Optimal Control of Partial Differential Equations [*]

Axel Kröner [†]

July 12, 2019

[*] Vorlesungsskriptum zur Vorlesung *Optimierung mit partiellen Differentialgleichungen* gelesen an der Humboldt Universität zu Berlin im Sommersemester 2019.

[†] Institut für Mathematik, Humboldt-Universität zu Berlin, Germany.

Contents

Part I.

Theory I

1. Literature

I **Optimal control of PDEs**

- Frédéric Bonnans. *Lecture notes: Optimal Control of Partial Differential Equations*, 2019
- Michael Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. *Optimization with PDE Constraints*, vol. 23 of *Mathematical Modelling: Theory and Applications*. Springer, 2009.
- Jean-Pierre Raymond. *Lecture notes: Optimal Control of Partial Differential Equations*, http://www.math.univ-toulouse.fr/~raymond/book-ficus.pdf
- Fredi Tröltzsch. *Optimal control of partial differential equations*, Graduate Studies in Mathematics, vol. 112, American Mathematical Society, Providence, RI, 2010, Theory, methods and applications, Translated from the 2005 German original by Jürgen Sprekels.

II **Additional references**

- Frédéric Bonnans and Alexander Shapiro. Perturbation Analysis of Optimization Problems. Springer, 2000.
- Eduardo Casas and Fredi Tröltzsch. Second order optimality conditions and their role in PDE control. Jahresbericht der Deutschen Mathematiker-Vereinigung, 117 (2015), 3-44. http://page.math.tu-berlin.de/~troeltz/arbeiten/casas_troeltzsch_survey.pdf
- Jaques Louis Lions. Optimal control of systems governed of partial differential equations, Springer, 1971.
- Pekka Neittaanmaki, Jürgen Sprekels, and Dan Tiba, *Optimization of Elliptic Systems: Theory and Applications*, Springer, 2006
- Juan Carlos De los Reyes, *Numerical PDE-Constrained Optimization*, SpringerBriefs in Optimization, Springer Verlag, 2015.
- Michael Ulbrich. Semismooth Newton methods for operator equations in function spaces. *SIAM J. Control Optim. 13*, 3 (2002), 805–842.

III **References on functional analysis and PDE theory**

- Hans-Wilhelm Alt. Linear Functional Analysis – An Application-Oriented Introduction, 2016 Original German edition with the title *Lineare Funktionalanalysis* (6th ed.) published by Springer-Verlag Berlin Heidelberg, 2012
- Haim Brezis. Functional Analysis, Sobolev Spaces and Partial Differential Equations, Springer Science & Business Media, 2010
- Lawrence C. Evans. Partial Differential Equations, Volume 19 de Graduate studies in mathematics, American Mathematical Soc., 2010 *Partial Differential Equations*, vol. 19 of *Graduate Studies in Mathematics.* American Mathematical Society, 1998
- David Gilbarg and Neil S. Trudinger. Elliptic partial differential equations of second order, Springer-Verlag Berlin-Heidelberg, 2001.

These lecture notes follow in main parts the monographs Bonnans [6], de Los Reyes [], Hinze, Pinnau, Ulbrich, and Ulbrich [15], Vexler [21], Raymond [19], and Tröltzsch [20].

2. Introduction

Contents

2.1. Characterization of control problems

Roughly speaking a control problem consists of:

- A controlled system, that is an input-output process,
- an observation of the output of the controlled system,
- an objective to be achieved.

The input can be given as a function in a boundary condition, an initial condition, a coefficient in the partial differential equation, or any parameter in the equation, and the output is the solution of the partial differential equation. The input is called the control variable, or the *control*, and the output is called the *state* of the system. An *observation* of the system is a mapping (very often a linear operator) depending on the state.

We can seek for various objectives:

(i) Minimize a criterion depending on the observation of the state and on the control variable. This is an *optimal control problem*. The unknown of this minimization problem is the control variable.

(ii) We can look for a control for which the observation belongs to some target. This corresponds to a *controllability problem*.

(iii) We can look for a control which stabilizes the state or an observation of the state of the system. This is a *stabilization problem.*

In this course we are interested in problems of type (i), i.e. optimal control problems of partial differential equations (sometimes we just write 'control problems').

2.2. Examples

In the following we consider optimal control problems with different control action:

2.2.1. Distributed control

In this first example we want to control the temperature (denoted by $y\colon \Omega \to \mathbb{R}$) in a domain $\Omega \subset \mathbb{R}^n$, $n \in \mathbb{N}$, such that it is close to a desired given temperature $y_\mathrm{d}\colon \Omega \to \mathbb{R}$. The control, denoted by $u\colon \Omega \to \mathbb{R}$, can act on the whole domain. More precisely, we consider the following problem

$$\min \quad J(u,y) := \frac{1}{2}\int_\Omega (y(x) - y_\mathrm{d}(x))^2 \mathrm{d}x + \frac{\alpha}{2}\int_\Omega u(x)^2 \mathrm{d}x, \quad \alpha \geq 0, \tag{2.1}$$

under the constraints

$$\begin{cases} -\Delta y = u & \text{in } \Omega, \\ \partial_n y = \beta(c - y) & \text{on } \partial\Omega. \end{cases} \tag{2.2}$$

Here the control is a distributed function, $\beta > 0$ describes the heat transfer and c the outside temperature. Equation (2.2) describes the relation between the control and the state and is called *state equation.* The functional J is called the *cost functional.* The term $\frac{\alpha}{2}\int_\Omega u(s)\mathrm{d}s$ represents the *control costs* and is sometimes called *regularization term.* Often one imposes additional constraints on the control and the state, e.g., *control constraints* of type

$$u_m \leq u(x) \leq u_M \text{ a.e. on } \partial\Omega, \quad u_m, u_M \in \mathbb{R}, \tag{2.3}$$

and *state constraints* of type

$$g(y) \in K \tag{2.4}$$

for $g\colon Y \to R$ with Banach space R and some closed and convex cone $K \subset R$.

2.2.2. Boundary control

In this example the control enters the equation as a boundary condition. We consider the problem

$$\min \quad J(u,y) := \frac{1}{2}\int_{\Omega} (y(x) - y_{\mathrm{d}}(x))^2 \mathrm{d}x + \frac{\alpha}{2}\int_{\partial\Omega} u(s)^2 \mathrm{d}s \tag{2.5}$$

under the constraints

$$\begin{cases} -\Delta y = 0 & \text{in } \Omega, \\ \partial_n y = \beta(u - y) & \text{on } \partial\Omega. \end{cases} \tag{2.6}$$

with heat transfer coefficient $\beta > 0$. This problem can be considered as a model problem for optimal cooling in metallurgy.

2.2.3. Instationary problem

In this example we consider a control problem with a distributed control and distributed cost functional:

$$\min \quad J(u,y) := \frac{1}{2}\int_0^T \int_{\Omega} (y(t,x) - y_{\mathrm{d}}(t,x))^2 \mathrm{d}x\mathrm{d}t + \frac{\alpha}{2}\int_0^T \int_{\Omega} u(t,x)^2 \mathrm{d}x\mathrm{d}t \tag{2.7}$$

under the constraints

$$\begin{cases} y_t - \Delta y = u & \text{in } (0,T)\times\Omega, \\ \partial_n y = \beta(c - y) & \text{on } (0,T)\times\partial\Omega, \\ y(0,x) = y_0(x) & \text{in } \Omega \end{cases} \tag{2.8}$$

for some $T > 0$. In this case the control $u\colon (0,T)\times\Omega \to \mathbb{R}$ and state $y\colon (0,T)\times\Omega \to \mathbb{R}$ depend on time and space. The initial condition $y_0\colon \Omega \to \mathbb{R}$ and the desired temperature distribution $y_{\mathrm{d}}\colon (0,T)\times\Omega \to \mathbb{R}$ are given.

Instead of the right hand side the control may enter the boundary or initial condition.

2.2.4. Identification of a source of pollution

Here, we consider a river or a lake with polluted water, occupying a two or three dimensional domain $\Omega \subset \mathbb{R}^n$, $n \in \{2, 3\}$. The control problem consists in finding the source of pollution (which is unknown). The concentration of pollutant $y(t, x)$ can be measured in a subset $\mathcal{O}$ of Ω, during the interval of time $[0, T]$ for $T > 0$. The concentration y is supposed to satisfy the equation

$$\begin{cases} y_t - \Delta y + V \cdot \nabla y + \sigma y = s(t)\delta_a & \text{in } (0,T) \times \Omega, \\ \dfrac{\partial y}{\partial n} = 0 & \text{on } \partial(0,T) \times \Omega, \\ y(x,0) = y_0, & \text{in } \Omega, \end{cases} \tag{2.9}$$

where $a \in K$ is the position of the source of pollution, K is a compact subset in Ω, $s(t)$ is the flow rate of pollution, and $V \in \mathbb{R}^n$ and $\sigma \in \mathbb{R}$ characterizing the flow. The initial concentration y_0 is supposed to be known or estimated (it could also be an unknown of the problem). The problem consists in finding $a \in K$ which minimizes

$$\int_0^T \int_{\mathcal{O}} (y(t,x) - y_{\mathrm{d}}(t,x))^2 \mathrm{d}x \mathrm{d}t, \tag{2.10}$$

where y is the solution of (2.9) and y_{d} corresponds to the measured concentration. In this problem the rate $s(t)$ is supposed to be known.

We can imagine other problems where the source of pollution is known but not accessible, and for which the rate $s(t)$ is unknown. In that case the problem consists in finding s satisfying some a priori bounds $s_0 \leq s(t) \leq s_1$ and minimizing

$$\int_0^T \int_{\mathcal{O}} (y(t,x) - y_{\mathrm{d}}(t,x))^2 \mathrm{d}x \mathrm{d}t. \tag{2.11}$$

In all four examples discussed above we have not defined in which sets the control and state variable live; this is of great importance for well-posedness and the development of numerical algorithms.

Before analyzing optimal control problems we recall some basic definitions and properties from functional analysis.

3. Finite dimensional setting

Let $J = J(u,y)$, $J : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ denote a cost functional to be minimized, and that an $n \times n$ matrix A, an $n \times m$ matrix B, and a nonempty set $U_{ad} \subset \mathbb{R}^m$ are given (where *ad* stands for admissible). We consider the optimization problem

$$\min J(u,y), \quad \text{s.t. } Ay = Bu, \quad u \in U_{ad}. \tag{3.1}$$

Often quadratic cost functionals are used, e.g.

$$J(y,u) := \tfrac{1}{2}|y - y_d|^2 + \frac{\alpha}{2}|u|^2. \tag{3.2}$$

Assume, that A is regular. Then,

$$y = A^{-1}Bu, \tag{3.3}$$

and

$$S\colon \mathbb{R}^m \to \mathbb{R}^n, \quad S := A^{-1}B \tag{3.4}$$

is well-defined. We can define the reduced cost functional

$$F(u) := J(u, Su) \tag{3.5}$$

and the reduced optimization problem

$$\min F(u), \quad u \in U_{ad}. \tag{3.6}$$

A vector $\bar{u} \in U_{ad}$ is called an *optimal control* for problem (3.1) if $F(u) \leq F(u)$ for all $u \in U_{ad}$.

Theorem 3.1. *Suppose that J is continuous on $\mathbb{R}^n \times U_{ad}$ and that the set U_{ad} is nonempty, bounded, and closed. If the matrix A is invertible, then (3.1) has at least one solution.*

Proof. Obviously, the continuity of J implies that F is also continuous on U_{ad}. Moreover, as a bounded and closed set in a finite-dimensional space, U_{ad} is compact. By the well-known Weierstrass theorem, F attains its minimum in U_{ad}. Hence, there is some $\bar{u} \in U_{ad}$ such that $F(\bar{u}) = \min F(u)$. □

We leave it as an exercise to derive the necessary optimality conditions

$$F'(\bar{u})(u - \bar{u}) \geq 0 \quad \text{for all } u \in U_{ad}. \tag{3.7}$$

Defining $\bar{p}$ as the solution of

$$A^\top \bar{p} = \nabla_y J(\bar{u}, \bar{y}), \quad \bar{y} := y[\bar{u}], \tag{3.8}$$

we have also

$$(B^\top p + \nabla_y J(\bar{y}, \bar{u}), u - \bar{u})_{\mathbb{R}^n} \geq 0 \quad \text{for all } u \in U_{ad}. \tag{3.9}$$

4. Basics

In this chapter we recall some basic results from linear functional analysis and linear PDE theory which will be of great importance in the sequel.

Contents

4.1. Linear and multilinear mappings

Let X, Y be Banach spaces (i.e. normed vector spaces in which every Cauchy sequence has a limit).

Definition 4.1. *Let $A\colon X \to Y$ be linear. Then A is* continuous *iff*

$$\|Ax\|_Y \le c\,\|x\|_X \quad \text{for all } x \in X. \tag{4.1}$$

We denote by $L(X,Y)$ the space of all linear continuous operators when endowed with the norm

$$\|A\|_{L(X,Y)} := \sup\{\|Ax\|_Y \ : \ \|x\|_X \le 1\}. \tag{4.2}$$

We often write $\|A\|$ for $\|A\|_{L(X,Y)}$. The space $L(X,Y)$ is Banach with the operator norm $\|\cdot\|$, cf. Alt [1, 5.3 Theorem, p. 142].

Theorem 4.2. *Let $E \subset X$ be a dense subspace (carrying the same norm as X). Then for all $A \in L(E,Y)$, there exists a unique extension $\mathcal{A} \in L(X,Y)$ with*

$$\mathcal{A}e = Ae \quad \text{for all } e \in E \tag{4.3}$$

and for $x \in X$ (the limit below exists):

$$\mathcal{A}x := \lim_k \{\mathcal{A}e_k \ : \ e_k \in E, e_k \to x\}. \tag{4.4}$$

Proof. For a proof we refer to Alt [1, E5.3, p.161]. □

Definition 4.3 (Multilinear mappings)**.** *Let $X := X_1 \times \cdots \times X_n$, $a\colon X \to Y$ multilinear, the X_i and Y being Banach spaces. Then a is continuous iff there exists $c > 0$ such that*

$$\|a(x_1,\dots,x_n)\|_Y \le c\,\|x_1\|_{X_1} \cdots \|x_n\|_{X_n} \text{for all } x = (x_1,\dots,x_n) \in X. \tag{4.5}$$

If E is a dense subset of X, $a\colon E \to Y$ is multilinear, and the above inequality holds for all $x \in E$, then a has a unique continuous multilinear extension to X.

Remark 4.4. *If the above multilinear mapping a is continuous, then it is of class C^∞, with derivative at $x \in X$ in direction $h \in X$ given by*

$$Da(x)h = \sum_i a(x_1,\dots,x_{i-1},h_i,x_{i+1},\dots,x_n) \tag{4.6}$$

Example 4.5. *Let p,q,r belong to $[1,\infty]$, such that $\frac{1}{r} = \frac{1}{p} + \frac{1}{q}$, and let Ω be a measurable subset of $\mathbb{R}^n$. By Hölder's inequality, the mapping $L^p(\Omega) \times L^q(\Omega) \to L^r(\Omega)$, $(f,g) \mapsto fg$ is of class C^1.*

4.2. Dual and bidual space

Definition 4.6 (Dual space)**.** *A linear form over X is a linear mapping $X \to \mathbb{R}$. We call* topological dual *(or, in short,* dual*) and denote by X^*, the set of continuous linear forms over X.*

*The action (*duality product*) of $x^* \in X^*$ over $x \in X$ is denoted by $\langle x^*, x\rangle_X$. The dual X^* is a Banach space, endowed with the dual norm*

$$\|x^*\|_{X^*} := \sup\{|\langle x^*, x\rangle_X| \ : \ \|x\|_X \leq 1\}. \tag{4.7}$$

Example 4.7. Consider for $U := C[0,1]$ the linear functional

$$f(u) := u\,(1/2) \ \text{ with } u \in U. \tag{4.8}$$

Since for all $u \in U$ we have

$$|f(u)| = |u(\tfrac{1}{2})| \leq \max_{t\in[0,1]} |u(t)| = 1 \cdot \|u\|_U\,, \tag{4.9}$$

we obtain that f is bounded with $\|f\|_{U^*} \leq 1$. For $u \equiv 1$ we have $|f(u)| = 1 = \|u\|_U$, and hence $\|f\|_{U^*} \geq 1$. Thus, we obtain $\|f\|_{U^*} = 1$.

Example 4.8. *Let $1 \leq p < \infty$ and $\frac{1}{p} + \frac{1}{q} = 1$ (with $q = \infty$ if $p = 1$). Then for $g \in L^q(\Omega)$, $\Omega \in \mathbb{R}^n$, $n \in \mathbb{N}$,*

$$T_g f := \int_\Omega f(x)g(x)\mathrm{d}x \quad \textit{for } f \in L^p(\Omega) \tag{4.10}$$

defines a linear functional $T_g \in L^p(\Omega)^*$, see Alt [1, 6.12 Theorem].

Definition 4.9 (Bidual, reflexive spaces)**.** *The bidual (dual of the dual) of X is denoted by X^{**}. With $x \in X$ associate the linear form over X^*:*

$$l_x(x^*) := \langle x^*, x\rangle_X. \tag{4.11}$$

The mapping $X \to X^{**}$, $x \mapsto l_x$ is isometric*: $\|l_x\|_{X^{**}} = \|x\|_X$. So, we can identify X with the image of l, which is a closed subspace of X^{**}.*

We say that X is reflexive *if l is onto: in that case we can identify X and X^{**}, i.e., X is the dual of X^*.*

Example 4.10. (i) By the Riesz representation theorem every Hilbert space is reflexive.

(ii) The spaces $L^p(\Omega)$ are reflexive for $1 < p < \infty$. One can show that $L^p(\Omega)^*$ can be identified with $L^q(\Omega)$ with *conjugate exponent* q of p given by the relation $\frac{1}{p} + \frac{1}{q} = 1$. More precisely, to every continuous linear functional $F \in L^p(\Omega)^*$ there corresponds a uniquely determined function $f \in L^q(\Omega)$ such that

$$F(u) = \int_\Omega f(x)u(x)\mathrm{d}x \quad \text{for all } u \in L^p(\Omega). \tag{4.12}$$

Repeating this argument we conclude that the bidual space $L^p(\Omega)^{**}$ can be identified with $L^p(\Omega)$ proving reflexivity. We remark that the continuity of the above functional is a consequence of Hölder's inequality

$$\int_\Omega |f(x)||u(x)|\mathrm{d}x \leq \left(\int_\Omega |f(x)|^q \mathrm{d}x\right)^{\frac{1}{q}} \left(\int_\Omega |u(x)|^p \mathrm{d}x\right)^{\frac{1}{p}}. \tag{4.13}$$

(iii) The spaces $L^1(\Omega)$ and $L^\infty(\Omega)$ are not reflexive.

4.3. Adjoint operator

Given $A \in L(X,Y)$, fix $y^* \in Y^*$. The mapping

$$l_{y^*} : X \to \mathbb{R}, \quad x \mapsto \langle y^*, Ax\rangle_Y \tag{4.14}$$

is continuous since $|l_{y^*}(x)| \leq \|y^*\| \, \|A\| \, \|x\|$, so that

$$\|l_{y^*}\|_{X^*} \leq \|y^*\| \, \|A\| \, . \tag{4.15}$$

Also $y^* \mapsto l_{y^*}$ is linear and (by the above inequality) continuous: so we may denote it as A^*y^*, with $A^* \in L(Y^*, X^*)$, and:

$$\langle A^*y^*, x\rangle_X = \langle y^*, Ax\rangle_Y \quad \text{for all } x \in X, y^* \in Y^*. \tag{4.16}$$

Remark 4.11. (i) We have $\|A^*\|_{L(Y^*,X^*)} = \|A\|_{L(X,Y)}$, see Alt [1, Section 12.1 Adjoint operators].

(ii) (4.16) may lead to the misconception that A^* is already explicitly determined by it (for instance, in terms of a matrix representation or via an integral operator). However, this is not to be expected, since a functional $f \in X^*$ may admit several completely different representations; cf. (4.18).

(iii) If X and Y are Hilbert spaces we have $A^* \in L(Y,X)$ and it is characterized by

$$(x, A^*y)_X = (Ax, y)_Y \quad \text{for all } x \in X, \; y \in Y. \tag{4.17}$$

In the following, we consider the explicit representation of continuous linear functionals to characterize dual spaces. Note that there may exist several possibilities to represent the same continuous linear functional; e.g.,

$$F(v) = \int_0^1 \ln(\exp(3v(x) - 5))dx + 5, \quad G(v) = 3v, \tag{4.18}$$

look quite different, but represent the same functional on $\mathbb{R}$.

