 + \tau(t_f - t_0), \qquad \tau \in [0, 1]$$

New state $\bar{x}$:

$$\bar{x}(\tau) := x(t(\tau)) = x(t_0 + \tau(t_f - t_0))$$

New control $\bar{u}$:

$$\bar{u}(\tau) := u(t(\tau)) = u(t_0 + \tau(t_f - t_0))$$

Transformation: Free Initial/Final Time to Fixed Time

Differentiation of *new state* w.r.t. *new time*:

$$\begin{aligned} \bar{x}'(\tau) &= \dot{x}(t_0 + \tau(t_f - t_0)) \cdot t'(\tau) \\ &= (t_f - t_0) \cdot f(t_0 + \tau(t_f - t_0), x(t_0 + \tau(t_f - t_0)), u(t_0 + \tau(t_f - t_0))) \\ &= (t_f - t_0) f(t_0 + \tau(t_f - t_0), \bar{x}(\tau), \bar{u}(\tau)) \end{aligned}$$

Either t_0 and/or t_f are *new real optimization variables* or introduce *additional differential equations*:

$$\begin{aligned} \dot{t}_0(\tau) &= 0, \qquad t_0(0) \text{ free}, \\ \dot{t}_f(\tau) &= 0, \qquad t_f(0) \text{ free} \end{aligned}$$

Transformation: Free Initial/Final Time to Fixed Time

Transformed problem:

Minimize

$$\varphi(\bar{x}(0),\bar{x}(1)) + \int_0^1 (t_f - t_0) f_0(t(\tau),\bar{x}(\tau),\bar{u}(\tau))d\tau$$

s.t.

$$\begin{aligned}
\bar{x}'(\tau) &= (t_f - t_0) f(t(\tau),\bar{x}(\tau),\bar{u}(\tau)) \quad \text{a.e. in } [0,1],\\
\psi(\bar{x}(0),\bar{x}(1)) &= 0_{n_\psi},\\
c(t(\tau),\bar{x}(\tau),\bar{u}(\tau)) &\le 0_{n_c} \quad \text{a.e. in } [0,1],\\
s(t(\tau),\bar{x}(\tau)) &\le 0_{n_s} \quad \text{in } [0,1],\\
\bar{u}(\tau) &\in \mathcal{U} \quad \text{a.e. in } [0,1].
\end{aligned}$$

Transformation: Non-Autonomous Problem to Autonomous Problem

Additional state:

$$\dot{T}(t) = 1, \qquad T(t_0) = t_0$$

Replace t in f_0, f, c, s by T:

Minimize

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

s.t.

$$\begin{aligned}
\dot{x}(t) &= f(T(t),x(t),u(t)) \quad \text{a.e. in } [t_0,t_f],\\
\dot{T}(t) &= 1, \quad T(t_0) = t_0,\\
\psi(x(t_0),x(t_f)) &= 0_{n_\psi},\\
c(T(t),x(t),u(t)) &\le 0_{n_c} \quad \text{a.e. in } [t_0,t_f],\\
s(T(t),x(t)) &\le 0_{n_s} \quad \text{in } [t_0,t_f],\\
u(t) &\in \mathcal{U} \quad \text{a.e. in } [t_0,t_f].
\end{aligned}$$

Transformation: Bolza-Problem to Mayer-Problem

Bolza objective functional:

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

Additional state z:

$$\dot{z}(t) = f_0(t,x(t),u(t)), \qquad z(t_0) = 0.$$

Then:

$$z(t_f) = z(t_0) + \int_{t_0}^{t_f} f_0(t,x(t),u(t))dt = \int_{t_0}^{t_f} f_0(t,x(t),u(t))dt.$$

Equivalent Mayer objective functional:

$$\varphi(x(t_0),x(t_f)) + z(t_f)$$

Transformation: Tschebyscheff-Problems

Tschebyscheff-Problem:

$$\text{Minimize} \quad \max_{t\in[t_0,t_f]} h(t,x(t),u(t)) \quad \text{s.t.} \quad (47)-(51) \tag{52}$$

Define

$$\alpha := \max_{t\in[t_0,t_f]} h(t,x(t),u(t))$$

Then:

$$h(t,x(t),u(t)) \le \alpha \qquad \text{a.e. in } [t_0,t_f] \tag{53}$$

This is an additional mixed control-state constraint!

Transformation: Tschebyscheff-Problems

The Tschebyscheff-problem is equivalent with

$$\text{Minimize} \quad \alpha \quad \text{s.t.} \quad (47)-(51),(53)$$

α is either an *additional optimization variable* or an *additional state* with

$$\dot\alpha(t) = 0, \qquad \alpha(t_0) \text{ free}$$

Questions

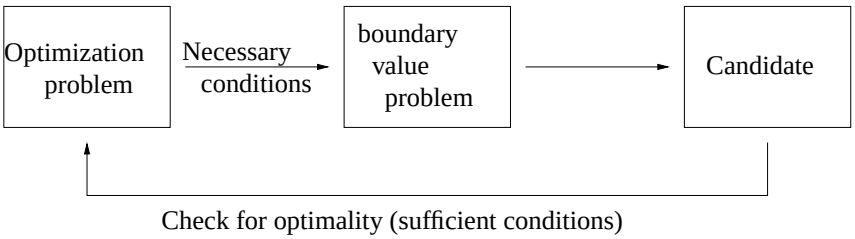
- Necessary optimality conditions (minimum principle) ?
  - Problems without state and control constraints
  - Problems with mixed control-state constraints
  - Problems with pure state constraints
- Solution approaches for optimal control problems ?

Solution Approaches I

Indirect approach

- based on the evaluation of necessary optimality conditions
- leads to boundary value problems
- numerical solution by multiple shooting methods

$OCP \rightarrow$ minimum principle $\rightarrow$ BVP $\rightarrow$ multiple shooting

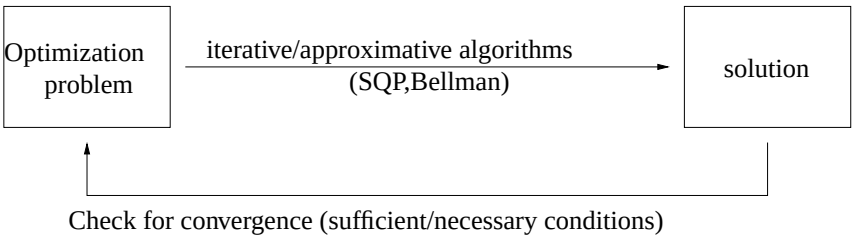


Solution Approaches II

Direct approach

- based on discretization of the optimal control problem
- leads to finite dimensional optimization problem
- numerical solution by sequential quadratic programming (SQP)

$$\boldsymbol{OCP} \quad \rightarrow \quad \text{discretization} \quad \rightarrow \quad \text{NLP} \quad \rightarrow \quad \text{SQP}$$



Questions

- Necessary optimality conditions (minimum principle) ?
  - Problems without state and control constraints
  - Problems with mixed control-state constraints
  - Problems with pure state constraints
-

7.1 Minimum Principles

Minimum Principle for Unconstrained Problems

Hamiltonian:

$$\mathcal{H}(t,x,u,\lambda,\ell_0) := \ell_0 f_0(t,x,u) + \lambda^\top f(t,x,u).$$

Theorem 7.2 (Local Minimum Principle for Unconstrained Problems). *Assumptions:*

(i) *Let φ, f_0, f, ψ be continuous w.r.t. all arguments and continuously differentiable w.r.t. x and u.*

(ii) *Let $\mathcal{U} \subseteq \mathbb{R}^{n_u}$ be a closed convex set with int$(\mathcal{U}) \neq \emptyset$.*

(iii) *Let $(\hat{x},\hat{u}) \in W^{1,\infty}([t_0,t_f],\mathbb{R}^{n_x}) \times L^\infty([t_0,t_f],\mathbb{R}^{n_u})$ be a (weak) local minimum.*

(iv) *no pure state constraints (49) and no mixed control-state constraints (48)!*

Minimum Principle for Unconstrained Problems

Theorem 7.2 (continued). *Assertion:*
There exist Lagrange multipliers $\ell_0 \in \mathbb{R}$, $\sigma \in \mathbb{R}^{n_\psi}$, $\lambda \in W^{1,\infty}([t_0,t_f],\mathbb{R}^{n_x})$ with:

(i) $\ell_0 \geq 0$, $(\ell_0,\sigma,\lambda) \neq \Theta$,

(ii) *Adjoint Differential Equation:*

$$\dot{\lambda}(t) = -\mathcal{H}'_x(t,\hat{x}(t),\hat{u}(t),\lambda(t),\ell_0)^\top \qquad \text{a.e. in } [t_0,t_f],$$

Minimum Principle for Unconstrained Problems

Theorem 7.2 (continued).

(iii) Transversality Conditions:

$$\begin{aligned} \lambda(t_0)^\top &= -\left(\ell_0 \varphi'_{x_0} + \sigma^\top \psi'_{x_0}\right), \\ \lambda(t_f)^\top &= \ell_0 \varphi'_{x_f} + \sigma^\top \psi'_{x_f}. \end{aligned}$$

(iv) Optimality Condition: A.e. in $[t_0, t_f]$ it holds

$$\mathcal{H}'_u(t, \hat{x}(t), \hat{u}(t), \lambda(t), \ell_0)(u - \hat{u}(t)) \geq 0 \qquad \forall u \in \mathcal{U}.$$

Example: Minimum Energy

Example 7.3 *(Minimum Energy without State Constraint).* Minimize

$$F(x_1, x_2, u) = \frac{1}{2} \int_0^1 u(t)^2 dt$$

subject to

$$\begin{aligned} \dot{x}_1(t) &= x_2(t), \qquad x_1(0) = x_1(1) = 0, \\ \dot{x}_2(t) &= u(t), \qquad x_2(0) = -x_2(1) = 1. \end{aligned}$$

We apply Theorem 7.2.

Example: Minimum Energy

Example 7.3 *(continued).* Hamiltonian:

$$\mathcal{H} = \ell_0 \frac{u^2}{2} + \lambda_1 x_2 + \lambda_2 u.$$

Adjoint Differential Equation:

$$\begin{aligned} \dot{\lambda}_1(t) &= -\mathcal{H}'_{x_1} = 0, \\ \dot{\lambda}_2(t) &= -\mathcal{H}'_{x_2} = -\lambda_1(t). \end{aligned}$$

Solution:

$$\lambda_1(t) = c_1, \qquad \lambda_2(t) = -c_1 t + c_2$$

for some constants c_1, c_2.

Example: Minimum Energy

Example 7.3 *(continued).* Transversality conditions (yield no additional informations):

$$-\sigma_1 = c_1 = \sigma_3, \qquad c_2 = -\sigma_2, \quad -c_1 + c_2 = \sigma_4.$$

Optimality condition ($\mathcal{U} = \mathbb{R} \Rightarrow \mathcal{H}'_u = 0$):

$$0 = \mathcal{H}'_u = \ell_0 \hat{u}(t) + \lambda_2(t)$$

Example: Minimum Energy

Example 7.3 *(continued).*

(i) *Case 1:* $\ell_0 = 0$ The optimality condition yields $\lambda_2(t) = 0$ a.e.. Hence, $c_1 = c_2 = 0$ and also $\lambda_1(t) = c_1 = 0$. Thus, $\ell_0 = 0, \lambda_1 \equiv 0, \lambda_2 \equiv 0$ which contradicts the condition $(\ell_0, \lambda_1, \lambda_2) \neq 0$.

(ii) *Case 2:* $\ell_0 = 1$ *(without loss of generality)* Optimality condition:

$$\hat{u}(t) = -\lambda_2(t) = c_1 t - c_2.$$

Example: Minimum Energy

Example 7.3 *(continued).* The constants c_1, c_2 are determined by the boundary conditions $x_1(1) = 0$ and $x_2(1) = -1$:

$$\left.\begin{array}{rcl} \frac{c_1}{6} - \frac{c_2}{2} & = & -1 \\ \frac{c_1}{2} - c_2 & = & -2 \end{array}\right\} \quad \Rightarrow \quad c_1 = 0,\ c_2 = 2.$$

Candidate for optimal solution:

$$\begin{array}{rclrcl} \hat{u}(t) & = & -2, & \sigma & = & (0,-2,0,2)^\top, \\ \hat{x}_1(t) & = & t(1-t), & \hat{x}_2(t) & = & 1-2t, \\ \lambda_1(t) & = & 0, & \lambda_2(t) & = & 2. \end{array}$$

Example

The follwing example shows, that the case $\ell_0 = 0$ may occur.

Example 7.4. Minimize

$$F(x,u) = \int_0^1 u(t)dt$$

subject to

$$\dot{x}(t) \quad = \quad u(t)^2, \qquad x(0) = x(1) = 0.$$

Obviously, only $u \equiv 0$ satisfies the constraints and thus is optimal. *The constraints fully determine the solution! No degree of freedom left!*

Example

Example 7.4 *(continued).* With $\mathcal{H}(x,u,\lambda,\ell_0) = \ell_0 u + \lambda u^2$ it follows

$$\begin{array}{rcl} 0 & = & \mathcal{H}'_u = \ell_0 + 2\lambda u, \\ \dot{\lambda} & = & 0 \qquad \Rightarrow \qquad \lambda = \text{const}. \end{array}$$

Now, assume that $\ell_0 \neq 0$. Then:

$$u = -\ell_0/(2\lambda) = \text{const} \neq 0.$$

This control implies $x(1) > 0$ and hence u is *not feasible!*

Hence, it must hold $\ell_0 = 0$.

Global Minimum Principle

The local minimum principle can be generalized.

Let $\mathcal{U} \subseteq \mathbb{R}^{n_u}$ be measurable (e.g. non-convex, empty interior, discrete set). Then:

Global Minimum Principle

For almost every $t \in [t_0, t_f]$ it holds

$$\mathcal{H}(t,\hat{x}(t),\hat{u}(t),\lambda(t),\ell_0) \leq \mathcal{H}(t,\hat{x}(t),u,\lambda(t),\ell_0) \qquad \forall u \in \mathcal{U} \tag{54}$$

resp.

$$\hat{u}(t) = \arg\min_{u\in\mathcal{U}} \mathcal{H}(t,\hat{x}(t),u,\lambda(t),\ell_0). \tag{55}$$

The optimal $\hat{u}$ minimizes the Hamiltonian!

Special Case

The global minimum principle for $\mathcal{U} = \mathbb{R}^{n_u}$ implies

$$\begin{aligned} \mathcal{H}'_u(t, \hat{x}(t), \hat{u}(t), \lambda(t), \ell_0) &= 0, \\ \mathcal{H}''_{uu}(t, \hat{x}(t), \hat{u}(t), \lambda(t), \ell_0) & \text{ positive semi-definite.} \end{aligned}$$

If in addition the strengthened Legendre-Clebsch condition

$$\mathcal{H}''_{uu}(t, \hat{x}(t), \hat{u}(t), \lambda(t), \ell_0) \text{ positive definite}$$

holds, then $\mathcal{H}'_u = 0$ can be solved for $\hat{u}$ by the implicit function theorem:

$$\hat{u}(t) = u^*(t, \hat{x}(t), \lambda(t), \ell_0).$$

Global Minimum Principle and Linear Problems

We apply the global minimum principle to a linear optimal control problem:

Problem 7.5 (Linear Optimal Control Problem). *Minimize*

$$\eta^\top x(t_f)$$

subject to

$$\begin{aligned} \dot{x}(t) &= A(t)x(t) + B(t)u(t), && \text{a.e. in } [t_0, t_f], \\ x(t_0) &= x_0, \\ u(t) &\in \mathcal{U}, && \text{a.e. in } [t_0, t_f]. \end{aligned}$$

Global Minimum Principle and Linear Problems

Global minimum principle:

$$\lambda(t)^\top (A(t)\hat{x}(t) + B(t)\hat{u}(t)) \leq \lambda(t)^\top (A(t)\hat{x}(t) + B(t)u) \qquad \forall u \in \mathcal{U}.$$

Thus:

$$\lambda(t)^\top B(t)\hat{u}(t) \leq \lambda(t)^\top B(t)u \qquad \forall u \in \mathcal{U}.$$

Adjoint equation and transversality condition:

$$\dot{\lambda}(t)^\top = -\lambda(t)^\top A(t), \qquad \lambda(t_f) = \eta.$$

Notice: $\ell_0 = 0$ would lead to $\lambda(t_f) = 0$ and thus $\lambda \equiv 0$ contradicting $(\ell_0, \lambda) \neq 0$. Hence, $\ell_0 = 1$!

Global Minimum Principle and Linear Problems

For the linear optimal control problem 7.5 the minimum principle is also *sufficient*.

Given:

- fundamental system Φ with $\dot{\Phi} = A\Phi$, $\Phi(t_0) = I$
- feasible control $u(t) \in \mathcal{U}$ and corresponding state x:

$$x(t) = \Phi(t) \left(x_0 + \int_{t_0}^{t} \Phi(\tau)^{-1} B(\tau) u(\tau) d\tau \right)$$

Observation: It holds $\lambda(t)^\top \Phi(t) = \text{const} = \lambda(t_0)^\top$ because

$$\frac{d}{dt} \left(\lambda(t)^\top \Phi(t) \right) = \underbrace{\dot{\lambda}(t)^\top}_{= -\lambda(t)^\top A(t)} \Phi(t) + \lambda(t)^\top \underbrace{\dot{\Phi}(t)}_{= A(t)\Phi(t)} = 0.$$

Global Minimum Principle and Linear Problems

Now:

- Let $(\hat{x}, \hat{u})$ satisfy the *global minimum principle*.
- Let (x, u) be an arbitrary feasible trajectory.

Then:

$$\begin{aligned}
&\lambda(t)^\top x(t) - \lambda(t)^\top \hat{x}(t) \\
&= \underbrace{\lambda(t)^\top \Phi(t)}_{=\lambda(t_0)^\top} \left(\int_{t_0}^{t} \Phi(\tau)^{-1} B(\tau) \left(u(\tau) - \hat{u}(\tau)\right) d\tau \right) \\
&= \int_{t_0}^{t} \underbrace{\lambda(t_0)^\top \Phi(\tau)^{-1}}_{=\lambda(\tau)^\top} B(\tau) \left(u(\tau) - \hat{u}(\tau)\right) d\tau.
\end{aligned}$$

Global Minimum Principle and Linear Problems

At $t = t_f$ it holds $\lambda(t_f) = \eta$ and exploitation of the minimum principle yields

$$\eta^\top x(t_f) - \eta^\top \hat{x}(t_f) = \int_{t_0}^{t_f} \underbrace{\lambda(\tau)^\top B(\tau) u(\tau) - \lambda(\tau)^\top B(\tau) \hat{u}(\tau)}_{\geq 0} d\tau \geq 0.$$

Hence, $(\hat{x}, \hat{u})$ *is optimal!*

Minimum Principle for Control-State Constraints

Augmented Hamiltonian:

$$\hat{\mathcal{H}}(t, x, u, \lambda, \eta, \ell_0) := \mathcal{H}(t, x, u, \lambda, \ell_0) + \eta^\top c(t, x, u).$$

Theorem 7.6 (Local Minimum Principle for Control-State Constraints). *Assumptions:*

(i) *Let φ, f_0, f, c, ψ be continuous w.r.t. all arguments and continuously differentiable w.r.t. x and u.*

(ii) *Let $\mathcal{U} = \mathbb{R}^{n_u}$.*

(iii) *Let $(\hat{x}, \hat{u}) \in W^{1,\infty}([t_0, t_f], \mathbb{R}^{n_x}) \times L^\infty([t_0, t_f], \mathbb{R}^{n_u})$ be a (weak) local minimum.*

(iv) *rank $\left(c_u'(t, \hat{x}(t), \hat{u}(t))\right) = n_c$ a.e. in $[t_0, t_f]$*

(v) *no pure state constraints (49)!*

Minimum Principle for Control-State Constraints

Theorem 7.6 (continued). *Assertion:*

There exits multipliers $\ell_0 \in \mathbb{R}$, $\lambda \in W^{1,\infty}([t_0, t_f], \mathbb{R}^{n_x})$, $\eta \in L^\infty([t_0, t_f], \mathbb{R}^{n_c})$, $\sigma \in \mathbb{R}^{n_\psi}$ with:

(i) $\ell_0 \geq 0$, $(\ell_0, \sigma, \lambda, \eta) \neq \Theta$,

(ii) *Adjoint Differential Equation:*

$$\dot{\lambda}(t)^\top = -\hat{\mathcal{H}}_x'(t, \hat{x}(t), \hat{u}(t), \lambda(t), \eta(t), \ell_0)$$

Minimum Principle for Control-State Constraints

Theorem 7.6 (continued).

(iii) Transversality Conditions:

$$\begin{aligned}\lambda(t_0)^\top &= -\left(\ell_0\varphi'_{x_0} + \sigma^\top\psi'_{x_0}\right),\\ \lambda(t_f)^\top &= \ell_0\varphi'_{x_f} + \sigma^\top\psi'_{x_f}.\end{aligned}$$

(iv) Optimality Condition: A.e. in $[t_0,t_f]$ it holds

$$\hat{\mathcal{H}}'_u(t,\hat{x}(t),\hat{u}(t),\lambda(t),\eta(t),\ell_0) = 0_{n_u}.$$

(v) Complementarity Condition: A.e. in $[t_0,t_f]$ it holds

$$\eta(t)^\top c(t,\hat{x}(t),\hat{u}(t)) = 0, \qquad \eta(t) \geq 0_{n_c}.$$

Minimum Principle for State Constraints

Definition 7.7 (Functions of Bounded Variation (BV)). $\mu : [t_0,t_f] \to \mathbb{R}$ is called of bounded variation, if there exists a constant C such that for any partition

$$\mathbb{G} := \{t_0 < t_1 < \ldots < t_m = t_f\}$$

of $[t_0,t_f]$ it holds

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

$BV([t_0,t_f],\mathbb{R})$: space of functions of bounded variation

$NBV([t_0,t_f],\mathbb{R})$: space of normalized functions of bounded variation, i.e. $\mu(t_0) = 0$ and μ is continuous from the right in (t_0,t_f).

Functions of bounded variation are differentiable almost everywhere, except at at most countably many points.

Minimum Principle for State Constraints

Theorem 7.8 (Local Minimum Principle for State Constraints). *Assumptions:*

(i) Let φ, f_0, f, s, ψ be continuous w.r.t. all arguments and continuously differentiable w.r.t. x and u.

(ii) Let $\mathcal{U} \subseteq \mathbb{R}^{n_u}$ be a closed and convex set with int$(\mathcal{U}) \neq \emptyset$.

(iii) Let $(\hat{x},\hat{u}) \in W^{1,\infty}([t_0,t_f],\mathbb{R}^{n_x}) \times L^\infty([t_0,t_f],\mathbb{R}^{n_u})$ be a (weak) local minimum.

(i) no mixed control-state constraints (48)!

Minimum Principle for State Constrained Problems III

Theorem 7.8 (continued). *Assertion:*
There exist multipliers $\ell_0 \in \mathbb{R}$, $\lambda \in BV([t_0,t_f],\mathbb{R}^{n_x})$, $\mu \in NBV([t_0,t_f],\mathbb{R}^{n_s})$, and $\sigma \in \mathbb{R}^{n_\psi}$ with:

(i) $\ell_0 \geq 0$, $(\ell_0,\sigma,\lambda,\mu) \neq \Theta$,

(ii) Adjoint Equation:

$$\lambda(t) = \lambda(t_f) + \int_t^{t_f} \mathcal{H}'_x[\tau]^\top d\tau + \int_t^{t_f} s'_x[\tau]^\top d\mu(\tau), \qquad \forall t \in [t_0,t_f].$$

(The latter integral is a Riemann-Stieltjes integral!)

Abbreviation: $\mathcal{H}'_x[t] := \mathcal{H}'_x(t,\hat{x}(t),\hat{u}(t),\lambda(t),\ell_0)$, similar for $s'_x[t]$

Minimum Principle for State Constraints

Theorem 7.8 (continued).

(iii) Transversality Conditions:

$$\begin{aligned}\boldsymbol{\lambda}(t_0)^\top &= -\left(\ell_0\varphi'_{x_0} + \boldsymbol{\sigma}^\top\psi'_{x_0}\right),\\ \boldsymbol{\lambda}(t_f)^\top &= \ell_0\varphi'_{x_f} + \boldsymbol{\sigma}^\top\psi'_{x_f}.\end{aligned}$$

(iv) Optimality Condition: A.e. in $[t_0, t_f]$ it holds

$$\mathcal{H}'_u(t,\hat{x}(t),\hat{u}(t),\boldsymbol{\lambda}(t),\ell_0)(u-\hat{u}(t)) \geq 0 \qquad \forall u \in \mathcal{U}.$$

(v) Complementarity Condition: $\boldsymbol{\mu}_i$ is monotonically increasing on $[t_0, t_f]$ and constant on intervals (t_1, t_2), $t_1 < t_2$ with $s_i(t,\hat{x}(t)) < 0$.

Remarks

Remark 7.9.

- *There is also a global minimum principle for theorem 7.8, where the optimality condition is given by*

$$\hat{u}(t) = \arg\min_{u\in\mathcal{U}} \mathcal{H}(t,\hat{x}(t),u,\boldsymbol{\lambda}(t),\ell_0).$$

Remarks

Remark 7.9 (continued).

- *$\boldsymbol{\lambda}$ is differentiable almost everywhere in $[t_0, t_f]$ with*

$$\dot{\boldsymbol{\lambda}}(t)^\top = -\mathcal{H}'_x(t,\hat{x}(t),\hat{u}(t),\boldsymbol{\lambda}(t),\ell_0) - \dot{\boldsymbol{\mu}}(t)^\top s'_x(t,\hat{x}(t)).$$

At points $t_j \in (t_0, t_f)$ where $\boldsymbol{\mu}$ (and possibly $\boldsymbol{\lambda}$) is not differentiable, the jump-conditions hold:

$$\boldsymbol{\lambda}(t_j) - \boldsymbol{\lambda}(t_j^-) = -s'_x(t_j,\hat{x}(t_j))^\top\left(\boldsymbol{\mu}(t_j) - \boldsymbol{\mu}(t_j^-)\right)$$

and

$$\mathcal{H}[t_j] - \mathcal{H}[t_j^-] = s'_t(t_j,\hat{x}(t_j))^\top\left(\boldsymbol{\mu}(t_j) - \boldsymbol{\mu}(t_j^-)\right).$$

For Further Reading

References

[PBGM64] Pontryagin, L. S., Boltyanskij, V. G., Gamkrelidze, R. V., and Mishchenko, E. F. *Mathematische Theorie optimaler Prozesse*, Oldenbourg, München, 1964.

[Hes64] Hestenes, M. R. *Variational theory and optimal control theory*. Computational Methods in Optimization Problems, pp. 1–22. 1964.

[Hes66] Hestenes, M. R. *Calculus of variations and optimal control theory*. John Wiley & Sons, New York, 1966.

[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.

[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.

[Neu76] Neustadt, L. W. *Optimization: A Theory of Necessary Conditions*. Princeton, New Jersey, 1976.

[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.

[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.

[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

[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.

7.2 Indirect Methods and Boundary Value Problems

Examples for Boundary Value Problems I

Necessary conditions for variational problems or optimal control problems lead to boundary value problems.

Example 7.10 *(Variational Problems).* Variational problem:

$$\min \int_a^b f(t,x(t),x'(t))dt \qquad \text{s.t.} \qquad x(a) = x_a,\ x(b) = x_b$$

Necessary for optimality: Euler-Lagrange differential equation ($f''_{x'x'} \neq 0$):

$$\begin{aligned} f'_x[t] - f''_{x't}[t] - f''_{x'x}[t]x'(t) - f''_{x'x'}[t]x''(t) &= 0, \\ x(a) = x_a, \qquad & x(b) = x_b, \end{aligned}$$

Two-point boundary value problem!