4.4. Punctual convergence by a density argument

Lemma 4.12. *Let X, Y be Banach spaces and A_k be a bounded sequence in $L(X, Y)$. Let $A_k x \to 0$ for all x in a dense subset E of X. Then $A_k x \to 0$ for all $x \in X$.*

Proof. Given $\varepsilon > 0$, use

$$\|A_k x\| \leq \|A_k(x-e)\| + \|A_k e\| \leq \sup_l \|A_l\| \, \|x-e\| + \|A_k e\| \tag{4.19}$$

with $e \in E$ such that $\sup_l \|A_l\| \, \|x-e\| \leq \varepsilon$. It follows that

$$\limsup_k \|A_k x\| \leq \varepsilon \tag{4.20}$$

The result follows. □

4.5. Weak and weak* convergence

In infinite dimensional spaces bounded and closed sets are no longer compact. To obtain compactness results, one has to use the concept of weak convergence.

Definition 4.13. *Let X be a Banach space. A sequence $(x_k) \subset X$* converges weakly *to $x \in X$, written $x_k \rightharpoonup x$, if*

$$\langle x^*, x_k \rangle_X \to \langle x^*, x \rangle_X \quad \text{as } k \to 0 \quad \text{for all } x^* \in X^*. \tag{4.21}$$

Proposition 4.14.
(i) Strong convergence $x_k \to x$ implies weak convergence $x_k \rightharpoonup x$.
(ii) In finite dimensional spaces X we have

$$x_k \to x \text{ in } X \text{ iff } x_k \rightharpoonup x \text{ in } X. \tag{4.22}$$

(iii) Let X be a Banach space and let $(x_k) \subset X$ be weakly convergent to $x \in X$. Then the weak limit x is unique. Furthermore, (x_k) is bounded.

Proof. Proof of (iii): See Alt [1, p. 228/9 and Theorem 8.13, p. 241]. □

Lemma 4.15. *Let X be a reflexive Banach space. Then, any bounded sequence has a weakly convergent subsequence.*

Testing over a dense subset

Lemma 4.16. *We have that $x_k \rightharpoonup \bar{x}$ iff x_k is bounded and, for some dense subset E of X^*:*

$$\langle x^*, x_k \rangle_X \to \langle x^*, \bar{x} \rangle_X \quad \text{for all } x^* \in E. \tag{4.23}$$

Proof. Apply lemma 4.12. □

Remark 4.17 (Necessity of the boundedness hypothesis)**.** A counterexample can be constructed as follows. In $X = l^2$ (space of summable square sequence) identified with its dual, let $x_k := ke_k$ (e_k is the kth element of natural basis) and E be the subspace of elements with finitely many nonzero coordinates. The hypothesis (4.23) of the lemma holds with $\bar{x} = 0$, but x_k does not weakly converge since it is unbounded.

A direct argument for the impossibility of weak convergence is: let

$$z := \sum_k e_k / k^{\frac{2}{3}}, \tag{4.24}$$

then $z \in X$ and $(z, x_k)_X = k^{\frac{1}{3}}$ is unbounded.

Theorem 4.18 (Transportation of weak convergence by linear operators)**.** *If $A \in L(X,Y)$ (X, Y Banach spaces), then*

$$x_k \rightharpoonup \bar{x} \in X \quad \textit{implies} \quad Ax_k \rightharpoonup A\bar{x} \in Y. \tag{4.25}$$

Proof. Indeed, for any $y^* \in Y^*$:

$$\langle y^*, Ax_k \rangle_Y = \langle A^* y^*, x_k \rangle_X \to \langle A^* y^*, \bar{x} \rangle_X = \langle y^*, A\bar{x} \rangle_Y. \tag{4.26}$$

□

Weak versus strong convergence

Strong convergence implies weak convergence. When does the converse hold ? Easy answer in the case of an Hilbert space

Lemma 4.19. *Let X be a Hilbert space. If $x_k \rightharpoonup \bar{x}$, then $x_k \to \bar{x}$ iff $\|x_k\|_X \to \|\bar{x}\|_X$.*

Proof. Use

$$\begin{aligned} \limsup \|x_k - \bar{x}\|_X^2 &= \limsup \|x\|_X^2 - 2 \lim (x_k, \bar{x})_X + \|\bar{x}\|_X^2 \\ &= \limsup \|x\|_X^2 - \|\bar{x}\|_X^2 . \end{aligned} \tag{4.27}$$

□

Example 4.20. $X = L^2(0, \pi)$, $x_k(t) = \sin kt$. Sequence of constant norm, weakly (but not strongly) converging to 0.

Remark 4.21. In the case of a L^p space, see the related Brézis-Lieb theorem 16.2.

Definition 4.22 (Weak* convergence)**.** *Let X be a Banach space. We say that the sequence x_k^* in X^* weakly* converges to $x^* \in X^*$ if*

$$\langle x_k^*, x \rangle \to \langle x^*, x \rangle, \qquad \textit{for all } x \in X. \tag{4.28}$$

Then, see Brézis [10, ch. 3, Cor. 3.30]:

Lemma 4.23. *Let X be a separable Banach space (i.e., there exists a dense sequence). Then any bounded sequence in X^* has a weakly* converging subsequence.*

Example 4.24. *Let $\Omega \subset \mathbb{R}^n$ measurable. Then $X = L^1(\Omega)$ is separable. So, any bounded sequence in $X^* = L^\infty(\Omega)$ has a weakly* converging subsequence.*

4.6. Closed convex sets

Lemma 4.25. *In a Banach space closed convex sets are weakly sequentially closed.*

Proof. Let X be a Banach space and $K \subset X$ be convex and closed. By the Hahn Banach theorem (see Brézis [5], ch.1) if $\bar{x} \notin K$, there exists $x^* \in X^*$ strictly separating $\bar{x}$ from K:

$$\langle x^*, \bar{x} \rangle < \inf\{\langle x^*, x \rangle;\ x \in K\}. \tag{4.29}$$

If $(x_k) \subset K$ weakly converges to $\hat{x}$, then

$$\langle x^*, \bar{x} \rangle < \inf\{\langle x^*, x \rangle;\ x \in K\} \leq \lim_k \langle x^*, x_k \rangle = \langle x^*, \hat{x} \rangle. \tag{4.30}$$

proving that $\hat{x} \neq \bar{x}$. □

Example 4.26. Let $X := L^2(0, 2\pi)$ and

$$A = \{v \in X \ :\ \|v\|_X = 1\}. \tag{4.31}$$

This set is sequentially closed but not weakly sequentially closed. Indeed, for a sequence (u_n) with

$$u_n(x) := \frac{1}{\sqrt{\pi}} \sin(nx) \tag{4.32}$$

we have $u_n \in A$ for all n but $u_n \rightharpoonup 0$ with $0 \notin A$: For $f \in X$

$$(f, u_n)_X = \int_0^{2\pi} f(x) u_n(x) \mathrm{d}x \tag{4.33}$$

defines the nth-Fourier coefficient associated with f with respect to the orthonormal system consisting of the function $\sin(nx)/\sqrt{\pi}$, $n \in \mathbb{N}$, in X. Due to the Bessel inequality

$$\sum_{k=1}^{\infty} |(f, u_n)_X|^2 \leq \|f\|_X^2 \tag{4.34}$$

we have

$$(f, u_n)_X \to 0 \quad \text{as } n \to \infty. \tag{4.35}$$

We know $0 = (f, 0)_X$ for all $f \in X$. Consequently, (u_n) converges weakly to the zero function. On the other hand we have

$$\|u_n\|_X^2 = \int_0^{2\pi} u_n(x)^2 \mathrm{d}x = 1. \tag{4.36}$$

That means all elements of the sequence live on the unit sphere, but the weak limit is zero.

Example 4.27. Let $\Omega \subset \mathbb{R}^n$, $n \in \mathbb{N}$, be open and bounded, $X := L^2(\Omega)$, $a, b \in \mathbb{R}$, $a \leq b$, and

$$A = \{v \in X \ : \ a \leq v(x) \leq b \text{ a e.}\}. \tag{4.37}$$

The set A is convex and closed in X (see exercise) hence A is also weakly closed.

Weakly* closed convex sets.

In practice we find sets of the form, for some $E \subset X \times \mathbb{R}$:

$$K := \{x^* \in X^*;\ \langle x^*, x\rangle_X \leq \alpha, \text{ for all } (x, \alpha) \in E\}. \tag{4.38}$$

Clearly this type of set is weak* sequentially closed.

Example 4.28. Let $X = L^1(\mathbb{R}^n)$, $f \in X$ and $f^* \in X^* = L^\infty(\mathbb{R}^n)$. We say that $f \geq 0$ if $f(x) \geq 0$ a.e., and that $f^* \geq 0$ if $\langle f^*, f\rangle_X \geq 0$, for all $f \geq 0$. Thus, a weak* limit of a non-negative sequence in X^* is non-negative.

Remark 4.29 (Case of probability measures)**.** Let $\Omega \subset \mathbb{R}^n$ be compact, and $X = C(\Omega)$ be the Banach space of continuous functions over Ω, endowed with the norm

$$\|f\| := \max\{|f(x)|; x \in \Omega\}. \tag{4.39}$$

The dual space is $X^* := M(\Omega)$, space of bounded Borel measures over Ω. The set $P(\Omega)$ of Borel probability measures over Ω is a sequentially weakly* closed subset of $M(\Omega)$. Since X is separable it follows that a (necessarily) bounded sequence of probability measures over Ω has a subsequence, weakly* converging to a probability measure.

Example 4.30. We can identify $L^1(\Omega)$ with a subset of $M(\Omega)$ (set of measures having a density). So, a bounded sequence in $L^1(\Omega)$ will have a weakly* converging subsequence in $M(\Omega)$, but not necessarily a weakly converging subsequence in $L^1(\Omega)$.

4.7. Compactness

Definition 4.31. *A linear operator $K\colon X \to Y$ between normed spaces is called* compact *if for every bounded sequence $(x_n) \subset X$ the sequence $(Kx_n) \subset Y$ has a convergent subsequence.*

Remark 4.32. Compact operators are bounded operators, see Alt [1, Sec. 10.1].

Lemma 4.33. *Let X be a reflexive Banach space, Y Banach space, and $T\colon X \to Y$ a linear mapping. Then T is compact iff ($x_n \rightharpoonup x$ in X implies $Tx_n \to Tx$ in Y).*

Proof. Let $x_n \rightharpoonup x$ weakly as $n \to \infty$. By Proposition 4.14(iii), the sequence (x_n) is bounded, and so by definition of compact operators there exists a $y \in Y$ such that $Tx_n \to y$ strongly in Y for a subsequence $n \to \infty$. For $y^* \in Y^*$ the map $y \mapsto \langle y^*, Ty\rangle_Y$ defines an element in X^*. Therefore,

$$\langle y^*, Tx_n\rangle_Y \to \langle y^*, Tx\rangle_Y \quad (\text{as } n \to \infty). \tag{4.40}$$

This yields that $Tx_n \rightharpoonup Tx$ weakly in Y. As strong convergence implies weak convergence, one must have $y = Tx$. Hence $Tx_n \to Tx$ converges strongly for a subsequence $n \to \infty$. On noting that all of the above argumentation can be applied to every subsequence of (x_n), it follows that the whole sequence converges strongly to Tx.

Reversely: We have that T is continuous, and so $T \in L(X;Y)$. Moreover, by Lemma 4.15 bounded sequences in reflexive spaces contain weakly convergent subsequences. □

4.8. Weakly lower semicontinuity

Theorem 4.34. *Let X be a Banach space. Then any continuous, convex functional $F\colon X \to \mathbb{R}$ is weakly lower semicontinuous , i.e.*

$$x_k \rightharpoonup x \quad \textit{implies} \quad \liminf_{k\to\infty} F(x_k) \geq F(x). \tag{4.41}$$

Proof. See Brezis [10, p. 61]. □

Example 4.35. *Let X be a Banach space. The functionals*

$$f_1(u) := \|u\|_X \;\; \textit{and} \; f_2(u) := \|u\|_X^2 \tag{4.42}$$

are convex and strict convex, respectively, hence, weakly lower semicontinuous.

4.9. Weak derivative

Let $\Omega \subset \mathbb{R}^n$, $n \in \mathbb{N}$, be an open subset. For any function $u \in C^1(\bar{\Omega})$ there holds

$$\int_\Omega Du(x)\varphi(x)\mathrm{d}x = (-1)\int_\Omega u(x)D^\alpha\varphi(x)\mathrm{d}x \quad \text{for all } \varphi \in \mathcal{D}(\Omega) := C_c^\infty(\Omega). \tag{4.43}$$

This motivates the following definition.

Definition 4.36 (Weak derivative)**.** *Let $\Omega \subset \mathbb{R}^n$ be open and let $u \in L^1_{loc}(\Omega)$. If there exists a function $w \in L^1_{loc}(\Omega)$ such that*

$$\int_\Omega w(x)\varphi(x)\mathrm{d}x = (-1)\int_\Omega u(x)D\varphi(x)\mathrm{d}x \quad \textit{for all } \varphi \in \mathcal{D}(\Omega), \tag{4.44}$$

then $Du := w$ is called the weak partial derivative of u.

Lemma 4.37 (Uniqueness of weak derivatives)**.** *The weak derivative is unique.*

Proof. Assuming there exist two different weak derivatives w and w'. Set $g := w - w'$. Then $g \in L^1_{\text{loc}}(\Omega)$ is nonzero and

$$\int_\Omega g(x)\varphi(x)\mathrm{d}x = 0, \quad \text{for all } \varphi \in \mathcal{D}(\Omega). \tag{4.45}$$

For $\eta > 0$ small, the following set has positive measure:

$$E := \{x \in \Omega \ : \ g(x) > \eta \ : \ |x| \leq 1/\eta\}. \tag{4.46}$$

The Lebesgue measure being regular (e.g. [17, Thm. 3.2, p. 76]), there exists $K \subset E$, compact of positive measure. Set, for $\alpha > 0$:

$$\psi_\alpha(x) := (1 - \operatorname{dist}(x, K)/\alpha)_+. \tag{4.47}$$

It converges punctually to the characteristic function of K (equal to 1 over K and 0 outside). By the dominated convergence theorem,

$$\int_\Omega g(x)\psi_\alpha(x)\mathrm{d}x \to \int_K g(x)\mathrm{d}x > \eta\,\mathrm{meas}(K) > 0. \tag{4.48}$$

So, fix α such that $\int_\Omega g(x)\psi_\alpha(x)\mathrm{d}x > 0$. Observe that ψ_α is continuous with compact support. Let $\varphi_\varepsilon \in \mathcal{D}(\Omega)$ be obtained by convolution (see Appendix 16.3) of ψ_α with a regularizing kernel. Then by the dominated convergence theorem

$$\lim_{\varepsilon \downarrow 0} \int_\Omega g(x)\varphi_\varepsilon(x)\mathrm{d}x = \int_\Omega g(x)\psi_\alpha(x)\mathrm{d}x > 0. \tag{4.49}$$

But this is in contradiction with (4.45). □

Example 4.38. The function $y(x) = |x|$ in $\Omega = (-1, 1)$ has the first-order weak derivative

$$y'(x) := \begin{cases} -1, & x \in (-1, 0), \\ 1, & x \in [0, 1). \end{cases} \tag{4.50}$$

We have for each $v \in \mathcal{D}(-1, 1)$ that

$$\begin{aligned} \int_{-1}^{1} |x| v'(x) \mathrm{d}x &= \int_{-1}^{0} (-x) v'(x) \mathrm{d}x + \int_{0}^{1} x v'(x) \mathrm{d}x \\ &= -xv(x)\Big|_{-1}^{0} - \int_{-1}^{0} (-1) v(x) \mathrm{d}x + xv(x)\Big|_{0}^{1} - \int_{0}^{1} (+1) v(x) \mathrm{d}x = -\int_{-1}^{1} y'(x) v(x) \mathrm{d}x. \end{aligned} \tag{4.51}$$

The value y' at $x = 0$ is immaterial, since an isolated point has zero measure.

Higher order derivatives.

With $\alpha = (\alpha_1, \ldots, \alpha_n)$ multiindex (vector of $\mathbb{N}^n$), of order $|\alpha| := \sum_i \alpha_i$, associate the monomial (symbol) $\xi^\alpha := \xi_1^{\alpha_1} \cdots \xi_n^{\alpha_n}$, and the differential operator:

$$D^\alpha f(x) := \frac{D^{|\alpha|} f(x)}{D^{\alpha_1} x_1 \ldots D^{\alpha_n} x_n}. \tag{4.52}$$

If f and g are locally integrable, we say that $g = D^\alpha f$ in the weak sense if for all $\varphi \in D(\Omega)$:

$$(-1)^{|\alpha|} \int_\Omega f(x) D^\alpha \varphi(x) \mathrm{d}x = \int_\Omega g(x) \varphi(x) \mathrm{d}x. \tag{4.53}$$

4.10. Integration by parts

Throughout these notes we consider regular domains.

Definition 4.39 (Regular domains). *Let $\Omega \subset \mathbb{R}^n$ be open and bounded. We say that for $k \in \mathbb{N}_0 \cup \{\infty\}$, $0 \leq \beta \leq 1$,*

$$\Omega \textit{ has a } C^{k,\beta}\text{-boundary}, \tag{4.54}$$

if for any $x \in \partial\Omega$ there exists a $r > 0$, $l \in \{1, ..., n\}$, $\sigma \in \{-1, 1\}$ and a function $\gamma \in C^{k,\beta}(\mathbb{R}^{n-1})$ such that

$$\Omega \cap B(x; r) = \{y \in B(x, r) : \sigma y_l < \gamma(y_1, ..., y_{l-1}, y_{l+1}, ..., y_n)\}, \tag{4.55}$$

where $B(x, r)$ denotes the open ball around x with radius r.

Instead of $C^{0,1}$-boundary we say also Lipschitz-boundary *and the corresponding domain a* Lipschitz-domain.

We recall the integration by parts formula.

Theorem 4.40 (Gaus-Green theorem). *Let $\Omega \subset \mathbb{R}^n$, $n \in \mathbb{N}$, be open and bounded with Lipschitz-boundary. Then for all $u, v \in C^1(\bar{\Omega})$ we have*

$$\int_\Omega u_{x_i}(x)v(x)\mathrm{d}x = -\int_\Omega u(x)v_{x_i}(x)\mathrm{d}x + \int_{\partial\Omega} u(x)v(x)n_i(x)dS(x). \tag{4.56}$$

4.11. Sobolev spaces

Definition 4.41 (Sobolev spaces)**.** *Let $\Omega \subset \mathbb{R}^n$ be open. For $m \in \mathbb{N}_0$ and $p \in [1,\infty]$ we define the* Sobolev space $W^{m,p}(\Omega)$ *by*

$$W^{m,p}(\Omega) := \left\{ u \in L^p(\Omega) \middle| \begin{array}{l} \textit{u has weak derivatives } D^\alpha u \in L^p(\Omega) \\ \textit{for all } |\alpha| \leq m \textit{ with multiindex } \alpha \in \mathbb{N}_0^n \end{array} \right\} \tag{4.57}$$

equipped with the norm

$$\begin{aligned} \|u\|_{W^{m,p}(\Omega)} &:= \left(\sum_{|\alpha| \leq m} \|D^\alpha u\|_{L^p(\Omega)}^p \right)^{\frac{1}{p}} \quad \textit{for } 1 \leq p < \infty, \\ \|u\|_{W^{m,\infty}(\Omega)} &:= \sum_{|\alpha| \leq m} \|D^\alpha u\|_{L^\infty(\Omega)} . \end{aligned} \tag{4.58}$$

The space of C^m functions over Ω with bounded $W^{m,p}(\Omega)$ norm is a dense subset of $W^{m,p}(\Omega)$.

Notations: For $p = 2$ we set $H^m(\Omega) := W^{m,2}(\Omega)$. For $y \in H^1(\Omega)$ we set

$$\nabla y(x) := \big(y_{x_1}(x), \quad \ldots, \quad y_{x_n}(x)\big)^\top, \tag{4.59}$$

where y_{x_i} for $i = 1$ to n denote the weak partial derivative.