Examples for Boundary Value Problems II

Example 7.11 *(Optimal Control Problems without State Constraints).* Local minimum principle: control law $u = u^*(t,x,\lambda)$ from $\mathcal{H}'_u = 0$ yields

$$\begin{aligned} \dot{x}(t) &= f(t,x(t),u^*(t,x(t),\lambda(t))), \\ \dot{\lambda}(t) &= -\mathcal{H}'_x(t,x(t),u^*(t,x(t),\lambda(t)),\lambda(t),\ell_0)^\top, \\ 0_{n_\psi} &= \psi(x(t_0),x(t_f)), \\ \lambda(t_0)^\top &= -\left(\ell_0\varphi'_{x_0}(x(t_0),x(t_f)) + \sigma^\top\psi'_{x_0}(x(t_0),x(t_f))\right), \\ \lambda(t_f)^\top &= \ell_0\varphi'_{x_f}(x(t_0),x(t_f)) + \sigma^\top\psi'_{x_f}(x(t_0),x(t_f)). \end{aligned}$$

Two-point boundary value problem!

2-Point Boundary Value Problem

The boundary value problems formally are of the subsequent form:

Problem 7.12 (Two-point Boundary Value Problem (BVP)). *For given functions $g : [a,b] \times \mathbb{R}^{n_y} \to \mathbb{R}^{n_y}$ and $r : \mathbb{R}^{n_y} \times \mathbb{R}^{n_y} \to \mathbb{R}^{n_y}$ find a solution y of the boundary value problem*

$$\begin{aligned} y'(t) &= g(t,y(t)), \\ r(y(a),y(b)) &= 0_{n_y} \end{aligned}$$

in the interval $[a,b]$.

Multiple-Point Boundary Value Problems I

In case of

- broken extremals (i.e. continuous, but not differentiable functions) for variational problems or
- switching conditions or pure state constraints for optimal control problems

the necessary conditions lead to

multiple-point boundary value problems

with conditions at

intermediate points τ_j, $j = 1, \ldots, k$.

Multiple-Point Boundary Value Problems II

Problem 7.13 (Multiple-point BVP). *For given functions $g_1, \ldots, g_m : [a,b] \times \mathbb{R}^{n_y} \to \mathbb{R}^{n_y}$ and $r_j : [a,b] \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_y} \to \mathbb{R}^{k_j}$, $j = 1, \ldots, m+1$ find a function y and switching points $a < \tau_1 < \ldots < \tau_m < b$ with*

$$y'(t) = \begin{cases} g_1(t, y(t)), & \text{if } a \le t < \tau_1, \\ g_j(t, y(t)), & \text{if } \tau_j \le t < \tau_{j+1},\ j = 1, \ldots, m-1, \\ g_m(t, y(t)), & \text{if } \tau_m \le t \le b \end{cases}$$

subject to intermediate transition conditions

$$r_j(\tau_j, y(\tau_j^+), y(\tau_j^-)) = 0_{k_j}, \qquad j = 1, \ldots, m,$$

and boundary condtions

$$r_{m+1}(y(a), y(b)) = 0_{k_{m+1}}.$$

Remarks

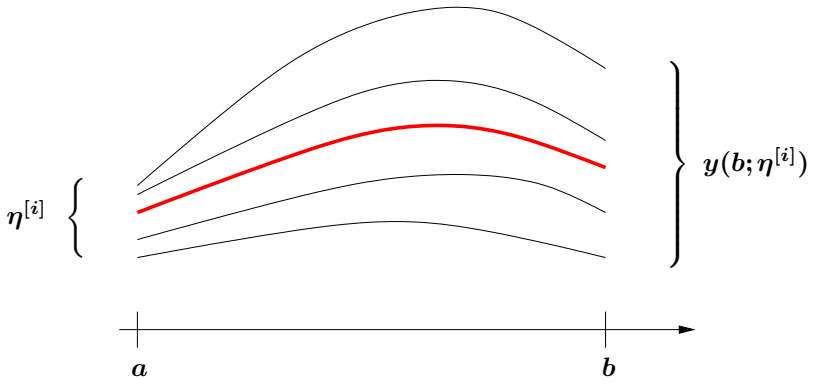
Remark 7.14.

- *By appropriate transformation techniques the multiple-point BVP can be transformed to an equivalent two-point BVP. Thus, it suffices to consider only the latter.*
- *Existence and uniqueness results for BVP's can be found in Ascher et al. [AMR95], Chapter 3. A BVP can have none, exactly one, or infinitely many solutions.*

7.2.1 Single Shooting

Idea of Single Shooting

We consider the two-point BVP 7.12.
Idea of the single shooting method:



Motivation

Given: initial guess η of initial value $y(a)$

Solution: $y(t;\eta)$ of IVP

$$y'(t) = g(t, y(t)), \qquad y(a) = \eta.$$

Restriction: boundary condition

$$F(\eta) := r(y(a;\eta), y(b;\eta)) = r(\eta, y(b;\eta)) = 0_{n_y} \tag{56}$$

Equation (56) is a nonlinear equation.

Single Shooting Algorithm

Application of Newton's method yields the single shooting method:

Algorithm: Single Shooting Method

(0) Choose initial guess $\eta^{[0]} \in \mathbb{R}^{n_y}$ and set $i = 0$.

(1) Solve IVP

$$y'(t) = g(t, y(t)), \qquad y(a) = \eta^{[i]}, \qquad a \le t \le b,$$

compute $F(\eta^{[i]})$, and calculate the Jacobian

$$F'(\eta^{[i]}) = r'_{y_a}(\eta^{[i]}, y(b;\eta^{[i]})) + r'_{y_b}(\eta^{[i]}, y(b;\eta^{[i]})) \cdot S(b),$$

Single Shooting Algorithm

where S solves the sensitivity differential equation

$$S'(t) = g'_y(t, y(t;\eta^{[i]})) \cdot S(t), \qquad S(a) = I_{n_y \times n_y}.$$

(2) If $F(\eta^{[i]}) = 0_{n_y}$ (or some more sophisticated stopping criterion), stop with success.

(3) Compute Newton-direction $d^{[i]}$ as solution of the linear equation

$$F'(\eta^{[i]})d = -F(\eta^{[i]}).$$

(4) Set $\eta^{[i+1]} = \eta^{[i]} + d^{[i]}$ and $i = i + 1$ and go to (1).

Remarks

Remark 7.15. *The Jacobian $F'(\eta^{[i]})$ in step (2) of the single shooting method can be approximated alternatively by finite differences:*

$$\frac{\partial}{\partial \eta_j} F(\eta) \approx \frac{F(\eta + he_j) - F(\eta)}{h}, \qquad j = 1, \ldots, n_y,$$

$e_j = j^{th}$ unity vector. This approach requires to solve n_y nonlinear IVP's! The sensitivity differential equation approach requires to solve n_y linear IVP's.

Convergence of Single Shooting

The single shooting method essentially is Newton's method. $\Rightarrow$ convergence results for Newton's method remain valid (locally superlinear, locally quadratic convergence)

Jacobian $F'(\eta^{[i]})$ is non-singular if

$$r'_{y_a}(\eta^{[i]}, y(b;\eta^{[i]})) \cdot S(a) + r'_{y_b}(\eta^{[i]}, y(b;\eta^{[i]})) \cdot S(b)$$

is non-singular.

Example

Example 7.16. Consider the subsequent optimal control problem: Minimize

$$\frac{5}{2}(x(1)-1)^2+\frac{1}{2}\int_0^1 u(t)^2+x(t)^3 dt$$

subject to

$$\dot{x}(t)=u(t)-r(t), \qquad x(0)=4$$

with $r(t)=15\exp(-2t)$.

Example

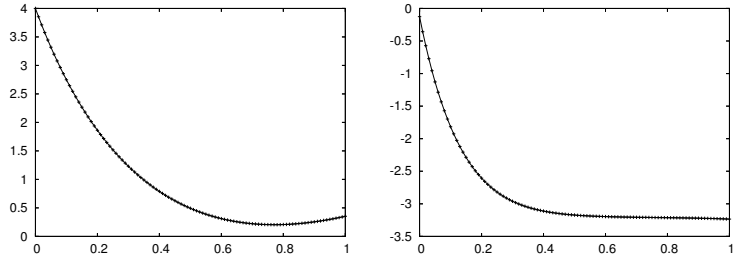
Example 7.16 *(continued).* The minimum principle results in the BVP

$$\begin{aligned} \dot{x}(t) &= -\lambda(t)-r(t), \\ \dot{\lambda}(t) &= -\frac{3}{2}x(t)^2, \\ x(0)-4 &= 0, \\ \lambda(1)-5(x(1)-1) &= 0. \end{aligned}$$

Example

Example 7.16 *(continued).* We apply the single shooting method with initial guess $\eta^{[0]}=(4,-5)^\top$ and obtain the following solution:

State x: Adjoint $\lambda=-u$:



7.2.2 Multiple Shooting

Motivation

Problem with single shooting:

Applicability and stability: Estimate

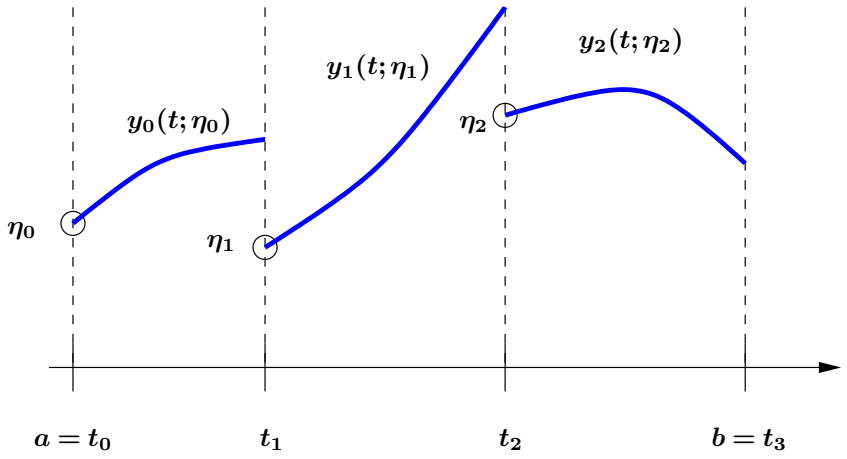
$$\|y(t;\eta_1)-y(t;\eta_2)\| \le \|\eta_1-\eta_2\|\exp(L(t-a)).$$

Idea: Multiple shooting by introducing additional multiple shooting nodes

$$a=t_0<t_1<\cdots<t_{N-1}<t_N=b.$$

Number and position of multiple shooting nodes depend on particular example!

Multiple Shooting Idea



Multiple Shooting Method I

In every interval $[t_j, t_{j+1})$, $j = 0, \ldots, N-1$ starting at η_j solve the IVP

$$y'(t) = g(t, y(t)), \qquad y(t_j) = \eta_j.$$

Solution in $[t_j, t_{j+1})$: $y_j(t;\eta_j)$

Piecewise defined function:

$$y(t;\eta_0,\ldots,\eta_{N-1}) := \begin{cases} y_j(t;\eta_j), & \text{if } t_j \le t < t_{j+1},\ j = 0,\ldots,N-1, \\ y_{N-1}(t_N;\eta_{N-1}), & \text{if } t = b \end{cases}$$

Multiple Shooting Method II

Continuity and boundary conditions:

$$F(\eta) := F(\eta_0,\ldots,\eta_{N-1}) := \begin{pmatrix} y_0(t_1;\eta_0) - \eta_1 \\ y_1(t_2;\eta_1) - \eta_2 \\ \vdots \\ y_{N-2}(t_{N-1};\eta_{N-2}) - \eta_{N-1} \\ r(\eta_0, y_{N-1}(t_N;\eta_{N-1})) \end{pmatrix} = 0_{N\cdot n_y}. \tag{57}$$

Equation (57) again is a nonlinear equation for the variables $\eta := (\eta_0,\ldots,\eta_{N-1})^\top \in \mathbb{R}^{N\cdot n_y}$.
Special case $N = 1$: single shooting

Sparsity

Structure of Jacobian: sparse!

$$F'(\eta) = \begin{pmatrix} S_0 & -I & & & \\ & S_1 & -I & & \\ & & \ddots & \ddots & \\ & & & S_{N-2} & -I \\ A & & & & B\cdot S_{N-1} \end{pmatrix} \tag{58}$$

with

$$\begin{aligned} S_j &:= \frac{\partial}{\partial \eta_j} y_j(t_{j+1};\eta_j), \\ A &:= r'_{y_a}(\eta_0, y_{N-1}(t_N;\eta_{N-1})), \\ B &:= r'_{y_b}(\eta_0, y_{N-1}(t_N;\eta_{N-1})). \end{aligned}$$

Multiple Shooting Algorithm

Application of Newton's method yields the following algorithm.

Algorithm: Multiple Shooting

(0) Choose initial guess $\eta^{[0]} = (\eta_0^{[0]}, \dots, \eta_{N-1}^{[0]})^\top \in \mathbb{R}^{N \cdot n_y}$ and set $i = 0$.

(1) For $j = 0, \dots, N-1$ solve the IVP's

$$y'(t) = g(t, y(t)), \qquad y(t_j) = \eta_j^{[i]}, \qquad t_j \le t \le t_{j+1},$$

compute $F(\eta^{[i]})$ and compute the sensitivity matrices $S_j = S(t_{j+1})$, where S solves the sensitivity differential equation

Multiple Shooting Algorithm

$$S'(t) = g_y'(t, y(t; \eta_j^{[i]})) \cdot S(t), \qquad S(t_j) = I_{n_y \times n_y}, \quad t_j \le t \le t_{j+1}.$$

Compute $F'(\eta^{[i]})$ according to (58).

(2) If $F(\eta^{[i]}) = 0_{N \cdot n_y}$ (or some more sophisticated stopping criterion), stop with success.

(3) Compute the Newton-direction $d^{[i]}$ as solution of the linear equation

$$F'(\eta^{[i]}) d = -F(\eta^{[i]}).$$

(4) Set $\eta^{[i+1]} = \eta^{[i]} + d^{[i]}$ and $i = i + 1$ and go to (1).

Remarks

Remark 7.17. *Newton's method can be globalized by introducing a step size α_i:*

$$x^{[i+1]} = x^{[i]} + \alpha_i d^{[i]}, \qquad i = 0, 1, 2, \dots.$$

The step size α_i may be computed using a line search for the function

$$\varphi(\alpha) := \frac{1}{2} \| F(x^{[i]} + \alpha d^{[i]}) \|_2^2$$

by, e.g. Armijo's rule.

Remarks

Remark 7.18. *The exact Jacobian $F'(\eta^{[i]})$, which is very expensive to evaluate ($\rightarrow$ sensitivity differential equation), can be replaced by a so-called update formula known from Quasi-Newton methods from unconstrained optimization. For instance the rank-1-update formula of Broyden may be used:*

$$J_+ = J + \frac{(z - Jd) d^\top}{d^\top d}, \qquad d = \eta^+ - \eta, \quad z = F(\eta^+) - F(\eta).$$

Remarks

Remark 7.19. *A difficult task from numerical point of view is the determination of a sufficiently good initial guess for the solution of the BVP. Especially if the BVP results from the minimum principle for optimal control problems, it is important to provide a good estimate of the switching structure (i.e. number and position of active, inactive, and singular subarcs) as well as for the state and adjoints at the multiple shooting nodes.*

Unfortunately, there is no general way to do this. Possible approaches are homotopy methods (i.e. easy to solve neighboring problems are solved and their solutions serve as initial guess) or direct discretization methods, which will be discussed later on.

For Further Reading

References

[1] Uri M. Ascher, Robert M.M. Mattheij, and Robert D. Russell. *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations*, volume 13 of *Classics In Applied Mathematics*. SIAM, Philadelphia, 1995.

7.3 Direct Discretization Methods

Direct Discretization Methods

Problem 7.20 (Optimal Control Problem). *For fixed time points $t_0 < t_f$ minimize*

$$\varphi(x(t_0), x(t_f))$$

subject to

$$\begin{aligned} \dot{x}(t) &= f(t, x(t), u(t)), \\ 0_{n_\psi} &= \psi(x(t_0), x(t_f)), \\ c(t, x(t), u(t)) &\leq 0_{n_c}, \\ s(t, x(t)) &\leq 0_{n_s}, \\ u(t) &\in \mathcal{U} := \{u \in \mathbb{R}^{n_u} \mid u_{min} \leq u \leq u_{max}\}. \end{aligned}$$

Idea

General structure of discretization methods:

- Parametrization of the control: The control u is replaced by a function u_h that depends on a finite number of parameters, e.g. a piecewise constant function on some grid. The index h indicates the dependence on some discretization parameter, e.g. the step size of a grid.
- Discretization method for the differential equation: The differential equation $\dot{x}(t) = f(t, x(t), u(t))$ is discretized by some discretization method, e.g. an one-step method.
- Optimizer: The resulting finite dimensional optimization problem after discretization is solved by a suitable optimization routine, e.g. SQP methods.

7.3.1 Euler Discretization

Full Discretization by Explicit Euler's Method

Grid

$$\mathbb{G}_h := \{t_0 < t_1 < \ldots < t_N = t_f\},$$

step sizes $h_i = t_{i+1} - t_i$, $i = 0, \ldots, N-1$, $h := \max_{i=0,\ldots,N-1} h_i$.

Idea:

- Discretize the differential equation using the explicit Euler's method
- feasibility of constraints only on grid

Full Discretization by Explicit Euler's Method

Problem 7.21 (Full Discretization by Explicit Euler's Method). *Find $x_h : \mathbb{G}_h \to \mathbb{R}^{n_x}$, $t_i \mapsto x_h(t_i) =: x_i$, $u_h : \mathbb{G}_h \to \mathbb{R}^{n_u}$, $t_i \mapsto u_h(t_i) =: u_i$ such that*

$$\varphi(x_0, x_N)$$

is minimized subject to

$$\begin{array}{rcll} x_{i+1} & = & x_i + h_i f(t_i, x_i, u_i), & i = 0, 1, \ldots, N-1, \\ \psi(x_0, x_N) & = & 0_{n_\psi}, & \\ c(t_i, x_i, u_i) & \leq & 0_{n_c}, & i = 0, 1, \ldots, N, \\ s(t_i, x_i) & \leq & 0_{n_s}, & i = 0, 1, \ldots, N, \\ u_i & \in & [u_{min}, u_{max}], & i = 0, 1, \ldots, N. \end{array}$$

Full Discretization by Explicit Euler's Method

Finite dimensional nonlinear optimization problem

$$\begin{array}{ll} \min\limits_{z} & F(z) \\ \text{s.t.} & z \in S, \quad G(z) \leq \Theta, \quad H(z) = \Theta \end{array}$$

$(n_x + n_u) \cdot (N+1)$ optimization variables

$$z := (x_0, x_1, \ldots, x_N, u_0, u_1, \ldots, u_N)^\top$$

Objective functional

$$F(z) := \varphi(x_0, x_N)$$

Full Discretization by Explicit Euler's Method

$(n_c + n_s) \cdot (N+1)$ inequality constraints

$$G(z) := \begin{pmatrix} c(t_0, x_0, u_0) \\ \vdots \\ c(t_N, x_N, u_N) \\ s(t_0, x_0) \\ \vdots \\ s(t_N, x_N) \end{pmatrix}$$

Full Discretization by Explicit Euler's Method

$n_x \cdot N + n_\psi$ equality constraints

$$H(z) := \begin{pmatrix} x_0 + h_0 f(t_0, x_0, u_0) - x_1 \\ \vdots \\ x_{N-1} + h_{N-1} f(t_{N-1}, x_{N-1}, u_{N-1}) - x_N \\ \psi(x_0, x_N) \end{pmatrix}$$

Set constraints

$$S := \mathbb{R}^{n_x \cdot (N+1)} \times [u_{min}, u_{max}]^{N+1}$$

Full Discretization by Explicit Euler's Method

Advantage:

sparse structure (exploitation possible)

Disadvantage:

large scale

Full Discretization by Explicit Euler's Method

Structure of optimization problem:

$$F'(z) \quad = \quad \left(\varphi'_{x_0} \,\middle|\, \Theta \,\middle|\, \cdots \,\middle|\, \Theta \,\middle|\, \varphi'_{x_f} \,\middle|\, \Theta \,\middle|\, \cdots \,\middle|\, \Theta\right)$$

Full Discretization by Explicit Euler's Method

$$G'(z) \quad = \quad \left(\begin{array}{ccc|ccc} c'_x[t_0] & & & c'_u[t_0] & & \\ & \ddots & & & \ddots & \\ & & c'_x[t_N] & & & c'_u[t_N] \\ \hline s'_x[t_0] & & & & & \\ & \ddots & & & \Theta & \\ & & s'_x[t_N] & & & \end{array}\right)$$

Full Discretization by Explicit Euler's Method

$$H'(z) \quad = \quad \left(\begin{array}{cccc|ccc} M_0 & -I_{n_x} & & & h_0 f'_u[t_0] & & \\ & \ddots & & & & \ddots & \\ & & M_{N-1} & -I_{n_x} & & & h_{N-1} f'_u[t_{N-1}] \\ \hline \psi'_{x_0} & & & \psi'_{x_f} & & \Theta & \end{array}\right)$$

with

$$M_i := I_{n_x} + h_i f'_x(t_i, x_i, u_i), \qquad i = 0, \ldots, N-1$$

Reduced Discretization by Explicit Euler's Method

Idea: eliminate the difference equations!

$$\begin{array}{rcl} x_0 & =: & X_0(x_0), \\ x_1 & = & x_0 + h_0 f(t_0, x_0, u_0) \\ & = & X_0(x_0) + h_0 f(t_0, X_0(x_0), u_0) \\ & =: & X_1(x_0, u_0), \\ x_2 & = & x_1 + h_1 f(t_1, x_1, u_1) \\ & = & X_1(x_0, u_0) + h_1 f(t_1, X_1(x_0, u_0), u_1) \\ & =: & X_2(x_0, u_0, u_1), \\ & \vdots & \\ x_N & = & x_{N-1} + h_{N-1} f(t_{N-1}, x_{N-1}, u_{N-1}) \\ & = & X_{N-1}(x_0, u_0, \ldots, u_{N-2}) + h_{N-1} f(t_{N-1}, X_{N-1}(x_0, u_0, \ldots, u_{N-2}), u_{N-1}) \\ & =: & X_{N-1}(x_0, u_0, \ldots, u_{N-1}). \end{array}$$

Reduced Discretization by Explicit Euler's Method

Problem 7.22 (Reduced Discretization by Explicit Euler's Method). *Find an initial value $x_0 \in \mathbb{R}^{n_x}$ and a grid function $u_h : \mathbb{G}_h \to \mathbb{R}^{n_u}$, $t_i \mapsto u_h(t_i) =: u_i$, such that*

$$\varphi(x_0, X_N(x_0, u_0, \ldots, u_{N-1}))$$

is minimized subject to

$$\begin{array}{rcll} \psi(x_0, X_N(x_0, u_0, \ldots, u_{N-1})) & = & 0_{n_\psi}, & \\ c(t_i, X_i(x_0, u_0, \ldots, u_{i-1}), u_i) & \leq & 0_{n_c}, & i = 0, 1, \ldots, N, \\ s(t_i, X_i(x_0, u_0, \ldots, u_{i-1})) & \leq & 0_{n_s}, & i = 0, 1, \ldots, N, \\ u_i & \in & [u_{min}, u_{max}], & i = 0, 1, \ldots, N. \end{array}$$

Reduced Discretization by Explicit Euler's Method

Finite dimensional nonlinear optimization problem

$$\begin{array}{ll} \min\limits_{z} & \boldsymbol{F}(z) \\ \text{s.t.} & z \in \boldsymbol{S}, \quad \boldsymbol{G}(z) \leq \boldsymbol{\Theta}, \quad \boldsymbol{H}(z) = \boldsymbol{\Theta} \end{array} \tag{59}$$

$\boldsymbol{n_x + n_u \cdot (N+1)}$ optimization variables

$$z := (x_0, u_0, u_1, \ldots, u_N)^\top$$

Objective functional

$$\boldsymbol{F}(z) := \varphi(x_0, \boldsymbol{X}_N(x_0, u_0, \ldots, u_{N-1}))$$

Reduced Discretization by Explicit Euler's Method

$\boldsymbol{(n_c + n_s) \cdot (N+1)}$ inequality constraints

$$\boldsymbol{G}(z) := \begin{pmatrix} \boldsymbol{c}(t_0, x_0, u_0) \\ \vdots \\ \boldsymbol{c}(t_N, \boldsymbol{X}_N(x_0, u_0, \ldots, u_{N-1}), u_N) \\ \boldsymbol{s}(t_0, x_0) \\ \vdots \\ \boldsymbol{s}(t_N, \boldsymbol{X}_N(x_0, u_0, \ldots, u_{N-1})) \end{pmatrix}$$

Reduced Discretization by Explicit Euler's Method

$\boldsymbol{n_\psi}$ equality constraints

$$\boldsymbol{H}(z) := \left(\; \psi(x_0, \boldsymbol{X}_N(x_0, u_0, \ldots, u_{N-1})) \; \right)$$

Set constraints

$$\boldsymbol{S} := \mathbb{R}^{n_x} \times [u_{min}, u_{max}]^{N+1}$$

Reduced Discretization by Explicit Euler's Method

Advantage:

low dimension compared to full discretization approach

Disadvantage:

dense (exploitation of structure not possible)

Reduced Discretization by Explicit Euler's Method

Structue of optimization problem:

$$\boldsymbol{G}'(z) \quad = \quad \left(\begin{array}{c|ccc} \boldsymbol{c}'_x[t_0] \cdot \frac{\partial x_0}{\partial x_0} & \boldsymbol{c}'_u[t_0] & & \\ \boldsymbol{c}'_x[t_1] \cdot \frac{\partial \boldsymbol{X}_1}{\partial x_0} & \boldsymbol{c}'_x[t_1] \cdot \frac{\partial \boldsymbol{X}_1}{\partial u_0} & \boldsymbol{c}'_u[t_1] & \\ \vdots & \vdots & \ddots & \ddots \\ \boldsymbol{c}'_x[t_N] \cdot \frac{\partial \boldsymbol{X}_N}{\partial x_0} & \boldsymbol{c}'_x[t_N] \cdot \frac{\partial \boldsymbol{X}_N}{\partial u_0} & \cdots & \boldsymbol{c}'_x[t_N] \cdot \frac{\partial \boldsymbol{X}_N}{\partial u_{N-1}} \quad \boldsymbol{c}'_u[t_N] \\ \hline \boldsymbol{s}'_x[t_0] \cdot \frac{\partial x_0}{\partial x_0} & & & \\ \boldsymbol{s}'_x[t_1] \cdot \frac{\partial \boldsymbol{X}_1}{\partial x_0} & \boldsymbol{s}'_x[t_1] \cdot \frac{\partial \boldsymbol{X}_1}{\partial u_0} & & \\ \vdots & \vdots & \ddots & \\ \boldsymbol{s}'_x[t_N] \cdot \frac{\partial \boldsymbol{X}_N}{\partial x_0} & \boldsymbol{s}'_x[t_N] \cdot \frac{\partial \boldsymbol{X}_N}{\partial u_0} & \cdots & \boldsymbol{s}'_x[t_N] \cdot \frac{\partial \boldsymbol{X}_N}{\partial u_{N-1}} \end{array} \right)$$

Reduced Discretization by Explicit Euler's Method

$$H'(z) \quad = \quad \left(\psi'_{x_0} + \psi'_{x_f} \cdot \frac{\partial X_N}{\partial x_0} \;\middle|\; \psi'_{x_f} \cdot \frac{\partial X_N}{\partial u_0} \;\middle|\; \cdots \;\middle|\; \psi'_{x_f} \cdot \frac{\partial X_N}{\partial u_{N-1}} \;\middle|\; \Theta \right)$$

Reduced Discretization by Explicit Euler's Method

Required: sensitivities

$$\frac{\partial X_i(x_0, u_0, \ldots, u_{i-1})}{\partial x_0}, \quad \frac{\partial X_i(x_0, u_0, \ldots, u_{i-1})}{\partial u_j}, \quad i, j = 0, \ldots, N$$

$\Rightarrow$ sensitivity analysis

7.4 Necessary Conditions and SQP Methods

Nonlinear Optimization Problem

Constinuously differentiable functions:

$$\begin{aligned} F &: \mathbb{R}^n \to \mathbb{R}, \\ G = (G_1, \ldots, G_m)^\top &: \mathbb{R}^n \to \mathbb{R}^m, \\ H = (H_1, \ldots, H_p)^\top &: \mathbb{R}^n \to \mathbb{R}^p \end{aligned}$$

Set

$$S := \{z \in \mathbb{R}^n \mid \underline{z} \le z \le \overline{z}\}, \qquad \underline{z} < \overline{z},\ \underline{z}, \overline{z} \in (\mathbb{R} \cup \{-\infty, \infty\})^n$$

Problem 7.23 (Optimization Problem). *Find $z \in \mathbb{R}^n$, such that $F(z)$ is minimized subject to the constraints*

$$\begin{aligned} G_i(z) &\le 0, \quad i = 1, \ldots, m, \\ H_j(z) &= 0, \quad j = 1, \ldots, p, \\ z &\in S. \end{aligned}$$

Definitions

Feasible set:

$$\Sigma := \{z \in S \mid G_i(z) \le 0,\ i = 1, \ldots, m,\ H_j(z) = 0,\ j = 1, \ldots, p\}.$$

Index set of active inequality constraints at z:

$$A(z) := \{i \mid G_i(z) = 0,\ 1 \le i \le m\}$$

Lagrangian:

$$L(z, \ell_0, \mu, \lambda) = \ell_0 F(z) + \mu^\top G(z) + \lambda^\top H(z)$$

7.4.1 Necessary Optimality Conditions

Fritz-John Conditions

The following first order necessary optimality conditions are due to Fritz John.

Theorem 7.24 (Fritz-John conditions). *Let $\hat{z}$ be a local minimum of problem 7.23. Let F, G_i, $i = 1, \ldots, m$, and H_j, $j = 1, \ldots, p$ be continuously differentiable. Then, there exist multipliers $\ell_0 \in \mathbb{R}$, $\mu = (\mu_1, \ldots, \mu_m)^\top \in \mathbb{R}^m$, $\lambda = (\lambda_1, \ldots, \lambda_p)^\top \in \mathbb{R}^p$ with $(\ell_0, \mu, \lambda) \neq 0$ and*

(a) Sign conditions:

$$\ell_0 \ge 0, \qquad \mu_i \ge 0, \qquad i = 1, \ldots, m. \tag{60}$$

Fritz-John Conditions

Theorem 7.24 (continued).

(b) Optimality condition:

$$L'_z(\hat{z}, \ell_0, \mu, \lambda)(z - \hat{z}) \geq 0 \qquad \forall z \in S. \tag{61}$$

(c) Complementarity condition:

$$\mu_i G_i(\hat{z}) = 0, \qquad i = 1, \ldots, m. \tag{62}$$

(d) Feasibility:

$$\hat{z} \in \Sigma. \tag{63}$$

Karush-Kuhn-Tucker (KKT) Conditions and Regularity Conditions

Regularity condition of Mangasarian-Fromowitz at $\hat{z}$:

(a) The gradients $\nabla H_j(\hat{z})$, $j = 1, \ldots, p$ are linearly independent.

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

$$\nabla G_i(\hat{z})^\top \hat{d} < 0 \text{ for } i \in A(\hat{z}),$$
$$\nabla H_j(\hat{z})^\top \hat{d} = 0 \text{ for } j = 1, \ldots, p.$$

Karush-Kuhn-Tucker (KKT) Conditions and Regularity Conditions

Linear Independence Constraint Qualification (LICQ) at $\hat{z}$:

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

(b) The gradients

$$\nabla G_i(\hat{z}),\ i \in A(\hat{z}), \qquad \nabla H_j(\hat{z}),\ j = 1, \ldots, p$$

are linearly independent.

LICQ guarantees uniqueness of Lagrange-multipliers!

Karush-Kuhn-Tucker (KKT) Conditions and Regularity Conditions

Theorem 7.25 (KKT Conditions). *If either the Mangasarian-Fromowitz condition or the Linear Independence Constraint Qualification is satisfied at a local minimum $\hat{z}$, then the Fritz-John conditions are satisfied with $\ell_0 = 1$.*

(x, λ, μ) is called KKT point, if it satisfies the necessary Fritz-John conditions with $\ell_0 = 1$.

7.4.2 Sequential Quadratic Programming (SQP)

Numerical Solution by SQP

Iteration:

$$z^{[k+1]} = z^{[k]} + \alpha_k d^{[k]}, \qquad k = 0, 1, 2, \ldots$$

Finding Search Direction

Determination of search direction $d^{[k]}$: Solve

Quadratic Program (QP)

$$\begin{aligned} \min_{d \in \mathbb{R}^n} \quad & \frac{1}{2} d^\top B_k d + F'(z^{[k]}) d \\ \text{s.t.} \quad & \\ & G(z^{[k]}) + G'(z^{[k]}) d \leq 0_m, \\ & H(z^{[k]}) + H'(z^{[k]}) d = 0_p, \\ & \underline{z} \leq z^{[k]} + d \leq \overline{z}. \end{aligned}$$

B_k: suitable update matrix (modified BFGS-update-formula, cf. Powell [Pow78])

Remark

Remark 7.26. *Instead of using an update matrix for B_k, it is also possible to use the Hessian of the Lagrangian, i.e. $B_k = L''_{zz}(z^{[k]}, \ell_0, \mu^{[k]}, \lambda^{[k]})$. But, the Hessian may be indefinite. In this case, QP is not convex anymore and becomes more difficult to solve.*

SQP Algorithm

Algorithm: Local SQP Method

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

(ii) If $(z^{[k]}, \mu^{[k]}, \lambda^{[k]})$ is a KKT point, stop with success.

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

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

SQP Algorithm

Remark 7.27.

- *The quadratic program can be solved by, e.g., an active set method, cf. Goldfarb and Idnani [GI83], Gill et al. [GM78, GMSW91].*
- *The iterates $z^{[k]}$ in general are infeasible for the nonlinear program.*
- *locally superlinear convergence*

Globalization of SQP

Exact ℓ_1-merit function:

$$\ell_1(z; \eta) := F(z) + \eta \left(\sum_{i=1}^m \max\{0, G_i(z)\} + \sum_{j=1}^p |H_j(z)| \right), \qquad \eta > 0$$

Penalty terms:

$$z \notin \Sigma \qquad \Rightarrow \qquad \sum_{i=1}^m \max\{0, G_i(z)\} + \sum_{j=1}^p |H_j(z)| > 0$$

Required: direction of descent $d^{[k]}$ with

$$\ell_1'(z^{[k]}; d^{[k]}; \eta) < 0$$

Globalization of SQP

One dimensional line search:

$$\min_{\alpha>0} \psi(\alpha) := \ell_1(z^{[k]} + \alpha d^{[k]}; \eta)$$

Choice of η: Direction of descent, if

$$\eta \geq \max\{\mu_1^{[k+1]}, \ldots, \mu_m^{[k+1]}, |\lambda_1^{[k+1]}|, \ldots, |\lambda_p^{[k+1]}|\}. \tag{64}$$

$\mu^{[k+1]}$, $\lambda^{[k+1]}$: Lagrange multipliers of quadratic program

Iterative adaption of η:

$$\eta_{k+1} := \max\{\eta_k, \max\{\mu_1^{[k+1]}, \ldots, \mu_m^{[k+1]}, |\lambda_1^{[k+1]}|, \ldots, |\lambda_p^{[k+1]}|\} + \varepsilon\}, \tag{65}$$

Global SQP Algorithm

Algorithmus: Globalized SQP Method

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

(ii) If $(z^{[k]}, \mu^{[k]}, \lambda^{[k]})$ is a KKT point, stop with success.

(iii) Quadratic program: Compute a KKT point $(d^{[k]}, \mu^{[k+1]}, \lambda^{[k+1]}) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p$ of the quadratic program.

(iv) Adapt η according to (65).

Global SQP Algorithm

(v) Armijo's rule: Determine the step size $\alpha_k = \max\{\beta^j \mid j = 0, 1, 2, \ldots\}$ with

$$\ell_1(z^{[k]} + \alpha_k d^{[k]}; \eta) \leq \ell_1(z^{[k]}; \eta) + \sigma\alpha_k \ell_1'(z^{[k]}; d^{[k]}; \eta).$$

(vi) Modified BFGS update: Compute B_{k+1} according to some update rule.

(vii) Set $z^{[k+1]} := z^{[k]} + \alpha_k d^{[k]}$, $k := k + 1$, and go to (ii).

Remarks

Remark 7.28.

- *Under appropriate assumptions, the global SQP method turns into the local SQP method after a finite number of steps. This means that the step length $\alpha_k = 1$ is accepted by Armijo's rule in this case. As mentioned above, the local method converges at least at a superlinear rate and hence, the global method does so as well.*
- *In praxis, it often happens that the quadratic program is infeasible. Powell [Pow78] overcame this remedy by relaxing the constraints of the quadratic program in such a way that the relaxed problem becomes feasible.*

For Further Reading

References

[GMW81] Gill, P. E., Murray, W. and Wright, M. H. *Practical Optimization*. Academic Press, London, 1981.

[Pol97] Polak, E. *Optimization. Algorithms and consistent approximations*. Applied Mathematical Sciences, Vol. 124, Springer, New York, 1997.

[NW99] Nocedal, J. and Wright, S. J. *Numerical optimization*. Springer Series in Operations Research, New York, 1999.

[GK02] Geiger, C. and Kanzow, C. *Theorie und Numerik restringierter Optimierungsaufgaben*. Springer, Berlin-Heidelberg-New York, 2002.

[Alt02] Alt, W. *Nichtlineare Optimierung: Eine Einführung in Theorie, Verfahren und Anwendungen*. Vieweg, Braunschweig/Wiesbaden, 2002.

[Han77] Han, S. P. *A Globally Convergent Method for Nonlinear Programming*. Journal of Optimization Theory and Applications, 22 (3); 297–309, 1977.

[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.

[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.

[Sch81] Schittkowski, K. *The Nonlinear Programming Method of Wilson, Han, and Powell with an Augmented Lagrangian Type Line Search Function. Part 1: Convergence Analysis, Part 2: An Efficient Implementation with Linear Least Squares Subproblems*. Numerische Mathematik, 383; 83–114, 115–127, 1981.

[Sch83] Schittkowski, K. *On the Convergence of a Sequential Quadratic Programming Method with an Augmented Lagrangian Line Search Function*. Mathematische Operationsforschung und Statistik, Series Optimization, 14 (2); 197–216, 1983.

[Bur89] Burke, J. V. and Han, S. P. *A robust sequential quadratic programming method*. Mathematical Programming, 43; 277–303, 1989.

[1] P. E. Gill and W. Murray. Numerically stable methods for quadratic programming. *Mathematical Programming*, 14:349–372, 1978.

[2] P. E. Gill, W. Murray, M. A. Saunders, and M. H. Wright. Inertia-controlling methods for general quadratic programming. *SIAM Review*, 33(1):1–36, 1991.

[3] D. Goldfarb and A. Idnani. A numerically stable dual method for solving strictly convex quadratic programs. *Mathematical Programming*, 27:1–33, 1983.

[Sch85] Schittkowski, K. *NLPQL: A Fortran subroutine for solving constrained nonlinear programming problems*. Annals of Operations Research, 5; 484–500, 1985.

[Kra88] Kraft, D. *A Software Package for Sequential Quadratic Programming*. DFVLR-FB-88-28, Oberpfaffenhofen, 1988.

[GMS02] Gill, P. E., Murray, W. and Saunders, M. A. *SNOPT: An SQP algorithm for large-scale constrained optimization*. SIAM Journal on Optimization, 12; 979–1006, 2002.

[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.

[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.

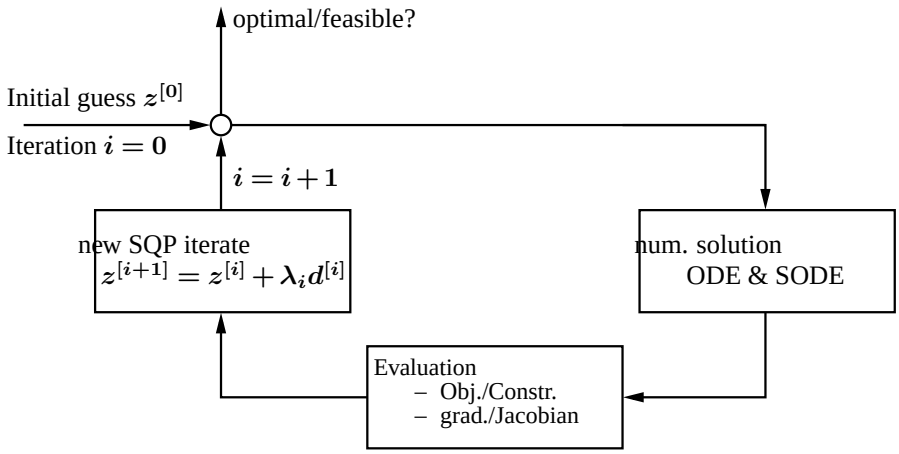
[Sch96] Schulz, V. H. *Reduced SQP Methods for Large-Scale Optimal Control Problems in DAE with Application to Path Planning Problems for Satellite Mounted Robots*. Ph.D. thesis, Interdisziplinäres Zentrum für Wissenschaftliches Rechnen, Universität Heidelberg, 1996.

[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.

[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.

7.5 Computing Gradients

Solution of Discretized Optimal Control Problem by SQP



Calculation of Gradients

Task: Compute

- gradient F' of the objective functional
- Jacobians G' and H' of the constraints

needed within the SQP method.

Full discretization approach: $\surd$

Overview

Reduced discretization approach:

- Sensitivity equation, if number of constraints *greater* than the number of variables
- Adjoint equation, if number of constraints *smaller* than the number of variables
- Algorithmic differentiation: „forward mode“ $\Leftrightarrow$ sensitivity differential equation „backward mode“ $\Leftrightarrow$ adjoint equation www.autodiff.org
- Approximation by finite differences

7.5.1 Sensitivity Equation Approach

Sensitivity Equation

explicit Euler's method:

$$\begin{aligned} \boldsymbol{X_0}(z) &= \boldsymbol{x_0}, && (66)\\ \boldsymbol{X_{i+1}}(z) &= \boldsymbol{X_i}(z) + \boldsymbol{h_i} \boldsymbol{f}(\boldsymbol{t_i}, \boldsymbol{X_i}(z), \boldsymbol{u_i}), \qquad \boldsymbol{i = 0, 1, \ldots, N-1} && (67) \end{aligned}$$

Reminder: $\boldsymbol{z} = (\boldsymbol{x_0}, \boldsymbol{u_0}, \ldots, \boldsymbol{u_N})^\top \in \mathbb{R}^n$ is the variable in (59).

We need the Sensitivities

$$\boldsymbol{S_i} := \frac{\partial \boldsymbol{X_i}(z)}{\partial \boldsymbol{z}}, \qquad \boldsymbol{i = 0, 1, \ldots, N.}$$

Sensitivity Equation

Differentiation of (67) w.r.t. $\boldsymbol{z}$ yields

$$\boldsymbol{S_{i+1}} = \boldsymbol{S_i} + \boldsymbol{h_i} \left(\boldsymbol{f'_x}(\boldsymbol{t_i}, \boldsymbol{X_i}(z), \boldsymbol{u_i}) \cdot \boldsymbol{S_i} + \boldsymbol{f'_u}(\boldsymbol{t_i}, \boldsymbol{X_i}(z), \boldsymbol{u_i}) \cdot \frac{\partial \boldsymbol{u_i}}{\partial \boldsymbol{z}} \right), \qquad \boldsymbol{i = 0, 1, \ldots, N-1.} \quad (68)$$

According to (66) it holds

$$\boldsymbol{S_0} = \frac{\partial \boldsymbol{X_0}(z)}{\partial \boldsymbol{z}} = \left(\boldsymbol{I_{n_x}} \,\middle|\, \boldsymbol{\Theta} \,\middle|\, \cdots \,\middle|\, \boldsymbol{\Theta} \right) \in \mathbb{R}^{\boldsymbol{n_x} \times (\boldsymbol{n_x} + (\boldsymbol{N}+1)\boldsymbol{n_u})}. \quad (69)$$

This approach is known as internal numerical differentiation (IND), cf. Bock [Boc87].

Gradient

Gradient of objective function: Chain rule

$$\frac{\boldsymbol{d}}{\boldsymbol{dz}} \boldsymbol{\varphi}(\boldsymbol{X_0}(z), \boldsymbol{X_N}(z)) = \boldsymbol{\varphi'_{x_0}}(\boldsymbol{X_0}(z), \boldsymbol{X_N}(z)) \cdot \boldsymbol{S_0} + \boldsymbol{\varphi'_{x_f}}(\boldsymbol{X_0}(z), \boldsymbol{X_N}(z)) \cdot \boldsymbol{S_N}.$$

Likewise for constraints!

Conclusions

Effort: Essentially an IVP of dimension $\boldsymbol{n_x \cdot (n+1)}$ has to be solved.

Notice: With incresing dimension $\boldsymbol{n = n_x + (N+1)n_u}$ of the variable $\boldsymbol{z}$ this becomes expensive!

Sensitivity Differential Equation

Remark 7.29.

- *In praxis, often a step-size selection strategy is used to solve the parametric IVP. In view of the sensitivity analysis it is necessary to solve the sensitivity differential equation using the same step-sizes. Otherwise, the gradient will not be exact but merely an approximation!*
- *The discussed method for sensitivity analysis can be extended to general one-step methods (explicit or implicit) and multiple-step methods.*
- *Similar strategies concerning sensitivity analysis were suggested by Caracotsios and Stewart [CS85], Maly and Petzold [MP96], and Brenan et al. [BCP96]. A comparison of the different approaches can be found in Feehery et al. [FTB97].*

7.5.2 Adjoint Equation Approach

Adjoint Method I

The adjoint method avoids the calculation of the sensitivities S_i.

Compute gradient w.r.t. z of a function

$$\Gamma(z) := \gamma(X_0(z), X_N(z), z)$$

Euler's method:

$$X_{i+1}(z) \quad = \quad X_i(z) + h_i f(t_i, X_i(z), u_i), \qquad i = 0, \ldots, N-1$$

Adjoint Method II

Auxiliary functional:

$$J(z) := \Gamma(z) + \sum_{i=0}^{N-1} \lambda_{i+1}^\top \left(X_{i+1}(z) - X_i(z) - h_i f(t_i, X_i(z), u_i) \right)$$

with multipliers λ_i, $i = 1, \ldots, N$.

Differentiation of J w.r.t. z: Eliminate terms with $S(t_i)$!

$$\begin{aligned} J'(z) \quad = \quad & \left(\gamma'_{x_0} - \lambda_1^\top - h_0 \lambda_1^\top f'_x[t_0] \right) \cdot S_0 + \left(\gamma'_{x_N} + \lambda_N^\top \right) \cdot S_N + \gamma'_z \\ & + \sum_{i=1}^{N-1} \left(\lambda_i^\top - \lambda_{i+1}^\top - h_i \lambda_{i+1}^\top f'_x[t_i] \right) \cdot S_i - \sum_{i=0}^{N-1} h_i \lambda_{i+1}^\top f'_u[t_i] \frac{\partial u_i}{\partial z}. \end{aligned}$$

Adjoint Method III

Adjoint equation: (backwards in time)

$$\lambda_i^\top - \lambda_{i+1}^\top - h_i \lambda_{i+1}^\top f'_x[t_i] = 0_{n_x}, \quad i = 0, \ldots, N-1 \tag{70}$$

Transversality condition at t_N:

$$\lambda_N^\top = -\gamma'_{x_N}(X_0(z), X_N(z), z). \tag{71}$$

Gradient:

$$J'(z) = \left(\gamma'_{x_0} - \lambda_0^\top \right) \cdot S_0 + \gamma'_z - \sum_{i=0}^{N-1} h_i \lambda_{i+1}^\top f'_u[t_i] \frac{\partial u_i}{\partial z}.$$

Adjoint Method IV

It remains to show that $J'(z) = \Gamma'(z)$:

Theorem 7.30. *It holds*

$$\Gamma'(z) = J'(z) = \left(\gamma'_{x_0} - \lambda_0^\top \right) \cdot S_0 + \gamma'_z - \sum_{i=0}^{N-1} h_i \lambda_{i+1}^\top f'_u[t_i] \frac{\partial u_i}{\partial z}. \tag{72}$$

Proof

Proof. Multiplication of the sensitivity equation

$$S_{i+1} - S_i - h_i f'_x[t_i] S_i - h_i f'_u[t_i] \frac{\partial u_i}{\partial z} = 0_{n_x}, \quad i = 0, \ldots, N-1$$

by $\lambda_{i+1}^\top$ from the left yields

$$-h_i \lambda_{i+1}^\top f'_u[t_i] \frac{\partial u_i}{\partial z} = -\lambda_{i+1}^\top S_{i+1} + \lambda_{i+1}^\top S_i + h_i \lambda_{i+1}^\top f'_x[t_i] S_i, \quad i = 0, \ldots, N-1$$

Proof

and hence

$$\begin{aligned}
J'(z) &= \left(\gamma'_{x_0} - \lambda_0^\top\right)\cdot S_0 + \gamma'_z \\
&\quad + \sum_{i=0}^{N-1}\left(\lambda_{i+1}^\top S_i + h\lambda_{i+1}^\top f'_x[t_i]S_i - \lambda_{i+1}^\top S_{i+1}\right) \\
&\overset{(70)}{=} \left(\gamma'_{x_0} - \lambda_0^\top\right)\cdot S_0 + \gamma'_z + \sum_{i=0}^{N-1}\left(\lambda_i^\top S_i - \lambda_{i+1}^\top S_{i+1}\right) \\
&= \left(\gamma'_{x_0} - \lambda_0^\top\right)\cdot S_0 + \gamma'_z + \lambda_0^\top S_0 - \lambda_N^\top S_N \\
&\overset{(71)}{=} \gamma'_{x_0} S_0 + \gamma'_z + \gamma'_{x_N} S_N \\
&= \Gamma'(z).
\end{aligned}$$

□

Conclusions

Result: Equation (72) provides a formula for the gradient of Γ, where Γ stands for each of the functions $F, G = (G_1, \ldots, G_m), H = (H_1, \ldots, H_p)$ in (59).

Effort: The adjoint equation has to be solved for each (!) component of F, G, H in order to calculate F', G', H' in (59). This amounts essentially to solving an IVP of dimension $n_x \cdot (2 + m + p)$. The trajectory $(X_i, i = 0, \ldots, N)$ has to be stored.

Notice: The effort does not depend on the number of variables z! The adjoint method is effective if only a few constraints are present.

Adjoint Method: Remark

Remark 7.31. *With $\mathcal{H}(t, x, u, \lambda) = \lambda^\top f(t, x, u)$ the adjoint equation can be written as*

$$\lambda_i^\top = \lambda_{i+1}^\top - h_i\left(-\mathcal{H}'_x(t_i, x_i, u_i, \lambda_{i+1})\right), \qquad i = 0, \ldots, N-1.$$

At first glance: similar to explicit Euler's method, backwards in time (step size $-h_i$), applied to adjoint differential equation $\dot\lambda^\top = -\mathcal{H}'_x$.

But: $\mathcal{H}$ is evaluated at (t_i, x_i, u_i) and not at $(t_{i+1}, x_{i+1}, u_{i+1})$! Thus, neither the explicit nor the implicit Euler's method occurs, but a mixed method (→ symplectic method, Hamiltonian systems).

For Further Reading

References

[1] Hans Georg Bock. Randwertproblemmethoden zur Parameteridentifizierung in Systemen nichtlinearer Differentialgleichungen. volume 183 of *Bonner Mathematische Schriften*, Bonn, 1987.

[2] Kathry E. Brenan, Stephen L. Campbell, and Linda R. Petzold. *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations*, volume 14 of *Classics In Applied Mathematics*. SIAM, Philadelphia, 1996.

[3] M. Caracotsios and W. E. Stewart. Sensitivity analysis of initial-boundary-value problems with mixed PDEs and algebraic equations. *Computers chem. Engng*, 19(9):1019–1030, 1985.

[4] William F. Feehery, John E. Tolsma, and Paul I. Barton. Efficient sensitivity analysis of large-scale differential-algebraic systems. *Applied Numerical Mathematics*, 25:41–54, 1997.

[5] Timothy Maly and Linda R. Petzold. Numerical Methods and Software for Sensitivity Analysis of Differential-Algebraic Systems. *Applied Numerical Mathematics*, 20(1):57–79, 1996.

7.6 Discrete Minimum Principle

Outline

Goals:

- derive discrete local minimum principle for Euler's method
- compare to continuous local minimum principle
- find relation between discrete and continuous multipliers
- approximation of continuous adjoint $\boldsymbol{\lambda}$ and multipliers $\boldsymbol{\eta}$ and $\boldsymbol{\mu}$

Notation

Continuous multipliers:

$$\begin{array}{rcl} \boldsymbol{\lambda} & \leftrightarrow & \boldsymbol{f(t,x,u) - \dot{x} = 0,} \\ \boldsymbol{\eta} & \leftrightarrow & \boldsymbol{c(t,x,u) \leq 0,} \\ \boldsymbol{\mu} & \leftrightarrow & \boldsymbol{s(t,x) \leq 0,} \\ \boldsymbol{\sigma} & \leftrightarrow & \boldsymbol{\psi(x(t_0), x(t_f)) = 0.} \end{array}$$

Discrete multipliers:

$$\begin{array}{rcl} \boldsymbol{\lambda_{i+1}} & \leftrightarrow & \boldsymbol{x_i + hf(t_i, x_i, u_i) - x_{i+1} = 0,} \\ \boldsymbol{\zeta_i} & \leftrightarrow & \boldsymbol{c(t_i, x_i, u_i) \leq 0,} \\ \boldsymbol{\nu_i} & \leftrightarrow & \boldsymbol{s(t_i, x_i) \leq 0,} \\ \boldsymbol{\kappa} & \leftrightarrow & \boldsymbol{\psi(x_0, x_N) = 0.} \end{array}$$

Auxiliary Functions

Hamiltonian/augmented Hamiltonian:

$$\begin{aligned} \mathcal{H}(t,x,u,\lambda,\ell_0) &= \lambda^\top f(t,x,u), \\ \hat{\mathcal{H}}(t,x,u,\lambda,\zeta,\ell_0) &= \mathcal{H}(t,x,u,\lambda,\ell_0) + \zeta^\top c(t,x,u). \end{aligned}$$

Auxiliary Functions

Lagrangian:

$$\begin{aligned} L(x,u,\lambda,\zeta,\nu,\kappa,\ell_0) &:= \ell_0 \varphi(x_0,x_N) + \kappa^\top \psi(x_0,x_N) \\ &\quad + \sum_{i=0}^{N-1} \lambda_{i+1}^\top (x_i + h_i f(t_i,x_i,u_i) - x_{i+1}) \\ &\quad + \sum_{i=0}^{N} \zeta_i^\top c(t_i,x_i,u_i) + \sum_{i=0}^{N} \nu_i^\top s(t_i,x_i) \\ &= \ell_0 \varphi(x_0,x_N) + \kappa^\top \psi(x_0,x_N) + \zeta_N^\top c(t_N,x_N,u_N) \\ &\quad + \sum_{i=0}^{N-1} \left(h_i \mathcal{H}(t_i,x_i,u_i,\lambda_{i+1},\ell_0) + \zeta_i^\top c(t_i,x_i,u_i) \right) \\ &\quad + \sum_{i=0}^{N-1} \lambda_{i+1}^\top (x_i - x_{i+1}) + \sum_{i=0}^{N} \nu_i^\top s(t_i,x_i) \end{aligned}$$

Auxiliary Functions

$$\begin{aligned} L(x,u,\lambda,\zeta,\nu,\kappa,\ell_0) &= \ell_0\varphi(x_0,x_N)+\kappa^\top\psi(x_0,x_N)+\zeta_N^\top c(t_N,x_N,u_N) \\ &\quad + \sum_{i=0}^{N-1} h_i\hat{\mathcal{H}}\left(t_i,x_i,u_i,\lambda_{i+1},\frac{\zeta_i}{h_i},\ell_0\right) \\ &\quad + \sum_{i=0}^{N-1}\lambda_{i+1}^\top(x_i-x_{i+1})+\sum_{i=0}^{N}\nu_i^\top s(t_i,x_i) \end{aligned}$$

Discrete Minimum Principle

Evaluation of the first order necessary conditions of Fritz-John leads to

Theorem 7.32 (Discrete Local Minimum Principle). *Assumptions:*

(i) Let $\varphi, f_0, f, c, s, \psi$ be continuously differentiable w.r.t. x and u.