Remark 4.42. *We write $W^{0,p}(\Omega) = L^p(\Omega)$ for all $p \in [1,\infty]$. The space $H^m(\Omega)$ is a Hilbert space with inner product*

$$(u,v)_{H^m(\Omega)} := \sum_{|\alpha| \leq m} (D^\alpha u, D^\alpha v)_{L^2(\Omega)}. \tag{4.60}$$

Definition 4.43. *Let $\Omega \subset \mathbb{R}^n$ be open. For $m \in \mathbb{N}_0$ and $p \in [1,\infty]$ we define*

$$W_0^{m,p}(\Omega) := \{\textit{the closure of } \mathcal{D}(\Omega) \textit{ in } W^{m,p}(\Omega)\} \tag{4.61}$$

with $\mathcal{D}(\Omega) := C_c^\infty(\Omega)$. The space is equipped with the same norm as $W^{m,p}(\Omega)$ namely (4.58)*.*

4.12. Representation of the dual in an abstract setting

Let X, Y be Banach spaces. Given $A \in L(X, Y)$, assume that

$$\|x\|_X = \|Ax\|_Y . \tag{4.62}$$

Lemma 4.44. *We have that*

$$X^* = \{A^\top y^*; y^* \in Y^*\}. \tag{4.63}$$

That is, any $x^* \in X^*$ is such that, for some $y^* \in Y^*$:

$$\langle x^*, x\rangle_Z = \langle y^*, Ax\rangle_Y. \tag{4.64}$$

Proof. Clearly the r.h.s. of (4.63) is contained in X^* and we next prove the converse. Set $W := AX$, and let $\mathcal{A} \in L(X, W)$ defined by $\mathcal{A}x := Ax$, for all $x \in X$. Then $\mathcal{A}$ is isometric. So, given $x^* \in X^*$, $w \mapsto \langle x^*, \mathcal{A}^{-1}w\rangle_X$ is a continuous linear form over W. By a corollary of the Hahn-Banach theorem, see Brézis [10, Chap. 1, Cor. 1.2], any continuous linear form over W (with the norm induced by Y) has a continuous extension to Y, say y^*, i.e., $\langle x^*, \mathcal{A}^{-1}w\rangle_X = \langle y^*, w\rangle_Y$, and since $\mathcal{A}$ is bijective:

$$\langle x^*, x\rangle_X = \langle y^*, Ax\rangle_Y, \quad \text{for all } x \in X. \tag{4.65}$$

The conclusion follows. □

Example 4.45. Let X_1 and X_2 be Banach spaces, subspaces of a vector space X_0. Let $X := X_1 \cap X_2$ be endowed with the intersection norm

$$\|x\|_X := \|x\|_{X_1} + \|x\|_{X_2} . \tag{4.66}$$

Obviously X is a Banach space. The previous lemma applies with $Ax := (J_1x, J_2x)$, where J_i is the injection of X into X_i, for $i = 1, 2$. Then $J_i^\top \in L(X_i^*, X^*)$, $J_i^\top x_i^*$ is the restriction of $x_i^* \in X_i^*$ to X, and

$$A^\top(x_1^*, x_2^*) = J_1^* x_1^* + J_2^* x_2^* \tag{4.67}$$

can be identified with the sum $x_1^* + x_2^*$. We have proved that

$$(X_1 \cap X_2)^* = X_1^* + X_2^*. \tag{4.68}$$

In addition, $\|J_i^*\| = \|J_i\| \leq 1$, so that for any $x^* \in X^*$:

$$\|x^*\|_{X^*} \leq \inf\{\|x^*\|_{X_1^*} + \|x_2^*\|_{X_2^*} \ : \ J_1^* x_1^* + J_2^* x_2^* = x^*\}. \tag{4.69}$$

4.13. Representation of the dual of $W^{m,p}(\Omega)$

.

Lemma 4.46. *With any $f \in W^{m,p}(\Omega)^*$ are associated $g_\alpha \in L^q(\Omega)$, $1/p + 1/q = 1$, $|\alpha| \leq m$, such that*

$$\langle f, u \rangle_{W^{m,p}} = \sum_{|\alpha| \leq m} \int_\Omega g_\alpha(x) D^\alpha u(x) \mathrm{d}x. \tag{4.70}$$

Proof. Apply lemma 3.2 and use $L^p(\Omega)^* = L^q(\Omega)$. □

Representation of the dual of $W_0^{m,p}(\Omega)$. Since $W_0^{m,p}(\Omega)$ is, by the definition, a closed subspace of $W^{m,p}(\Omega)$, by the previous lemma and the Hahn Banach theorem, with any $f \in W_0^{m,p}(\Omega)^*$, are associated $g_\alpha \in L^q(\Omega)$, $1/p + 1/q = 1$, $|\alpha| \leq m$, such that

$$\langle f, u \rangle_{W_0^{m,p}} = \sum_{|\alpha| \leq m} g_\alpha(x) D^\alpha u(x) \mathrm{d}x. \tag{4.71}$$

Since $\mathcal{D}(\Omega)$ is a dense subset of $W_0^{m,p}(\Omega)$, we deduce that, in a weak sense (of distributions, see [1]),

$$f = \sum_{|\alpha| \leq m} (-1)^{|\alpha|} D^\alpha g_\alpha. \tag{4.72}$$

Example 4.47. Any $f \in H_0^1(\Omega)^*$ is of the form

$$f = g_0 + \sum_{i=1}^{n} \frac{\partial g_i}{\partial x_i}; \quad g_0, \ldots, g_n \quad \text{in} \quad L^2(\Omega). \tag{4.73}$$

Note, the decomposition is not unique.

We write $H^{-1}(\Omega) := (H_0^1(\Omega))^*$.

4.14. Trace

Rather than entering into the detail of geometric hypotheses on the boundary, let us assume that Ω is bounded with a C^1 boundary, or is a half-space. Then the following properties hold, see [15]:

1. $C^\infty(\bar{\Omega}) \cap H^1(\Omega)$ is a dense subset of $H^1(\Omega)$. Consider the trace mapping

$$\tau \colon C^1(\bar{\Omega}) \cap H^1(\Omega) \to C(\partial\Omega). \tag{4.74}$$

2. For some $c > 0$,

$$\|\tau v\|_{L^2(\partial\Omega)} \leq c \|v\|_{H^1(\Omega)}, \tag{4.75}$$

so that τ has a unique continuous extension (still denoted by τ) from $H^1(\Omega)$ into $L^2(\partial\Omega)$, which has dense range, and kernel equal to $H^1_0(\Omega)$.

3. The range of τ, denoted by $H_\tau(\partial\Omega)$, is endowed with the trace norm

$$\|z\|_\tau := \min_{u \in H^1(\Omega)} \|y\|_{H^1(\Omega)}; \quad \tau y = z. \tag{4.76}$$

Since $H^1(\Omega)$ is a Hilbert space, we may write

$$H^1(\Omega) = H^1_0(\Omega) \oplus H^1_0(\Omega)^\perp. \tag{4.77}$$

The solution of the above problem is the unique $y \in H^1_0(\Omega)^\perp$, with trace z, or $y = \tau^\dagger z$, where $\tau^\dagger$ is the (continuous) pseudo-inverse (inverse of minimum norm) of τ, and so,

$$\|z\|_\tau = \left\|\tau^\dagger z\right\|_{H^1(\Omega)}. \tag{4.78}$$

$H_\tau(\partial\Omega)$ is a Hilbert space.

Remark 4.48 (More on $\tau^\dagger$). Given $z \in H_\tau(\partial\Omega)$, we have that $y = \tau^\dagger z$ is the unique solution of $\tau y = z$ that is orthogonal to $H^1_0(\Omega)$, i.e.:

$$\int_\Omega (y(x)v(x) + \nabla y(x) \cdot \nabla v(x))\mathrm{d}x = 0, \quad \text{for all } v \in H^1_0(\Omega), \tag{4.79}$$

or equivalently

$$y - \Delta y = 0 \text{ in } \Omega; \quad y = z \quad \text{on} \quad \partial\Omega. \tag{4.80}$$

Remark 4.49 (Inclusions of boundary spaces)**.** It turns out that $H_\tau(\partial\Omega)$ is isomorphic to $H^{\frac{1}{2}}(\partial\Omega)$. We have the (strict when $n > 1$) dense and compact inclusions

$$H^1(\partial\Omega) \subset H^{\frac{1}{2}}(\partial\Omega) \subset L^2(\partial\Omega). \tag{4.81}$$

In particular, a bounded sequence in $H^{\frac{1}{2}}(\partial\Omega)$ has a strongly converging subsequence in $L^2(\partial\Omega)$.

About the normal derivative. We know that if Ω, u, v are smooth:

$$\int_{\partial\Omega} \partial_n u(x)v(x)\mathrm{d}x = \int_\Omega v(x)\Delta u(x)\mathrm{d}x + \int_\Omega \nabla v(x)\cdot\nabla u(x)\mathrm{d}x. \tag{4.82}$$

Call $b(u,v)$ the r.h.s.: we can identify it with a continuous bilinear form over $H^1_\Delta(\Omega)\times H^1(\Omega)$, where

$$H^1_\Delta(\Omega) := \{v \in H^1(\Omega); \Delta v \in L^2(\Omega)\}. \tag{4.83}$$

Then define the normal derivative as the operator $N : H^1_\Delta(\Omega) \to H^1(\Omega)^*$, such that

$$\langle N(u), v\rangle_{H^1(\Omega)} := b(u,v). \tag{4.84}$$

Now take $u \in H^1_\Delta(\Omega)$, $z \in H^{\frac{1}{2}}(\partial\Omega)$ and $v = v_z$, $v_z := \tau^\dagger z$. Then

$$z \mapsto \int_\Omega v_z(x)\Delta u(x)\mathrm{d}x + \int_\Omega \nabla v_z(x)\cdot\nabla u(x)\mathrm{d}x \tag{4.85}$$

is a continuous linear form over $H^{\frac{1}{2}}(\partial\Omega)$, which we may call normal derivative of u. This normal derivative operator $u \mapsto \partial_n u$ is a continuous linear mapping:

$$H^1_\Delta(\Omega) \to H^{\frac{1}{2}}(\partial\Omega)^* \tag{4.86}$$

defined by: for all $u \in H^1_\Delta(\Omega)$:

$$\langle \partial_n u, z\rangle_{H^{\frac{1}{2}}(\partial\Omega)} := \int_\Omega v_z(x)\Delta u(x)\mathrm{d}x + \int_\Omega \nabla v_z(x)\cdot\nabla u(x)\mathrm{d}x. \tag{4.87}$$

Remark 4.50. *When $u \in H^2(\Omega)$, one can identify its trace with an element of the space $H^{\frac{3}{2}}(\partial\Omega)$, and its normal derivative with an element of the space $H^{\frac{1}{2}}(\partial\Omega)$, see [15].*

Remark 4.51. *$W^{m,p}_0(\Omega)$ contains exactly all $u \in W^{m,p}(\Omega)$ such that $D^\alpha u = 0$ for $|\alpha| \le m-1$ on $\partial\Omega$ with the interpretation of $D^\alpha u|_{\partial\Omega}$ in the sense of traces. For $p = 2$ we write $H^m_0(\Omega)$.*

4.15. Poincare inequality

Proposition 4.52. *Let $\Omega \subset \mathbb{R}^n$ be open and bounded. Then there exists a constant $C > 0$ with*

$$\int_\Omega |u(x)|^2 \mathrm{d}x \leq c \int_\Omega |\nabla u(x)|^2 \mathrm{d}x \quad \textit{for all } u \in H_0^1(\Omega). \tag{4.88}$$

Proof. See Alt [1, Sec. 6.7]. □

Remark 4.53. *We associate with $H_0^1(\Omega)$, defined in (4.61), the to $\|\cdot\|_{W^{1,2}(\Omega)}$ equivalent norm namely $\|\nabla \cdot\|_{L^2(\Omega)}$.*

4.16. Sobolev embedding theorem

Definition 4.54. *Let X and Y be normed vector spaces. A linear operator $A\colon X \to Y$ is called a* continuous embedding *($X \hookrightarrow Y$), if it is continuous, linear, and injective.*

Theorem 4.55. *Let $\Omega \subset \mathbb{R}^n$ be open, bounded with Lipschitz-boundary.*

(i) Let $m \in \mathbb{N}$, $1 \leq p < \infty$. For all $k \in \mathbb{N}_0$, $0 \leq \beta \leq 1$ with

$$m - \frac{n}{p} \geq k + \beta \quad \textit{and} \quad 0 < \beta < 1 \tag{4.89}$$

one has the continuous embedding

$$W^{m,p}(\Omega) \hookrightarrow C^{k,\beta}(\bar{\Omega}). \tag{4.90}$$

That means, there exists a $C > 0$ such that for all $u \in W^{m,p}(\Omega)$ possibly after a modification of measure zero $u \in C^{k,\beta}(\bar{\Omega})$ and

$$\|u\|_{C^{k,\beta}(\bar{\Omega})} \leq C \, \|u\|_{W^{m,p}(\Omega)} \, . \tag{4.91}$$

The embedding is compact if the inequality in (4.89) is strict.

(ii) For $m_1 \geq 0, m_2 \geq 0$ and $1 \leq p_1 < \infty$, $1 \leq p_2 < \infty$. If

$$m_1 - \frac{n}{p_1} \geq m_2 - \frac{n}{p_2}, \qquad \textit{and } m_1 \geq m_2, \tag{4.92}$$

then we have the continuous embedding

$$W^{m_1,p_1}(\Omega) \hookrightarrow W^{m_2,p_2}(\Omega). \tag{4.93}$$

The embedding is compact if the inequalities in (4.92) are satisfied strictly.

Proof. See Alt [1, p.333 and p.338]. □

Example 4.56. *For $n \leq 3$ we have the continuous embedding $H^1(\Omega) \hookrightarrow L^6(\Omega)$ and the compact embedding $H^2(\Omega) \subset C(\bar{\Omega})$.*

4.17. Linear elliptic PDEs

We consider the elliptic boundary value problem

$$\begin{cases} -\Delta y = f & \text{in } \Omega, \\ y = 0 & \text{in } \partial\Omega \end{cases} \tag{4.94}$$

for $f \in L^2(\Omega)$, where $\Omega \subset \mathbb{R}^n$ is an open, bounded set. This admits in particular discontinuities in the right hand side f, e.g. a source term which acts only on parts of Ω. Since a classical solution $y \in C^2(\Omega) \cap C^0(\bar{\Omega})$ exists at best for continuous right hand sides, we need a generalized solution concept.

We assume that $y \in C^2(\Omega) \cap C^1(\bar{\Omega})$ is a classical solution of (4.94). Then we have $y \in H^1_0(\Omega)$ by Remark 4.51. Multiplying by $v \in \mathcal{D}(\Omega)$ and integrating over Ω we have

$$\int_\Omega -\Delta y(x)v(x)\mathrm{d}x = \int_\Omega f(x)v(x)\mathrm{d}x \quad \text{for all } v \in \mathcal{D}(\Omega). \tag{4.95}$$

We can easily check that (4.94) and (4.95) are equivalent. Integration by parts gives (omitting the argument x)

$$-\int_\Omega y_{x_i x_i} v \mathrm{d}x = \int_\Omega y_{x_i} v_{x_i} \mathrm{d}x - \int_{\partial\Omega} y_{x_i} v n_i \mathrm{d}S(x) = \int_\Omega y_{x_i} v_{x_i} \mathrm{d}x. \tag{4.96}$$

Note, that the boundary integral vanishes since $v|_{\partial\Omega} = 0$. Thus (4.94) is equivalent to

$$\int_\Omega \nabla y \cdot \nabla v \mathrm{d}x = \int_\Omega f v \mathrm{d}x \quad \text{for all } v \in \mathcal{D}(\Omega). \tag{4.97}$$

Definition 4.57. *Let $f \in L^2(\Omega)$. A function $y \in H^1_0(\Omega)$ is called a weak solution of the boundary value problem (4.94) if it satisfies the* variational formulation *or* weak formulation

$$(\nabla y, \nabla v)_{L^2(\Omega)} = (f, v)_{L^2(\Omega)} \quad \textit{for all } v \in H^1_0(\Omega). \tag{4.98}$$

To allow the treatment of more general equations we introduce the notation

$$V := H^1_0(\Omega), \quad a(y,v) := (\nabla y, \nabla v)_{L^2(\Omega)}, \quad F(v) := (f,v)_{L^2(\Omega)}. \tag{4.99}$$

Then, $a \colon V \times V \to \mathbb{R}$ is a bilinear form, $F \in V^*$ is a linear functional on V and (4.98) can be written as

$$\text{Find } y \in V\colon\ a(y,v) = F(v) \quad \text{for all } v \in V. \tag{4.100}$$

4.18. Lax-Milgram - Existence and uniqueness

Theorem 4.58 (Riesz). *With any linear continuous form L on a Hilbert space H we can associate $y \in H$ such that*

$$L(v) = (y, v)_H, \quad \textit{for all } v \textit{ in } H. \tag{4.101}$$

Theorem 4.59 (Lax-Milgram). *Let V be a real Hilbert space with inner product $(\cdot,\cdot)_V$ and let $a\colon V \times V \to \mathbb{R}$ be a bilinear form that satisfies with constants $\alpha_0 > 0$ and $\beta_0 > 0$*

$$\begin{aligned} |a(y,v)| &\leq \alpha_0 \|y\|_V \|v\|_V && \textit{for all } y, v \in V && \textit{(boundedness)},\\ a(y,y) &\geq \beta_0 \|y\|_V^2 && \textit{for all } y \in V && \textit{(coercivity)}. \end{aligned} \tag{4.102}$$

Then for any bounded linear functional $F \in V^*$ the variational equation (4.100) has a unique solution $y \in V$. Moreover, y satisfies

$$\|y\|_V \leq \frac{1}{\beta_0} \|F\|_{V^*}. \tag{4.103}$$

Proof. See Alt [1, p. 164]. □

Remark 4.60. *Given $y \in V$, the mapping $v \mapsto a(y,v)$ defines a linear form on V that we may denote by $\mathcal{A}[y]$. Since $a(\cdot,\cdot)$ is continuous, we have that for all y, v in V:*

$$|\mathcal{A}[y]v| = |a(y,v)| \leq c \|y\|_V \|v\|_V, \tag{4.104}$$

and so $\mathcal{A}[y]$ is a continuous linear form. In addition

$$A\colon V \to V^*, \quad y \mapsto \mathcal{A}[y] \tag{4.105}$$

is linear and continuous since

$$\|A\| = \sup\{\|\mathcal{A}[y]\| \mid \|y\|_V \leq 1\} = \sup\{|a(y,v)| \mid \|y\|_V \leq 1, \ \|v\|_V \leq 1\} = C, \tag{4.106}$$

cf. Bonnans [7].

Remark 4.61. *In view of the previous remark we see that the equation (4.100) is equivalent to*

$$Ay = F \textit{ in } V^*. \tag{4.107}$$

Furthermore, the operator A satisfies

$$A^{-1} \in L(V^*, V), \quad \left\|A^{-1}\right\|_{V^*,V} \leq \frac{1}{\beta_0}. \tag{4.108}$$

Theorem 4.62 (Existence and uniqueness). *Let $\Omega \subset \mathbb{R}^n$ be open and bounded. Then the bilinear form a in (4.99) is bounded and V-coercive for $V = H_0^1(\Omega)$ and the associated operator $A \in L(V, V^*)$ in (4.105) has a bounded inverse. In particular, (4.94) has for all $f \in L^2(\Omega)$ a unique weak solution $y \in H_0^1(\Omega)$ given by (4.98) and satisfies*

$$\|y\|_{H_0^1(\Omega)} \leq c \|f\|_{L^2(\Omega)}, \tag{4.109}$$

where c depends on Ω but not on f.

Proof. It can be easily checked that a is *bilinear*. The *boundedness* follows from

$$|a(y, v)| = |(\nabla y(x), \nabla v(x))_{L^2(\Omega)}| \leq \|\nabla y\|_{L^2(\Omega)} \|\nabla v\|_{L^2(\Omega)}. \tag{4.110}$$

a is V-coercive:

$$a(y, y) = \int_\Omega \nabla y \cdot \nabla y \, \mathrm{d}x = \|\nabla y\|_{L^2(\Omega)}^2. \tag{4.111}$$

For the *linear form F* we have

$$\|F\|_{V^*} = \sup_{\|v\|_V=1} F(v) = \sup_{\|v\|_V=1} (f, v)_{L^2(\Omega)} \leq \sup_{\|v\|_V=1} \|f\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} \leq C \|f\|_{L^2(\Omega)}. \tag{4.112}$$

Estimate (4.109) follows from (4.103) and (4.112). □

Under more assumptions on the data we can show higher regularity of the solution.

Theorem 4.63 (Higher regularity). *Let $\Omega \subset \mathbb{R}^n$ be open, bounded with C^2-boundary. Then for any $f \in L^2(\Omega)$ the weak solution $y \in H_0^1(\Omega)$ of the Dirichlet problem (4.94) satisfies $y \in H^2(\Omega)$ and we have the estimate*

$$\|y\|_{H^2(\Omega)} \leq C(\|y\|_{H^1(\Omega)} + \|f\|_{L^2(\Omega)}), \tag{4.113}$$

where C does not depend on f.

Proof. See, e.g., Alt [1, A12.3]. □

4.19. Boundary conditions of Robin type

We consider equations of type

$$\begin{cases} -\Delta y + c_0 y = f & \text{in } \Omega, \\ \dfrac{\partial y}{\partial n} + \alpha y = g & \text{on } \partial\Omega, \end{cases} \tag{4.114}$$

where $f \in L^2(\Omega)$ and $g \in L^2(\partial\Omega)$ are given and $c_0 \in L^\infty(\Omega)$ and $\alpha \in L^\infty(\partial\Omega)$ are non-negative coefficients.