(ii) Let $(\hat{x},\hat{u})$ be a local minimum of problem 7.21.

Then there exist multipliers $\ell_0\in\mathbb{R}$, $\kappa\in\mathbb{R}^{n_\psi}$, $\lambda=(\lambda_0,\dots,\lambda_N)^\top\in\mathbb{R}^{n_x(N+1)}$, $\zeta=(\zeta_0,\dots,\zeta_N)^\top\in\mathbb{R}^{n_c(N+1)}$, $\nu=(\nu_0,\dots,\nu_N)^\top\in\mathbb{R}^{n_s(N+1)}$ with:

(i) $\ell_0\geq 0$, $(\ell_0,\kappa,\lambda,\zeta,\nu)\neq\Theta$,

Discrete Minimum Principle

Theorem 7.32 (continued).

(ii) Discrete adjoint equation: For $i=0,\dots,N-1$ it holds

$$\begin{aligned} \lambda_i &= \lambda_{i+1}+h_i\hat{\mathcal{H}}_x'\left(t_i,\hat{x}_i,\hat{u}_i,\lambda_{i+1},\frac{\zeta_i}{h_i},\ell_0\right)^\top+s_x'(t_i,\hat{x}_i)^\top\nu_i \\ &= \lambda_N+\sum_{j=i}^{N-1}h_j\hat{\mathcal{H}}_x'\left(t_j,\hat{x}_j,\hat{u}_j,\lambda_{j+1},\frac{\zeta_j}{h_j},\ell_0\right)^\top+\sum_{j=i}^{N-1}s_x'(t_j,\hat{x}_j)^\top\nu_j. \end{aligned}$$

Discrete Minimum Principle

Theorem 7.32 (continued).

(iii) Discrete transversality conditions:

$$\lambda_0 = -\left(\ell_0\varphi_{x_0}'(\hat{x}_0,\hat{x}_N)^\top+\psi_{x_0}'(\hat{x}_0,\hat{x}_N)^\top\kappa\right),$$

$$\begin{aligned} \lambda_N &= \ell_0\varphi_{x_f}'(\hat{x}_0,\hat{x}_N)^\top+\psi_{x_f}'(\hat{x}_0,\hat{x}_N)^\top\kappa \\ &\quad +c_x'(t_N,\hat{x}_N,\hat{u}_N)^\top\zeta_N+s_x'(t_N,\hat{x}_N)^\top\nu_N. \end{aligned}$$

Discrete Minimum Principle

Theorem 7.32 (continued).

(iii) Discrete optimality conditions: For all $i=0,\dots,N-1$ and all $u\in\mathcal{U}$ it holds

$$\hat{\mathcal{H}}_u'\left(t_i,\hat{x}_i,\hat{u}_i,\lambda_{i+1},\frac{\zeta_i}{h_i},\ell_0\right)(u-\hat{u}_i)\geq 0.$$

Furthermore, it holds $\zeta_N^\top c_u'(t_N,\hat{x}_N,\hat{u}_N)(u-\hat{u}_N)\geq 0$ for all $u\in\mathcal{U}$.

Discrete Minimum Principle

Theorem 7.32 (continued).

(iv) Discrete complementarity conditions: It holds

$$\begin{aligned} \zeta_i &\geq 0_{n_c}, \quad i = 0, \ldots, N, \\ \nu_i &\geq 0_{n_s}, \quad i = 0, \ldots, N, \\ \zeta_i^\top c(t_i, \hat{x}_i, \hat{u}_i) &= 0, \quad i = 0, \ldots, N, \\ \nu_i^\top s(t_i, \hat{x}_i) &= 0, \quad i = 0, \ldots, N. \end{aligned}$$

Approximation of Adjoints I

Problems with mixed control-state constraints and without pure state constraints ($n_s = 0$, $s \equiv 0$):
Adjoint differential equation (cf. Theorem 7.6):

$$\dot{\lambda}(t) = -\hat{\mathcal{H}}'_x(t, \hat{x}(t), \hat{u}(t), \lambda(t), \eta(t), \ell_0)^\top.$$

Discrete adjoint equation:

$$\frac{\lambda_{i+1} - \lambda_i}{h_i} = -\hat{\mathcal{H}}'_x\left(t_i, \hat{x}_i, \hat{u}_i, \lambda_{i+1}, \frac{\zeta_i}{h_i}, \ell_0\right)^\top, \qquad i = 0, \ldots, N-1.$$

Relation:

$$\frac{\zeta_i}{h_i} \approx \eta(t_i), \qquad i = 0, \ldots, N-1$$

Approximation of Adjoints II

Problems with pure state constraints and without mixed control-state constraints ($n_c = 0$, $c \equiv 0$):
Adjoint equation in integral form (cf. Theorem 7.8):

$$\lambda(t) = \lambda(t_f) + \int_t^{t_f} \mathcal{H}'_x[\tau]^\top d\tau + \sum_{i=1}^{n_s} \int_t^{t_f} s'_{i,x}[\tau] d\mu_i(\tau), \qquad \forall t \in [t_0, t_f],$$

Continuous transversality condition:

$$\lambda(t_f)^\top = \ell_0 \varphi'_{x_f}(x(t_0), x(t_f)) + \sigma^\top \psi'_{x_f}(x(t_0), x(t_f)).$$

Approximation of Adjoints III

Discrete adjoint equation and transversality condition: For $i = 0, \ldots, N-1$

$$\begin{aligned} \lambda_i &= \lambda_N + \sum_{j=i}^{N-1} h_j \mathcal{H}'_x(t_j, \hat{x}_j, \hat{u}_j, \lambda_{j+1}, \ell_0)^\top + \sum_{j=i}^{N-1} s'_x(t_j, \hat{x}_j)^\top \nu_j \\ &= \ell_0 \varphi'_{x_f}(\hat{x}_0, \hat{x}_N)^\top + \psi'_{x_f}(\hat{x}_0, \hat{x}_N)^\top \kappa \\ &\quad + \underbrace{\sum_{j=i}^{N-1} h_j \mathcal{H}'_x(t_j, \hat{x}_j, \hat{u}_j, \lambda_{j+1}, \ell_0)^\top}_{\approx \int_t^{t_f} \mathcal{H}'_x[\tau]^\top d\tau} + \underbrace{\sum_{j=i}^{N} s'_x(t_j, \hat{x}_j)^\top \nu_j}_{\approx \int_t^{t_f} s'_x[\tau]^\top d\mu(\tau)}. \end{aligned}$$

Approximation of Adjoints IV

Interpretation:

- $\kappa \approx \sigma$
- Approximation of continuous multiplier μ:

$$\nu_i \approx \mu(t_{i+1}) - \mu(t_i), \qquad i = 0, \ldots, N-1$$

ν_N is interpreted as jump height at $t = t_f$, i.e.

$$\nu_N \approx \mu(t_f) - \mu(t_f^-).$$

Notice: The continuous multiplier μ may jump at t_f!

Approximation of Adjoints V

- Furthermore, we interprete

$$\lambda_N = \underbrace{\ell_0 \varphi'_{x_f}(\hat{x}_0, \hat{x}_N)^\top + \psi'_{x_f}(\hat{x}_0, \hat{x}_N)^\top \kappa}_{\approx \lambda(t_f)} + s'_x(t_N, \hat{x}_N)^\top \nu_N$$

as leftsided limit $\lambda(t_f^-)$ of the continuous adjoint λ.

Notice: The continuous adjoint λ may jump at t_f (caused by μ)!

Approximation of Adjoints VI

The complementarity conditions yield

$$0_{n_s} \leq \nu_i \approx \mu(t_{i+1}) - \mu(t_i) \qquad \leftrightarrow \qquad \text{monotonicity of } \mu$$

Since $\nu_i^\top s(t_i, x_i) = 0$ it follows

$$s(t_i, x_i) < 0_{n_s} \qquad \Rightarrow \qquad 0_{n_s} = \nu_i \approx \mu(t_{i+1}) - \mu(t_i),$$

which corresponds to μ being constant on inactive subarcs.

Remarks

Remark 7.33.

- *The variable u_N only occurs in the constraint $c(t_N, x_N, u_N) \leq 0_{n_c}$ and has no impact on the objective function.*
- *In the continuous case, a global minimum principle (54) resp. (55) holds. In the discrete case there is in general no global minimum principle!*

  Only an approximate minimum principle holds, cf. Mordukhovich [Mor88].

  Under additional convexity-like conditions also a discrete global minimum principle can be proved, cf. Ioffe and Tihomirov [IT79], Section 6.4, p. 277.

For Further Reading

References

[1] A. D. Ioffe and V. M. Tihomirov. Theory of extremal problems. volume 6 of *Studies in Mathematics and its Applications*, Amsterdam, New York, Oxford, 1979. North-Holland Publishing Company.

[2] B. S. Mordukhovich. An approximate maximum principle for finite-difference control systems. *U.S.S.R. Computational Mathematics and Mathematical Physics*, 28(1):106–114, 1988.

7.7 Convergence

Convergence of Euler's discretization

Optimal control problem:

$$\begin{aligned} \text{Minimize} \quad & \varphi(x(t_f)) + \int_0^{t_f} f_0(x(t), u(t))\,dt \\ \text{s.t.} \quad & \dot{x}(t) - f(x(t), u(t)) = 0_{n_x}, \\ & x(0) - \xi = 0_{n_x}, \\ & \psi(x(t_f)) = 0_{n_\psi}, \\ & c(x(t), u(t)) \leq 0_{n_c} \end{aligned}$$

Convergence of Euler's discretization

Discretization:

$$\begin{aligned} \text{Minimize} \quad & \varphi(x_N) + h \sum_{i=0}^{N-1} f_0(x_i, u_i) \\ \text{s.t.} \quad & \frac{x_{i+1} - x_i}{h} - f(x_i, u_i) = 0_{n_x}, \qquad i = 0, \ldots, N-1, \\ & x_0 - \xi = 0_{n_x}, \\ & \psi(x_N) = 0_{n_\psi}, \\ & c(x_i, u_i) \leq 0_{n_c}, \qquad i = 0, \ldots, N-1 \end{aligned}$$

Assumptions I

Assumptions:

(i) Let f_0, f, c, φ, ψ be continuously 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.

Assumptions II

(iii) Uniform rank condition for c: There exists a constant $\alpha > 0$ with

$$\|c'_{\mathcal{J}(t),u}[t]^\top d\| \geq \alpha \|d\| \qquad \forall d \in \mathbb{R}^{|\mathcal{J}(t)|} \text{ a.e. in } [0, t_f].$$

Notation:

$$\begin{aligned} \mathcal{J}(t) := \{i \mid c_i(\hat{x}(t), \hat{u}(t)) = 0\} \quad &: \text{ index set of active constr. at } t, \\ c_{\mathcal{J}(t)}[t] \quad &: \text{ active constraints at } t. \end{aligned}$$

Assumptions III

(iv) Surjectivity of linearized equality constraints: The BVP

$$\dot{y}(t) - \tilde{A}(t) y(t) - \tilde{B}(t) v(t) = 0_{n_x}, \quad y(0) = 0_{n_x}, \quad \psi'_{x_f}(\hat{x}(t_f)) y(t_f) = h$$

has a solution for every $h \in \mathbb{R}^{n_\psi}$, where

$$\begin{aligned} \tilde{A}(t) &= f'_x[t] - f'_u[t] c'_{\mathcal{J}(t),u}[t]^\top \left(c'_{\mathcal{J}(t),u}[t] c'_{\mathcal{J}(t),u}[t]^\top \right)^{-1} c'_{\mathcal{J}(t),x}[t], \\ \tilde{B}(t) &= f'_u[t] \left(I_{n_u} - c'_{\mathcal{J}(t),u}[t]^\top \left(c'_{\mathcal{J}(t),u}[t] c'_{\mathcal{J}(t),u}[t]^\top \right)^{-1} c'_{\mathcal{J}(t),x}[t] \right). \end{aligned}$$

Assumptions IV

(v) Coercivity:

There exists $\beta > 0$ with

$$d^\top \hat{\mathcal{H}}''_{uu}[t] d \geq \beta \|d\|^2$$

for all $d \in \mathbb{R}^{n_u}$ with

$$c'_{J^+(t),u}[t] d = 0_{|\mathcal{J}^+(t)|}.$$

Notation:

$$\mathcal{J}^+(t) := \{i \in \mathcal{J}(t) \mid \eta_i(t) > 0\} \quad : \quad \eta_i \text{ denotes multiplier for } c_i$$

Assumptions V

(vi) Riccati differential equation: The Riccati differential equation

$$\begin{aligned}\dot{Q}(t) = &-Q(t) f'_x[t] - f'_x[t]^\top Q(t) - \hat{\mathcal{H}}''_{xx}[t] \\ &+ \left[\begin{pmatrix} \hat{\mathcal{H}}''_{ux}[t] \\ c'_{J^+(t),x}[t] \end{pmatrix}^\top + Q(t) \begin{pmatrix} f'_u[t]^\top \\ \Theta \end{pmatrix}^\top \right] \begin{pmatrix} \hat{\mathcal{H}}''_{uu}[t] & c'_{J^+(t),u}[t]^\top \\ c'_{J^+(t),u}[t] & \Theta \end{pmatrix}^{\cdot} \\ &\cdot \left[\begin{pmatrix} f'_u[t]^\top \\ \Theta \end{pmatrix} Q(t) + \begin{pmatrix} \hat{\mathcal{H}}''_{ux}[t] \\ c'_{J^+(t),x}[t] \end{pmatrix} \right]\end{aligned}$$

has a bounded solution Q on $[0, t_f]$.

Assumptions VI

Q satisfies the rank assumption:

$$d^\top (\Gamma - Q(t_f))\, d \geq 0 \qquad \forall d \in \mathbb{R}^{n_x} \ : \ \psi'_{x_f}(\hat{x}(t_f)) d = 0_{n_\psi},$$

where

$$\Gamma := \left(\varphi(\hat{x}(t_f)) + \sigma_f^\top \psi(\hat{x}(t_f))\right)''_{xx}.$$

Convergence of Euler's discretization

Unter the above assumptions the following convergence result holds:

Theorem 7.34. *Let the assumptions (i)-(vi) be fulfilled. Then, for every sufficiently small step size $h > 0$ there exist a locally unique KKT point $(x_h, u_h, \lambda_h, \zeta_h, \kappa_0, \kappa_f)$ of the discretized optimal control problem and*

$$\begin{aligned}\max\{&\|x_h - \hat{x}\|_{1,\infty}, \|u_h - \hat{u}\|_\infty, \|\lambda_h - \lambda\|_{1,\infty}, \\ &\|\kappa_0 - \sigma_0\|, \|\kappa_f - \sigma_f\|, \|\eta_h - \eta\|_\infty\} \quad = \quad \mathcal{O}(h),\end{aligned}$$

where λ_h denotes the discrete adjoint, η_h the discrete multiplier for the mixed control-state constraints, κ_0 the discrete multiplier for the initial condition, and κ_f the discrete multiplier for the final condition.

The lenghtly and difficult proof can be found in Malanowski et al. [MBM97].

Remarks

Remark 7.35.

- *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.[DHM00, DHV00b].*

Convergence of Runge-Kutta Discretizations

Optimal control problem:

$$\begin{aligned} \text{Minimize} \quad & \varphi(x(1)) \\ \text{s.t.} \quad & \dot{x}(t) = f(x(t), u(t)), \qquad t \in [0,1] \\ & x(0) = \xi. \end{aligned}$$

Convergence of Runge-Kutta Discretizations

Discretization by Runge-Kutta:

$$\begin{aligned} \text{Minimize} \quad & \varphi(x_N) \\ \text{s.t.} \quad & \frac{x_{k+1} - x_k}{h} = \sum_{i=1}^{s} b_i f(\eta_i, u_{ki}), \quad k = 0, \dots, N-1, \\ & \eta_i = x_k + h \sum_{j=1}^{s} a_{ij} f(\eta_j, u_{kj}), \quad i = 1, \dots, s, \\ & x_0 = \xi. \end{aligned}$$

Notice: An own optimization variable u_{ki} is introduced at every stage!

Assumptions I

Assumptions:

- Smoothness: (notice: u is at least continuous!) The optimal control problem has a solution

$$(\hat{x}, \hat{u}) \in W^{p,\infty}([0,1], \mathbb{R}^{n_x}) \times W^{p-1,\infty}([0,1], \mathbb{R}^{n_u}), \qquad p \geq 2.$$

  The first p derivatives of f and φ are locally lipschitz-continuous in some neighborhood of $(\hat{x}, \hat{u})$.

Assumptions II

- Coercivity: There exists $\alpha > 0$ with

$$\mathcal{B}(x,u) \geq \alpha \|u\|_2^2 \qquad \forall (x,u) \in \mathcal{M},$$

  where

$$\begin{aligned} \mathcal{B}(x,u) \quad = \quad \frac{1}{2} \Bigg(& x(1)^\top V x(1) + \int_0^1 x(t)^\top Q(t) x(t) + 2x(t)^\top S(t) u(t) \\ & + u(t)^\top R(t) u(t) dt \Bigg) \end{aligned}$$

Assumptions III

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

$$\mathcal{M} = \left\{ (x,u) \in W^{1,2}([0,1], \mathbb{R}^{n_x}) \times L^2([0,1], \mathbb{R}^{n_u}) \;\middle|\; \dot{x} = Ax + Bu,\ x(0) = 0 \right\}.$$

OCP Order Conditions

OCP order conditions: *(≠ IVP order conditions for Runge-Kutta methods)*

Order	conditions ($c_i = \sum a_{ij}$, $d_j = \sum b_i a_{ij}$)
$p = 1$	$\sum b_i = 1$
$p = 2$	$\sum d_i = \frac{1}{2}$
$p = 3$	$\sum c_i d_i = \frac{1}{6}, \sum b_i c_i^2 = \frac{1}{3}, \sum \frac{d_i^2}{b_i} = \frac{1}{3}$
$p = 4$	$\sum b_i c_i^3 = \frac{1}{4}, \sum b_i c_i a_{ij} c_j = \frac{1}{8}, \sum d_i c_i^2 = \frac{1}{12}, \sum d_i a_{ij} c_j = \frac{1}{24},$ $\sum \frac{c_i d_i^2}{b_i} = \frac{1}{12}, \sum \frac{d_i^3}{b_i^2} = \frac{1}{4}, \sum \frac{b_i c_i a_{ij} d_j}{b_j} = \frac{5}{24}, \sum \frac{d_i a_{ij} d_j}{b_j} = \frac{1}{8}$

Convergence of Runge-Kutta Discretizations

Theorem 7.36. *Let the smoothness condition, the coercivity condition, and $b_i > 0$, $i = 1, \dots, s$ hold. Let the Runge-Kutta method fulfill the OCP order conditions up to order p, cf. table.*

Then, for sufficiently small step sizes h there exists a strict local minimum of the discretized optimal control problem.

Convergence of Runge-Kutta Discretizations

Theorem 7.36 (continued). *If $\frac{d^{p-1}\hat{u}}{dt^{p-1}}$ is of bounded variation, then*

$$\max_{0 \le k \le N} \{\|x_k - \hat{x}(t_k)\| + \|\lambda_k - \lambda(t_k)\| + \|u^*(x_k, \lambda_k) - \hat{u}(t_k)\|\} = \mathcal{O}(h^p).$$

If $\frac{d^{p-1}\hat{u}}{dt^{p-1}}$ is Riemann integrable, then

$$\max_{0 \le k \le N} \{\|x_k - \hat{x}(t_k)\| + \|\lambda_k - \lambda(t_k)\| + \|u^*(x_k, \lambda_k) - \hat{u}(t_k)\|\} = o(h^{p-1}).$$

Herein, $u^*(x_k, \lambda_k)$ denotes a local minimum of the Hamiltonian $\mathcal{H}(x_k, u, \lambda_k)$ w.r.t. u.

The lengthly and difficult proof can be found in Hager [Hag00].

Example

The subsequent example shows, that the condition $b_i > 0$, $i = 1, \dots, s$ is essential.

Example 7.37 *(Hager [Hag00], p. 272).* Consider

$$\min \quad \frac{1}{2} \int_0^1 u(t)^2 + 2x(t)^2 dt \quad \text{s.t.} \quad \dot{x}(t) = \frac{1}{2} x(t) + u(t),\ x(0) = 1.$$

Optimal solution:

$$\hat{x}(t) = \frac{2\exp(3t) + \exp(3)}{\exp(3t/2)(2 + \exp(3))}, \quad \hat{u}(t) = \frac{2(\exp(3t) - \exp(3))}{\exp(3t/2)(2 + \exp(3))}.$$

Example

Example 7.37 *(continued).* modified Euler's method:

0	0	0
1/2	1/2	0
	0	1

Heun's method:

0	0	0
1	1	0
	1/2	1/2

Example

Example 7.37 *(continued).* Error of the state $\boldsymbol{x}$ for the norm $\|\cdot\|_\infty$ for Heun's method:

N	error of $\boldsymbol{x}$	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 at second order!

Example

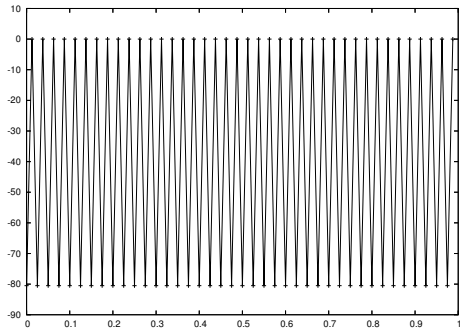
Example 7.37 *(continued).* Error of the state $\boldsymbol{x}$ for the norm $\|\cdot\|_\infty$ for the modified Euler's method:

N	error of $\boldsymbol{x}$	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

Example

Example 7.37 *(continued).*

- *No convergence!*
- Numerical solution for $N = 40$: strong oscillations in control $\boldsymbol{u}$ between 0 at $t_i + h/2$ and approximately $-1/(2h)$ at t_i:



Example

Example 7.37 *(continued).* Error of the state $\boldsymbol{x}$ for the norm $\|\cdot\|_\infty$ for the modified Euler's method with *piecewise constant control approximation*:

N	error of $\boldsymbol{x}$	order
10	0.3358800781952942E-03	–
20	0.8930396513584515E-04	1.9111501
40	0.2273822819465199E-04	1.9736044
80	0.5366500129055929E-05	2.0830664
160	0.1250729642299220E-05	2.1012115
320	0.7884779272826492E-06	0.6656277

Modified Euler's method with piecewise constant control approximation converges at second order!

For Further Reading

References

[1] Kazimierz Malanowski, Christof Büskens, and Helmut Maurer. Convergence of Approximations to Nonlinear Optimal Control Problems. In Anthony Fiacco, editor, *Mathematical programming with data perturbations*, volume 195, pages 253–284. Dekker. Lecture Notes in Pure and Applied Mathematics, 1997.

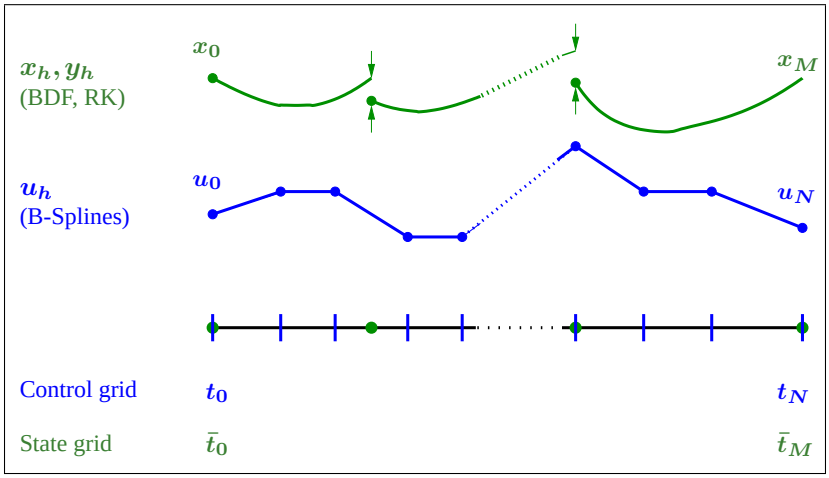
[2] A. L. Dontchev, W. W. Hager, and K. Malanowski. Error Bounds for Euler Approximation of a State and Control Constrained Optimal Control Problem. *Numerical Functional Analysis and Optimization*, 21(5 & 6):653–682, 2000.

[3] A. L. Dontchev, W. W. Hager, and V. M. Veliov. Second-Order Runge-Kutta Approximations in Control Constrained Optimal Control. *SIAM Journal on Numerical Analysis*, 38(1):202–226, 2000.

[4] William W. Hager. Runge-Kutta methods in optimal control and the transformed adjoint system. *Numerische Mathematik*, 87:247–282, 2000.

7.8 Direct Shooting Method

Direct Multiple Shooting Method



B-Splines

Elementary B-splines of order $k \in \mathbb{N}$: Recursively defined by

$$
\begin{aligned}
B_i^1(t) &:= \begin{cases} 1, & \text{if } \tau_i \le t < \tau_{i+1} \\ 0, & \text{otherwise} \end{cases}, \\
B_i^k(t) &:= \frac{t-\tau_i}{\tau_{i+k-1}-\tau_i} B_i^{k-1}(t) + \frac{\tau_{i+k}-t}{\tau_{i+k}-\tau_{i+1}} B_{i+1}^{k-1}(t)
\end{aligned}
\tag{73}
$$

where

$$
\tau_i := \begin{cases} t_0, & \text{if } 1 \le i \le k, \\ t_{i-k}, & \text{if } k+1 \le i \le N+k-1, \\ t_N, & \text{if } N+k \le i \le N+2k-1. \end{cases}
$$

Auxiliary grid:

$$
\mathbb{G}_u^k := \{\tau_i \mid i = 1, \ldots, N+2k-1\} \tag{74}
$$

Porperties of B-Splines

Properties:

- B-Splines date back to de Boor [DB78].
- Evaluation of recursion (73) is well-conditioned, cf. Deuflhard and Hohmann [DH91].
- Elementary B-Splines B_i^k, $i = 1, \ldots, N + k - 1$ are piecewise polynomials of degree $k - 1$. They define basis of space of splines

$$\left\{ s(\cdot) \in C^{k-2}([t_0, t_f], \mathbb{R}) \,\middle|\, s(\cdot)|_{[t_j, t_{j+1}]} \in \mathcal{P}^{k-1}([t_j, t_{j+1}]) \right\}.$$

Porperties of B-Splines

- For $k \geq 2$ it holds $B_i^k(\cdot) \in C^{k-2}([t_0, t_f], \mathbb{R})$. For $k \geq 3$ it holds the recursion

$$\frac{d}{dt} B_i^k(t) = \frac{k-1}{\tau_{i+k-1} - \tau_i} B_i^{k-1}(t) - \frac{k-1}{\tau_{i+k} - \tau_{i+1}} B_{i+1}^{k-1}(t).$$

- The elementary B-Splines have local support and

$$B_i^k(t) \begin{cases} > 0, & \text{if } t \in (\tau_i, \tau_{i+k}), \\ = 0, & \text{otherwise.} \end{cases}$$

Some Elementary B-Splines

$k = 1$: piecewise constant functions

$k = 2$: continuous, piecewise linear functions

$$B_i^2(t) = \begin{cases} \dfrac{t - \tau_i}{\tau_{i+1} - \tau_i}, & \text{if } \tau_i \leq t < \tau_{i+1}, \\ \dfrac{\tau_{i+2} - t}{\tau_{i+2} - \tau_{i+1}}, & \text{if } \tau_{i+1} \leq t < \tau_{i+2}, \\ 0, & \text{otherwise.} \end{cases}$$

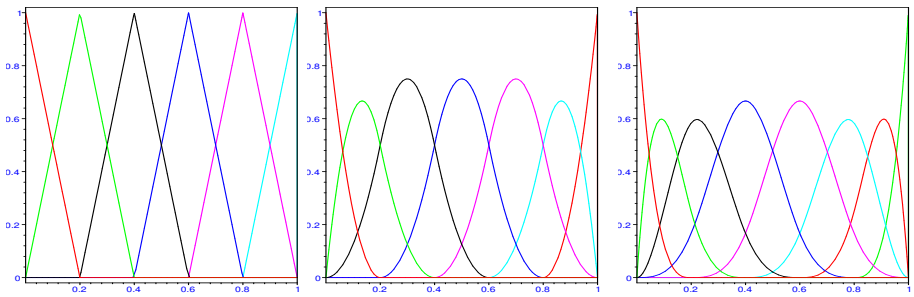
Some Elementary B-Splines

$k = 3$: continuously differentiable, piecewise quadratic functions

$$B_i^3(t) = \begin{cases} \dfrac{(t - \tau_i)^2}{(\tau_{i+2} - \tau_i)(\tau_{i+1} - \tau_i)}, & \text{if } t \in [\tau_i, \tau_{i+1}), \\ \dfrac{(t - \tau_i)(\tau_{i+2} - t)}{(\tau_{i+2} - \tau_i)(\tau_{i+2} - \tau_{i+1})} + \dfrac{(\tau_{i+3} - t)(t - \tau_{i+1})}{(\tau_{i+3} - \tau_{i+1})(\tau_{i+2} - \tau_{i+1})}, & \text{if } t \in [\tau_{i+1}, \tau_{i+2}), \\ \dfrac{(\tau_{i+3} - t)^2}{(\tau_{i+3} - \tau_{i+1})(\tau_{i+3} - \tau_{i+2})}, & \text{if } t \in [\tau_{i+2}, \tau_{i+3}), \\ 0, & \text{otherwise.} \end{cases}$$

Some Elementary B-Splines

B-splines of order $k = 2$ (left), $k = 3$ (middle), and $k = 4$ (right) for $[t_0, t_f] = [0, 1]$ and $N = 5$ on equidistant grid:



Control Approximations

Discretized control: Linear combination of elementary B-splines

$$u_h(t) = \sum_{i=1}^{N+k-1} c_i \, B_i^k(t) \tag{75}$$

with coefficients

$$c := (c_1, \ldots, c_{N+k+1}) \in \mathbb{R}^{n_u (N+k-1)}.$$

Notice: u_h is fully determined by finitely many parameters c_i!

Notation:

$$u_h(t) = u_h(t; c) = u_h(t; c_1, \ldots, c_{N+k-1})$$

c_i are called de Boor points

Advantages

B-splines possess two advantages from numerical point of view:

- It is easy to construct arbitrarily smooth control approximations.
- B-splines possess local support, i.e. the de Boor point c_i influences the value $u_h(t)$ only if $t \in [\tau_i, \tau_{i+k}]$. This property leads to a sparse structure of the gradient of the objective functional and the Jacobian of the constraints. Exploitation of this sparsity leads to fast algorithms.

State Approximation

Goal: derive reduced discretization ($\longrightarrow$ reduced discretization by Euler's method)

Ansatz:

- Replace u in the differential equation by the approximation u_h in (75).

State Approximation

- In every state grid interval $[T_i, T_{i+1}]$, $i = 0, \ldots, M-1$ starting at guesses X_i, $i = 0, \ldots, M-1$ solve the IVP's

  $$\dot{x}(t) = f(t, x(t), u_h(t; c)), \qquad x(T_i) = X_i$$

  by some one-step method, solution: X^i.

  Notice: Control grid points in $\mathbb{G}_u$ are supposed to be grid points of the one-step method! *Otherwise, there will be difficulties for nonsmooth control approximations!*

State Approximation

- Replace the continuous state in the optimal control problem by the discretized state.

State Approximation

Variables:

$$z := (X_0, \dots, X_{M-1}, c)^\top \in \mathbb{R}^{n_x M + n_u (N+k-1)}$$

State approximation: Putting together the piecewise solutions $X^i(t;z)$, $i = 0, \dots, M-1$ yields the approximation

$$X(t;z) := \begin{cases} X^i(t;z), & \text{if } t \in [T_i, T_{i+1}),\ i = 0, \dots, M-1, \\ X^{M-1}(T_M;z), & \text{if } t = t_f. \end{cases}$$

Reduced Direct Multiple Shooting Method

Introducing X and u_h into the optimal control problem yields the following finite dimensional optimization problem:

Problem 7.38 (Reduced Direct Multiple Shooting Method). *Find* $z = (X_0, \dots, X_{M-1}, c)^\top \in \mathbb{R}^{n_x M + n_u (N+k-1)}$ *such that*

$$\varphi(X_0, X(T_M;z))$$

is minimized subject to

$$\begin{aligned} \psi(X_0, X(T_M;z)) &= 0_{n_\psi}, & \\ c(t_i, X(t_i;z), u_h(t_i;c)) &\le 0_{n_c}, & i = 0,1,\dots,N, \\ s(t_i, X(t_i;z), u_h(t_i;c)) &\le 0_{n_s}, & i = 0,1,\dots,N, \\ u_h(t_i;c) &\in [u_{min}, u_{max}], & i = 0,1,\dots,N, \end{aligned}$$

Reduced Direct Multiple Shooting Method

Problem 7.38 (continued). *and continuity conditions at the nodes* T_i*:*

$$X^i(T_{i+1};z) - X_{i+1} = 0_{n_x}, \qquad i = 0, \dots, M-1.$$

Reduced Direct Multiple Shooting Method

Remark 7.39.

- *Of course, an individual B-spline with different order may be used for each component of the control* u*.*
- *In this approach we first chose a parametrization of the control and afterwards discretized the differential equation using that control approximation.*

  The other way around is also possible: First choose a discretization scheme for the differential equation (e.g. some Runge-Kutta method) and consider the required function values of u *therein as optimization variables.*

Reduced Direct Multiple Shooting Method

Modified Euler's method

0	0	0
$1/2$	$1/2$	0
	0	1

requires function values of u at t_i and $t_i + h/2$.

Instead of a priori using a piecewise constant control approximation, it could be a good idea not to fix the control approximation too early.

Hence, u_i and $u_{i+1/2}$ are considered as independent optimization variables.

But: An earlier example showed especially for modified Euler's method that this strategy may fail!

7.9 Grid Refinement

Grid Refinement

So far: Solution on a fixed grid
Goal: Grid refinement depending on discretization error
Assumptions:

- We use a Runge-Kutta method of order p for the discretization of the differential equation of the optimal control problem. In addition, we have a second method of order $p+1$ (ideally, an imbedded Runge-Kutta method).
- On the grid
$$\mathbb{G} := \{t_0 < t_1 < \ldots < t_N = t_f\}$$
a numerical solution x_i, $i = 0, \ldots, N$ of the discretized optimal control problem is given.

Grid Refinement

Estimation of the local error: (cf. section on automatic step-size selection)

$$\text{err} := \|\eta(t+h) - \bar{\eta}(t+h)\|$$

Notation: $\eta(t+h)$: solution of one step of the Runge-Kutta method of order p $\bar{\eta}(t+h)$: solution of one step of the Runge-Kutta method of order $p+1$

Grid Refinement

Estimation of local error for existing solution x_i:
In every grid interval $[t_i, t_{i+1}]$, $i = 0, \ldots, N-1$ it holds

$$\text{err}_i := \|x_{i+1} - \bar{\eta}_{i+1}\|, \qquad i = 0, \ldots, N-1,$$

where $\bar{\eta}_{i+1} = x_i + (t_{i+1} - t_i)\bar{\Phi}(t_i, x_i, t_{i+1} - t_i))$, $i = 0, \ldots, N-1$.

Grid Refinement

Possible goals:

- equally distributed local error:
$$\frac{\max\limits_{i=0,\ldots,N-1} \text{err}_i}{\min\limits_{i=0,\ldots,N-1} \text{err}_i} \quad \rightarrow \quad \min$$
- reducing the maximum error:
$$\max_{i=0,\ldots,N-1} \text{err}_i \quad \rightarrow \quad \min$$

Grid Refinement

The subsequent heuristic intends to minimize the maximum error.
Algorithm: Grid Refinement

(0) Let tol > 0, $h_{min} > 0$ and some initial grid $\mathbb{G} = \{t_0 < t_1 < \ldots < t_N = t_f\}$ with $N \in \mathbb{N}$ be given.

(1) Solve the discretized optimal control problem on $\mathbb{G}$ and compute the local errors err_i, $i = 0, \ldots, N-1$.

(2) If $\max\limits_{i=0,\ldots,N-1} \text{err}_i \leq$ tol, then stop (the discretization error is within the tolerance).

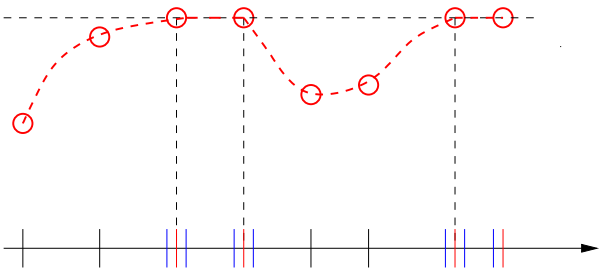
(3) For $i = 0, \ldots, N-1$ add the grid point $\frac{t_i + t_{i+1}}{2}$ to $\mathbb{G}$, whenever $\text{err}_i >$ tol and $|t_{i+1} - t_i| > h_{min}$ hold for the interval $[t_i, t_{i+1}]$.

(4) Let N be the number of intervals of the refined grid. Go to (2).

Grid Refinement

Additional modifications:

Add grid points close to numerical end points of active control or state constraints, since at these points the largest error has to be expected.



Alternative Approaches

- Betts and Huffman [BH98] discuss a generalization of our approach. Their approach is also based on the estimation of the local error of the employed one-step method and may add more than one grid points per refinement step within each interval.
- Further investigations of refinement strategies for state and control constraints can be found in Betts and Huffman [BH98] and Büskens [Büs98].

Alternative Approaches

- The grid refinement strategies of Laurent-Varin et al. [LVBB+04] are based on the discretization of the necessary conditions (minimum principle) by Runge-Kutta methods and the application of Newton's method. Newton's method allows to estimate the error to the exact solution numerically. This approach estimated the error in the necessary conditions; in particular, also the adjoint is included.

For Further Reading

References

[1] John T. Betts and W. P. Huffman. Mesh Refinement in Direct Transcription Methods for Optimal Control. *Optimal Control Applications and Methods*, 19:1–21, 1998.

[2] Christof Büskens. *Optimierungsmethoden und Sensitivitätsanalyse für optimale Steuerprozesse mit Steuer- und Zustandsbeschränkungen*. PhD thesis, Fachbereich Mathematik, Westfälische Wilhems-Universität Münster, 1998.

[3] J. Laurent-Varin, F. Bonnans, N. Berend, C. Talbot, and M. Haddou. On the refinement of discretization for optimal control problems. *IFAC Symposium on Automatic Control in Aerospace, St. Petersburg*, 2004.

7.10 Dynamic Programming

7.10.1 The Discrete Case

Dynamic Programming for Discrete Optimal Control Problems

Problem 7.40 (Discrete Optimal Control Problem). *Minimize*

$$F(x,u) = \sum_{j=0}^{N} f_0(t_j, x(t_j), u(t_j)),$$

w.r.t. the grid functions $x : \mathbb{G} \to \mathbb{R}^{n_x}$ *and* $u : \mathbb{G} \to \mathbb{R}^{n_u}$ *subject to*

$$\begin{array}{rcll} x(t_{j+1}) & = & f(t_j, x(t_j), u(t_j)), & j = 0, 1, \ldots, N-1, \\ x(t_j) & \in & X(t_j), & j = 0, 1, \ldots, N, \\ u(t_j) & \in & U(t_j, x(t_j)), & j = 0, 1, \ldots, N. \end{array}$$

Examples: inventory problems, discretized optimal control problems

Dynamic Programming for Discrete Optimal Control Problems

Remark 7.41. *Usually, the sets* $X(t_j)$ *are given in terms of inequalities and equalities:*

$$X(t_j) = \{x \in \mathbb{R}^{n_x} \mid g(t_j, x) \leq 0,\ h(t_j, x) = 0\}.$$

Accordingly, the sets $U(t_j, x)$ *often are given by*

$$U(t_j, x) = \{u \in \mathbb{R}^{n_u} \mid \tilde{g}(t_j, x, u) \leq 0,\ \tilde{h}(t_j, x, u) = 0\}.$$

Important special case: box constraints

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

Optimality Principle of Bellman

Notation:

- fixed time point $t_k \in \{t_0, t_1, \ldots, t_N\}$
- $\mathbb{G}_k := \{t_j \mid j = k, k+1, \ldots, N\}$
- $\hat{x} \in X(t_k)$ feasible

Optimality Principle of Bellman

Consider the family of discrete optimal control problems:

Problem 7.42 (Discrete Optimal Control Problem $P(t_k, \hat{x})$). *Minimize*

$$\sum_{j=k}^{N} f_0(t_j, x(t_j), u(t_j))$$

w.r.t. the grid functions $x : \mathbb{G}_k \to \mathbb{R}^{n_x}$ *and* $u : \mathbb{G}_k \to \mathbb{R}^{n_u}$ *subject to*

$$\begin{array}{rcll} x(t_{j+1}) & = & f(t_j, x(t_j), u(t_j)), & j = k, 1, \ldots, N-1, \\ x(t_k) & = & \hat{x}, & \\ x(t_j) & \in & X(t_j), & j = k, k+1, \ldots, N, \\ u(t_j) & \in & U(t_j, x(t_j)), & j = k, k+1, \ldots, N. \end{array}$$

Optimal Value Function

Definition 7.43 (Optimal Value Function). Let $t_k \in \mathbb{G}$. For $\hat{x} \in X(t_k)$ let $V(t_k, \hat{x})$ denote the optimal objective function value of problem $P(t_k, \hat{x})$. For $\hat{x} \notin X(t_k)$ define $V(t_k, \hat{x}) = \infty$.

The function $V : \mathbb{G} \times \mathbb{R}^{n_x} \to \mathbb{R}$, $(t_k, \hat{x}) \mapsto V(t_k, \hat{x})$ is called optimal value function of the discrete optimal control problem.

Optimality Principle of Bellman

It holds

Theorem 7.44 (Optimality Principle of Bellman). *Let* $\hat{x}(\cdot)$ *and* $\hat{u}(\cdot)$ *be optimal for problem 7.40.*
Then, $\hat{x}|_{\mathbb{G}_k}$ *and* $\hat{u}|_{\mathbb{G}_k}$ *are optimal for* $P(t_k, \hat{x}(t_k))$.

Optimality Principle of Bellman: Proof

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

$$\sum_{j=k}^{N} f_0(t_j, \tilde{x}(t_j), \tilde{u}(t_j)) < \sum_{j=k}^{N} f_0(t_j, \hat{x}(t_j), \hat{u}(t_j))$$

and $\tilde{x}(t_k) = \hat{x}(t_k)$.

Optimality Principle of Bellman: Proof

Hence, $x : \mathbb{G} \to \mathbb{R}^{n_x}$ and $u : \mathbb{G} \to \mathbb{R}^{n_u}$ defined by

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

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

are feasible for problem 7.40 and satisfy

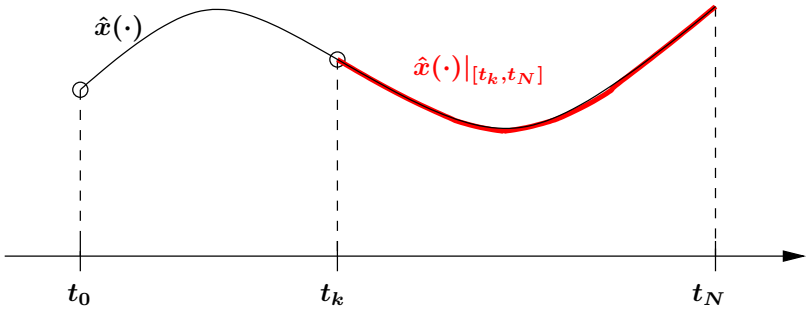
$$\sum_{j=0}^{k-1} f_0(t_j, \hat{x}(t_j), \hat{u}(t_j)) + \sum_{j=k}^{N} f_0(t_j, \tilde{x}(t_j), \tilde{u}(t_j)) < \sum_{j=0}^{N} f_0(t_j, \hat{x}(t_j), \hat{u}(t_j)).$$

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

Optimality Principle of Bellman

The decisions in the time periods $k, k+1, \ldots, N$ of problem 7.40 for given x_k are independent of the decisions in the period $t_0, t_1, \ldots, t_{k-1}$:

Bellman's Optimality Principle: Remaining optimal trajectories remain optimal



Optimality Principle of Bellman

Essential assumptions:

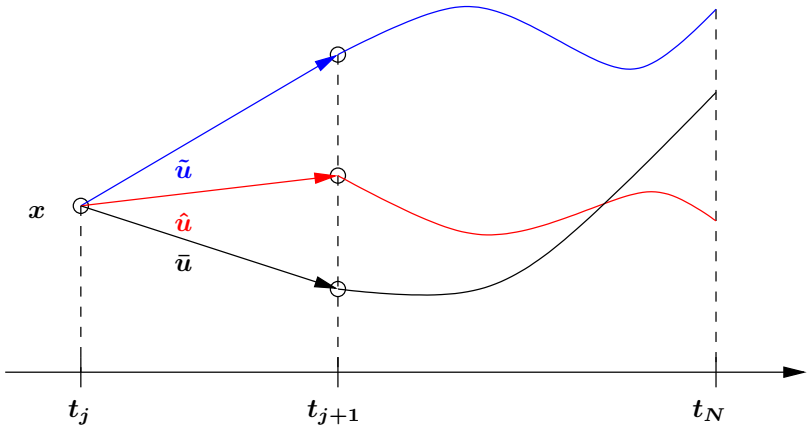
Stagewise dependencies:

- $x(t_{j+1})$ depends only on $t_j, x(t_j), u(t_j)$!
- The objective function is separable!
- The constraints refer only to the grid point t_j!

The optimality principle does not hold, if, e.g., $\varphi(x(t_0), x(t_N))$ is to be minimized or $\psi(x(t_0), x(t_N)) = 0$ is supposed to hold!

Dynamic Programming Method

Bellman's Method of Dynamic Programming: Recursion of optimal value function



Recursion for Optimal Value Function

Optimal value function for $P(t_N, x)$):

$$V(t_N, x) = \min_{u \in U(t_N, x)} f_0(t_N, x, u). \tag{76}$$

Assumption: The optimal value function $V(t_{j+1}, x)$ is known for every $x \in \mathbb{R}^{n_x}$.
Optimality principle:

$$V(t_j, x) = \min_{u \in U(t_j, x)} \{f_0(t_j, x, u) + V(t_{j+1}, f(t_j, x, u)\}, \quad j = 0, 1, \ldots, N-1. \tag{77}$$

This is a recursion for the optimal value function; backwards in time!
Optimal initial value $\hat{x}(t_0)$ of problem 7.40:

$$\hat{x}(t_0) = \arg \min_{x \in X(t_0)} V(t_0, x). \tag{78}$$

Dynamic Programming Method (Version I)

(i) Backward

1. Let $V(t_N, x)$ be given by (76).
2. For $j = N-1, \ldots, 0$: Compute $V(t_j, x)$ as in (77).

(ii) Forward

1. Let $\hat{x}(t_0)$ be given by (78).
2. For $j = 0, 1, \ldots, N-1$: Determine

$$\hat{u}(t_j) = \arg \min_{u \in U(t_j, \hat{x}(t_j))} \{f_0(t_j, \hat{x}(t_j), u) + V(t_{j+1}, f(t_j, \hat{x}(t_j), u))\}$$

and let $\hat{x}(t_{j+1}) = f(t_j, \hat{x}(t_j), \hat{u}(t_j))$.

3. Determine $\hat{u}(t_N) = \arg \min_{u \in U(t_N, \hat{x}(t_N))} f_0(t_N, \hat{x}(t_N), u)$.

Dynamic Programming Method (Version II)

(i) Backward

1. Let $V(t_N, x)$ be given by (76) and $u^*(t_N, x)$ the corresponding optimal control.
2. For $j = N-1, \ldots, 0$: Compute $V(t_j, x)$ as in (77). Let $u^*(t_j, x)$ be the corresponding optimal control at t_j and x (feedback control!).

(ii) Forward

1. Let $\hat{x}(t_0)$ be given by (78).
2. For $j = 0, 1, \ldots, N-1$: Determine $\hat{u}(t_j) = u^*(t_j, \hat{x}(t_j))$ and
$$\hat{x}(t_{j+1}) = f(t_j, \hat{x}(t_j), \hat{u}(t_j)).$$
3. Determine $\hat{u}(t_N) = u^*(t_N, \hat{x}(t_N))$.

Remarks

Remark 7.45. *Both versions yield an optimal solution of the discrete optimal control problem 7.40. Version II is more preferable for calculations by hand, since it produces an optimal feedback control law u^* as a function of time and state (⟶ suitable for controllers!).*

Version I is more preferable for computer implementations, since it is not necessary to store the feedback control law u^*. Hence, version I needs less memory, but the result of the algorithm is only the optimal trajectory of x and u as a function of time (open-loop control).

Remarks

Remark 7.46. *The main disadvantage of the dynamic programming method is the so-called curse of dimensions.*

In the worst case it is necessary to investigate every discrete trajectory. Nevertheless, the method works well for low dimensions or special problems like assignment problems, knapsack problems, inventory problems with integer data.

For Further Reading

References

[Bel57] Bellman, R. E. *Dynamic Programming*. University Press, Princeton, New Jersey, 1957.

[BD71] Bellman, R. E. and Dreyfus, S. E. *Applied Dynamic Programming*. University Press, Princeton, New Jersey, 1971.

[BG93] Bomze, I. M. and Grossmann, W. *Optimierung - Theorie und Algorithmen*. BI-Wissenschaftsverlag, Mannheim, 1993.

[NM02] Neumann, K. and Morlock, M. *Operations Research*. Carl Hanser Verlag, München Wien, 2002.

[Win04] Winston, W. L. *Operations Research: Applications and Algorithms*. Brooks/Cole–Thomson Learning, Belmont, 4th edition, 2004.

7.10.2 The Continuous Case

Optimal Control Problems and Hamilton-Jacobi-Bellman Equation

We consider optimal control problems starting at $t_* \in [t_0, t_f]$ in $x_* \in \mathbb{R}^{n_x}$:

Problem 7.47 (Optimal Control Problem $P(t_*, x_*)$). *Find $x \in W^{1,\infty}([t_*, t_f], \mathbb{R}^{n_x})$ and $u \in L^\infty([t_*, t_f], \mathbb{R}^{n_u})$, such that*

$$F(u; t_*, x_*) := \varphi(x(t_f)) + \int_{t_*}^{t_f} f_0(t, x(t), u(t))dt \tag{79}$$

is minimized subject to

$$\begin{array}{rcll} \dot{x}(t) & = & f(t, x(t), u(t)) & \text{a.e. in } [t_*, t_f], \\ x(t_*) & = & x_*, & \\ u(t) & \in & \mathcal{U} & \text{a.e. in } [t_*, t_f]. \end{array}$$

Optimal Value Function

Definition 7.48 (Optimal Value Function). For $(t_*, x_*) \in [t_0, t_f] \times \mathbb{R}^{n_x}$ the optimal value function is defined as

$$V(t_*, x_*) := \inf_{u:[t_*,t_f]\to\mathcal{U}} F(u; t_*, x_*).$$

$\hat{u} : [t_*, t_f] \to \mathcal{U}$ is called optimal for$P(t_*, x_*)$ if $V(t_*, x_*) = F(\hat{u}; t_*, x_*)$.

Recursion of Optimal Value Function

Similar as in the discrete case, there is also a continuous version of of Bellman's recursion:

Theorem 7.49 (Optimality Principle of Bellman). *For every $(t_*, x_*) \in [t_0, t_f] \times \mathbb{R}^{n_x}$ the optimal value function satisfies the recursion*

$$V(t_*, x_*) = \inf_{u:[t_*,s]\to\mathcal{U}} \left(\int_{t_*}^{s} f_0(\tau, x(\tau), u(\tau))d\tau + V(s, x(s)) \right) \quad \forall s \in [t_*, t_f),$$

where x is given by $\dot{x}(\tau) = f(\tau, x(\tau), u(\tau))$, $x(t_*) = x_*$ and it holds $V(t_f, x_*) = \varphi(x_*)$.

Recursion of Optimal Value Function

Theorem 7.49 (continued). *If $\hat{u}$ is optimal, then*

$$V(t_*, x_*) = \int_{t_*}^{s} f_0(\tau, \hat{x}(\tau), \hat{u}(\tau))d\tau + V(s, \hat{x}(s)), \qquad \forall s \in [t_*, t_f), \tag{80}$$

where $\hat{x}$ denotes the state corresponding to $\hat{u}$.

Proof

Proof. First we show „$\leq$“:

Let $s \in [t_*, t_f)$, $u : [t_*, t_f] \to \mathcal{U}$, and $\tilde{u} : [s, t_f) \to \mathcal{U}$ be arbitrary. Define

$$\bar{u}(\tau) = \begin{cases} u(\tau), & \text{if } \tau \in [t_*, s), \\ \tilde{u}(\tau), & \text{if } \tau \in [s, t_f). \end{cases}$$

Obviously, $\bar{u}(\tau) \in \mathcal{U}$ for all $\tau \in [t_*, t_f)$.

The corresponding $\bar{x}$ of $\dot{x}(\tau) = f(\tau, x(\tau), \bar{u}(\tau))$, $x(t_*) = x_*$ satisfies

$$\bar{x}(\tau) = \begin{cases} x(\tau), & \text{if } \tau \in [t_*, s), \\ \tilde{x}(\tau), & \text{if } \tau \in [s, t_f). \end{cases}$$

Proof

By the definition of V it holds

$$\begin{aligned} V(t_*, x_*) &\leq \int_{t_*}^{t_f} f_0(\tau, \bar{x}(\tau), \bar{u}(\tau)) d\tau + \varphi(\bar{x}(t_f)) \\ &= \int_{t_*}^{s} f_0(\tau, x(\tau), u(\tau)) d\tau + F(\tilde{u}; s, x(s)). \end{aligned}$$

Passing to the infimum over all $\tilde{u} : [s, t_f] \to \mathcal{U}$ and all $u : [t_*, t_f] \to \mathcal{U}$ yields

$$V(t_*, x_*) \leq \inf_{u:[t_*,t_f]\to\mathcal{U}} \left(\int_{t_*}^{s} f_0(\tau, x(\tau), u(\tau)) d\tau + V(s, x(s)) \right). \tag{81}$$

Proof

Now, let $\varepsilon > 0$ arbitrary and $u : [t_*, t_f] \to \mathcal{U}$ with $F(u; t_*, x_*) \leq V(t_*, x_*) + \varepsilon$. Then,

$$\begin{aligned} V(t_*, x_*) + \varepsilon &\geq F(u; t_*, x_*) = \int_{t_*}^{s} f_0(\tau, x(\tau), u(\tau)) d\tau + F(u; s, x(s)) \\ &\geq \int_{t_*}^{s} f_0(\tau, \bar{x}(\tau), u(\tau)) d\tau + V(s, x(s)) \\ &\overset{(81)}{\geq} V(t_*, x_*). \end{aligned}$$

For optimal $\hat{u}$ above inequality holds with $\varepsilon = 0$. □

Recursion of Optimal Value Function

Using Bellman's recursion it is possible to derive a partial differential equation for the optimal value function.