For a classical solution y of (4.114) we have for any test function $v \in C^1(\bar{\Omega})$ after integration by parts

$$\begin{aligned} (f, v)_{L^2(\Omega)} &= (-\Delta y + c_0 y, v)_{L^2(\Omega)} \\ &= (\nabla y, \nabla v)_{L^2(\Omega)} + (c_0 y, v)_{L^2(\Omega)} - \int_{\partial\Omega} \partial_n y v \mathrm{d}x \quad \text{for all } v \in C^1(\bar{\Omega}). \end{aligned} \tag{4.115}$$

Inserting the boundary condition $\frac{\partial y}{\partial n} = -\alpha y + g$ we arrive at

$$(\nabla y, \nabla v)_{L^2(\Omega)} + (c_0 y, v)_{L^2(\Omega)} + (\alpha y, v)_{L^2(\partial\Omega)} = (f, v)_{L^2(\Omega)} + (g, v)_{L^2(\partial\Omega)} \quad \text{for all } v \in H^1(\Omega). \tag{4.116}$$

The extension to $H^1(\Omega)$ is possible, since for $y \in H^1(\Omega)$ both sides are continuous with respect to $v \in H^1(\Omega)$ and since $C^1(\bar{\Omega})$ is dense in $H^1(\Omega)$.

Definition 4.64. *A function $y \in H^1(\Omega)$ is called a* weak solution *of the problem* (4.114)*, if it satisfies the* variational formulation *or* weak formulation (4.116)*.*

To apply the general theory we set

$$\begin{aligned} V &:= H^1(\Omega), && (4.117) \\ a(y, v) &:= (\nabla y, \nabla v)_{L^2(\Omega)} + (c_0 y, v)_{L^2(\Omega)} + (\alpha y, v)_{L^2(\partial\Omega)} \quad \text{for all } y, v \in V, && (4.118) \\ F(v) &:= (f, v)_{L^2(\Omega)} + (g, v)_{L^2(\partial\Omega)}. && (4.119) \end{aligned}$$

Theorem 4.65 (Existence and uniqueness for the Robin boundary problem)**.** *Let $\Omega \subset \mathbb{R}^n$ open and bounded with Lipschitz boundary and let*

$$\begin{cases} c_0 \in L^\infty(\Omega) \text{ and } \alpha \in L^\infty(\partial\Omega) \text{ be nonnegative,} \\ \|c_0\|_{L^2(\Omega)} + \|\alpha\|_{L^2(\partial\Omega)} > 0. \end{cases} \tag{4.120}$$

Then the bilinear form a defined in (4.117) is bounded and V-coercive for $V = H^1(\Omega)$ and the associated operator $A \in L(V, V^*)$ in (4.105) has a bounded inverse. In particular (4.114) has for all $f \in L^2(\Omega)$ and $g \in L^2(\partial\Omega)$ a unique weak solution $y \in H^1(\Omega)$ given by (4.116) and satisfies

$$\|y\|_{H^1(\Omega)} \leq c(\|f\|_{L^2(\Omega)} + \|g\|_{L^2(\partial\Omega)}) \tag{4.121}$$

with $c > 0$ depending on Ω, c_0, and α but not on f and g.

Proof. This follows from the Lax-Milgram theorem. The boundedness of $a(y, v)$ and $F(v)$ follows from the trace theorem. The V-coercivity is a consequence of the generalized Poincare inequality which says that for $\Gamma_1 \subset \partial\Omega$ be a measurable set such that $|\Gamma_1| > 0$. Then there exists a constant $c(\Gamma_1) > 0$ independent from $y \in H^1(\Omega)$ with

$$\|y\|^2_{H^1(\Omega)} \leq c(\Gamma_1) \left(\int_\Omega |\nabla y|^2 \mathrm{d}x + \left(\int_{\Gamma_1} y^2 \mathrm{d}x \right) \right) \tag{4.122}$$

for all $y \in H^1(\Omega)$, see [20, p.35] and [22]. □

4.20. Existence and uniqueness for semilinear elliptic equations

We consider uniformly elliptic equations of type

$$\begin{cases} Ly + d(x,y) = f & \text{in } \Omega, \\ \dfrac{\partial y}{\partial n} + \alpha y + b(x,y) = g & \text{on } \partial\Omega, \end{cases} \tag{4.123}$$

where the operator L is given by

$$Ly = -\sum_{i,j=1}^{n} (a_{ij} u_{x_i})_{x_j} + c_0 y, \quad a_{ij}, c_0 \in L^\infty(\Omega), \quad c_0 \geq 0, \quad a_{ij} = a_{ji}, \tag{4.124}$$

and L is assumed to be uniformly elliptic in the sense that for $\theta > 0$

$$\sum_{i,j=1}^{n} a_{ij}\xi_i\xi_j \geq \theta \left\|\xi\right\|^2 \quad \text{for almost all } x \in \Omega \text{ and all } \xi \in \mathbb{R}^n. \tag{4.125}$$

Moreover, we assume that the function $0 \leq \alpha \in L^\infty(\partial\Omega)$ and that $d \colon \Omega \times \mathbb{R} \to \mathbb{R}$ and $b \colon \partial\Omega \times \mathbb{R} \to \mathbb{R}$ satisfy

$$\begin{cases} \text{(i) } d(x,\dot{\ }) \text{ is continuous and monotone increasing for a.a. } x \in \Omega, \\ \text{(ii) } b(x,\cdot) \text{ is continuous and monotone increasing for a.a. } x \in \partial\Omega, \\ \text{(iii) } d(\cdot,y) \in L^\infty(\Omega),\ b(\cdot,y) \in L^\infty(\partial\Omega) \text{ for all } y \in \mathbb{R}. \end{cases} \tag{4.126}$$

Using the definition of a as in (4.99) we define a weak formulation by

$$\begin{cases} \text{Find } y \in V := H^1(\Omega): \\ \quad a(y,v) + (d(\cdot,y),v)_{L^2(\Omega)} + (b(\cdot,y),v)_{L^2(\partial\Omega)} = (f,v)_{L^2(\Omega)} + (g,v)_{L^2(\partial\Omega)} \\ \hfill \text{for all } v \in V \end{cases} \tag{4.127}$$

with the bilinear form (4.118).

Theorem 4.66 (Existence and uniqueness for the semilinear elliptic problem)**.** *Let $\Omega \subset \mathbb{R}^n$ be open and bounded with Lipschitz boundary and let*

$$\begin{cases} c_0 \in L^\infty(\Omega) \text{ and } \alpha \in L^\infty(\partial\Omega) \text{ be nonnegative}, \\ \left\|c_0\right\|_{L^2(\Omega)} + \left\|\alpha\right\|_{L^2(\partial\Omega)} > 0 \end{cases} \tag{4.128}$$

and let (4.125) *and* (4.126) *be satisfied. Moreover, let*

$$r > \frac{n}{2}, \quad s > n-1, \quad n \geq 2. \tag{4.129}$$

Then (4.123) *has for any* $f \in L^r(\Omega)$ *and* $g \in L^s(\partial\Omega)$ *a unique weak solution* $y \in H^1(\Omega) \cap C(\bar{\Omega})$. *There exists a* $c > 0$ *such that*

$$\|y\|_{H^1(\Omega)} + \|y\|_{C(\bar{\Omega})} \leq c(\|f - d(\cdot, 0)\|_{L^r(\Omega)} + \|g - b(\cdot, 0)\|_{L^s(\partial\Omega)}) \tag{4.130}$$

with c *depending on* Ω, c_0, *and* α *but not on* f, g, b, d.

Proof. For a proof see, e.g., Tröltzsch [20, Section 4.2], for the continuity Murthy and Stampacchia [18]. □

5. Existence of optimal controls

In this chapter we consider optimal control problems for linear and semilinear elliptic equations and prove existence of solutions.

Contents

5.1. The linear-quadratic case

We consider the linear-quadratic optimization problem

$$\begin{cases} \min\limits_{(y,u)\in U\times Y} J(u,y) := \frac{1}{2}\left\|Qy - y_{\mathrm{d}}\right\|_H^2 + \dfrac{\alpha}{2}\left\|u\right\|_U^2 \\ \text{subject to} \quad Ay + Bu = g, \quad u \in U_{\mathrm{ad}}, \quad y \in Y_{\mathrm{ad}}, \end{cases} \tag{5.1}$$

with H, U are Hilbert spaces, Y, Z are Banach spaces and

$$y_{\mathrm{d}} \in H, \quad g \in Z, \quad A \in \mathcal{L}(Y,Z), \quad B \in \mathcal{L}(U,Z), \quad Q \in \mathcal{L}(Y,H) \tag{5.2}$$

and let the following assumptions hold:

Assumption 5.1.

1. *$\alpha \geq 0$, $U_{ad} \subset U$ is nonempty, convex, closed and in the case $\alpha = 0$ bounded.*
2. *$Y_{ad} \subset Y$ is convex and closed, such that (5.1) has a feasible point[1].*
3. *A has a bounded inverse.*

[1] A point $(u,y) \in U_{\mathrm{ad}} \times Y_{\mathrm{ad}}$ is *feasible* if it satisifies the state equation.

Definition 5.2. *The pair $(\bar{u}, \bar{y}) \in U_{ad} \times Y_{ad}$ is called a solution for (5.1), if $A\bar{y} + B\bar{u} = g$ and*

$$J(\bar{u}, \bar{y}) \leq J(u, y) \quad \text{for all } (u, y) \in U_{ad} \times Y_{ad} \text{ with } Ay + Bu = g. \tag{5.3}$$

Theorem 5.3 (Existence and uniqueness)**.** *Let Assumption 5.1 hold. Then problem (5.1) has a non-empty set of solutions. If $\alpha > 0$ there exists a unique solution.*

Proof. At first we assume that Y is reflexiv, since this proof can be easily extended to nonlinear problems. The modifications for general Y will be mentioned at the end.

On exsistence:

(i) Let the feasible set be defined by

$$F_{\text{ad}} := \{(u, y) \in U \times Y \;:\; (u, y) \in U_{\text{ad}} \times Y_{\text{ad}},\; Ay + Bu = g\}. \tag{5.4}$$

Since $J \geq 0$ and F_{ad} non-empty, the infimum

$$J^* := \inf_{(u,y) \in F_{\text{ad}}} J(u, y) \tag{5.5}$$

exists and hence we find a minimizing sequence $(u_k, y_k) \subset F_{\text{ad}}$ with

$$\lim_{k \to \infty} J(u_k, y_k) = J^*. \tag{5.6}$$

(ii) By assumption either U_{ad} is bounded or $\alpha > 0$. In the latter case boundedness follows from

$$J(u_N, y_N) \geq J(u_k, y_k) \geq \frac{\alpha}{2} \|u_k\|_U^2 \quad \text{for all } k \geq N \tag{5.7}$$

and some $N \in \mathbb{N}$. Since

$$B \in \mathcal{L}(U, Z), \quad A^{-1} \in \mathcal{L}(Z, Y), \tag{5.8}$$

the sequence

$$y_k = A^{-1}(g - Bu_k) \tag{5.9}$$

is bounded. Thus,

$$(u_k, y_k) \subset F_{\text{ad}} \cap (\bar{B}_U(r) \times \bar{B}_Y(r)) =: M \tag{5.10}$$

for $r > 0$ large enough, where $(\bar{B}_U(r) \times \bar{B}_Y(r))$ denotes the closed ball of radius r in Y and U around zero.

(iii) By assumption $U_{\mathrm{ad}} \times Y_{\mathrm{ad}}$ is closed and convex, hence F_{ad} is closed and convex. Thus, the set M is bounded, convex, and closed. Therefore, there exists a weakly convergent subsequence $(u_{k_i}, y_{k_i}) \subset (u_k, y_k)$ and some $(\bar{u}, \bar{y}) \in F_{\mathrm{ad}}$ with

$$F_{\mathrm{ad}} \ni (u_{k_i}, y_{k_i}) \rightharpoonup (\bar{u}, \bar{y}), \quad (i \to \infty). \tag{5.11}$$

(iv) Finally, $Y \times U \to \mathbb{R}$, $(u, y) \mapsto J(u, y)$ is obviously continuous and convex. We conclude by Theorem 4.34 that

$$J^* = \lim \inf_{l \to \infty} J(u_{k_l}, y_{k_l}) \geq J(\bar{u}, \bar{y}) \geq J^*, \tag{5.12}$$

where the last inqeuality follows from $(\bar{u}, \bar{y}) \in F_{\mathrm{ad}}$. Thus, $(\bar{u}, \bar{y})$ is a solution of (5.1).

On uniqueness: If $\alpha > 0$, then $u \mapsto J(u, A^{-1}(g - Bu))$ is strictly convex, which contradicts the existence of more than one minimizer.

General Y: If Y is not reflexive, we can still select a sequence $(u_{k_i}) \subset (u_k)$ since U is reflexive. Since

$$y_{k_i} = A^{-1}(g - Bu_{k_i}), \quad A^{-1}B \in \mathcal{L}(U, Y), \tag{5.13}$$

also the subsequence (u_{k_i}, y_{k_i}) converges weakly in $Y \times U$ and we obtain as above $F_{\mathrm{ad}} \ni (u_{k_i}, y_{k_i}) \rightharpoonup (\bar{u}, \bar{y})$ as $i \to \infty$. □

Reduced problem

Since $Ay + Bu = g$ implies $y = A^{-1}(g - Bu)$ problem (5.1) is equivalent to

$$\min_{u \in U} F(u) \quad \text{s.t.} \quad u \in \hat{U}_{\mathrm{ad}} \tag{5.14}$$

with

$$F(u) := J(u, A^{-1}(g - Bu)), \quad \hat{U}_{\mathrm{ad}} := \{u \in U \ : \ u \in U_{\mathrm{ad}},\ A^{-1}(g - Bu) \in Y_{\mathrm{ad}}\}. \tag{5.15}$$

One can easily verify that F is continuous and convex, $\hat{U}_{\mathrm{ad}}$ is closed and convex. We proceed as above. Setting $\bar{y} = A^{-1}(g - B\bar{u})$, we obtain a solution of (5.1).

5.2. Application: Distributed control

We consider a distributed control problem for elliptic equations.

$$\begin{cases} \displaystyle\min_{(u,y)\in L^2(\Omega)\times H_0^1(\Omega)} J(u,y) := \frac{1}{2}\|y-y_{\mathrm{d}}\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u\|^2_{L^2(\Omega)}\,, \\ -\Delta y = \gamma u, \quad \text{in } \Omega, \quad y = 0 \ \text{ on } \partial\Omega, \\ u_m \le u \le u_M \quad \text{a.e. in } \Omega, \end{cases} \tag{5.16}$$

where $\gamma \in L^\infty(\Omega)\setminus\{0\}$, $\gamma \ge 0$, $y_{\mathrm{d}} \in L^2(\Omega)$, $u_m, u_M \in L^2(\Omega)$, $u_m \le u_M$, and $\alpha \ge 0$.

We make the following choice for the set of admissible controls and state space:

$$U := L^2(\Omega), \quad U_{\mathrm{ad}} := \{u \in U \ : \ u_m \le u \le u_M \text{ a.e.}\}, \quad Y_{\mathrm{ad}} := Y := H_0^1(\Omega). \tag{5.17}$$

The weak formulation of the state equations can be written in the form

$$\text{Find } y \in Y\colon\ a(y,v) = (\gamma u, v)_{L^2(\Omega)} \quad \text{for all } v \in Y \tag{5.18}$$

with $a(y,v) := (\nabla y, \nabla v)_{L^2(\Omega)}$, or short

$$Ay + Bu = 0, \tag{5.19}$$

where $A \in \mathcal{L}(Y,Y^*)$ represents a (see (4.107)), and $B \in L(U,Y^*)$ is given by $Bu = -(\gamma u,\cdot)_{L^2(\Omega)}$. By Theorem 4.62 the operator $A \in L(Y,Y^*)$ has a bounded inverse. Thus, Assumption 5.1 is satisfied with $Z = Y^*$. By setting $H := U$, $g = 0$ and $Q = I_{Y,U}$ with the trivial, continuous embedding

$$I_{Y,U}\colon Y \to U, y \mapsto y, \tag{5.20}$$

(5.16) is equivalent to (5.1).

5.3. The nonlinear case

The existence result for linear quadratic problems can be extended to nonlinear problems of type

$$\min_{(u,y)\in U\times Y} J(u,y) \quad \text{subject to} \quad e(u,y)=0, \quad u\in U_{\text{ad}}, \quad y\in Y_{\text{ad}}, \tag{5.21}$$

where $J\colon U\times Y\to\mathbb{R}$, $e\colon U\times Y\to Z$ are continuous with Banach space Z and reflexive Banach spaces U and Y.

Similarly as above, existence can be shown under the following assumptions.

Assumption 5.4.

1. *$U_{ad}\subset U$ is convex, bounded, closed, and non-empty.*
2. *$Y_{ad}\subset Y$ is convex and closed, such that (5.21) has a feasible point.*
3. *The state equation $e(u,y)=0$ has a bounded* control-to-state operator $u\in U_{ad}\mapsto y[u]\in\ Y$.
4. *$U\times Y\ni(u,y)\mapsto e(u,y)\in Z$ is continuous w.r.t. weak convergence in $U\times Y$.*
5. *J is (sequentially) weakly lower semicontinuous.*

Remark 5.5. *We observe that in the linear-quadratic case under the Assumption 5.1 and if U_{ad} is bounded, these conditions are satisfied: 3. is satisfied since $A^{-1}B\in\mathcal{L}(U,Y)$, 4. by Theorem 4.18 and 5. by convexity and continuity of the cost functional given in (5.1).*

Theorem 5.6 (Existence for nonlinear optimal control problems)**.** *Let Assumption 5.4 hold. Then problem (5.21) has a solution $(\bar u,\bar y)\in U_{ad}\times Y_{ad}$.*

Proof. The proof is similar to the one of Theorem 5.3. We denote the feasible set of (5.21) by F_{ad}. Assumption 5.4 1. and 3. imply the existence of a bounded minimizing sequence $(u_k,y_k)\subset F_{\text{ad}}$. Since U,Y are reflexiv, we can extract a weakly converging subsequence $(u_{k_i},y_{k_i})\rightharpoonup(\bar u,\bar y)$. By Assumption 5.4 1., 2., and 4. the feasible set F_{ad} of (5.21) is weakly sequentially closed and thus $(\bar u,\bar y)\in F_{\text{ad}}$. Now Assumption 5.4 5. can be used to show hat $(\bar u,\bar y)$ solves (5.21). □

5.4. Application: Robin boundary control

Verification of a continuity assumption for a semilinear PDE

Lemma 5.7. *Let $\Omega \subset \mathbb{R}^n$, $n \in \{1,2,3\}$, open and bounded subset. Furthermore, let $y \in H^1(\Omega)$ and (y_k) a sequence in $H^1(\Omega)$. Then convergence $y_k \to y$ in $L^5(\Omega)$ implies $y_k^3 \to y^3$ in $L^{5/3}(\Omega)$.*

Proof. We first observe that $y_k^3 \in L^{5/3}(\Omega)$ and $y^3 \in L^{5/3}(\Omega)$ obviously holds, since $y_k, y \in H^1(\Omega) \hookrightarrow L^5(\Omega)$. Next we prove for $a, b \in \mathbb{R}$

$$|b^3 - a^3| \leq 3(|a|^2 + |b|^2)|b - a|. \tag{5.22}$$

For appropiate $t \in [0,1]$ the mean value theorem yields

$$|b^3 - a^3| = 3|(a + t(b-a))^2(b-a)| \leq 3\max(a^2, b^2)|b-a| \leq 3(a^2+b^2)|b-a|. \tag{5.23}$$

Therefore, we have

$$\begin{aligned}\left\|y_k^3 - y_k^3\right\|_{L^{5/3}(\Omega)} &\leq 3\left\|(y_k^2 + y^2)|y_k - y|\right\|_{L^{5/3}(\Omega)} \\ &\leq 3\left\|(y_k^2|y_k - y|)\right\|_{L^{5/3}(\Omega)} + 3\left\|y^2|y_k - y|\right\|_{L^{5/3}(\Omega)}.\end{aligned} \tag{5.24}$$

Now the Hölder inequality with $p = 3/2$ and $q = 3$ yields

$$\left\|v^2 w\right\|_{L^{5/3}(\Omega)} = \left\||v|^{\frac{10}{3}}|w|^{\frac{5}{3}}\right\|_{L^1(\Omega)}^{3/5} \leq \|v\|_{L^5(\Omega)}^2 \|w\|_{L^5(\Omega)}. \tag{5.25}$$

This shows

$$\begin{aligned}\left\|y_k^3 - y^3\right\|_{L^{5/3}(\Omega)} &\leq 3\left\|y_k^2|y_k - y|\right\|_{L^{5/3}(\Omega)} + 3\left\|y^2|y_k - y|\right\|_{L^{5/3}(\Omega)} \\ &\leq 3(\|y_k\|_{L^5(\Omega)}^2 + \|y\|_{L^5(\Omega)}^2)\,\|y_k - y\|_{L^5(\Omega)} \\ &\longrightarrow 6\,\|y\|_{L^5(\Omega)}^2 \cdot 0 = 0.\end{aligned} \tag{5.26}$$

□

To check Assumption 5.4 often compact embedding from Y to a suitable space $\hat{Y}$ is used to convert weak convergence in Y to strong convergence in $\hat{Y}$.