Theorem 7.50 (Hamilton-Jacobi-Bellman Equation). *Let $\hat{u}$ be optimal. If the optimal value function is differentiable at $(t, x) \in [t_0, t_f) \times \mathbb{R}^{n_x}$ and $\hat{u}$ is continuous at t, then it holds*

$$\frac{\partial V}{\partial t}(t, x) + \mathcal{H}\left(t, x, \hat{u}(t), \frac{\partial V}{\partial x}(t, x)\right) = 0$$

where $\mathcal{H}$ denotes the Hamiltonian.

Proof

Proof. From (80) it follows

$$\frac{V(t + h, \hat{x}(t + h)) - V(t, \hat{x}(t))}{h} = -\frac{1}{h} \int_{t}^{t+h} f_0(\tau, \hat{x}(\tau), \hat{u}(\tau)) d\tau.$$

Passing to the limit $h \to 0$ yields the assertion. □

Recursion of Optimal Value Function

The solvability of the Hamilton-Jacobi-Bellman equation is even sufficient for optimality:

Theorem 7.51 (Sufficient Optimality Condition). *Let $W : [t_0, t_f] \times \mathbb{R}^{n_x} \to \mathbb{R}$ be continuously differentiable with*

$$\begin{aligned} W(t_f, x) &= \varphi(t_f, x), \qquad \forall x \in \mathbb{R}^{n_x}, \\ \frac{\partial W}{\partial t}(t, x) + \inf_{u \in \mathcal{U}} \mathcal{H}\left(t, x, u, \frac{\partial W}{\partial x}(t, x)\right) &= 0, \qquad \forall (t, x) \in [t_0, t_f) \times \mathbb{R}^{n_x}. \end{aligned}$$

Then:

$$W(t, x) \leq V(t, x) \qquad \forall (t, x) \in [t_0, t_f) \times \mathbb{R}^{n_x}.$$

If in addition $\hat{u}$ assumes the infimum for a.e. $t \in [t_0, t_f]$, then $\hat{u}$ is optimal for $P(t, x)$ and $W(t, x) = V(t, x)$.

Proof

Proof. Let $(t,x) \in [t_0,t_f) \times \mathbb{R}^{n_x}$ and $u : [t,t_f] \to \mathcal{U}$ arbitrary. Then:

$$\begin{aligned} F(u;t,x) &= \int_t^{t_f} f_0(\tau,x(\tau),u(\tau))d\tau + \varphi(x(t_f)) \\ &= \int_t^{t_f} f_0(\tau,x(\tau),u(\tau))d\tau + W(t,x) + \underbrace{\int_t^{t_f} \frac{d}{d\tau} W(\tau,x(\tau))d\tau}_{=W(t_f,x(t_f))-W(t,x)=\varphi(x(t_f))-W(t,x)} \\ &= \int_t^{t_f} f_0(\tau,x(\tau),u(\tau))d\tau + W(t,x) \\ &\quad + \int_t^{t_f} \left[\frac{\partial W}{\partial t}(\tau,x(\tau)) + \frac{\partial W}{\partial x}(\tau,x(\tau)) f(\tau,x(\tau),u(\tau)) \right] d\tau \\ &\geq W(t,x). \end{aligned}$$

Passing to the infimum over all u yields $V(t,x) \geq W(t,x)$.

Proof

If $\hat{u}$ yields the infimum of the Hamiltonian w.r.t. u, then analogously it follows

$$V(t,x) = F(\hat{u};t,x) = W(t,x) \leq V(t,x)$$

and $\hat{u}$ is optimal. □

Construction of Optimal Feedback Controls

Using the Hamilton-Jacobi-Bellman equation it is possible to construct feedback controls:

- Determine $u^*(t,x,\lambda)$ from the (global) minimum principle:

$$u^*(t,x,\lambda) = \arg\min_{u \in \mathcal{U}} \mathcal{H}(t,x,u,\lambda).$$

- Solve the Hamilton-Jacobi-Bellman equation (if possible).
- Compute the optimal state $\hat{x}$ by solving

$$\dot{x}(t) = f(t,x(t),u^*(t,x(t),W'_x(t,x(t)))), \qquad x(t_0) = x_0$$

  and the optimal control $\hat{u}$ according to

$$\hat{u}(t) = u^*(t,\hat{x}(t),W'_x(t,\hat{x}(t))).$$

Remarks

Remark 7.52.

- *Unfortunately, the optimal value function often is not differentiable, especially in the presence of control and state constraints.*

  Exception: convex linear-quadratic optimal control problems
- *The computational effort for solving the Hamilton-Jacobi-Bellman equation (partial differential equation) is very high. Hence, the method is only numerically reasonable for low state dimensions,*
- *With $\lambda(t) := \frac{\partial V}{\partial x}(t,\hat{x}(t))$ the adjoint λ can be interpreted as sensitivity of the optimal value function w.r.t. $\hat{x}$. In economical sciences λ is known as shadow price.*

8 Examples and Applications Revisited

Contents

The Brachistochrone-Problem

Reachable Set for Brachistochrone-Problem

differential inclusion (coordinate change for $y_2(\cdot)$)

$$\begin{aligned} x'(t) &\in F(t, x(t)) \quad \text{for a.e. } t \in [0,1] \\ x(0) &\in X_0 \end{aligned}$$

with the data

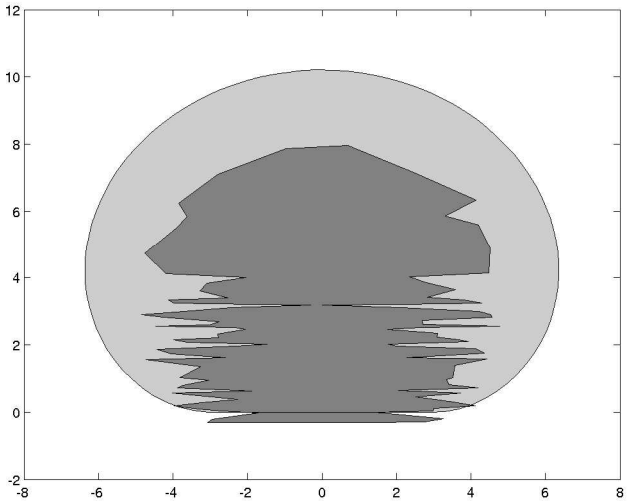
$$F(t,x) = \{f(t,x,u) \,|\, u \in \mathcal{U}\}, \quad f(t,x,u) = \begin{pmatrix} \sqrt{2\gamma x_2}\cos(u) \\ \sqrt{2\gamma x_2}\sin(u) \end{pmatrix},$$

$$\mathcal{U} = [-\pi, +\pi], \qquad X_0 = \{\begin{pmatrix} 0 \\ 1 \end{pmatrix}\}$$

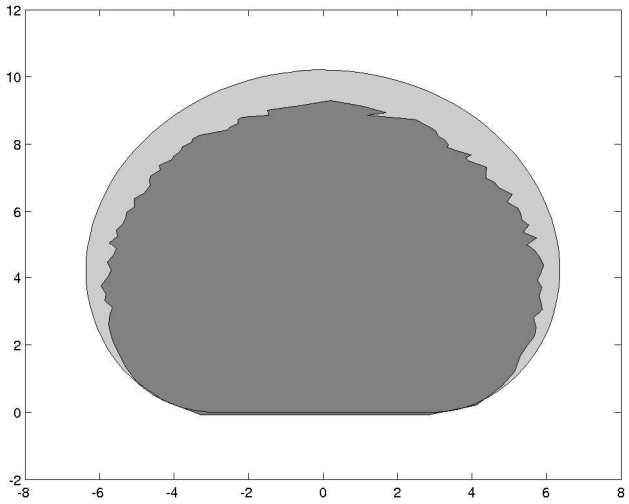
calculations by I. A. Chahma show the convergence of (nonlinear) Euler's method for $N = 4, 8, 16, 32$ (reference set uses $N = 64$)

The Brachistochrone-Problem

Approximation of Reachable Set with $N = 4$



Approximation of Reachable Set with $N = 8$



Hager's Counter Example

Optimal Control Problem
Minimize

$$J(x,u) = \frac{1}{2}\int_0^1 (u(t)^2 + 2x(t)^2)dt$$

subject to

$$\begin{aligned} x'(t) &= \frac{1}{2}x(t) + u(t) \quad \text{(a.e. } t \in [0,1]\text{)}, \\ x(0) &= 1 \end{aligned}$$

The optimal solution is

$$x^*(t) = \frac{2e^{3t} + e^3}{e^{\frac{3}{2}t}(2+e^3)}, \quad u^*(t) = \frac{2(e^{3t} - e^3)}{e^{\frac{3}{2}t}(2+e^3)}.$$

Hager's Counter Example

Disturbance of Optimal Control
disturbed differential inclusion

$$\begin{aligned} x'(t) &\in A(t)x(t) + B(t)U(t) \quad \text{for a.e. } t \in [0,1] \\ x(0) &\in X_0 \end{aligned}$$

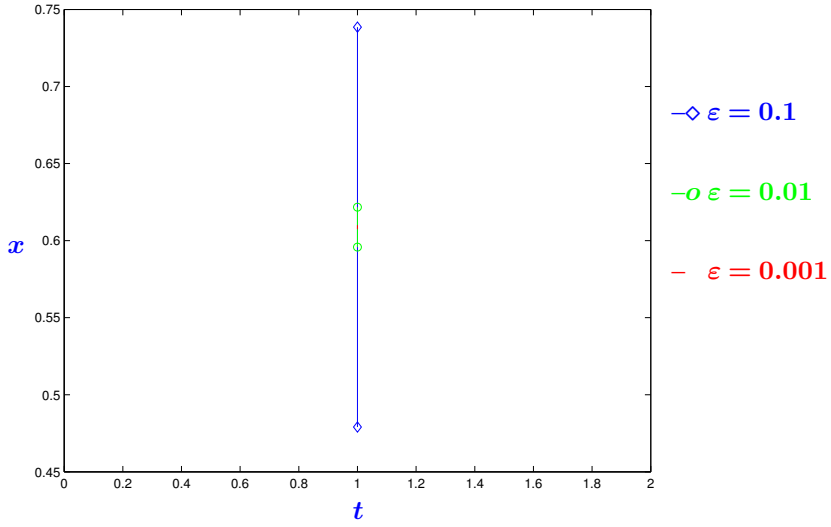
with the data

$$\begin{aligned} A(t) &= \left(\tfrac{1}{2}\right), & B(t) &= (1), \\ U(t) &= [u^*(t) - \varepsilon, u^*(t) + \varepsilon], & X_0 &= \{1\} \end{aligned}$$

$\mathcal{R}(t_f, 0, X_0) \subset \mathbb{R}$ contains the end-value $x^*(1)$
calculations use $\varepsilon = 0.1, 0.01, 0.001$

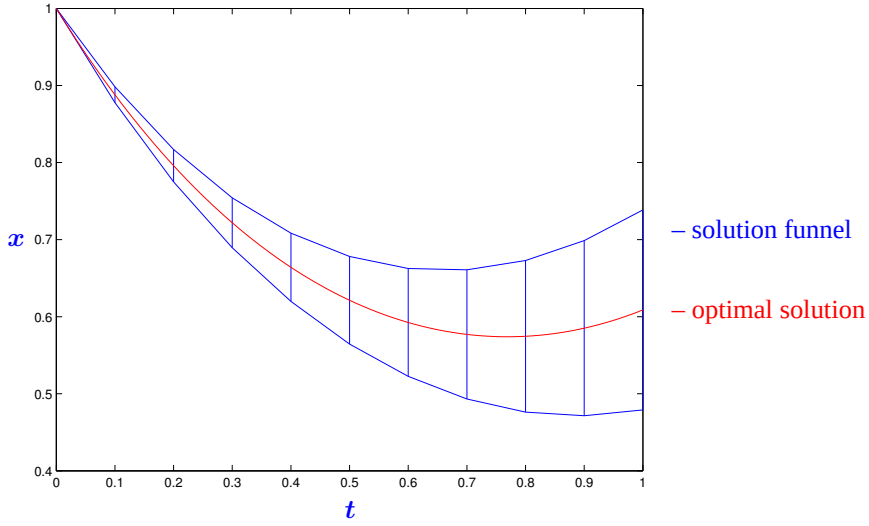
Hager's Counter Example

Reachable Set for $\varepsilon = 0.1, 0.01, 0.001$, 1D



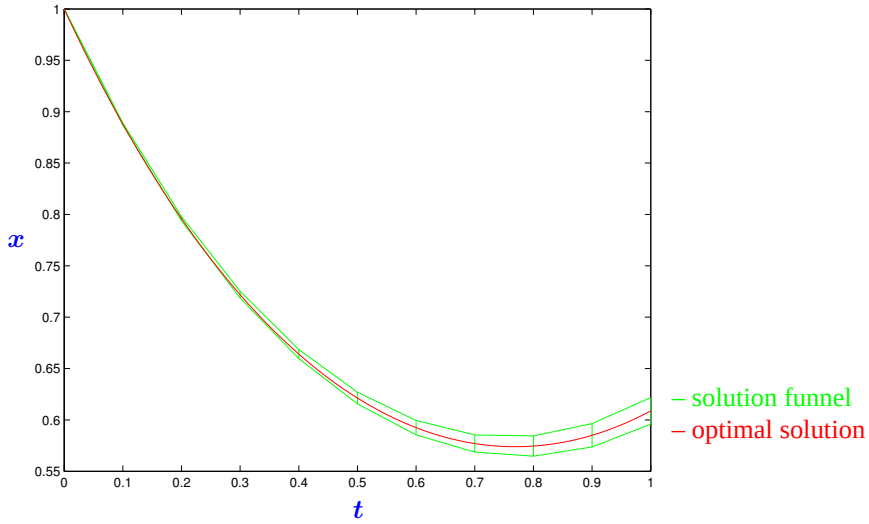
Hager's Counter Example

Solution Funnel for $\varepsilon = 0.1$, 1D



Hager's Counter Example

Solution Funnel for $\varepsilon = 0.01$, 1D



Differential Inclusion Motivated by Uncertainty

Uncertain Parameters in Oscillation
forced undamped oscillation

$$\begin{aligned} y''(t) + 4y(t) &= \sin(t) \cdot u(t) && \text{for a.e. } t \in [0,10], \\ \|u(t) - u_0\|_2 &\le \varepsilon_1 && \text{for a.e. } t \in [0,10], \\ \|y(0)\|_p &\le \varepsilon_2 \end{aligned}$$

u_0 theoretical amplitude of the force,
y_0 theoretical starting value,
$u(t)$ uncertain, bounded amplitude of the force,
ε_1 error bounds for amplitude, ε_2 error bounds for starting value, $p \in \{2, \infty\}$ chooses the norm

Differential Inclusion Motivated by Uncertainty
corresponding differential inclusion

Uncertain Parameters in Oscillation
differential inclusion

$$\begin{aligned} x'(t) &\in A(t)x(t) + B(t)U && \text{for a.e. } t \in [0,10] \\ x(0) &\in X_0 \end{aligned}$$

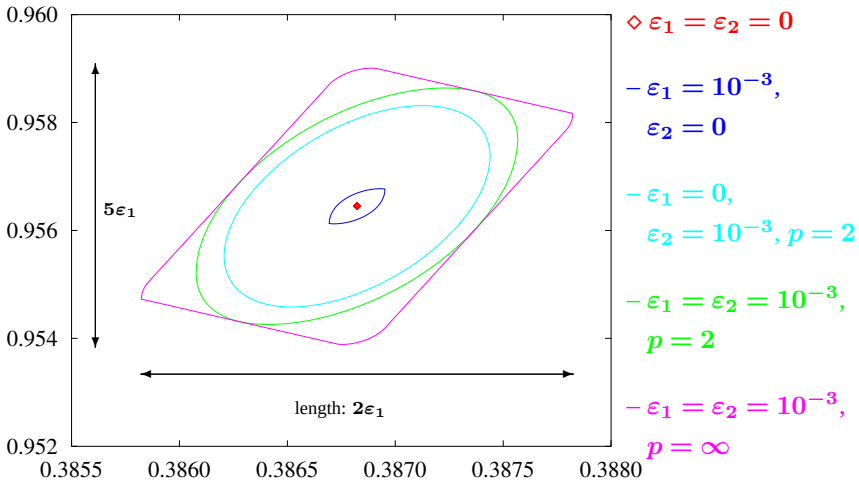
with the data

$$A(t) = \begin{pmatrix} 0 & 1 \\ -4 & 0 \end{pmatrix}, \qquad B(t) = \begin{pmatrix} 0 \\ \sin(t) \end{pmatrix},$$
$$U = [u_0 - \varepsilon_1, u_0 + \varepsilon], \qquad X_0 = B_{\varepsilon_2}(x_0) \text{ or } [-\varepsilon_2, \varepsilon_2]^2$$

calculations use $u_0 = 3$, $\varepsilon_1 \in \{0, 10^{-3}\}$, $x_0 = \binom{0}{0}$, $\varepsilon_2 \in \{0, \varepsilon_1\}$

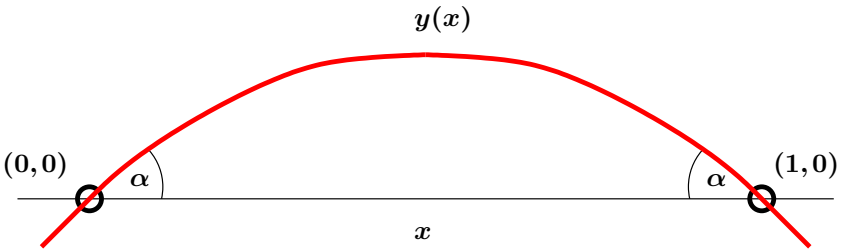
Differential Inclusion Motivated by Uncertainty

Reachable Set for Different Uncertainties



Minimum-Energy Problem

A rod is fixed at the points $(0,0)$ and $(1,0)$ in the (x,y)-plane in such a way, that it assumes an angle α w.r.t. the x-axis:



What curve yields a minimum of the rod's bending energy?

Minimum-Energy Problem

Reachable Set
differential inclusion

$$x'(t) \in A(t)x(t) + B(t)U \quad \text{for a.e. } t \in [0,1]$$
$$x(0) \in X_0$$

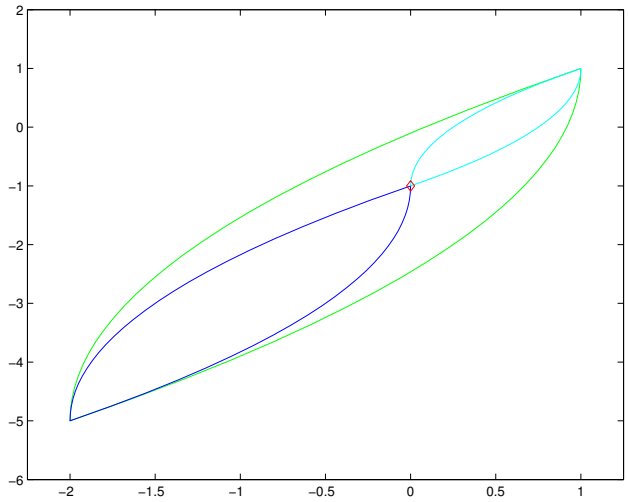
with the data

$$A(t) = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad B(t) = \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$
$$U = [u_1, u_2], \qquad X_0 = \{x_0\}$$

calculations use $U \subset [-6, 0]$,
$x_0 = \binom{0}{0}, \binom{0}{1}$, i.e. $\alpha = 0, \frac{\pi}{4}$

Minimum-Energy Problem

Reachable Set for several control sets $\mathcal{U}$



$\diamond\, \mathcal{U} = \{-2\}$

$-\mathcal{U} = [-6, -2]$

$-\mathcal{U} = [-2, 0]$

$-\mathcal{U} = [-6, 0]$

Minimum-Energy Problem

Modification: additional constraint $y(x) \leq y_{max}$

Minimum-Energy Problem

Minimize

$$J(y_1, y_2, u) = \int_0^1 u(x)^2 dx$$

subject to

$$\begin{aligned} y_1'(x) &= y_2(x), \qquad y_1(0) = y_1(1) = 0, \\ y_2'(x) &= u(x), \qquad y_2(0) = -y_2(1) = \tan\alpha. \end{aligned}$$

and the state constraint

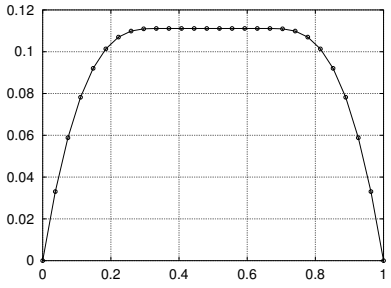
$$y_1(x) - y_{max} \leq 0.$$

Minimum-Energy Problem

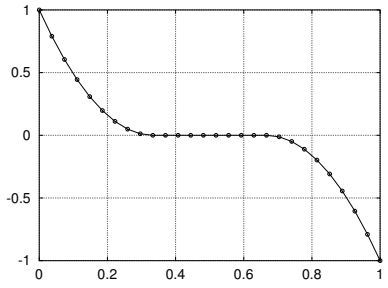
Numerical solution

for $y_{max} = 1/9$, $\alpha = 45°$, equidistant grid with $N = 27$, classical Runge-Kutta method:

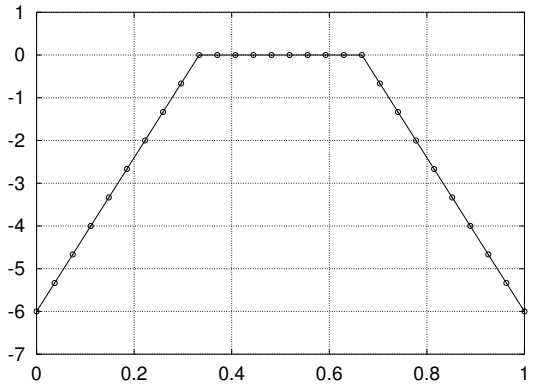
state $y_1(x)$:



state $y_2(x)$:



control $u(x)$:



Vertical Ascent of a Rocket

Reachable Set for Rocket Problem
differential inclusion (without end conditions and fuel consumption)

$$x'(t) \in A(t)x(t) + B(t)U + C(t) \quad \text{for a.e. } t \in [0,10]$$
$$x(0) \in X_0$$

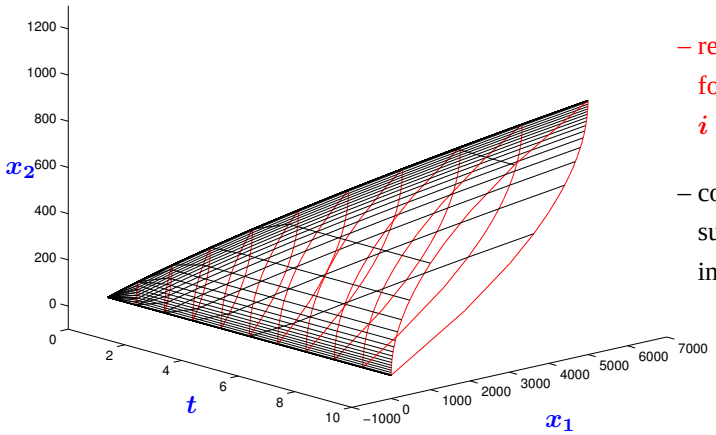
with the data

$$A(t) = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad B(t) = \begin{pmatrix} 0 \\ \frac{1}{m} \end{pmatrix}, \quad C(t) = \begin{pmatrix} 0 \\ -g \end{pmatrix},$$
$$U = [0, u_{\max}], \qquad X_0 = \{\begin{pmatrix} 0 \\ 0 \end{pmatrix}\}$$

calculations use $u_{\max} = 100$, normalized mass $m = 1$, $g = 9.81$

Vertical Ascent of a Rocket

Solution Funnel, 2D, Viewpoint (48, 18)

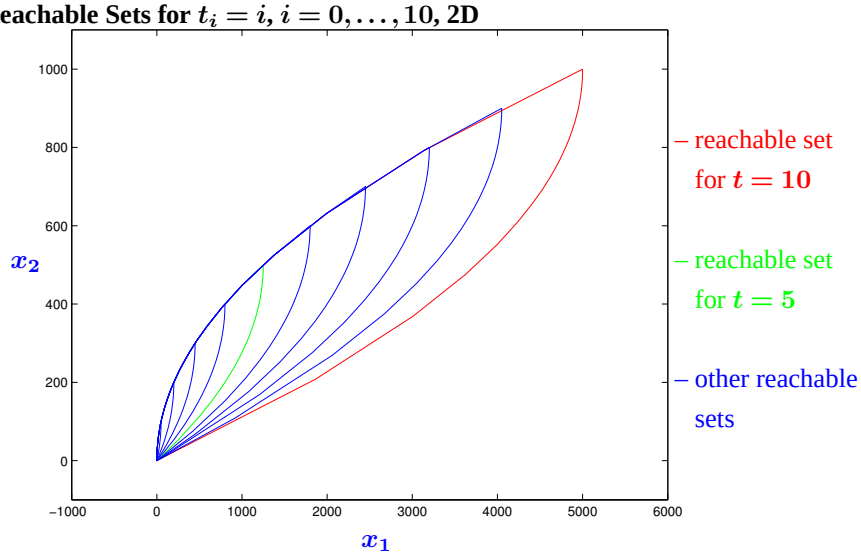


– reachable sets for $t_i = i$, $i = 0, \ldots, 10$

– corresponding supporting points in same direction

Vertical Ascent of a Rocket

All Reachable Sets for $t_i = i$, $i = 0, \ldots, 10$, 2D



Vertical Ascent of a Rocket

Reachable Set

differential inclusion (reverse time $s(t) = 10 - t$, $y(t) = x(s(t))$)

$$y'(t) \in \widetilde{A}(t)y(t) + \widetilde{B}(t)U + \widetilde{C}(t) \quad \text{for a.e. } t \in [0, 10]$$
$$y(0) \in Y_0$$

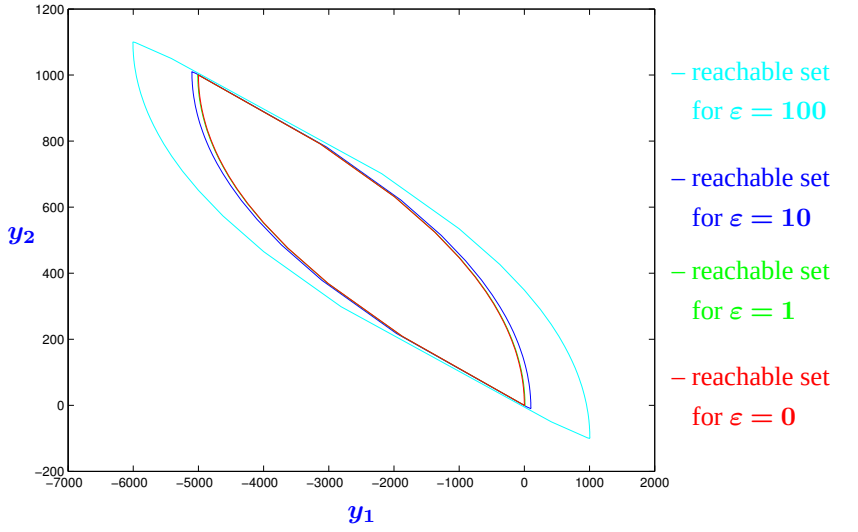
with the data

$$\widetilde{A}(t) = -A(t), \qquad \widetilde{B}(t) = -B(t), \qquad \widetilde{C}(t) = -C(t),$$
$$U = [0, u_{\max}], \qquad Y_0 = B_\varepsilon(\begin{pmatrix} 5000 \\ 1000 \end{pmatrix})$$

calculations use $u_{\max} = 100$, normalized mass $m = 1$, $g = 9.81$,
one former end value $x(10) = \binom{5000}{1000}$ and $\varepsilon = 1000, 100, 10, 1, 0$

Vertical Ascent of a Rocket

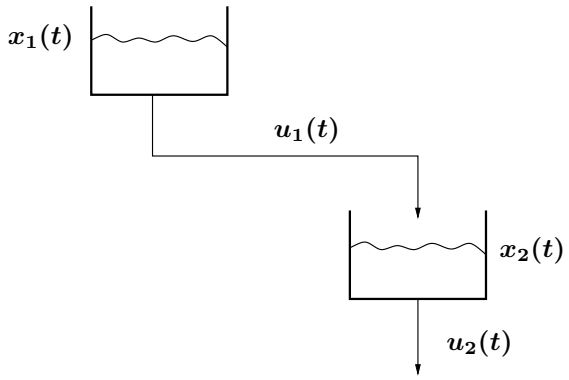
Reachable Sets for $\varepsilon = 100, 10, 1, 0$



System of two Water Boxes

Given:

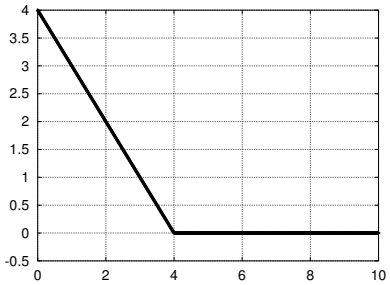
- 2 water boxes, water level $x_i(t) \geq 0$ at time t in box $i = 1, 2$
- outflow rates $0 \leq u_i(t) \leq 1$, $i = 1, 2$



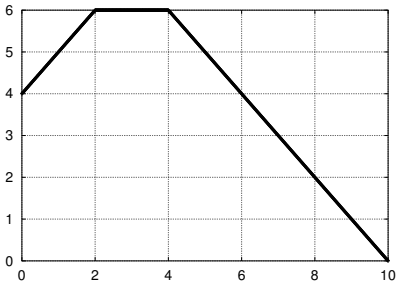
Solution

Numerical solution for $N = 999$ and piecewise constant control approximation.

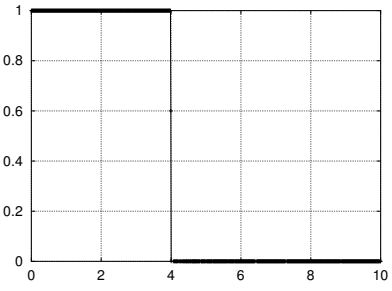
state x_1:



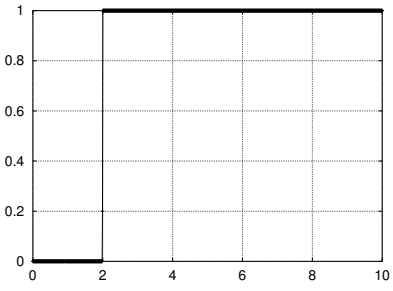
state x_2:



control u_1:



control u_2:



System of two Water Boxes

Reachable Set

artificial equation: $x_3'(t) = (10 - t) \cdot u_1(t) + t \cdot u_2(t)$

differential inclusion

$$x'(t) \in B(t)U \quad \text{for a.e. } t \in [0, 10]$$
$$x(0) \in X_0$$

with the data

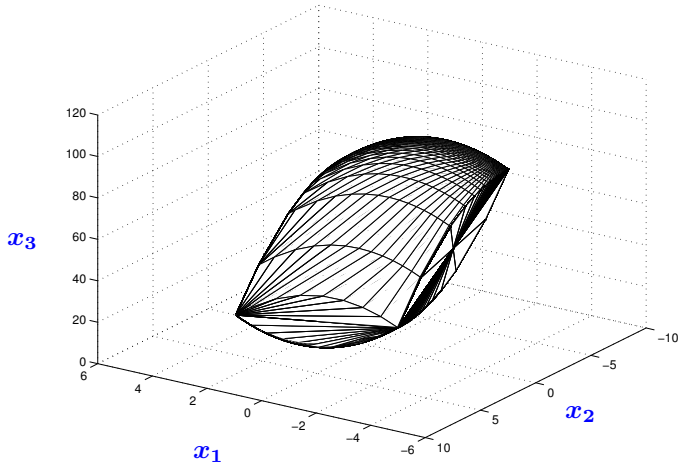
$$B(t) = \begin{pmatrix} -1 & 0 \\ 1 & -1 \\ 10-t & t \end{pmatrix}, \quad U = [0,1]^2, \quad X_0 = \{\begin{pmatrix} 5 \\ 0 \\ 0 \end{pmatrix}\}$$

$\mathcal{R}(t_f, 0, X_0) = X_0 + \int_0^{t_f} B(t)U\,dt$, calculations use $t_f = 10$

pure quadrature problem, reachable set = Aumann integral, no state constraints here!

System of two Water Boxes

Aumann Integral for $I = [0, 10]$, 3D



Climate Change Model

Model Parameters

WBGU scenario:

(WBGU = German Advisory Council on Global Change) simple model of climate change assessment:

$F(\cdot)$: cumulation of CO_2 emissions caused by humankind

$C(\cdot)$: carbon concentration

$T(\cdot)$: global mean temperature

$E(\cdot)$: CO_2 emission profile controlling the allowed CO_2 emissions

Questions:

- What are the admissible emissions in the year t?
- What are the feasible concentrations $C(t)$ in that year?
- Is it possible to realize a global mean temperature T^* in year t?

Climate Change Model

Reachable Set
control problem:

$$\begin{aligned} F'(t) &= E(t), \\ C'(t) &= B \cdot F(t) + \beta \cdot E(t) - \sigma \cdot (C(t) - C_1), \\ T'(t) &= \mu \cdot \ln\left(\frac{C(t)}{C_1}\right) - \alpha \cdot (T(t) - T_1), \\ E'(t) &= u(t)E(t), \quad |u(t)| \leq u_{\max} \end{aligned}$$

with state constraints

$$\begin{aligned} T_1 \leq\ & T(t) \leq T_{max}, \\ 0 \leq\ & T'(t) \leq T'_{crit}(T(t)), \end{aligned}$$

$$T'_{crit}(T(t)) = \begin{cases} T'_{\max} & \text{if } T_1 \leq T(t) \leq T_{\max} - 1, \\ T'_{\max}\sqrt{T_{\max} - T(t)} & \text{if } T_{\max} - 1 \leq T(t) \leq T_{\max}. \end{cases}$$

Climate Change Model

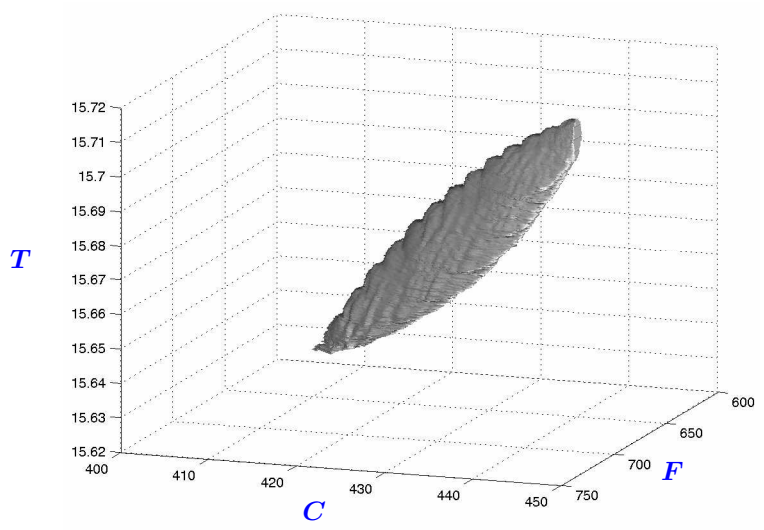
Further Model Parameters
$u_{\max}$: rate of emissions change
$T_{\max}$: maximal global mean temperature
$T_1 = 14.6$: minimal global mean temperature (preindustrial times)
$T'(\cdot)$: rate of temperatur change
$\mathcal{I} = [0, t_f]$: $t = 0$ means this year, $t_f = 30, 200$: years of forecast

further constants: B, β, σ, μ, α
starting values: $F(0) = 426$, $C(0) = 360$, $T(0) = 15.3$, $E(0) = 7.9$

calculations by I. A. Chahma show the result of (nonlinear) set-valued Euler's method for $N = 60$, comparison with optimal control software of C. Büskens

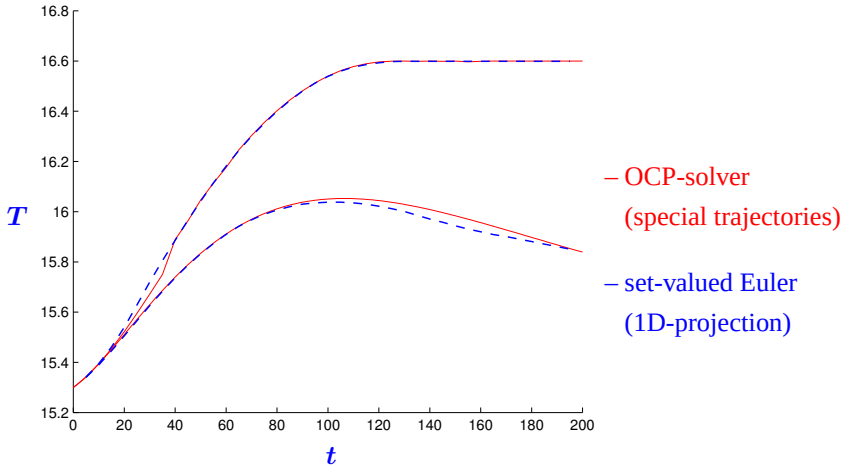
Climate Change Model

3D-projection on $F - C - T$-axes from 4D-reachable set, $t_f = 30$



Climate Change Model

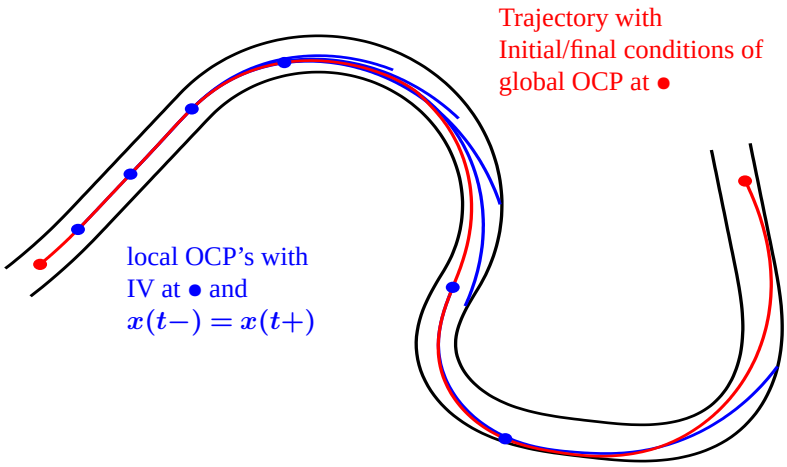
1D-projection on $T(\cdot)$ from 4D-reachable set, $t_f = 200$



Elch-Test

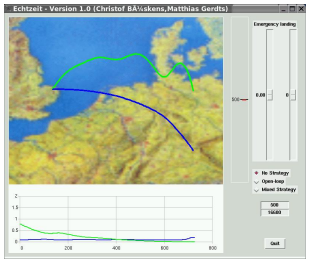
Slalom

Moving Horizon



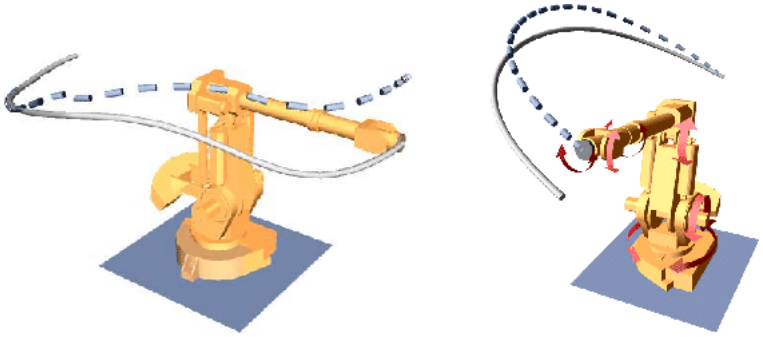
Realtime: BMW (10 %, 30 % Perturbation)

Emergency Landing Manoeuvre



- Scenario: propulsion system breakdown
- Goal: maximization of range w.r.t. current position
- Controls: lift coefficient, angle of bank
- no thrust available; no fuel consumption (constant mass)

Robots [Knauer'01,http://www.math.uni-bremen.de/∼knauer]



A Appendix

A.1 Matrix Norms

Matrix Norms

Definition A.1. A *matrix norm* is a mapping $\|\cdot\| : \mathbb{R}^{m\times n} \to [0,\infty)$ such that for all $\boldsymbol{A} \in \mathbb{R}^{m\times n}$

(i) $\|\boldsymbol{A}\| > 0$, if $\boldsymbol{A} \neq \boldsymbol{0}$

(ii) $\|\mu \cdot \boldsymbol{A}\| = |\mu| \cdot \|\boldsymbol{A}\|$

(iii) $\|\boldsymbol{A}+\boldsymbol{B}\| \leq \|\boldsymbol{A}\| + \|\boldsymbol{B}\|$ for all $\boldsymbol{B} \in \mathbb{R}^{m\times n}$

It is *submultiplicative*, if

$$\|\widetilde{\boldsymbol{A}} \cdot \widetilde{\boldsymbol{B}}\| \leq \|\widetilde{\boldsymbol{A}}\| \cdot \|\widetilde{\boldsymbol{B}}\|$$

for all $\widetilde{\boldsymbol{A}} \in \mathbb{R}^{p\times m}$, $\widetilde{\boldsymbol{B}} \in \mathbb{R}^{m\times n}$.
It is *compatible with the vector norms* $\|\cdot\|_a$ on $\mathbb{R}^n$ and $\|\cdot\|_b$ on $\mathbb{R}^m$, if

$$\|\boldsymbol{A}\boldsymbol{x}\|_b \leq \|\boldsymbol{A}\| \cdot \|\boldsymbol{x}\|_a$$

for all $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, $\boldsymbol{x} \in \mathbb{R}^n$.
See for further information: [Sto93, Lem97].

lub-Norm

Definition A.2. The *lub-norm* (lub = least upper bound) in $\mathbb{R}^{m\times n}$ is defined as

$$\mathbf{lub}(\boldsymbol{A}) := \max_{\substack{\boldsymbol{x}\in\mathbb{R}^n \\ \boldsymbol{x}\neq\boldsymbol{0}}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|}{\|\boldsymbol{x}\|}.$$

Proposition A.3 (see [Lem97, Sto93]). *A lub-norm in $\mathbb{R}^{m\times n}$ is a submultiplicative matrix-norm which is compatible with its defining vector norms $\|\cdot\|$ on $\mathbb{R}^n$ resp. $\mathbb{R}^m$.*
If $\|\boldsymbol{A}\|$ is another matrix-norm compatible with the vector norm $\|\cdot\|$, then $\mathbf{lub}(\boldsymbol{A})$ is the smallest matrix norm, i.e. $\mathbf{lub}(\boldsymbol{A}) \leq \|\boldsymbol{A}\|$.
It fulfills $\mathbf{lub}(\boldsymbol{I_n}) = \mathbf{lub}(\boldsymbol{U}) = \mathbf{1}$ for the identity matrix $\boldsymbol{I_n}$ and all unitary matrices $\boldsymbol{U}$.

Examples for Matrix Norms

Example A.4 *(see [Sto93, Lem97]).* The following mappings are matrix norms:

(i) $\|\boldsymbol{A}\|_{\mathrm{Z}} = \max\limits_{i=1,\ldots,m} \sum\limits_{k=1}^{n} |\boldsymbol{a_{ik}}|$ (row-sum-norm)

(ii) $\|\boldsymbol{A}\|_{\mathrm{S}} = \max\limits_{k=1,\ldots,n} \sum\limits_{i=1}^{m} |\boldsymbol{a_{ik}}|$ (column-sum-norm)

(iii) $\|\boldsymbol{A}\|_{\mathrm{F}} = \sqrt{\sum\limits_{i,k=1}^{n} |\boldsymbol{a_{ik}}|^{\mathbf{2}}}$ (Frobenius or Schur norm)

(iv) $\|\boldsymbol{A}\|_{\infty} = \max\limits_{i,k=1,\ldots,n} |\boldsymbol{a_{ik}}|$ (maximum norm)

(v) $\|\boldsymbol{A}\|_{\mathrm{sp}} = \max\limits_{\boldsymbol{x}\neq\boldsymbol{0}_{\mathbb{R}^n}} \sqrt{\frac{\boldsymbol{x}^\top \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}}} = \sqrt{\boldsymbol{\lambda}_{\max}(\boldsymbol{A}^\top \boldsymbol{A})}$ (spectral norm),
where $\boldsymbol{\lambda}_{\max}(\boldsymbol{A}^\top \boldsymbol{A})$ is the eigenvalue of $\boldsymbol{A}^\top \boldsymbol{A}$ with maximal absolute value.

Examples for Matrix Norms

Example A.4 *(continued).* (i) coincides with $\mathbf{lub}_{\infty}(\boldsymbol{A})$, the lub-norm of $\|\cdot\|_{\infty}$
(ii) coincides with $\mathbf{lub}_{\mathbf{1}}(\boldsymbol{A})$, the lub-norm of $\|\cdot\|_{\mathbf{1}}$
(v) coincides with $\mathbf{lub}_{\mathbf{2}}(\boldsymbol{A})$, the lub-norm of $\|\cdot\|_{\mathbf{2}}$
(iii) and (v) are compatible with the Euclidean vector norm $\|\cdot\|_{\mathbf{2}}$,
(i), (ii), (iii) and (v) are submultiplicative (not (iv)).

A.2 Measurable Functions

Measurable Functions

Definition A.5. Let $\mathcal{I} = [t_0, t_f]$ and $f : \mathcal{I} \to \overline{\mathbb{R}}$, $\overline{\mathbb{R}} := [-\infty, \infty]$. Then, $f(\cdot)$ is *measurable*, if for each $t \in \mathbb{R}$ the set

$$f^{-1}((-\infty, s]) = \{t \in \mathcal{I} \,|\, f(t) \leq s\}$$

is measurable.
$f : \mathcal{I} \to \overline{\mathbb{R}}^n$ is *measurable*, if for each closed set $S \subset \mathbb{R}^n$ the inverse image

$$f^{-1}(S) = \{t \in \mathcal{I} \,|\, f(t) \in S\}$$

is measurable.

Definition A.6. Let $\mathcal{I} = [t_0, t_f]$ and $f : \mathcal{I} \to \mathbb{R}^n$. Then, $f(\cdot)$ is *simple*, if there exists a finite partition $(\mathcal{I}_i)_{i=1,\ldots,k}$ of $\mathcal{I}$ and values $f_i \in \mathbb{R}^n$, $i = 1, \ldots, k$, with

$$f(t) = \sum_{i=1}^{k} \chi_{\mathcal{I}_i}(t) f_i \quad (t \in \mathcal{I}).$$

Hereby, $\chi_{\mathcal{I}_i}(t) = 1$, if $t \in \mathcal{I}_i$ and otherwise equals zero.

Corollary A.7. *Let $\mathcal{I} = [t_0, t_f]$ and $f : \mathcal{I} \to \mathbb{R}^n$ be a simple function.*
Then, $f(\cdot)$ is measurable, if each set $\mathcal{I}_i$, $i = 1, \ldots, k$, in Definition A.6 is measurable.

Measurable Functions

Proposition A.8. *Let $\mathcal{I} = [t_0, t_f]$ and $f : \mathcal{I} \to \overline{\mathbb{R}}^n$ with components $f_i(\cdot)$, $i = 1, \ldots, n$. Then, $f(\cdot)$ is measurable, if and only if each $f_i(\cdot)$, $i = 1, \ldots, n$, is measurable.*

Proof. see [Coh80, remarks following Proposition 2.6.3] □

Proposition A.9. *Let $\mathcal{I} = [t_0, t_f]$, $A \in \mathbb{R}^{m \times n}$, $\mu \in \mathbb{R}$ and $f, g : \mathcal{I} \to \mathbb{R}^n$. Then, the functions $A \cdot f$, $\mu \cdot f$ and $f + g$ are also measurable.*

Proof. follows from [Coh80, Proposition 2.6.1] and [Coh80, remarks following Proposition 2.6.3] □

Proposition A.10. *Let $\mathcal{I} = [t_0, t_f]$ and $f : \mathcal{I} \to \overline{\mathbb{R}}^n$ be measurable. Then, the function $f|_{\widetilde{\mathcal{I}}}(\cdot)$ is measurable for all measurable subsets $\widetilde{\mathcal{I}} \subset \mathcal{I}$.*

Proof. cf. [Coh80, remarks after Proposition 2.1.6] □

Proposition A.11. *Let $\mathcal{I} = [t_0, t_f]$ and $f : \mathcal{I} \to \overline{\mathbb{R}}^n$ be measurable. Consider $(\mathcal{I}_k)_{k \in \mathbb{N}}$ with measurable $\mathcal{I}_k \subset \mathcal{I}$ for $k \in \mathbb{N}$ and $\bigcup_{k \in \mathbb{N}} \mathcal{I}_k = \mathcal{I}$ such that all functions $(f|_{\mathcal{I}_k}(\cdot))_{k \in \mathbb{N}}$ are measurable.*
Then, the function $f(\cdot)$ itself is measurable.

Proof. cf. [Coh80, remarks after Proposition 2.1.6] □

Proposition A.12. *Let $\mathcal{I} = [t_0, t_f]$ and $f : \mathcal{I} \to \overline{\mathbb{R}}^n$ be measurable. Then, there exists a sequence of simple, measurable functions $f_m : \mathcal{I} \to \mathbb{R}^n$ with*

$$f(t) = \lim_{m \to \infty} f_m(t) \quad (\textit{for all } t \in \mathcal{I}).$$

Proof. cf. [Coh80, remarks after Proposition 2.1.7] □

Measurable Functions

Proposition A.13. *Let $\mathcal{I} = [t_0, t_f]$ and $f_m : \mathcal{I} \to \overline{\mathbb{R}}^n$ be measurable for all $m \in \mathbb{N}$. Then, the function*

$$f(t) := \sup_{m \in \mathbb{N}} f_m(t)$$

is also measurable.

Proof. see [Coh80, Proposition 2.1.4] □

Measurable Functions

Theorem A.14 (Lebesgue's Dominated Convergence Theorem). *Let $\mathcal{I} = [t_0, t_f]$ and $g : \mathcal{I} \to \overline{\mathbb{R}}_+^n$ with $\overline{\mathbb{R}}_+ := [0, \infty]$ be integrable. Let $f_m : \mathcal{I} \to \overline{\mathbb{R}}^n$, $m \in \mathbb{N}$, be measurable and $f : \mathcal{I} \to \overline{\mathbb{R}}^n$ be defined as*

$$f(t) = \lim_{m \in \mathbb{N}} f_m(t)$$

with the estimations

$$|f_m(t)| \leq g(t) \quad (m \in \mathbb{N}, \text{a.e. } t \in \mathcal{I})$$

Then, the functions $f_m(\cdot)$, $m \in \mathbb{N}$, and $f(\cdot)$ are integrable with

$$\int_{\mathcal{I}} f(t)dt = \lim_{m \in \mathbb{N}} \int_{\mathcal{I}} f_m(t)dt.$$

Proof. see [Coh80, Theorem 2.4.4] □

A.3 Functions with Bounded Variation and Absolutely Continuous Functions

Functions with Bounded Variation

Definition A.15. Let $\mathcal{I} = [t_0, t_f]$ and $f : \mathcal{I} \to \mathbb{R}^n$. Then, $f(\cdot)$ has *bounded variation* on $\mathcal{I}$, if for all partitions

$$t_0 < t_1 < \ldots < t_{N-1} < t_N = t_f, \qquad N \in \mathbb{N},$$

the sum

$$\sum_{i=0}^{N-1} \|f(t_{i+1}) - f(t_i)\| \leq C$$

is bounded by a constant C which is independent from the partition. The infimum of such constants is called *variation* of $f(\cdot)$, namely $\bigvee_{t_0}^{t_f} \mathrm{f}(\cdot)$ or $\bigvee_{\mathcal{I}} \mathrm{f}(\cdot)$.

Proposition A.16. *Let $\mathcal{I} = [t_0, t_f]$ and $f : \mathcal{I} \to \mathbb{R}^n$ have bounded variation. Then, $f(\cdot)$ is continuous except on a set of at most countable points, hence also measurable.*

Proof. cf. [dB03, 4.3, Corollary to Theorem 3] □

For more details on absolutely continuous functions, see e.g. [dB03, Nat81, Nat55].

Absolutely Continuous Functions

Definition A.17. Let $\mathcal{I} = [t_0, t_f]$ and $f : \mathcal{I} \to \mathbb{R}^n$. Then, $f(\cdot)$ is *absolutely continuous*, if for every $\varepsilon > 0$ there exists $\delta > 0$ such that for all pairwise disjoint subintervals $(]a_i, b_i[)_{i=1,\ldots,N} \subset [t_0, T]$, $N \in \mathbb{N}$, with total length

$$\sum_{i=1}^{N} (b_i - a_i) \leq \delta$$

follows:

$$\sum_{i=1}^{N} \|f(b_i) - f(a_i)\| \leq \varepsilon$$

Proposition A.18. *Let $\mathcal{I} = [t_0, t_f]$ and $f : \mathcal{I} \to \mathbb{R}^n$. Then, $f(\cdot)$ is absolutely continuous, if and only if for all $t \in \mathcal{I}$*

$$f(t) = f(t_0) + \int_{t_0}^{t} f'(\tau)d\tau.$$

Proof. cf. [dB03, 9.3, Corollary 3 and following remarks] □

For more details on absolutely continuous functions, see e.g. [dB03, Nat81, Nat55].

A.4 Additional Results

Ordered Sets/Lemma of Zorn

Definition A.19. Let $(\mathcal{M}, \preceq)$ be an ordered, nonempty set. Then, $\mathcal{N} \subset \mathcal{M}$ is *totally ordered*, if for all $n, \widetilde{n} \in \mathcal{N}$ we have $n \preceq m$ or $m \preceq n$.

Lemma A.20 (Zorn). $(\mathcal{M}, \preceq)$ *ordered, nonempty set which has an upper bound for every totally ordered subset* $\mathcal{N}$*, i.e.*

$$\exists m \in \mathcal{N} \text{ such that for all } n \in \mathcal{N}: \ m \preceq n,$$

then $\mathcal{M}$ *has a maximal element* $m^0 \in \mathcal{M}$*, i.e. for all* $m \in \mathcal{M}$ *with* $m^0 \preceq m$ *follows immediately* $m \preceq m^0$*.*

Proposition of Hahn-Banach

Proposition A.21 (Hahn-Banach). *Let* X *be a real linear space,* $Z \subset X$ *a linear subspace and let*

(i) $p : X \to \mathbb{R}$ *sublinear, i.e.*

$$\begin{aligned} p(x+y) &\leq p(x)+p(y) && \text{for all } x, y \in X, \\ p(\alpha \cdot x) &= \alpha \cdot p(x) && \text{for all } \alpha \geq 0,\ x \in X \end{aligned}$$

(ii) $f : Z \to \mathbb{R}$ *linear*

(iii) $f(z) \leq p(z)$ *for all* $z \in Z$

Then, there exists a linear continuation $F : X \to \mathbb{R}$ *with*

$$\begin{aligned} F(z) &= f(z) && \text{for all } z \in Z, \text{ i.e. } F|_Z = f, \\ F(x) &\leq p(x) && \text{for all } x \in X \end{aligned}$$

B References

References

[AC84] J.-P. Aubin and A. Cellina. *Differential Inclusions*, volume 264 of *Grundlehren der mathematischen Wissenschaften*. Springer-Verlag, Berlin–Heidelberg–New York–Tokyo, 1984.

[AF90] J.-P. Aubin and H. Frankowska. *Set-Valued Analysis*, volume 2 of *Systems & Control: Foundations and Applications*. Birkhäuser, Boston–Basel–Berlin, 1990.