Example 5.8. *Let $\Omega \subset \mathbb{R}^n$, $n \leq 3$, be open and bounded with Lipschitz-boundary. We show Assumption 5.4 4. for*

$$U \times Y = L^2(\Omega) \times H^1(\Omega) \to Z = Y^*, \quad (u, y) \mapsto e(y,u) := -\Delta y + y^3 - u. \tag{5.27}$$

Here, we only analyze the nonlinear term. The embedding

$$Y = H^1(\Omega) \subset L^5(\Omega) \text{ is compact for } n = 2, 3\ , \tag{5.28}$$

see Theorem 4.55. Therefore, weak convergence $y_k \rightharpoonup y$ in Y implies strong convergence of $y_k \to y$ in $L^5(\Omega)$ and thus strong convergence

$$y_k^3 \to y^3 \text{ in } L^{5/3}(\Omega) = L^{5/2}(\Omega)^* \subset Z \quad \text{continuous} \tag{5.29}$$

by Lemma 5.7.

[2]

The last embedding above follows from the following consideration: Since

$$i\colon H^1(\Omega) \to L^{5/2}(\Omega) \text{ is a dense continuous embedding,} \tag{5.30}$$

there exists a continuous embedding[3]

$$i^*\colon L^{5/2}(\Omega)^* \to H^1(\Omega)^*.^{4} \tag{5.31}$$

We consider the following semilinear elliptic optimal control problem.

$$\begin{cases} \underset{y\in H^1(\Omega), u\in L^2(\partial\Omega)}{\text{Min}} J(y,u) := \frac{1}{2}\|y - y_{\text{d}}\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u\|^2_{L^2(\partial\Omega)}\,, \\ -\Delta y + y^3 = 0 \text{ in } \Omega, \quad \partial_n y + y = u \text{ on } \partial\Omega, \\ u_m \le u \le u_M \text{ a.e. on } \partial\Omega, \end{cases} \tag{5.32}$$

where $\Omega \subset \mathbb{R}^n$, $n = 2$ or $n = 3$, is open and bounded with Lipschitz boundary and $y_{\text{d}} \in L^2(\Omega)$, $u_m, u_M \in L^n(\partial\Omega)$, $u_m \le u_M$. Let

$$U := L^n(\partial\Omega), \quad Y := H^1(\Omega) \tag{5.33}$$

and

$$U_{\text{ad}} := \{u \in U \ :\ u_m \le u \le u_M \text{ a.e.}\}, \quad Y_{\text{ad}} := Y. \tag{5.34}$$

Note, that here we consider $L^n(\partial\Omega)$ as control space, which is related to the fact that we want to apply Theorem 4.66 for well-posedness of the state equation (with $f := 0$ and $g := u$). We verify Assumption 5.4:

[2] We identify elements $v \in L^{5/3}(\Omega)$ with the mapping $T_v\colon L^{5/2}(\Omega) \to \mathbb{R}$, $w \mapsto \int_\Omega v(x)w(x)\mathrm{d}x \in L^{5/2}(\Omega)^*$, cf. Example 4.8.

[3] Injectivity can be shown as follows: Let $h \in L^{5/2}(\Omega)^*$ such that $i^*h = 0$. Let $\hat{h} \in L^{5/2}(\Omega)$ and $(v^k) \subset H^1(\Omega)$ such that $iv_k \to \hat{h}$ in $L^{5/2}(\Omega)$. Then $0 = \langle i^*h, v_k\rangle_{H^1(\Omega)} = \langle h, iv_k\rangle_{,L^{5/2}(\Omega)} \to \langle h, \hat{h}\rangle_{L^{5/2}(\Omega)}$. Since it holds for all $\hat{h} \in L^{5/2}(\Omega)$ we have $h = 0$.

[4] Here, i^* may be interpreted as the restriction of T_v to elements in $H^1(\Omega)$.

1. $U_{\mathrm{ad}} \subset U$ is bounded, closed, and convex.
2. Follows directly.
3. If we consider weak solutions according to (4.127) then the PDE-constraint is an operator

   $$e\colon Y \times U \to Z := Y^*, \quad (y,u) \mapsto a(y,\cdot) + (y^3,\cdot)_{L^2(\Omega)} - (u,\cdot)_{L^2(\partial\Omega)}. \tag{5.35}$$

   where $a(y,v) := (\nabla y, \nabla v)_{L^2(\Omega)} + (y,v)_{L^2(\partial\Omega)}$. [5]. We know by Theorem 4.66 that there exists a unique bounded solution operator $U_{\mathrm{ad}} \to Y$, $u \mapsto y[u]$.
4. Moreover, $(y,u) \in Y \times U \mapsto e(y,u) \in Z$ is continuous under weak convergence, since the nonlinear term $y \in Y \mapsto y^3 \in Y$ is by Example 5.8 sequentially weakly continuous.
5. Finally, the objective function $J\colon Y \times U \to \mathbb{R}$ is continuous, convex and thus sequentially lower semicontinuous.

Thus, Assumption 5.4 is verified and therefore (5.32) has a solution by Theorem 5.6.

[5] Note, we have

$$e\colon Y \times U \to Z := Y^*, \quad (y,u) \mapsto \langle A + y^3 + Bu, \cdot\rangle_Y \tag{5.36}$$

with $A\colon Y \to Y^*$ given as in (4.105) and $B\colon L^2(\partial\Omega) \to H^1(\Omega)^*$ with $\langle Bu,\cdot\rangle_Y := \int_{\partial\Omega} u \cdot \mathrm{d}x$. Using $H^1(\Omega) \to L^6(\Omega)$ continuous for $n \leq 3$ and thus $y^3 \in L^2(\Omega)$ we have $\langle y^3,\cdot\rangle_Y = (y^3,\cdot)_{L^2(\Omega)}$, which gives (5.35).

6. Differentiability and adjoint representation

Before we can derive optimality conditions we need to introduce derivatives for mappings between Banach spaces.

Contents

6.1. Differentiability in Banach spaces

We extend the notion of differentiability to operators between Banach spaces.

Definition 6.1. *Let X and Y be Banach spaces and $F\colon U \to Y$ be an operator with $U \subset X$ non-empty open subset.*

(i) F is called directionally differentiable *at $x \in U$ if the limit*

$$dF(x,h) = \lim_{t\to 0^+} \frac{F(x+th)-F(x)}{t} \in Y \tag{6.1}$$

exists for all $h \in X$. In this case, $dF(x,h)$ is called directional derivative of F in the direction H.

(ii) F is called Gâteaux differentiable *at $x \in U$ if F is directional differentiable at x and*

$$F'(x)\colon X \to Y, \quad h \mapsto dF(x,h) \text{ is bounded and linear,} \tag{6.2}$$

i.e. $F'(x) \in L(X,Y)$.

(iii) *F is called* Fréchet differentiable *at $x \in U$ if F is Gâteaux differentiable at x and if the following approximation conditions holds*

$$\|F(x+h) - F(x) - F'(x)h\|_Y = o(\|h\|_X) \quad \textit{for} \ \|h\|_X \to 0. \tag{6.3}$$

(iv) *If F is directional-/G-/F-differentiable at every $x \in V$, $V \subset U$ open, then F is called directionally/G-/F-differentiable on V.*

Definition 6.2 (Higher derivatives)**.** *(i) If F is G-differentiable in a neighborhood V of x and $F'\colon V \to L(X,Y)$ is itself G-differentiable at x, then F is called twice G-differentiable at x. We write $F''(x) \in L(X, L(X,Y))$ for the second derivative of F at x.*

(ii) Accordingly we define F-differentiability of order k. We say that F is k-times continuously F-differentiable if F is k-times F differentiable and $F^{(k)}$ is continuous.

One shows easily that F-differentiability of F at x implies continuity of F at x (exercise).

Remark 6.3. *We collect some additional facts.*

1. *The chain rule holds for F-differentiable operators: Let $H(x) := G(F(x))$ and assume that F and G are F-differentiable at x and $F(x)$, respectively. Then H is F-differentiable at x with $H'(x) = G'(F(x))F'(x)$.*

   Furthermore, if F is G-differentiable at x and G is F-differentiable at $F(x)$, then H is G-differentiable and the chain rule holds. As a consequence, also the sum rule holds for F- and G-differentials.

2. *If F is G-differentiable on a neighborhood of x and F' is continuous at x then F is F-differentiable at x.*

3. *If $F\colon X \times Y \to Z$ is F-differentiable at (x,y) then $F(\cdot,y)$ and $F(x,\cdot)$ are F-differentiable at x and y, respectively. These derivatives are called partial derivatives and denoted by $F_x(x,y)$ and $F_y(x,y)$, respectively. There holds (since F is F-differentiable)*

$$F'(x,y)(h_x,h_y) = F_x(x,y)h_x + F_y(x,y)h_y. \tag{6.4}$$

Example 6.4. Let H be a Hilbert space with scalar product $(\cdot,\cdot)_H$ and norm $\|\cdot\|_H$. We set

$$f(u) := \|u\|_H^2. \tag{6.5}$$

Then

$$\begin{aligned}\lim_{t\to 0}\frac{1}{t}(f(u+th)-f(u)) &= \lim_{t\to 0}\frac{1}{t}(\|u+th\|_H^2-\|u\|_H^2)\\ &= \lim_{t\to 0}\frac{1}{t}(2t(u,h)_H+t^2\,\|h\|_H^2)\\ &= 2(u,h)_H.\end{aligned}\tag{6.6}$$

Hence, $u \mapsto f(u)$ is Gâteaux-differentiable with derivative $f'(u)h = 2(u,h)_H$.

By Riesz theorem, the dual space H^* can be identified with H, i.e. the Riesz representation of

$$f'(u) = (2u,\cdot)_H \tag{6.7}$$

is $2u \in H$. Often we write

$$\nabla f(u) = 2u \in H \tag{6.8}$$

and call $\nabla f(u)$ in this case the *gradient of f at u.*

6.2. Implicit function theorem

To derive properties of the control-to-state operator sometimes one can apply the implicit function theorem in Banach spaces.

Theorem 6.5 (Implicit function theorem)**.** *Let X, Y, Z be Banach spaces and let*

$$\left\{ \begin{array}{l} F\colon G \to Z \text{ be a continuously } F\text{-differentiable map} \\ \text{from an open set } G \subset X \times Y \text{ to } Z. \end{array} \right. . \tag{6.9}$$

Let $(\bar{x}, \bar{y}) \in G$ be such that $F(\bar{x}, \bar{y}) = 0$ and that

$$F_y(\bar{x}, \bar{y}) \in L(Y, Z) \text{ has a bounded inverse.} \tag{6.10}$$

Then there exists an open neighborhood $U_X(\bar{x}) \times U_Y(\bar{y}) \subset G$ of $(\bar{x}, \bar{y})$ and a unique continuous function $w\colon U_X(\bar{x}) \to Y$ such that

(i) $w(\bar{x}) = \bar{y}$*,*

(ii) for all $x \in U_X(\bar{x})$ there exists exactly one $y \in U_y(\bar{y})$ with $F(x, y) = 0$, namely $y = w(x)$.

Moreover, the mapping $w\colon U_X(\bar{x}) \to Y$ is continuously F-differentiable with derivative

$$w'(x) = -F_y(x, w(x))^{-1} F_x(x, w(x)). \tag{6.11}$$

Proof. For a proof see, e.g., [23, Thm. 4 B]. □

6.3. Sensitivities and adjoints

In the following we consider again problems of type

$$\min_{(u,y)\in U\times Y} J(u,y) \quad \text{subject to} \quad e(u,y)=0, \quad (u,y)\in W_{\text{ad}}, \tag{6.12}$$

where $J\colon U\times Y \to \mathbb{R}$ is the objective function, $e\colon U\times Y \to Z$ is an operator between Banach spaces, and $W_{\text{ad}} \subset W := U\times Y$ is a nonempty closed set.

We assume that J and e are continuously F-differentiable and that the state equation

$$e(u,y)=0 \tag{6.13}$$

possesses for each $u\in U$ a unique solution $y[u]\in Y$. Thus we have a control-to-state operator

$$U\to Y, \quad u\mapsto y[u]. \tag{6.14}$$

Furthermore, we assume that

$$e_y(u,y[u])\in L(Y,Z) \text{ is continuously invertible.} \tag{6.15}$$

Then the implicit function theorem 6.5 implies that $y[\cdot]$ is continuously F-differentiable.

Linearized state equation. An equation for the derivative $y'[u]$, the *linearized equation* is obtained by differentiating the equation $e(u,y[u])=0$ with respect to u:

$$e_y(u,y[u])y'[u]+e_u(u,y[u])=0. \tag{6.16}$$

Sensitivity approach. For $u\in U$ and direction $s\in U$, the chain rule yields for sensitivity of F:

$$dF(u,s)=\langle F'(u),s\rangle_U=\langle J_y(u,y[u]),y'[u]s\rangle_Y+\langle J_u(u,y[u]),s\rangle_U. \tag{6.17}$$

In this expression, the sensitivity $y'[u]s$ appears. Differentiating $e(u,y[u])=0$ in the direction s yields

$$e_y(u,y[u])y'[u]s+e_u(u,y[u])s=0. \tag{6.18}$$

Hence, the sensitivity $\delta_s y := y'[u]s$ is given as the solution of the linearized state equation

$$e_y(u,y[u])\delta_s y=-e_u(u,y[u])s. \tag{6.19}$$

Therefore, to compute the directional derivative $dF(u,s) = \langle F(u), s\rangle_U$ via the sensitivity approach the following steps are required:

1. Compute the sensitivity $\delta_s y = dy(u,s)$ by solving
$$e_y(u,y[u])\delta_s y = -e_u(u,y[u])s. \tag{6.20}$$
2. Compute $dF(u,s) = \langle F'(u), s\rangle_U$ via
$$dF(u,s) = \langle J_y(u,y[u]), \delta_s y\rangle_Y + \langle J_u(u,y[u]), s\rangle_U. \tag{6.21}$$
This procedure is expensive if the whole derivative $F'(u)$ is required, since this means that for a basis B of U, all the directional derivatives
$$dJ(u,v), \quad v \in B \tag{6.22}$$
have to be computed. Each of them requires the solution of one linearized state equation (6.20) with $s = v$. This is an effort that grows linearly in the dimension of U.

The adjoint approach. We derive a more efficient way of representing the derivative of F.

Lemma 6.6. *Let $p = p[u] \in Z^*$ be the adjoint state, i.e. solution of*
$$e_y(u,y[u])^* p = -J_y(u,y[u]). \tag{6.23}$$
Then we have the representation
$$F'(u) = e_u(u,y[u])^* p[u] + J_u(u,y[u]). \tag{6.24}$$

Proof. From
$$\begin{aligned}\langle F'(u), s\rangle_U &= \langle J_y(u,y[u]), y'[u]s\rangle_Y + \langle J_u(u,y[u]), s\rangle_U \\ &= \langle y'[u]^* J_y(u,y[u]), s\rangle_U + \langle J_u(u,y[u]), s\rangle_U,\end{aligned} \tag{6.25}$$
we have
$$F'(u) = y'[u]^* J_y(u,y[u]) + J_u(u,y[u]). \tag{6.26}$$
Therefore, not the operator $y'[u] \in \mathcal{L}(U,Y)$, but only the vector $y'[u]^* J_y(u,y[u]) \in U^*$ is really required. Since by (6.16)
$$y'[u]^* J_y(u,y[u]) = -e_u(u,y[u])^* e_y(u,y[u])^{-*} J_y(u,y[u]), \tag{6.27}$$

[1] it follows that

$$y'[u]^* J_y(u, y[u]) = e_u(u, y[u])^* p[u], \tag{6.28}$$

where the adjoint state $p = p[u] \in Z^*$ solves

$$e_y(u, y[u])^* p = -J_y(u, y[u]). \tag{6.29}$$

□

The derivative $F'(u)$ in any direction can thus be computed via the adjoint approach as follows:

1. Compute the adjoint state by solving the adjont equation

$$e_y(u, y[u])^* p = -J_y(u, y[u]). \tag{6.30}$$

2. Compute $F'(u)$ via

$$F'(u) = e_u(u, y[u])^* p + J_u(u, y[u]). \tag{6.31}$$

[1] Let for Banach spaces X and Y the operator $A \in L(X, Y)$ be invertible. Then we have for all $y^* \in Y^*$ and $y \in Y$ that $\langle (A^{-1})^* A^* y^*, y \rangle_Y = \langle A^* y^*, A^{-1} y \rangle_{X^*, X} = \langle y^*, y \rangle_Y$. Hence, the *-operation and taking the inverse commutes.

6.4. Application to a linear-quadratic optimal control problem

We consider the following optimal control problem

$$\begin{cases} \min\limits_{(u,y)\in U\times Y} J(u,y) := \tfrac{1}{2}\left\|Qy - y_{\mathrm{d}}\right\|_H^2 + \dfrac{\alpha}{2}\left\|u\right\|_U^2, \\ \text{subject to} \quad Ay + Bu = g, \quad u \in U_{\mathrm{ad}}, \quad y \in Y_{\mathrm{ad}}, \end{cases} \tag{6.32}$$

with H, U are Hilbert spaces, Y, Z are Banach spaces and $y_{\mathrm{d}} \in H$, $g \in Z$, $A \in L(Y,Z)$, $B \in L(U,Z)$, $Q \in L(Y,H)$ and let Assumptions 5.1 hold. We obtain the form (6.12) by setting

$$e(u,y) := Ay + Bu - g, \quad W_{\mathrm{ad}} = Y_{\mathrm{ad}} \times U_{\mathrm{ad}}. \tag{6.33}$$

By assumption there exists a continuous affine linear control-to-state operator

$$U \to Y, \quad u \mapsto y[u] = A^{-1}(g - Bu). \tag{6.34}$$

For the derivatives we have for $s_y \in Y$ and $s_u \in U$[2]

$$\begin{aligned} \langle J_y(u,y), s_y\rangle_Y &= (Qy - y_{\mathrm{d}}, Qs_y)_H, & e_y(u,y)s_y &= As_y, \\ \langle J_u(u,y), s_u\rangle_U &= \alpha(u, s_u)_U, & e_u(u,y)s_u &= Bs_u. \end{aligned} \tag{6.35}$$

Therefore,

$$\begin{aligned} J_y(u,y) &= (Qy - y_{\mathrm{d}}, Q\cdot)_H, & e_y(u,y) &= A, \\ J_u(u,y) &= \alpha(u,\cdot)_U, & e_u(u,y) &= B. \end{aligned} \tag{6.36}$$

If we choose the Riesz representation $U^* = U$, $H^* = H$, then

$$J_y(u,y) = (Qy - y_{\mathrm{d}}, Q\cdot)_H = \langle Qy - y_{\mathrm{d}}, Q\cdot\rangle_H = \langle Q^*(Qy - y_{\mathrm{d}}), \cdot\rangle_Y = Q^*(Qy - y_{\mathrm{d}}), \tag{6.37}$$

$$J_u(u,y) = \alpha(u,\cdot)_U = \alpha u. \tag{6.38}$$

The reduced objective function is

$$F(u) = J(u, y[u]) = \tfrac{1}{2}\left\|Q(A^{-1}(g - Bu)) - y_{\mathrm{d}}\right\|_H^2 + \frac{\alpha}{2}\left\|u\right\|_U^2. \tag{6.39}$$

[2] Note, that here we do not identify Y with Y^*. We use the fact that $\langle u,v\rangle_Y = (u,v)_H$ for $u \in L^2(\Omega)$ and $v \in H^1(\Omega)$.

For evaluation of F, we first solve the state equation

$$Ay + Bu = g \tag{6.40}$$

to obtain $y = y[u]$ and then we evaluate $J(u, y)$. In the following let $y = y[u]$.

Sensitivity approach. For $s \in U$ we obtain $dF(u, s) = \langle F'(u), s\rangle_U$ by first solving the linearized state equation

$$A\delta_s y = -Bs \tag{6.41}$$

for $\delta_s y$ and then setting

$$dF(u, s) = (Qy - y_{\mathrm{d}}, Q\delta_s y)_H + \alpha(u, s)_U. \tag{6.42}$$

Adjoint approach. We obtain $F'(u)$ by first solving the adjoint equation

$$F'(u) = B^*p + \alpha(u, \cdot) \quad (= B^*p + \alpha u \text{ if } U^* = U). \tag{6.43}$$

Application to distributed control. We consider the following elliptic control problem

$$\begin{cases} \text{Min} \quad J(u, y) = \dfrac{1}{2} \|y - y_{\mathrm{d}}\|^2_{L^2(\Omega)} + \dfrac{\alpha}{2} \|u\|^2_{L^2(\Omega)}, \\ -\Delta y = \gamma u \text{ in } \Omega, \\ \partial_n y = \dfrac{\beta}{\kappa}(y_a - y) \text{ on } \partial\Omega, \\ u_m \le u(x) \le u_M \text{ a.e. in } \Omega. \end{cases} \tag{6.44}$$

The appropiate spaces are $U = L^2(\Omega)$ and $Y = H^1(\Omega)$, and we assume

$$\begin{aligned} u_m, u_M \in U, &\qquad y_{\mathrm{d}} \in L^2(\Omega), \quad \alpha > 0, \\ y_a \in L^2(\partial\Omega), &\quad \gamma \in L^\infty(\Omega) \setminus \{0\}, \quad \gamma \ge 0. \end{aligned} \tag{6.45}$$

The coefficient γ weigths the control and y_a can be interpreted as the surrounding temperature in the case of the heat equation, $\beta > 0$ and $\kappa > 0$ are given coefficients. The weak formulation of the state equation is

$$y \in Y, \quad a(y, v) = (\gamma u, v)_{L^2(\Omega)} + ((\beta/\kappa)y_a, v)_{L^2(\partial\Omega)} \quad \text{for all } v \in Y \tag{6.46}$$

with $a(y, v) := (\nabla y, \nabla v)_{L^2(\Omega)} + ((\beta/\kappa)y, v)_{L^2(\partial\Omega)}$.

Now, let $Z = Y^*$, $H = L^2(\Omega)$ and

(i) $A \in L(Y, Y^*)$ the operator induced by a,

(ii) $B \in L(U, Y^*)$, $Bu = -(\gamma u, \cdot)_{L^2(\Omega)}$,

(iii) $g \in Y^*$, $g = ((\beta/\kappa)y_a, \cdot)_{L^2(\partial\Omega)}$,

(iv) $U_{\text{ad}} = \{u \in U \ : \ u_m \leq u \leq u_M \text{ a.e. in } \Omega\}$, $Y_{\text{ad}} = Y$,

(v) $Q \in L(Y, H)$, $Qy = y$.

Then, we arrive at a linear quadratic problem of the form (6.32).

Computation of the adjoints. Note that all spaces are Hilbert spaces and thus reflexive. In particular, we identify the dual of $U = L^2(\Omega)$ with U working with $\langle \cdot, \cdot \rangle_U = (\cdot, \cdot)_U$. We do the same with $H = L^2(\Omega)$. For A^* we obtain using the symmetry of a and reflexivity that

$$\begin{aligned} \langle A^*v, w\rangle_Y &= \langle v, Aw\rangle_Z = \langle Aw, v\rangle_Y \\ &= a(w,v) = a(v,w) = \langle Av, w\rangle_Y \quad \text{for all } v, w \in Y. \end{aligned} \tag{6.47}$$

Hence, we have $A^* = A$. For B^* we have

$$\begin{aligned} (B^*v, w)_U = \langle B^*v, w\rangle_U = \langle v, Bw\rangle_Z = \langle Bw, v\rangle_Y = (-\gamma w, v)_{L^2(\Omega)} \\ = -(\gamma v, w)_U \quad \text{for all } v \in Y, w \in U. \end{aligned} \tag{6.48}$$

Hence $B^*v = -\gamma v$. Finally, for Q^* we obtain

$$\langle Q^*v, w\rangle_Y = \langle v, Qw\rangle_H = (v, w)_{L^2(\Omega)}. \tag{6.49}$$

Therefore, $Q^*v = (v, \cdot)_{L^2(\Omega)}$. This means that

$$J_y(u, y) = (Q^*(Qy - y_{\text{d}}), \cdot)_{L^2(\Omega)} = (y - y_{\text{d}}, \cdot)_{L^2(\Omega)}. \tag{6.50}$$

Taking all together, the adjoint equation thus reads

$$Ap = -(y - y_{\text{d}}, \cdot)_{L^2(\Omega)}, \tag{6.51}$$

which is the weak form of

$$\begin{aligned} -\Delta p &= -(y - y_{\text{d}}) && \text{in } \Omega, \\ \partial_n p + \frac{\beta}{\kappa} p &= 0, && \text{on } \partial\Omega. \end{aligned} \tag{6.52}$$

The adjoint gradient representation then is

$$F'(u) = B^*p[u] + J_u(u, y[u]) = -\gamma p + \alpha u. \tag{6.53}$$

6.5. A Lagrangian-based view of the adjoint approach

We define the Lagrangian $L\colon U \times Y \times Z^* \to \mathbb{R}$ by

$$L(u, y, p) := J(u, y) + \langle p, e(u, y)\rangle_Z. \tag{6.54}$$

Inserting $y = y[u]$ gives for arbitrary $p \in Z^*$

$$F(u) = J(u, y[u]) = J(u, y[u]) + \langle p, e(u, y[u])\rangle_Z = L(u, y[u], p). \tag{6.55}$$

Differentiating this, we obtain

$$\langle F'(u), s\rangle_U = \langle L_y(u, y[u], p), y'[u]s\rangle_Y + \langle L_u(u, y[u], p), s\rangle_U. \tag{6.56}$$

We choose a special $p = p[u]$ such that

$$L_y(u, y[u], p) = 0, \tag{6.57}$$

which is the adjoint equation; in fact

$$\begin{aligned} \langle L_y(u, y, p), d\rangle_Y &= \langle J_y(u, y), d\rangle_Y + \langle p, e_y(u, y)d\rangle_Z \\ &= \langle J_y(u, y) + e_y(u, y)^* p, d\rangle_Y. \end{aligned} \tag{6.58}$$

Therefore,

$$L_y(u, y[u], p) = J_y(u, y[u]) + e_y(u, y[u])^* p. \tag{6.59}$$

Now choosing $p = p[u]$ according to (6.57), we obtain from (6.56) that

$$F'(u) = L_u(u, y[u], p[u]) = J_u(u, y[u]) + e_u(u, y[u])^* p[u] \tag{6.60}$$

is the adjoint gradient representation.

7. Optimality conditions

Contents

7.1. General setting: Necessary optimality conditions

We consider the problem

$$\min_{w \in W} J(w), \quad \text{s.t.} \quad w \in C, \tag{7.1}$$

where W is a Banach space, $J \colon W \to \mathbb{R}$, and $C \subset W$ is nonempty, closed, and convex.

Theorem 7.1. *Let W be a Banach space and $C \subset W$ be nonempty and convex. Furthermore, let $J \colon V \to \mathbb{R}$ for open set $V \supset C$. Let $\bar{w}$ be a local solution of (7.1) at which J is Gâteaux-differentiable. Then the following optimality condition holds:*

$$\bar{w} \in C, \quad \langle J'(\bar{w}), w - \bar{w} \rangle_{W^*,W} \geq 0 \quad \text{for all } w \in C. \tag{7.2}$$

Furthermore, there holds: (i) If J is convex on C, the condition (7.2) is necessary and sufficient for global optimality.

(ii) If, in addition, J is strictly convex on C, then there exist at most one solution of (7.1), or, equivalently, of (7.2).

(iii) If W is reflexive, C is closed, and J is convex and continuous with

$$\lim_{w \in C, \|w\|_W \to \infty} J(w) = \infty, \tag{7.3}$$

then there exists a (global = local) solution of (7.1).

Proof. Let $w \in C$ be arbitrary. By convexity of C we have

$$w(t) = \bar{w} + t(w - \bar{w}) \in C \text{ for all } t \in [0, 1]. \tag{7.4}$$

Now by optimality of $\bar{w}$ we have

$$J(\bar{w} + t(w - \bar{w})) - J(\bar{w}) \geq 0 \quad \text{for all } t \in [0, 1] \tag{7.5}$$

and thus

$$\langle J'(\bar{w}), w - \bar{w}\rangle_{W^*,W} = \lim_{t \to 0^+} \frac{J(\bar{w} + t(w - \bar{w})) - J(\bar{w})}{t} \geq 0. \tag{7.6}$$

Since $w \in C$ was arbitrary, the proof of (7.2) is complete.

(i) Let J be convex. Hence, for all $t \in (0, 1]$ we have

$$J(\bar{w} + t(w - \bar{w})) \leq (1 - t)J(\bar{w}) + tJ(w), \tag{7.7}$$

and thus

$$J(w) - J(\bar{w}) \geq \frac{J(\bar{w} + t(w - \bar{w})) - J(\bar{w})}{t} \overset{t \to 0^+}{\to} \langle J'(\bar{w}), w - \bar{w}\rangle_{W^*,W}. \tag{7.8}$$

Then

$$J(w) - J(\bar{w}) \geq \langle J'(\bar{w}), w - \bar{w}\rangle_{W^*,W} \quad \text{for all } w \in C. \tag{7.9}$$

Now from (7.2) and (7.9) it follows that

$$J(w) - J(\bar{w}) \geq \langle J'(\bar{w}), w - \bar{w}\rangle_{W^*,W} \geq 0 \quad \text{for all } w \in C. \tag{7.10}$$

Thus $\bar{w}$ is optimal.

(ii) If J is strictly convex and $\bar{w}_1, \bar{w}_2$ are two global solutions, the point

$$\tfrac{1}{2}(\bar{w}_1 + \bar{w}_2) \in C \tag{7.11}$$

would be a better solution, unless $\bar{w}_1 = \bar{w}_2$.

(iii) Finally, let the assumption of the last assertion hold and let

$$(w_k) \subset C \text{ be a minimizing sequence.} \tag{7.12}$$

Then (w_k) is bounded (otherwise $J(w_k) \to \infty$) and thus (w_k) contains a weakly convergent subsequence $(w_k) \rightharpoonup \bar{w}$. Since C is convex and closed, it is weakly sequentially closed and thus $\bar{w} \in C$. From the continuity and convexity of J we conclude that J is weakly sequentially lower semicontinuous and thus

$$J(\bar{w}) \leq \lim_{k \to \infty} J(w_k) = \inf_{w \in C} J(w). \tag{7.13}$$

Thus, $\bar{w}$ solves the minimization problem. □

Lemma 7.2. *Let $C \subset W$ be a nonempty closed convex subset of the Hilbert space W and denote by $P\colon W \to C$ the projection onto C, i.e.*

$$P(w) \in C, \quad \|P(w) - w\|_W = \min_{v \in C} \|v - w\|_W \qquad \textit{for all } w \in W. \tag{7.14}$$

Then:

- *(i) P is well-defined.*
- *(ii) For all $w, z \in W$ there holds*

$$z = P(w) \Leftrightarrow z \in C, \quad (w - z, v - z)_W \leq 0 \quad \forall v \in C. \tag{7.15}$$

- *(iii) P is nonexpansive, i.e.*

$$\|P(v) - P(w)\|_W \leq \|v - w\|_W \quad \forall v, w \in W. \tag{7.16}$$

- *(iv) P is monotone, i.e.*

$$(P(v) - P(w), v - w)_W \geq 0 \quad \forall v, w \in W. \tag{7.17}$$

  Furthermore, equality holds if and only if $P(v) = P(w)$.

- *(v) For all $w \in C$ and $d \in W$, the function*

$$\varphi(t) := \frac{1}{t} \|P(w + td) - w\|_W, \quad t \geq 0, \tag{7.18}$$

  is nonincreasing.

Proof. See [15, p. 67]. □

Lemma 7.3. *Let W be a Hilbert space $C \subset W$ be nonempty, closed, and convex. Furthermore, let P denote the projection onto C. Then, for all $y \in W$ and for all $\gamma > 0$, the following conditions are equivalent:*

$$w \in C, \quad (y, v - w)_W \geq 0 \quad \textit{for all } v \in C, \tag{7.19}$$

$$w - P(w - \gamma y) = 0. \tag{7.20}$$

Proof. Let (7.19) hold. Then setting $w_\gamma = w - \gamma y$ we have

$$(w_\gamma - w, v - w)_W = -\gamma (y, v - w)_W \leq 0 \quad \forall v \in C. \tag{7.21}$$

By Lemma 7.2(ii) this implies $w = P(w_\gamma)$ as asserted in (7.20).

Conversely, let (7.20) hold. Then using the same notation we have $w = P(w_\gamma) \in C$. Furthermore, Lemma 7.2(ii) yields

$$(y, v - w)_W = -\frac{1}{\gamma}(w_\gamma - w, v - w)_W \geq 0, \quad v \in C. \tag{7.22}$$

□

Corollary 7.4. *Let W be a Hilbert space and $C \subset W$ be nonempty, closed, and convex. Furthermore, let $J\colon V \to \mathbb{R}$ be defined on an open neighborhood of C. Let $\bar{w}$ be a local solution of (7.1) at which J is Gâteaux-differentiable. Then the following optimality condition holds for arbitrary but fixed $\gamma > 0$:*

$$\bar{w} = P(\bar{w} - \gamma \nabla J(\bar{w})). \tag{7.23}$$

Here, $\nabla J(w) \in W$ denotes the Riesz-representation of $J'(w) \in W^*$.

Moreover, in the Hilbert space setting (7.23) is equivalent to (7.2) if C is nonempty, closed, convex and therefore in this case Theorem 7.1 holds with (7.23) instead of (7.2).

Proof. This follows from Theorem 7.1 and Lemma 7.3. □

7.2. Control-constrained problems: Necessary optimality conditions

We consider a general possibly nonlinear problem of the form

$$\min_{(u,y)\in U\times Y} J(u,y) \quad \text{subject to } e(u,y)=0, \quad u\in U_{\mathrm{ad}}. \tag{7.24}$$

Assumption 7.5. *(i) $U_{ad}\subset U$ is nonempty, convex, and closed.*

(ii) $J\colon U\times Y\to\mathbb{R}$ and $e\colon U\times Y\to Z$ are continuously Fréchet differentiable and U,Y,Z are Banach spaces.

(iii) For all $u\in V$ in a neighborhood $V\subset U$ of U_{ad}, the state equation $e(u,y)=0$ has a unique solution $y=y[u]\in Y$.

(iv) $e_y(u,y[u])\in L(Y,Z)$ has a bounded inverse for all $u\in V\supset U_{ad}$.

Under the assumptions the mapping $u\in V\mapsto y[u]\in Y$ is continuously F-differentiable by the implicit function theorem.

Obviously, the general linear-quadratic optimization problem

$$\begin{cases} \displaystyle\min_{(u,y)\in U\times Y} J(u,y) := \tfrac{1}{2}\left\|Qy-y_{\mathrm{d}}\right\|_H^2 + \frac{\alpha}{2}\left\|u\right\|_U^2, \\ \text{subject to } Ay+Bu=g, \quad u\in U_{\mathrm{ad}} \end{cases} \tag{7.25}$$

is a special case of (7.24), where H,U are Hilbert spaces, Y,Z are Banach spaces and $y_{\mathrm{d}}\in H$, $g\in Z$, $A\in L(Y,Z)$, $B\in L(U,Z)$, $Q\in L(Y,H)$. Moreover, Assumption 5.1 ensures Assumption 7.5, since $e_y(u,y)=A$.

We formulate the reduced problem of (7.25)

$$\min_{u\in U} F(u) \quad \text{s.t.} \quad u\in U_{\mathrm{ad}}, \quad F(u):=J(u,y[u]). \tag{7.26}$$

Theorem 7.6. *Let Assumption 7.5 hold. If $\bar{u}$ is a local solution of the reduced problem 7.26 then $\bar{u}$ satisfies the variational inequality*

$$\bar{u}\in U_{ad} \quad \text{and} \quad \langle F'(\bar{u}), u-\bar{u}\rangle_U \geq 0 \quad \text{for all } u\in U_{ad}. \tag{7.27}$$

Proof. We can directly apply Theorem 7.1. □

Depending on the structure of U_{ad} the variational inequality (7.27) can be expressed in a more convenient form. We show this for the case of box constraints.

Theorem 7.7. *Let $U = L^2(\Omega)$, $u_m, u_M \in L^2(\Omega)$, $u_m \leq u_M$, and U_{ad} be given by*

$$U_{ad} = \{u \in U \;:\; u_m \leq u \leq u_M \text{ a.e.}\}. \tag{7.28}$$

We work with $U^* = U$ write $\nabla F(u)$ for the derivative to emphasize that this is the Riesz representation. Then the following conditions are equivalent:

(i) $\bar{u} \in U_{ad}$,

$$(\nabla F(\bar{u}), u - \bar{u})_U \geq 0 \quad \text{for all } u \in U_{ad}. \tag{7.29}$$

(ii) $\bar{u} \in U_{ad}$,

$$\nabla F(\bar{u})(x) \begin{cases} = 0, & \text{if } u_m(x) < \bar{u}(x) < u_M(x), \\ \geq 0, & \text{if } u_m(x) = \bar{u}(x) < u_M(x), \\ \leq 0, & \text{if } u_m(x) < \bar{u}(x) = u_M(x). \end{cases} \tag{7.30}$$

(iii) There are $\bar{\lambda}_{u_m}, \bar{\lambda}_{u_M} \in U^* = L^2(\Omega)$ with

$$\begin{aligned} &\nabla F(\bar{u}) + \bar{\lambda}_{u_M} - \bar{\lambda}_{u_m} = 0, \\ &\bar{u} \geq u_m, \quad \lambda_{u_m} \geq 0, \quad \lambda_{u_m}(\bar{u} - u_m) = 0, \\ &\bar{u} \leq u_M, \quad \lambda_{u_M} \geq 0, \quad \lambda_{u_M}(u_M - \bar{u}) = 0. \end{aligned} \tag{7.31}$$

(iv) For any $\gamma > 0$: $\bar{u} = P_{U_{ad}}(\bar{u} - \gamma \nabla F(u))$, with

$$P_{U_{ad}}(u) = \min(\max(u_m, u), u_M). \tag{7.32}$$

Proof. (ii) ⇒ (i): If $\nabla F(\bar{u})$ satisfies (ii) then it is obvious that $\nabla F(\bar{u})^\top (u - \bar{u}) \geq 0$ a.e. for all $u \in U_{\mathrm{ad}}$ and thus

$$(\nabla F(\bar{u}), u - \bar{u})_U = \int_\Omega \nabla F(\bar{u})(u - \bar{u}) \mathrm{d}x \geq 0 \quad \text{for all } u \in U_{\mathrm{ad}}. \tag{7.33}$$

(i) ⇒ (ii): Clearly, (ii) is the same as

$$\nabla F(\bar{u})(x) \begin{cases} \geq 0 & \text{a.e. on } I_{u_m} := \{u \;:\; u_m(x) \leq u(x) < u_M(x)\}, \\ \leq 0 & \text{a.e. on } I_{u_M} := \{u \;:\; u_m(x) < u(x) \leq u_M(x)\}. \end{cases} \tag{7.34}$$

Assume this is not true. Then, without loss of generality, there exists a set $M \subset I_{u_m}$ of positive measure with $\nabla F(\bar{u})(x) < 0$ on M. Now, choose $u = \bar{u} + 1_M(u_M - \bar{u})$. Then $u \in U_{\mathrm{ad}}$, $u - \bar{u} > 0$ on M and $u - \bar{u} = 0$ elsewhere. Hence, we get the contradiction

$$(\nabla F(\bar{u}), u - \bar{u})_U = \int_M \underbrace{\nabla F(\bar{u})}_{<0} \underbrace{(u_M - \bar{u})}_{>0} \mathrm{d}x < 0. \tag{7.35}$$

(ii) ⇒ (iii): Let

$$\bar\lambda_{u_m} = \max(\nabla F(\bar u), 0), \quad \bar\lambda_{u_M} = \max(-\nabla F(\bar u), 0). \tag{7.36}$$

Then $u_m \le \bar u \le u_M$ and $\bar\lambda_{u_M}, \bar\lambda_{u_m} \ge 0$ hold trivially. Furthermore,

$$\bar u(x) > u_m(x) \Rightarrow \nabla F(\bar u(x)) \le 0 \Rightarrow \bar\lambda_{u_m}(x) = 0, \tag{7.37}$$
$$\bar u(x) < u_M(x) \Rightarrow \nabla F(\bar u(x)) \ge 0 \Rightarrow \bar\lambda_{u_M}(x) = 0. \tag{7.38}$$

(iii) ⇒ (ii):

$$\begin{aligned}
u_m(x) < \bar u(x) < u_M(x) &\Rightarrow \bar\lambda_{u_m}(x) = \lambda_{u_M}(x) = 0 \Rightarrow \nabla F(\bar u)(x) = 0,\\
u_m(x) = \bar u(x) < u_M(x) &\Rightarrow \bar\lambda_{u_M}(x) = 0 \Rightarrow \nabla F(\bar u(x)) = \bar\lambda_{u_m}(x) \ge 0,\\
u_m(x) < \bar u(x) = u_M(x) &\Rightarrow \bar\lambda_{u_m}(x) = 0 \Rightarrow \nabla F(\bar u(x)) = -\bar\lambda_{u_M}(x) \le 0.
\end{aligned} \tag{7.39}$$

(ii) ⇔ (iv): This is easily verified. Alternatively, one can use Lemma 7.3 to prove equivalence of (i) and (iv). □

Next, we use the adjoint representation of the derivative

$$F'(u) = e_u(u, y[u])^* p[u] + J_u(u, y[u]), \tag{7.40}$$

where the adjoint state $p[u] \in Z^*$ solves the adjoint equation

$$e_y(u, y[u])^* p = -J_y(u, y[u]). \tag{7.41}$$

We recall the definition of the Lagrange function associated with (7.24)

$$L\colon U \times Y \times Z^* \to \mathbb{R}, \quad L(u, y, p) = J(u, y) + \langle p, e(u, y)\rangle_Z. \tag{7.42}$$

The representation (7.40) of $F'(\bar u)$ yields the following corollary of Theorem 7.6.