[AH74] G. Alefeld and J. Herzberger. *Einführung in die Intervallrechnung*, volume 12 of *Reihe Informatik*. B.I.-Wissenschaftsverlag, Mannheim–Wien–Zürich, 1974.

[AH83] G. Alefeld and J. Herzberger. *Introduction to interval computations*. Computer Science and Applied Mathematics. Academic Pres, New York et al., 1983. extended and revised ed.

[AMR95] Uri M. Ascher, Robert M.M. Mattheij, and Robert D. Russell. *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations*, volume 13 of *Classics In Applied Mathematics*. SIAM, Philadelphia, 1995.

[Art89] Z. Artstein. Piecewice linear approximations of set-valued maps. *J. Approx. Theory*, 56:41–47, 1989.

[Art94] Z. Artstein. First order approximations for differential inclusions. *Set-Valued Anal.*, 2(1–2):7–17, 1994.

[Art95] Zvi Artstein. A Calculus for Set-Valued Maps and Set-Valued Evolution Equations. *Set-Valued Anal.*, 3:213–261, 1995.

[Aub91] J.-P. Aubin. *Viability theory*. Systems & Control: Foundations & Applications. Birkhäuser, Boston, MA et. al., 1991.

[Aum65] R. J. Aumann. Integrals of Set-Valued Functions. *J. Math. Anal. Appl.*, 12(1):1–12, 1965.

[Bai95] R. Baier. *Mengenwertige Integration und die diskrete Approximation erreichbarer Mengen*, volume 50 of *Bayreuth. Math. Schr.* Mathematisches Institut der Universität Bayreuth, 1995.

[Bai05] R. Baier. Selection Strategies for Set-Valued Runge-Kutta Methods. In Z. Li, L. G. Vulkov, and J. Wasniewski, editors, *Numerical Analysis and Its Applications, Third International Conference, NAA 2004, Rousse, Bulgaria, June 29 - July 3, 2004, Revised Selected Papers*, volume 3401 of *Lecture Notes in Computer Science*, pages 149–157, Berlin–Heidelberg, 2005. Springer.

[Bal82] E. I. Balaban. On the approximate evaluation of the Riemann integral of many-valued mapping. *U.S.S.R. Comput. Maths. Math. Phys.*, 22(2):233–238, 1982.

[BCP96] Kathry E. Brenan, Stephen L. Campbell, and Linda R. Petzold. *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations*, volume 14 of *Classics In Applied Mathematics*. SIAM, Philadelphia, 1996.

[Bel70] R. Bellman. *Introduction to Matrix Analysis*. McGraw-Hill Book Company, New York–St. Louis–San Francisco–Düsseldorf–London–Mexico–Panama–Sydney–Toronto, 1970. 2nd ed.

[BF34] T. Bonnesen and W. Fenchel. *Theorie der konvexen Körper*. Ergebnisse der Mathematik und ihrer Grenzgebiete, Band 3, Heft 1. Chelsea Publishing Company, Bronx–New York, 1934. reprint: Chelsea Publishing Company, Bronx–New York, 1971.

[BF87a] V. I. Blagodatskikh and A. F. Filippov. Differential inclusions and optimal control. In *Topology, Ordinary Differential Equations, Dynamical Systems*, volume 1986, issue 4 of *Proc. Steklov Inst. Math.*, pages 199–259. AMS, Providence, Rhode Island, 1987.

[BF87b] T. Bonnesen and W. Fenchel. *Theory of convex bodies*. BCS Associates, Moscow–Idaho, 1987. English translation.

[BF01a] R. Baier and E. Farhi. Differences of Convex Compact Sets in the Space of Directed Sets. Part I: The Space of Directed Sets. *Set-Valued Anal.*, 9(3):217–245, 2001.

[BF01b] R. Baier and E. Farhi. Differences of Convex Compact Sets in the Space of Directed Sets. Part II: Visualization of Directed Sets. *Set-Valued Anal.*, 9(3):247–272, 2001.

[BH98] John T. Betts and W. P. Huffman. Mesh Refinement in Direct Transcription Methods for Optimal Control. *Optimal Control Applications and Methods*, 19:1–21, 1998.

[BJ70] H. T. Banks and Marc Q. Jacobs. A Differential Calculus for Multifunctions. *J. Math. Anal. Appl.*, 29(2):246–272, 1970.

[BL94a] R. Baier and F. Lempio. Approximating reachable sets by extrapolation methods. In P. J. Laurent, A. Le Méhauteé, and L. L. Schumaker, editors, *Curves and Surfaces in Geometric Design*, pages 9–18, Wellesley, 1994. A K Peters.

[BL94b] R. Baier and F. Lempio. Computing Aumann's integral. In *[?]*, pages 71–92, 1994.

[Boc87] Hans Georg Bock. Randwertproblemmethoden zur Parameteridentifizierung in Systemen nichtlinearer Differentialgleichungen. volume 183 of *Bonner Mathematische Schriften*, Bonn, 1987.

[Bri70] T. F. Bridgland, Jr. Trajectory integrals of set valued functions. *Pac. J. Math.*, 33(1):43–68, 1970.

[Büs98] Christof Büskens. *Optimierungsmethoden und Sensitivitätsanalyse für optimale Steuerprozesse mit Steuer- und Zustandsbeschränkungen*. PhD thesis, Fachbereich Mathematik, Westfälische Wilhems-Universität Münster, 1998.

[But87] J. C. Butcher. *The Numerical Analysis of Ordinary Differential Equations*. John Wiley & Sons, Chichester–New York–Brisbane–Toronto–Singapore, 1987.

[BY98] J.-D. Boissonnat and M. Yvinec. *Algorithmic Geometry*. Cambridge University Press, Cambridge, 1998.

[Cha03] I. A. Chahma. Set-valued discrete approximation of state-constrained differential inclusions. *Bayreuth. Math. Schr.*, 67:3–162, 2003.

[Che94] F. L. Chernousko. *State Estimation for Dynamic Systems*. CRC Press, Boca Raton, Fl.–Ann Arbor–London–Tokyo, 1994.

[Cla83] F. H. Clarke. *Optimization and Nonsmooth Analysis*, volume 178 of *Canadian Mathematical Society Series of Monographs and Advanced Texts*. John Wiley & Sons, New York–Chicester–Brisbane–Toronto–Singapore, 1983.

[CLSW98] F. H. Clarke, Yu. S. Ledyaev, R. J. Stern, and P. R. Wolenski. *Nonsmooth Analysis and Control Theory*, volume 178 of *Graduate Texts in Mathematics*. Springer, New York–Berlin–Heidelberg–Barcelona–Budapest–Hong Kong–London–Milan–Paris–Santa Clara–Singapore–Tokyo, 1998.

[Coh80] D. L. Cohn. *Measure Theory*. Birkhäuser, Boston–Basel–Stuttgart, 1980.

[CR02] V. V. Chistyakov and A. Rychlewicz. On the extension and generation of set-valued mappings of bounded variation. *Studia Math.*, 153(3):235–396, 2002.

[CS85] M. Caracotsios and W. E. Stewart. Sensitivity analysis of initial-boundary-value problems with mixed PDEs and algebraic equations. *Computers chem. Engng*, 19(9):1019–1030, 1985.

[CV77] C. Castaing and M. Valadier. *Convex Analysis and Measurable Multifunctions*, volume 580 of *Lecture Notes in Math.* Springer-Verlag, Berlin–Heidelberg–New York, 1977.

[DB78] C. De Boor. *A practical guide to splines*, volume 27 of *Applied Mathematical Sciences*. Springer, New York, 1978.

[dB03] G. de Barra. *Measure Theory and Integration*. Ellis Horwood Series in Mathematics and Its Applications. Ellis Horwood Limited, Publisher, Chichester, 2003. 2nd updated ed.

[Deb67] G. Debreu. Integration of correspondences. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Held at the Statistical Laboratory, University of California, June 21–July 18, 1965 and December 27, 1965–January 7, 1966*, Volume II: Contributions to Probability Theory, Part 1, pages 351–372, Berkeley–Los Angeles, 1967. University of California Press.

[Dei92] K. Deimling. *Multivalued Differential Equations*, volume 1 of *de Gruyter Series in Nonlinear Analysis and Applications*. Walter de Gruyter, Berlin–New York, 1992.

[Den98] D. Dentcheva. Differentiable Selections and Castaing Representations of Multifunctions. *J. Math. Anal. Appl.*, 223(2):371–396, 1998.

[Den00] D. Dentcheva. Regular Castaing Representations of Multifunctions with Applications to Stochastic Programming. *SIAM J. Optim.*, 10(3):732–749, 2000.

[Den01] D. Dentcheva. Continuity of Multifunctions Characterized by Steiner Selections. *Nonlinear Anal.*, 47:1985–1996, 2001.

[DF89] A. L. Dontchev and E. Farkhi. Error Estimates for Discretized Differential Inclusions. *Computing*, 41(4):349–358, 1989.

[DF90] T. D. Donchev and E. M. Farkhi. Moduli of smoothness of vector valued functions of a real variable and applications. *Numer. Funct. Anal. Optim.*, 11(5 & 6):497–509, 1990.

[DH91] Peter Deuflhard and Andreas Hohmann. *Numerische Mathematik*. de Gruyter, Berlin, 1991.

[DHM00] A. L. Dontchev, W. W. Hager, and K. Malanowski. 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.

[DHV00a] A. L. Dontchev, W. W. Hager, and V. M. Veliov. Second-Order Runge-Kutta Approximations in Control Constrained Optimal Control. *SIAM J. Numer. Anal.*, 38(1):202–226, 2000.

[DHV00b] A. L. Dontchev, W. W. Hager, and V. M. Veliov. Second-Order Runge-Kutta Approximations in Control Constrained Optimal Control. *SIAM Journal on Numerical Analysis*, 38(1):202–226, 2000.

[Dim80] N. Dimitrova. Über die Distributivgesetze der erweiterten Intervallarithmetik. *Computing*, 24(1):33–49, 1980.

[DKRV97] P. Diamond, P. Kloeden, A. Rubinov, and A. Vladimirov. Comparative Properties of Three Metrics in the Space of Compact Convex Sets. *Set-Valued Anal.*, 5(3):267–289, 1997.

[DLZ86] V. F. Demyanov, C. Lemaréchal, and J. Zowe. Approximation to a Set-Valued Mapping, I: A Proposal. *Appl. Math. Optim.*, 14:203–214, 1986.

[Dom87] G. Dommisch. On the existence of lipschitz-continuous and differentiable selections for multifunctions. In J. Guddat, H. Th. Jongen, B. Kummer, and F. Nožička, editors, *Parametric Optimization and Related Topics. Volume 35 of Mathematical Research*, pages 60–74. Akademie-Verlag, Berlin, 1987.

[DR95] V. F. Demyanov and A. M. Rubinov. *Constructive nonsmooth analysis*, volume 7 of *Approximation and Optimization*. Peter Lang, Frankfurt am Main–Berlin–Bern–New York–Paris–Wien, 1995.

[DV93] Doitchinov B. D. and V. M. Veliov. Parametrizations of integrals of set-valued mappings and applications. *J. Math. Anal. Appl.*, 179(2):483–499, 1993.

[dvOS97] M. de Berg, M. van Kreveld, M. Overmars, and O. Schwarzkopf. *Computational Geometry. Algorithms and Applications*. Springer, Berlin–Heidelberg–New York–Barcelona–Budapest–Hong Kong–London–Milan–Paris–Santa Clara–Singapore–Tokyo, 1997.

[ET77] I. Ekeland and R. Témam. *Convex analysis and variational problems. Unabridged, corrected republication of the 1976 English original*, volume 28 of *Classics in Applied Mathematics*. SIAM, Philadelphia, PA, 1977.

[Fer94] R. Ferretti. Discrete time high-order schemes for viscosity solutions of Hamilton-Jacobi-Bellman equations. *Numer. Math.*, 67(3):315–344, 1994.

[Fil88] A. F. Filippov. *Differential Equations with Discontinuous Righthand Sides*. Mathematics and Its Applications (Soviet Series). Kluwer Academic Publishers, Dordrecht–Boston–London, 1988.

[FTB97] William F. Feehery, John E. Tolsma, and Paul I. Barton. Efficient sensitivity analysis of large-scale differential-algebraic systems. *Applied Numerical Mathematics*, 25:41–54, 1997.

[Gau90] S. Gautier. Affine and eclipsing multifunctions. *Numer. Funct. Anal. Optim.*, 7 & 8:679–699, 1990.

[GI83] D. Goldfarb and A. Idnani. A numerically stable dual method for solving strictly convex quadratic programs. *Mathematical Programming*, 27:1–33, 1983.

[GM78] P. E. Gill and W. Murray. Numerically stable methods for quadratic programming. *Mathematical Programming*, 14:349–372, 1978.

[GM92] S. Gautier and R. Morchadi. A selection of convex-compact-valued multi-functions with remarkable properties: The steiner selection. *Numer. Funct. Anal. Optim.*, 13(5&6):513–522, 1992.

[GMSW91] P. E. Gill, W. Murray, M. A. Saunders, and M. H. Wright. Inertia-controlling methods for general quadratic programming. *SIAM Review*, 33(1):1–36, 1991.

[GO97] J. E. Goodman and J. O'Rourke, editors. *Handbook of Discrete and Computational Geometry*. CRC Press Series on Discrete Mathematics and Its Applications. CRC Press, Boca Raton, FL–New York, 1997.

[Gra03] G. Grammel. Towards Fully Discretized Differential Inclusions. *Set-Valued Anal.*, 11(1):1–8, 2003.

[Grü03] B. Grünbaum. *Convex Polytopes*, volume 221 of *Graduate Texts in Mathematics*. Springer, New York–Berlin–Heidelberg–Hong Kong–London–Milan–Paris–Tokyo, 2nd edition, 2003.

[Gug77] H. W. Guggenheimer. *Applicable Geometry. Global and Local Convexity*. Applied Mathematics Series. Robert E. Krieger Publishing Co., Inc., Huntington, NY, 1977.

[GW93] P. M. Gruber and J. M. Wills. *Handbook of Convex Geometry. Volume A*. North-Holland, Amsterdam, 1993.

[Had50] H. Hadwiger. Minkowskische Addition und Subtraktion beliebiger Punktmengen und die Theoreme von Erhard Schmidt. *Mathematische Zeitschrift*, 53(3):210–218, 1950.

[Hag00] William W. Hager. Runge-Kutta methods in optimal control and the transformed adjoint system. *Numerische Mathematik*, 87:247–282, 2000.

[Hau27] F. Hausdorff. *Mengenlehre*. W. de Gruyter, Berlin, 1927.

[Hau91] F. Hausdorff. *Set theory*. Chelsea, New York, 1991. English translation.

[Her71] H. Hermes. On continuous and measurable selections and the existence of solutions of generalized differential equations. *Proc. Amer. Math. Soc*, 29(3):535–542, 1971.

[Hör54] P. L. Hörmander. Sur la fonction d'appui des ensembles convexes dans un espace localement convexe. *Ark. Mat.*, 3(12):181–186, 1954.

[HUL93] J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms I*, volume 305 of *Grundlehren der mathematischen Wissenschaften*. Springer-Verlag, Berlin–Heidelberg–New York–London–Paris–Tokyo–Hong Kong–Barcelona–Budapest, 1993.

[IT79] A. D. Ioffe and V. M. Tihomirov. Theory of extremal problems. volume 6 of *Studies in Mathematics and its Applications*, Amsterdam, New York, Oxford, 1979. North-Holland Publishing Company.

[JKDW01] L. Jaulin, M. Kieffer, O. Didrit, and É. Walter. *Applied Interval Analysis*. Springer, London–Berlin–Heidelberg–New York–Barcelona–Hong Kong–Milan–Paris-Singapore–Tokyo, 2001.

[Kau77a] E. Kaucher. Algebraische Erweiterungen der Intervallrechnung unter Erhaltung der Ordnungs- und Verbandsstrukturen. *Comput. Suppl.*, 1:65–79, 1977.

[Kau77b] E. Kaucher. Über Eigenschaften und Anwendungsmöglichkeiten der erweiterten Intervallrechnung und des hyperbolischen Fastkörpers über $\mathbb{R}$. *Comput. Suppl.*, 1:81–94, 1977.

[Kau80] E. Kaucher. Interval Analysis in the Extended Interval Space $\mathbb{R}$. *Comput. Suppl.*, 2:33–49, 1980.

[KBV90] V. Křivan, Č. Budějovice, and I. Vrkoč. Absolutely continuous selections from absolutely continuous set valued map. *Czechoslovak Math. J.*, 40(115):503–513, 1990.

[KF87] A. B. Kurzhanski and T. F. Filippova. On a description of the set of viable trajectories of a differential inclusion. *Sov. Math., Dokl.*, 34:30–33, 1987.

[Kis91] M. Kisielewicz. *Differential Inclusions and Optimal Control*, volume 44 of *Mathematics and Its Applications*. PWN - Polish Scientific Publishers, Warszawa–Dordrecht–Boston–London, 1991.

[KK94] M. Krastanov and N. Kirov. Dynamic interactive system for analysis of linear differential inclusions. In *[?]*, pages 123–130, 1994.

[Kle63] V. Klee, editor. *Convexity. Proceedings of symposia in pure mathematics. Vol. VIII. Held at the University of Washington Scattle, Washington June 13-15, 1961*, Providence, Rhode Island, 1963. AMS.

[KML93] A. Kastner-Maresch and F. Lempio. Difference methods with selection strategies for differential inclusions. *Numer. Funct. Anal. Optim.*, 14(5&6):555–572, 1993.

[Kul77] U. Kulisch. Ein Konzept für eine allgemeine Theorie der Rechnerarithmetik. *Comput. Suppl.*, 1:95–105, 1977.

[KV97] A. B. Kurzhanski and I. Vályi. *Ellipsoidal Calculus for Estimation and Control*. Systems & Control: Foundations & Applications. Birkhäuser, Boston–International Institute for Applied Systems Analysis, 1997.

[Lei80] K. Leichtweiß. *Konvexe Mengen*. Hochschultext. Springer-Verlag, Berlin–Heidelberg–New York, 1980.

[Lei85] K. Leichtweiß. *Vypuklye mnozhestva*. Glavnaya Redaktsiya Fiziko-Matematicheskoj Literatury. Nauka, Moskva, 1985. Russian translation by V. A. Zalgaller and T. V. Khachaturova.

[Lei98] K. Leichtweiß. *Affine Geometry of Convex Bodies*. Johann Ambrosius Barth Verlag, Heidelberg–Leipzig, 1998.

[Lem97] F. Lempio. *Numerische Mathematik I*, volume 51 of *Bayreuth. Math. Schr.* Mathematisches Institut der Universität Bayreuth, 1997.

[Lem98] F. Lempio. *Numerische Mathematik II*, volume 55 of *Bayreuth. Math. Schr.* Mathematisches Institut der Universität Bayreuth, 1998.

[LVBB+04] J. Laurent-Varin, F. Bonnans, N. Berend, C. Talbot, and M. Haddou. On the refinement of discretization for optimal control problems. *IFAC Symposium on Automatic Control in Aerospace, St. Petersburg*, 2004.

[LZ91] C. Lemaréchal and J. Zowe. The Eclipsing Concept to Approximate a Multi-Valued Mapping. *Optimization*, 22(1):3–37, 1991.

[Mar77] J. T. Marti. *Konvexe Analysis*, volume 54 of *Lehrbücher und Monographien aus dem Gebiet der Exakten Wissenschaften, Mathematische Reihe*. Birkhäuser, Basel–Stuttgart, 1977.

[Mar79] S. M. Markov. Calculus for interval functions of a real variable. *Computing*, 22(4):325–337, 1979.

[Mar80] S. M. Markov. Some applications of extended interval arithmetic to interval iterations. *Comput. Suppl.*, 2:69–84, 1980.

[Mar95] S. M. Markov. On directed interval arithmetic and its applications. *J. UCS*, 1(7):514–526, 1995.

[Mar98] S. M. Markov. On the Algebra of Intervals and Convex Bodies. *J. UCS*, 4(1):34–47, 1998.

[Mar00] S. M. Markov. On the Algebraic Properties of Convex Bodies and Some Applications. *J. Convex Anal.*, 7(1):129–166, 2000.

[MBM97] Kazimierz Malanowski, Christof Büskens, and Helmut Maurer. Convergence of Approximations to Nonlinear Optimal Control Problems. In Anthony Fiacco, editor, *Mathematical programming with data perturbations*, volume 195, pages 253–284. Dekker. Lecture Notes in Pure and Applied Mathematics, 1997.

[Mic56] E. Michael. Continuous selections I. *Ann. of Math. (2)*, 63(2):361–382, 1956.

[Min10] H. Minkowski. *Geometrie der Zahlen. I*. B. G. Teubner, Leipzig–Berlin, 1910. reprint: Chelsea Publishing Company, New York, 1953.

[Min11] H. Minkowski. *Theorie der konvexen Körper, insbesondere Begründung ihres Oberflächenbegriffs*, volume II of *Gesammelte Abhandlungen*, pages 131–229. B. G. Teubner, Leipzig–Berlin, 1911. reprint: Chelsea Publishing Company, New York, 1967.

[Moo66] R. E. Moore. *Interval Analysis*. Prentice-Hall, Englewood Cliffs, N.J., 1966.

[Mor88] B. S. Mordukhovich. An approximate maximum principle for finite-difference control systems. *U.S.S.R. Computational Mathematics and Mathematical Physics*, 28(1):106–114, 1988.

[MP96] Timothy Maly and Linda R. Petzold. Numerical Methods and Software for Sensitivity Analysis of Differential-Algebraic Systems. *Applied Numerical Mathematics*, 20(1):57–79, 1996.

[Nat55] I. P. Natanson. *Theory of functions of real variable*. Ungar, New York, 1955.

[Nat81] I. P. Natanson. *Theorie der Funktionen einer reellen Veränderlichen*, volume VI of *Mathematische Lehrbücher und Monographien. I. Abt.* Akademie-Verlag, Berlin, 1981. 5. Auflage, herausgegeben von Karl Bögel.

[Nic78] K. Nickel. Intervall-Mathematik. *Z. Angew. Math. Mech.*, 58(6):T 72–T 85, 1978.

[Nik88] M. S. Nikol'skiĭ. A method of approximating an attainable set for a differential inclusion. *Comput. Math. Math. Phys.*, 28(4):192–194, 1988. Translated from Zh. Vychisl. Mat. Mat. Fiz. 28 (1988), no. 8, pp. 1252–1254 (Russian).

[Ole65] C. Olech. A note concerning set-valued measurable functions. *Bull. Polish Acad. Sci. Math.*, 13:317–1965, 1965.

[O'R98] J. O'Rourke. *Computational Geometry in C*. Cambridge University Press, Cambridge, 1998. 2nd ed.

[Ort69] H.-J. Ortolf. Eine Verallgemeinerung der Intervallarithmetik. Technical report, GMD-Ber., Bonn, 1969.

[Pic03] K. Pichard. Unified Treatment of Algebraic and Geometric Difference by a New Difference Scheme and its Continuity Properties. *Set-Valued Anal.*, 11(2):111–132, 2003.

[Pol75] E. S. Polovinkin. *Riemannian Integral of Set-Valued Function*, pages 405–410. Lecture Notes in Computer Science 27, Optimization Techniques, IFIP Technical Conference, Novosibirsk, July 1–7, 1974. Springer-Verlag, Berlin–Heidelberg–New York, 1975.

[Pol83] E. S. Polovinkin. On integration of multivalued mappings. *Dokl. Akad. Nauk SSSR*, 28(1):223–228, 1983.

[Pon67] L. S. Pontryagin. Linear differential games. ii. *Sov. Math., Dokl.*, 8(4):910–912, 1967.

[Pow78] M. J. D. Powell. A fast algorithm for nonlinearily constrained optimization calculation. In G.A. Watson, editor, *Numerical Analysis*, volume 630 of *Lecture Notes in Mathematics*, Berlin-Heidelberg-New York, 1978. Springer.

[PS88] F. P. Preparata and M. I. Shamos. *Computational Geometry. An Introduction*. Texts and Monographs in Computer Science. Springer, New York et al., 1988. corr. and expanded 2. print.

[PU02] D. Pallaschke and R. Urbański. *Pairs of compact convex sets*, volume 548 of *Mathematics and Its Applications*. Kluwer Academic Publishers, Dordrecht, 2002.

[Råd52] H. Rådström. An embedding theorem for spaces of convex sets. *Proc. Amer. Math. Soc*, 3:165–169, 1952.

[Rat80] H. Ratschek. Representation of Interval Operations by Coordinates. *Computing*, 24(2–3):93–96, 1980.

[Roc72] R. T. Rockafellar. *Convex Analysis*, volume 28 of *Princeton Mathematical Series*. Princeton University Press, Princeton, NJ, 2nd edition, 1972.

[Roc76] R. T. Rockafellar. Integral functionals, normal integrands and measurable selections. In A. Dold and B. Eckmann, editors, *Nonlinear Operators and the Calculus of Variations. Volume 543 of Lecture Notes in Math.*, pages 157–207. Springer-Verlag, Berlin–Heidelberg–New York, 1976.

[Roy86] H. L. Royden. *Real Analysis*. Macmillan Publishing Company/Collier Macmillan Publisher, New York–London, 1986. 3rd ed.

[Sch68] Fred C. Schweppe. Recursive state estimation: Unknown but bounded errors and system inputs. *IEEE Trans. Automat. Control*, AC-13(1):22ff, 1968.

[Sch93] R. Schneider. *Convex Bodies: The Brunn-Minkowski Theory*, volume 44 of *Encyclopedia of Mathematics and Applications*. Cambridge University Press, Cambridge, 1993.

[Sil97] D. B. Silin. On Set-Valued Differentiation and Integration. *Set-Valued Anal.*, 5(2):107–146, 1997.

[SP88] B. Sendov and V. Popov. *The averaged moduli of smoothness*. Pure and Applied Mathematics. John Wiley & Sons, Chicester–New York–Brisbane–Toronto–Singapore, 1988.

[Ste73] Hans J. Stetter. Analysis of Discretization Methods for Ordinary Differential Equations. volume 23 of *Springer Tracts in Natural Philosophy*. Springer-Verlag Berlin Heidelberg New York, 1973.

[Sto93] J. Stoer. *Introduction to Numerical Analysis*. Texts in applied mathematics, 12. Springer, New York et.al., 3rd edition, 1993.

[Sv65] L. M. Sonneborn and F. S. van Vleck. The bang-bang principle for linear control problems. *SIAM J. Control, Ser. A*, 2(2):151–159, 1965.

[Val64] F. A. Valentine. *Convex Sets*. Robert E. Krieger Publishing Company, Huntington, N. Y., 1964. reprint.

[Vel89a] V. M. Veliov. Discrete approximations of integrals of multivalued mappings. *C. R. Acad. Bulgare Sci.*, 42(12), 1989.

[Vel89b] V. M. Veliov. Second order discrete approximations to strongly convex differential inclusions. *Systems Control Lett.*, 13:263–269, 1989.

[Vel92] V. M. Veliov. Second Order Discrete Approximation to Linear Differential Inclusions. *SIAM J. Numer. Anal.*, 29(2):439–451, 1992.

[Wal90] Wolfgang Walter. *Gewöhnliche Differentialgleichungen*. Springer, Berlin-Heidelberg-New York, 4 edition, 1990.

[Web94] R. Webster. *Convexity*. Oxford Science Publications. Oxford University Press, Oxford–New York–Tokyo, 1994.

[Wol90] P. R. Wolenski. The exponential formula for the reachable set of a Lipschitz differential inclusion. *SIAM J. Control Optim.*, 28(5):1148–1161, 1990.

[Zie98] G. M. Ziegler. *Lectures on Polytopes*, volume 152 of *Graduate Texts in Mathematics*. Springer, New York–Berlin–Heidelberg–London–Paris–Tokyo–Hong Kong–Barcelone–Budapest, 1998. revised 1st ed., corr. 2. print.