Corollary 7.8. *Let $(\bar u, \bar y)$ a minimum of the problem (7.24) and let the Assumption 7.5 hold. Then there exists an adjoint state (or Lagrange multiplier) $\bar p \in Z^*$ such that the following optimality conditions hold*

$$e(\bar u, \bar y) = 0, \tag{7.43}$$
$$e_y(\bar u, \bar y)^* \bar p = -J_y(\bar u, \bar y), \tag{7.44}$$
$$\bar u \in U_{ad}, \quad \langle J_u(\bar u, \bar y) + e_u(\bar u, \bar y)^* \bar p, u - \bar u\rangle_U \ge 0 \quad \textit{for all } u \in U_{ad}. \tag{7.45}$$

Using the Lagrange function we can write (7.43)–(7.45) in the compact form

$$L_p(\bar u, \bar y, \bar p) = e(\bar u, \bar y) = 0, \tag{7.46}$$
$$L_y(\bar u, \bar y, \bar p) = 0, \tag{7.47}$$
$$\bar u \in U_{ad}, \quad \langle L_u(\bar u, \bar y, \bar p), u - \bar u\rangle_U \ge 0 \quad \textit{for all } u \in U_{ad}. \tag{7.48}$$

Proof. We have only to combine (7.27), (7.40),and (7.41). □

To avoid dual operators, one can also use the equivalent variational form

$$\begin{aligned} &\langle e(\bar{u},\bar{y}), p\rangle_{Z^*} = 0 && \text{for all } p \in Z^*,\\ &\langle L_y(\bar{u},\bar{y},\bar{p}), v\rangle_Y = 0 && \text{for all } v \in Y, \\ &\bar{u} \in U_{\mathrm{ad}}, \quad \langle L_u(\bar{u},\bar{y},\bar{p}), u-\bar{u}\rangle_U \geq 0 && \text{for all } u \in U_{\mathrm{ad}}. \end{aligned} \tag{7.49}$$

7.3. Application: The linear-quadratic case

We consider the following optimal control problem:

$$\begin{cases} \min \quad J(u,y) := \dfrac{1}{2}\,\|Qy - y_{\mathrm{d}}\|_H^2 + \dfrac{\alpha}{2}\,\|u\|_U^2 \\ \qquad Ay + Bu = g, \quad u \in U_{\mathrm{ad}} \end{cases} \tag{7.50}$$

under Assumption 7.5. Then

$$e(u,y) = Ay + Bu - g, \quad e_y(u,y) = A, \quad e_u(u,y) = B \tag{7.51}$$

and Corollary 7.8 is applicable. We only have to compute L_y and L_u for the Lagrange function

$$\begin{aligned} L(u,y,p) &= J(u,y) + \langle p, Ay + Bu - g\rangle_Z \\ &= \tfrac{1}{2}(Qy - y_{\mathrm{d}}, Qy - y_{\mathrm{d}})_H + \frac{\alpha}{2}(u,u)_U + \langle p, Ay + Bu - g\rangle_Z. \end{aligned} \tag{7.52}$$

We have with the identification $H^* = H$ and $U^* = U$

$$\begin{aligned} \langle L_y(\bar u, \bar y, \bar p), v\rangle_Y &= (Q\bar y - y_{\mathrm{d}}, Qv)_H + \langle \bar p, Av\rangle_Z \\ &= \langle Q^*(Q\bar y - y_{\mathrm{d}}) + A^*\bar p, v\rangle_Y \quad \text{for all } v \in Y \end{aligned} \tag{7.53}$$

and

$$\begin{aligned} (L_u(\bar u, \bar y, \bar p), w)_U &= \alpha(\bar u, w)_U + \langle \bar p, Bw\rangle_Z \\ &= (\alpha \bar u + B^*\bar p, w)_U \quad \text{for all } w \in U. \end{aligned} \tag{7.54}$$

Thus (7.43)-(7.45) take the form

$$\begin{aligned} &A\bar y + B\bar u = g, \\ &A^*\bar p = -Q^*(Q\bar y - y_{\mathrm{d}}), \\ &\bar u \in U_{\mathrm{ad}}, \quad (\alpha \bar u + B^*\bar p, u - \bar u)_U \geq 0 \quad \text{for all } u \in U_{\mathrm{ad}}. \end{aligned} \tag{7.55}$$

Distributed control of linear elliptic equations

$$\begin{cases} \min \quad J(u,y) := \dfrac{1}{2}\,\|y - y_{\mathrm{d}}\|_{L^2(\Omega)}^2 + \dfrac{\alpha}{2}\,\|u\|_{L^2(\Omega)}^2, \quad \alpha > 0 \\ \qquad -\Delta y = \gamma u \quad \text{in } \Omega, \\ \qquad\quad y = 0 \quad \text{on } \partial\Omega, \\ \qquad\quad u_m \leq u \leq u_M \quad \text{in } \Omega \end{cases} \tag{7.56}$$

where

$$\gamma \in L^\infty(\Omega) \setminus \{0\}, \quad \gamma \geq 0, \quad u_m, u_M \in L^2(\Omega), \quad u_m \leq u_M. \tag{7.57}$$

We have already observed in Section 5.2 and satisfies the Assumption 7.5 with

$$U = H = L^2(\Omega), \quad Y = H_0^1(\Omega), \quad Z = Y^*, \quad g = 0, \quad Q = I_{Y,H}, \tag{7.58}$$

$U_{\mathrm{ad}} = \{u \in U \ : \ u_m \leq u \leq u_M \text{ a.e.}\}$ and

$$\begin{aligned} A \in L(Y, Y^*), \quad \langle Ay, v\rangle_Y = a(y, v) = \int_\Omega \nabla y \cdot \nabla v \mathrm{d}x, \\ B \in L(U, Y^*), \quad \langle Bu, v\rangle_Y = -(\gamma u, v)_{L^2(\Omega)}. \end{aligned} \tag{7.59}$$

Hence, the optimality system is given by (7.55). Moreover, we have $A^* = A$. In fact, as a Hilbert space Y is refexive and $Z^* = Y^{**}$ can be identified with Y through

$$\langle p, y^*\rangle_{Y^*} = \langle y^*, p\rangle_Y \quad \text{for all } y^* \in Y^* \text{ and } p \in Y = Y^{**}. \tag{7.60}$$

This yields

$$\begin{aligned} \langle A^* v, w\rangle_Y = \langle v, Aw\rangle_Z = \langle Aw, v\rangle_Y \\ = a(w, v) = a(v, w) = \langle Av, w\rangle_Y \quad \text{for all } v, w \in Y \end{aligned} \tag{7.61}$$

and thus $A^* = A$.

Theorem 7.9. *If $(\bar{u}, \bar{y})$ is a local solution of (11.1) then there exists $\bar{p} \in H_0^1(\Omega)$, $\bar{\lambda}_{u_m}, \bar{\lambda}_{u_M} \in L^2(\Omega)$ such that the following optimality conditions hold in the weak sense,*

$$\begin{aligned} &-\Delta\bar{y} = \gamma\bar{u}, \quad \bar{y}|_{\partial\Omega} = 0, \\ &-\Delta\bar{p} = -(\bar{y} - y_d), \quad \bar{p}|_{\partial\Omega} = 0, \\ &\alpha\bar{u} - \gamma\bar{p} + \bar{\lambda}_{u_M} - \bar{\lambda}_{u_m} = 0, \\ &\bar{u} \geq u_m, \quad \bar{\lambda}_{u_m} \geq 0, \quad \bar{\lambda}_{u_m}(\bar{u} - u_m) \quad = 0, \\ &\bar{u} \leq u_M, \quad \bar{\lambda}_{u_M} \geq 0, \quad \bar{\lambda}_{u_M}(u_M - \bar{u}) \quad = 0. \end{aligned} \tag{7.62}$$

Proof. The assertion follows from Theorem 7.7. □

7.4. Equivalent pointwise formulations of the optimality conditions

The optimality condition for (11.1) is the given as the variational inequality

$$\int_\Omega (\gamma(x)p(x) + \alpha\bar{u}(x))(u(x) - \bar{u}(x))\mathrm{d}x \geq 0 \quad \text{for all } u \in U_{\text{ad}} \tag{7.63}$$

where p is the corresponding adjont state.

Lemma 7.10. *A necessary and sufficient optimality for the variational inequality (7.63) to be satisfied is that for almost every $x \in \Omega$*

$$\bar{u}(x) = \begin{cases} u_m(x) & \text{if } \gamma(x)p(x) + \alpha\bar{u}(x) > 0, \\ \in [u_m(x), u_M(x)] & \text{if } \gamma(x)p(x) + \alpha\bar{u}(x) = 0, \\ u_M(x) & \text{if } \gamma(x)p(x) + \alpha\bar{u}(x) < 0. \end{cases} \tag{7.64}$$

An equivalent formulation is given by the pointwise variational inequality in $\mathbb{R}$

$$(\gamma(x)p(x) + \alpha\bar{u}(x))(v - \bar{u}(x)) \geq 0 \quad \forall v \in [u_m(x), u_M(x)], \quad \text{for a.e. } x \in \Omega. \tag{7.65}$$

Proof. First we show that the variational inequality implies (7.64). Let $\bar{u}, u_m, u_M$ be fixed representatives in the equivalence class $L^\infty(\Omega)$. Assume that (7.64) is false. We consider the measurable sets

$$A_+ := \{x \in \Omega \ : \ \gamma(x)p(x) + \alpha\bar{u}(x) > 0\}, \tag{7.66}$$
$$A_- := \{x \in \Omega \ : \ \gamma(x)p(x) + \alpha\bar{u}(x) < 0\}. \tag{7.67}$$

By assumption there exists $E_+ \subset A_+(\bar{u})$ with positive measure such that $\bar{u}(x) > u_m(x)$ for all $x \in E_+$ or analog there exists $E_- \subset A_-(\bar{u})$ with positive measure such that $\bar{u}(x) < u_M(x)$ for all $x \in E_-$. In the first case we define

$$u(x) = \begin{cases} u_m(x) & \text{for } x \in E_+, \\ \bar{u}(x) & \text{for } x \in \Omega \setminus E_+. \end{cases} \tag{7.68}$$

Then

$$\int_\Omega (\gamma(x)p(x)+\alpha\bar{u}(x))(u(x)-\bar{u}(x))\mathrm{d}x = \int_{E_+} (\gamma(x)p(x)+\alpha\bar{u}(x))(u_m(x)-\bar{u}(x))\mathrm{d}x < 0, \tag{7.69}$$

since the first factor is postive on E_+ and the second negative which contradicts (7.63). For the other case we proceed accordingly.

Next we show that (7.64) implies (7.65). For almost all $x \in A_+(\bar{u})$ we have that $\bar{u}(x) = u_m(x)$, and thus $v - \bar{u}(x) \geq 0$ for all $v \in [u_m(x), u_M(x)]$. Hence, by the definition of $A_+(\bar{u})$ we have

$$(\gamma(x)p(x) + \alpha\bar{u}(x))(v - \bar{u}(x)) \geq 0 \quad \text{for a.a. } x \in A_+(\bar{u}). \tag{7.70}$$

Similarly we show that the inequality holds almost everywhere in $A_-(\bar{u})$.

Finally we show that (7.65) implies (7.63). Let $u \in U_{\rm ad}$ be arbitrary chosen. Since $\bar{u} \in [u_m(x), u_M(x)]$ for almost every $x \in \Omega$ we may put $v := u(x)$ in (7.65) to find that

$$(\gamma(x)p(x) + \alpha\bar{u}(x))(u(x) - \bar{u}(x)) \geq 0 \quad \text{for a.a. } x \in \Omega. \tag{7.71}$$

By integration we obtain the variational inequality (7.63). □

The pointwise inequality (7.65) can be rewritten by rearrangement of the terms; we obtain the form

$$(\gamma(x)p(x) + \alpha\bar{u}(x))\bar{u}(x) \leq (\gamma(x)p(x)v + \alpha\bar{u}(x))v \quad \text{for all } v \in [u_m(x), u_M(x)] \tag{7.72}$$

for almost all $x \in \Omega$.

Lemma 7.11. *A control $u \in U_{ad}$ is optimal for (11.1) if and only if it satisfies, together with the adjoint state, one of the following two minimum conditions for almost all $x \in \Omega$:*

(i) Weak minimum principle*:*

$$\min_{v \in [u_m(x), u_M(x)]} (\gamma(x)p(x) + \alpha\bar{u}(x))v = (\gamma(x)p(x) + \alpha\bar{u}(x))\bar{u}(x). \tag{7.73}$$

(ii) Minimum principle

$$\min_{v \in [u_m(x), u_M(x)]} \left(\gamma(x)p(x)v + \frac{\alpha}{2}v^2\right) = \gamma(x)p(x)\bar{u}(x) + \frac{\alpha}{2}\bar{u}(x)^2. \tag{7.74}$$

Proof. The weak minimum principle is a reformulation of (7.72). We prove the minimum principle. A real number $\bar{v}$ solves for x the quadratic optimization problem in $\mathbb{R}$

$$\min_{v \in [u_m(x), u_M(x)]} g(v) := \gamma(x)p(x) + \frac{\alpha}{2}v^2, \tag{7.75}$$

iff the variational inequality

$$g(\bar{v})(v - \bar{v}) \geq 0 \quad \text{for all } v \in [u_m(x), u_M(x)] \tag{7.76}$$

is satisfied, i.e.

$$(\gamma(x)p(x) + \alpha\bar{v})(v - \bar{v}) \geq 0 \quad \text{for all } v \in [u_m(x), u_M(x)]. \tag{7.77}$$

The minimum condition follows from taking $\bar{v} = \bar{u}(x)$. □

On the regularity of the optimal control

In many situations one can show that the optimal control has better regularity.

Lemma 7.12. *Let $1 \leq r \leq \infty$ and let $u_m, u_M \in W^{1,r}(\Omega)$. Then for the projection on $U_{ad} := \{u \in L^2(\Omega) : u_m \leq u \leq u_M \text{ a.e. in } \Omega\}$ we have*

$$P_{U_{ad}}(u) \in W^{1,r}(\Omega) \quad \text{for all } u \in W^{1,r}(\Omega). \tag{7.78}$$

Corollary 7.13 (Improved regularity)**.** *Let $\Omega \subset \mathbb{R}^n$ be a bounded domain. Then for problem (11.1) we have*

$$\bar{u} = P\left(-\frac{1}{\alpha}\bar{p}\right) \in H^1(\Omega), \quad \bar{p} = p[\bar{u}]. \tag{7.79}$$

7.5. Application: The nonlinear case

We consider the following optimal control problem:

$$\begin{cases} \min \quad J(u,y) := \frac{1}{2}\|y-y_\mathrm{d}\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u\|^2_{L^2(\Omega)}, \\ -\Delta y + y + y^3 = \gamma u \quad \text{in } \Omega, \\ \partial_n y = 0 \quad \text{on } \partial\Omega, \\ u_m \le u \le u_M \quad \text{a.e. on } \Omega,\ u_m, u_M \in L^2(\Omega) \end{cases} \tag{7.80}$$

with $\Omega \subset \mathbb{R}^n$, $n \le 3$ open and bounded.

By the theory of monotone operators one can show, see Theorem 4.66, that there exists a unique bounded solution operator of the equation

$$u \in U := L^2(\Omega) \to y \in Y := H^1(\Omega). \tag{7.81}$$

Let $A\colon H^1(\Omega) \to H^1(\Omega)^*$ be the operator associated with the bilinear form $a(y,v) = \int_\Omega (\nabla y \cdot \nabla v + yv)\,\mathrm{d}x$ for the linear operator $-\Delta y + y$ and let

$$N\colon Y \to Z,\ y \mapsto y^3. \tag{7.82}$$

Then the weak formulation of the state equation can be written in the form

$$e(u,y) := Ay + N(y) - \gamma u = 0. \tag{7.83}$$

By the Sobolev embedding Theorem 1.14 one has for $n \le 3$ the continuous embedding

$$H^1(\Omega) \hookrightarrow L^6(\Omega). \tag{7.84}$$

Moreover, the mapping $N\colon L^6(\Omega) \to L^2(\Omega)$, $y \mapsto y^3$ is continuously Fréchét differentiable with

$$N'(y)v = 3y^2 v. \tag{7.85}$$

This follows from the following technical result:

Lemma 7.14. *Let $\omega \subset \mathbb{R}^n$ be measurable. Then, for all $p_i, p \in [1,\infty]$ with $\frac{1}{p_1} + \cdots + \frac{1}{p_k} = \frac{1}{p}$ and all $u_i \in L^{p_i}(\Omega)$, there holds $u_1 \ldots u_k \in L^p(\Omega)$ and*

$$\|u_1 \cdots u_k\|_{L^p(\Omega)} \le \|u_1\|_{L^{p_1}(\Omega)} \cdots \|u_k\|_{L^{p_k}(\Omega)} \tag{7.86}$$

Proof. The proofs follows by induction, see [15, p. 76]. □

We now return to the proof of the F-differentiability of N: We just have to apply the Lemma with $p_1 = p_2 = p_3 = 6$ and $p = 2$:

$$\begin{aligned}\left\|(y+h)^3 - y^3 - 3y^2h\right\|_{L^2(\Omega)} &= \left\|3yh^2 + h^3\right\|_{L^2(\Omega)} \le 3\left\|y\right\|_{L^6(\Omega)}\left\|h\right\|^2_{L^6(\Omega)} + \left\|h\right\|^3_{L^6(\Omega)} \\ &= O(\|h\|^2_{L^6(\Omega)}) = o(\|h\|_{L^6(\Omega)}).\end{aligned} \tag{7.87}$$

This shows the F differentiability of N with derivative N'. Furthermore, to prove the continuity of N', we estimate

$$\begin{aligned}\left\|(N'(y+h) - N'(y))v\right\|_{L^2(\Omega)} &= 3\left\|((y+h)^2 - y^2)v\right\|_{L^2(\Omega)} = 3\left\|(2y+h)hv\right\|_{L^2(\Omega)} \\ &= 3\left\|2y+h\right\|_{L^6(\Omega)}\left\|h\right\|_{L^6(\Omega)}\left\|v\right\|_{L^6(\Omega)}.\end{aligned} \tag{7.88}$$

Hence,

$$\left\|N'(y+h) - N'(y)\right\|_{L(L^6(\Omega),L^2(\Omega))} \le 3\left\|2y+h\right\|_{L^6(\Omega)}\left\|h\right\|_{L^6(\Omega)} \to 0 \quad (\|h\|_{L^6(\Omega)} \to 0). \tag{7.89}$$

Therefore, $e\colon Y \times U \to Y^* =: Z$ is continuously Fréchet differentiable with

$$e_y(u,y)v = Av + 3y^2v, \quad e_u(u,y)w = -\gamma w. \tag{7.90}$$

Finally, $e_y(u,y) \in L(Y,Z)$ has a bounded inverse since for any $y \in Y$ the equation

$$Av + 3y^2v = f \tag{7.91}$$

has a bounded solution operator $f \in Z \to v \in Y$ by the Lax Milgram lemma. In fact, $A + 3y^2\,\mathrm{id} \in L(Y,Z)$ and corresponds to the bounded and coercive bilinear form $(v,w) \in Y \times Y \mapsto a(v,w) + (3y^2v,w)_{L^2(\Omega)(\Omega)}$.

Hence, Assumption 7.5 is satisfied. The optimality conditions are now similar to the linear quadratic problem (11.1): Let $(\bar u, \bar y) \in U \times Y$ be an optimal solution. Then by Corollary 7.8 the optimality system is of the form (7.49) and reads as

$$\begin{aligned} a(\bar y, v) + (\bar y^3, v)_{L^2(\Omega)} - (\gamma \bar u, v)_{L^2(\Omega)} &= 0 \quad \text{for all } v \in Y, \\ (\bar y - y_\mathrm{d}, v)_{L^2(\Omega)} + a(\bar p, v) + (\bar p, 3\bar y^2 v)_{L^2(\Omega)} &= 0 \quad \text{for all } v \in Y, \\ u_m \le \bar u \le u_M, \quad (\alpha \bar u - \gamma \bar p, u - \bar u)_{L^2(\Omega)} &\ge 0, \quad \text{for all } u \in U,\ u_m \le u \le u_M. \end{aligned} \tag{7.92}$$

Now, the adjoint equation (7.92) is just the weak formulation

$$-\Delta \bar p + \bar p + 3\bar y^2 \bar p = -(\bar y - y_\mathrm{d}), \quad \partial_n \bar p|_{\partial\Omega} = 0. \tag{7.93}$$

By Theorem 7.7 we have

Theorem 7.15. *If* $(\bar{u}, \bar{y})$ *is an optimal solution of* (11.1) *then there exists a* $\bar{p} \in H^1(\Omega)$, $\bar{\lambda}_{u_m}, \bar{\lambda}_{u_M} \in L^2(\Omega)$ *such that the following optimality system holds in the weak sense*

$$\begin{cases} -\Delta \bar{y} + \bar{y} + \bar{y}^3 = \gamma \bar{u}, & \partial_n \bar{y}|_{\partial\Omega} = 0, \\ -\Delta \bar{p} + \bar{p} + 3\bar{y}^2 \bar{p} = -(\bar{y} - y_d), & \partial_n \bar{p}|_{\partial\Omega} = 0, \\ \alpha \bar{u} - \gamma \bar{p} + \bar{\lambda}_{u_M} - \bar{\lambda}_{u_m} = 0, & \\ \bar{u} \geq u_m, \quad \bar{\lambda}_{u_m} \geq 0, \quad \bar{\lambda}_{u_m}(\bar{u} - u_m) = 0, & \\ \bar{u} \leq u_M, \quad \bar{\lambda}_{u_M} \geq 0, \quad \bar{\lambda}_{u_M}(u_M - \bar{u}) = 0. & \end{cases} \tag{7.94}$$

8. Dirichlet boundary control

Let $\Omega \subset \mathbb{R}^n$ with C^2 boundary.

$$\begin{cases} \min\limits_{(u,y)\in U\times Y} J(u,y) := \frac{1}{2}\left\|y - y_{\mathrm{d}}\right\|^2_{L^2(\Omega)} + \dfrac{\alpha}{2}\left\|u\right\|^2_{L^2(\partial\Omega)}, \\ \text{s.t. } y - \Delta y = f \quad \text{in } \Omega, \quad y = u \quad \text{on } \Omega, \\ u \in U_{\mathrm{ad}} \end{cases} \tag{8.1}$$

In view of the cost function, we are a priori looking for u in the space $L^2(\partial\Omega)$. But then we do not know if the state equation is well-posed.

Formal analysis: The Lagrangian of the problem is:

$$L(u,y,p) := J(u,p) + \int_\Omega p(x)(f(x) + \Delta y(x) - y(x))\mathrm{d}x + \int_{\partial\Omega} q(x)(y(x) - u(x))\mathrm{d}x. \tag{8.2}$$

So the costate equation is (omitting 'x')

$$0 = L_y z = J_y z + \int_\Omega p(\Delta z - z) + \int_{\partial\Omega} qz. \tag{8.3}$$

Integrating by parts, we obtain

$$0 = J_y z + \int_\Omega (\Delta p - p)z + \int_{\partial\Omega} (p\partial_n z + (q - \partial_n p)z. \tag{8.4}$$

We deduce that $q = \partial_n p$, and p is solution of

$$p - \Delta p = y - y_d; \quad p = 0 \text{ on } \partial\Omega. \tag{8.5}$$

Then (again formally) the reduced cost is characterized by

$$F'(u)v = L_u v = \int_{\partial\Omega} (\alpha u - \partial_n p)v\mathrm{d}x. \tag{8.6}$$

Rigorous derivation. Let

$$w - \Delta w = 0 \quad \text{in } \Omega, \quad w = v \quad \text{on } \Omega. \tag{8.7}$$

Consider the operator

$$\Lambda : L^2(\Omega) \to H^{\frac{1}{2}}(\partial\Omega), \tag{8.8}$$

defined as follows: Given $\phi \in L^2(\Omega)$, $\Lambda(\phi) := -\partial_n w$, where w is solution of[1]

$$w - \Delta w = \phi; \quad w = 0 \text{ on } \partial\Omega. \tag{8.9}$$

Observe that (8.9) has a unique solution in $H^2(\Omega)$. By remark 4.50, Λ is welldefined and is a continuous operator. Since the inclusion of $H^{\frac{1}{2}}(\partial\Omega)$ into $L^2(\partial\Omega)$ is compact, we may redefine Λ as a compact operator $L^2(\Omega) \to L^2(\partial\Omega)$. Its tranpose $\Lambda^* \in L(L^2(\partial\Omega), L^2(\Omega))$ is also compact. So there exists

$$\hat{z} := \Lambda^* v \in L^2(\Omega). \tag{8.10}$$

If $\hat{z}$ is a smooth solution of the state equation (it is enough that $\hat{z} \in H^2(\Omega)$), then for any $\phi \in L^2(\Omega)$:

$$\begin{aligned} \int_\Omega \hat{z}\phi = (\Lambda^* v, \phi)_{L^2(\Omega)} = (v, \Lambda\phi)_{L^2(\partial\Omega)} = -(z, \partial_n w)_{L^2(\partial\Omega)} = -\int_\Omega (\nabla z \cdot \nabla w + z\Delta w) \\ = -\int_{\partial\Omega} w\partial_n z + \int_\Omega (w\Delta z + z(\phi - w)) = \int_\Omega z\phi. \end{aligned} \tag{8.11}$$

Since this holds for any $\phi \in L^2(\Omega)$, this means that $z = \Lambda^* u$.

Therefore for any $v \in L^2(\partial\Omega)$ we can define

$$z[v] \in L^2(\Omega) \text{ as } z[v] := \Lambda^* v. \tag{8.12}$$

Definition 8.1. *A function $y \in L^2(\Omega)$ is a solution to the state equation in (8.16) if, and only if,*

$$\int_\Omega y\phi = \langle f, w\rangle_{H_0^1(\Omega)} - \int_{\partial\Omega} u\partial_n w \tag{8.13}$$

for all $\phi \in L^2(\Omega)$, where w is the solution to

$$-\Delta w = \phi \textit{ in } \Omega, w = 0 \textit{ on } \partial\Omega. \tag{8.14}$$

We have already stated the costate equation, which has a unique solution $p \in H^2(\Omega) \cap H_0^1(\Omega)$.

[1] We use the fact that $w \in H^2(\Omega)$.

The derivative of the reduced cost function is given by

$$\begin{aligned} F'(u)v &= (\alpha u, v)_{L^2(\partial\Omega)} + (y - y_d, z)_{L^2(\Omega)} = (\alpha u, v)_{L^2(\partial\Omega)} + (y - y_d, \Lambda^* v)_{L^2(\Omega)} \\ &= (\alpha u, v)_{L^2(\partial\Omega)} + (\Lambda(y - y_d), v)_{L^2(\Omega)} = (\alpha u, v)_{L^2(\partial\Omega)} - \int_{\partial\Omega} \partial_n p v. \end{aligned} \tag{8.15}$$

To remember:

Distributed control:

$$\begin{cases} \min\limits_{(u,y)\in L^2(\Omega)\times H_0^1(\Omega)} J(u,y) := \frac{1}{2}\left\|y - y_{\mathrm{d}}\right\|_{L^2(\Omega)}^2 + \dfrac{\alpha}{2}\left\|u\right\|_{L^2(\partial\Omega)}^2, \\ \text{s.t. } y - \Delta y = f + u \quad \text{in } \Omega, \quad y = 0 \quad \text{on } \Omega, \\ \text{gives } p - \Delta p = y - y_{\mathrm{d}}, \quad p = 0 \quad \text{on } \Omega, \\ \alpha u + p = 0 \end{cases} \tag{8.16}$$

Neumann boundary control:

$$\begin{cases} \min\limits_{(u,y)\in L^2(\partial\Omega)\times H_0^1(\Omega)} J(u,y), \\ \text{s.t. } y - \Delta y = f \quad \text{in } \Omega, \quad \partial_n y = u \quad \text{on } \Omega, \\ \text{gives } p - \Delta p = y - y_{\mathrm{d}}, \quad \partial_n p = 0 \quad \text{on } \Omega, \\ \alpha u + p = 0 \text{ on } \partial\Omega, \end{cases} \tag{8.17}$$

Dirichlet boundary control:

$$\begin{cases} \min\limits_{(u,y)\in L^2(\partial\Omega)\times H_0^1(\Omega)} J(u,y), \\ \text{s.t. } y - \Delta y = f \quad \text{in } \Omega, \quad y = u \quad \text{on } \Omega, \\ \text{gives } p - \Delta p = (y - y_{\mathrm{d}}), \quad p = 0 \quad \text{on } \Omega, \\ \alpha u - \partial_n p = 0 \text{ on } \partial\Omega, \end{cases} \tag{8.18}$$

Note the different sign in the in the optimality condition in the last case.

Part II.

Numerics

9. Optimization algorithms

In this chapter we consider optimization algorithms for solving optimization problems with PDEs.

We will consider the algorithm in function space. In this chapter we assume, that the PDEs (state, adjoint equation, etc.) can be solved and describe the algorithms in an infinite dimensional setting. These algorithms are then applied to discretized problems. Since these algorithm are formulated in infinite dimensional spaces, they often have good convergence properties for high dimensional discretizations.

Contents

9.1. Gradient methods for optimization problems

We consider a unconstrained optimization problem

$$\min F(u), \quad u \in U \tag{9.1}$$

with Hilbert space U and Fréchet differentiable function $F\colon U \to \mathbb{R}$.

All iterative methods have a starting point with a value u_0 in U and produce a sequence (u_n) which should approximate a local minimum $\bar{u}$.

We distinguish three types of convergence rates.

Definition 9.1. *A sequence (u_n) converges q-linear with rate $0 < \gamma < 1$ towards $\bar{u} \in U$ if there exists a $N \in \mathbb{N}$, such that*

$$\|u_{n+1} - \bar{u}\|_U \leq \gamma \|u_n - \bar{u}\|_U \quad \forall n \geq N \tag{9.2}$$

is satisfied.

Definition 9.2. *A sequence (u_n) converges q-superlinear towards $\bar{u} \in U$ if $u_n \to \bar{u}$ for $n \to \infty$ and if there exists a sequence $(c_n) \subset \mathbb{R}$ with*

$$\|u_{n+1} - \bar{u}\|_U \leq c_n \|u_n - \bar{u}\|_U \quad \forall n \in \mathbb{N} \quad \textit{and } c_n \to 0,\ n \to \infty \tag{9.3}$$

is satisfied.

Definition 9.3. *A sequence (u_n) converges q-quadratic towards $\bar{u} \in U$ if $u_n \to \bar{u}$ for $n \to \infty$ and there exists a $C > 0$ and a $N \in \mathbb{N}$, such that*

$$\|u_{n+1} - \bar{u}\|_U \leq C \|u_n - \bar{u}\|_U^2 \quad \forall n \geq N \tag{9.4}$$

is satisfied.

The general descent methods have the following structure:

1. Set $u_0 \in U$ and set $k = 0$.

2. Check stopping criterion for u_k.

3. Determine the descent direction $s_k \in U$, i.e. $F'(u_k)(s_k) < 0$.

4. Determine the step size $\sigma_k > 0$ with sufficiently big descent

$$F(u_k) - F(u_k + \sigma_k s_k). \tag{9.5}$$

5. Set $u_{k+1} = u_k + \sigma_k s_k$.

6. Set $k = k + 1$ and goto Step 2.

Remark 9.4. (i) Different descent methods differ in the choice of the descent direction, the choice of the step size and the definition of the stopping criterion.

(ii) Choosing the steepest descent $s_k = -\nabla F(u_k)$ we obtain the *gradient method.* This is possible in a Hilbert space by the definition

$$(\nabla F(u), \delta u)_U = F'(u)(\delta u) \quad \forall \delta u \in U. \tag{9.6}$$

The existence and uniqueness of $\nabla F(u)$ by this definition is given by Riesz.

(iii) We consider $\nabla F(u) \neq 0$, then the direction

$$s = -\frac{1}{\|\nabla F(u)\|_U} \nabla F(u) \tag{9.7}$$

is the solution to the minimization problem

$$\text{Min} \quad (\nabla F(u), d)_U, \quad \text{s.t. } \|d\|_U = 1. \tag{9.8}$$

(iv) If one considers a different scalar product, we obtain the gradient method w.r.t. this new product. The choice of the scalar product is often a crucial point which may have big impact on the convergence properties of the gradient method.

We consider the typical step size rules for the choice of σ_k:

- *Minimization rule*: One chooses σ_k as the minimum of

  $$\text{Min} \quad g_k(\sigma), \quad \sigma \geq 0 \tag{9.9}$$

  with $g_k(\sigma) = F(u_k + \sigma s_k)$. This minimum can only in special situation determined exactly (e.g. if F is quadratic and convex).

- *Armijo-rule*: Choose $\gamma, \beta \in (0,1)$ Then one finds σ_k as the biggest number in

  $$\{1, \beta, \beta^2, \dots\}, \tag{9.10}$$

  so that the following condition is satisfied:

  $$F(u_k + \sigma_k s_k) - F(u_k) \leq \gamma \sigma_k (\nabla F(u_k), s_k)_U. \tag{9.11}$$

  One can show, that for every descent direction s_k there exists $\sigma_k = \beta^l$ with sufficient large l, such that the Armijo-condition is fullfilled.

In a gradient descent method with $s_k = -\nabla F(u_k)$ the Armijo-rule is equivalent to

$$F(u_k + \sigma_k s_k) - F(u_k) \leq -\gamma \sigma_k \|\nabla F(u_k)\|_U^2 \tag{9.12}$$

For the gradient method with Armijo step size rule we have the following result:

Theorem 9.5. *Let $F\colon U \to \mathbb{R}$ be a continuous Fréchet differentiable and a $u_0 \in U$. Let (u_n) be given by the gradient method with Armijo-rule with $\nabla F(u_n) \neq 0$ for all $n \in \mathbb{N}$. Then every limit point $\bar{u}$ of (u_n) is a stationary point of F.*

The theorem gives no existence of limit points and no convergence of the whole sequence.

Remark 9.6. The gradient method typically has q-linear convergence. If F is quadratic and convex, one can show that the sequence (u_n) converges and that there exists $0 < \gamma < 1$ such that

$$\|u_{n+1} - \bar{u}\|_U \leq \gamma \, \|u_n - \bar{u}\|_U \quad \forall n \in \mathbb{N}. \tag{9.13}$$

Example 9.7. For $U = \mathbb{R}^n$ with

$$F(u) = \frac{1}{2} u^\top A u - b^\top u \tag{9.14}$$

with a positive definite matrix $A \in \mathbb{R}^{n \times n}$ and a vector $b \in U$ there holds:

$$\gamma \leq \frac{\kappa - 1}{\kappa + 1}, \quad \kappa = \operatorname{cond}_2(A). \tag{9.15}$$

For $\kappa \gg 1$ it leads typical to $\gamma \approx 1$ and consequently to slow convergence.

9.2. Application

Let $\Omega \subset \mathbb{R}^n$ open and bounded. We consider the problem

$$\begin{cases} \min\limits_{(u,y)\in Y\times U} J(u,y) = \frac{1}{2}\left\|y - y_{\mathrm{d}}\right\|^2_{L^2(\Omega)} + \dfrac{\alpha}{2}\left\|u\right\|^2_{L^2(\partial\Omega)}, \quad \alpha > 0, \quad \text{s.t.} \\ \qquad -\Delta y + y^3 = f + u \quad \text{in } \Omega, \\ \qquad\quad \partial_n y + y = 0 \quad \text{on } \partial\Omega. \end{cases} \tag{9.16}$$

with adjoint equation

$$\begin{aligned} -\Delta p + 3y^2 p &= -(y - y_{\mathrm{d}}) && \text{in } \Omega, \\ \partial_n p + p &= 0 && \text{on } \partial\Omega. \end{aligned} \tag{9.17}$$

We have

$$F'(u)(\delta u) = (\alpha u - p, \delta u)_U \tag{9.18}$$

and hence

$$\nabla F(u) = \alpha u - p. \tag{9.19}$$

We obtain the following method:

1. Choose a $u_0 \in U$ and set $k = 0$.
2. Solve the state equation and determine $y_k = y_k[u_k]$ and $F(u_k) = J(u_k, y_k)$.
3. Solve the adjoint equation and obtain $p_k \in Y$.
4. Determine $\nabla F(u_k)$ as
$$\nabla F(u_k) = \alpha u_k - p_k. \tag{9.20}$$
5. Check the stopping criterion for u_k:
$$\left\|\nabla F(u_k)\right\|_U < \varepsilon \quad \text{then} \quad \text{Stopp.} \tag{9.21}$$
6. Set the search direction $s_k = -\nabla F(u_k)$.
7. Determine the step size $\sigma_k > 0$ after the Armijo-rule.

Remark 9.8. In step 7 possibly the state equation has to be solved several times to check the Armijo-condition.

9.3. Constraints on the controls

In case of additional constraints on the controls, i.e.

$$u_m \leq u \leq u_M \quad \text{a.e.} \tag{9.22}$$

for u_m and u_M in U, one determines the step size σ_k by

$$F(P_{[u_m,u_M]}(u_k + \sigma_k s_k)) = \min_{\sigma>0} F(P_{[u_m,u_M]}(u_k + \sigma s_k)) \tag{9.23}$$

and we set

$$u_{k+1} := P_{[u_m,u_M]}(u_k + \sigma_k s_k). \tag{9.24}$$

In this case the determination of the step size is a nontrivial task. The exact step size can usually no longer be calculated, therefore an acceptable step size has to be computed numerically, e.g., by the method of bisection: Starting from a small initial step size s_0 (e.g. the step size used in the previous step), one takes consecutively $s = \frac{s_0}{2}, \frac{s_0}{4}, \frac{s_0}{8}, \ldots$ until an s is found such that $F(P_{[u_m,u_M]}(u_k + \sigma s_k))$ is sufficiently smaller than the previous value $F(u_k)$.

9.4. Newton-method

We consider the Newton-method in Banach spaces.

Motivation of the Newton method by quadratic model function. A motivation for the Newton method is the minimization of a quadratic model function. Let u_k be the current iterate. If $F\colon U \to \mathbb{R}$ is twice continuously Fréchet differentiable, we have

$$F(u_k + \delta u) = F(u_k) + F'(u_k)(\delta u) + \frac{1}{2}F''(u_k)(\delta u, \delta u) + o(\|\delta u\|_U^2). \tag{9.25}$$

We consider the local model

$$m_k(p) := F(u_k) + F'(u_k)(p) + \frac{1}{2}F''(u_k)(p, p). \tag{9.26}$$

The necessay optimality condition for a minimum of δu of $m_k(p)$ is given as

$$m_k'(\delta u)(\tau u) = 0 \quad \forall \tau u \in U. \tag{9.27}$$

There holds

$$m_k'(\delta u)(\tau u) = F'(u_k)(\tau u) + F''(u_k)(\delta u, \tau u). \tag{9.28}$$

Hence, $\delta u \in U$ must solve the equation

$$F''(u_k)(\delta u, \tau u) = -F'(\delta u)(\tau u) \quad \forall \tau u \in U. \tag{9.29}$$

The variational equation is equivalent to the Newton equation

$$\nabla^2 F(u_k)\delta u = -\nabla F(u_k). \tag{9.30}$$

This idea leads to the following Newton-Line-Search-algorithm.

1. Choose an initial value $u_0 \in U$ and set $k = 0$.
2. Determine the Newton direction as the solution of
$$\nabla^2 F(u_k)\delta u = -\nabla F(u_k). \tag{9.31}$$
3. Determine the step size σ_k (e.g. with Armijo-rule).
4. Set $u_{k+1} = u_k + \sigma_k \delta u$.
5. Check the stopping criterion.
6. Set $k = k + 1$ and goto 2.

Remark 9.9. (i) If the direction δu is no descent direction, one cannot perform line-search (e.g. Armijo) in general. If the Newton direction is no descent direction normally one sets $\delta u = -\nabla F(u_k)$.

(ii) If $\nabla^2 F(u_k)$ is positiv definite, δu is a descent direction. There holds

$$(\nabla F(u_k), \delta u)_U = -F''(u_k)(\delta u, \delta u) < 0. \tag{9.32}$$

(iii) For the Armijo-rule one starts with $\sigma = 1$, to have the possibility to make a full Newton step.

Motivation of the Newton method by linearization. A different motivation for the Newton method is the iterative solution of the stationary equation

$$\nabla F(\bar{u}) = 0. \tag{9.33}$$

We consider the general system

$$G(z) = 0 \tag{9.34}$$

with operator $G\colon X \to X$ and a Banach space X. If G is Fréchet differentiable, then

$$G(x_k + \delta x) = G(x_k) + G'(x_k)(\delta x) + o(\|\delta x\|_X). \tag{9.35}$$

This motivates the Newton equation for the update δx:

$$G'(x_k)(\delta x) = -G(x_k). \tag{9.36}$$

When choosing $G(q) = \nabla F(q)$ we obtain the Newton equation

$$\nabla^2 F(u_k)\delta u = -\nabla F(u_k). \tag{9.37}$$

Theorem 9.10 (Newton-Kantorovich). *Let X be Banach space, $D \subset X$ open and convex. Let $F\colon D \to X$ continuously Fréchét-differentiable on D. Let $x_0 \in D$ be given, so that $F'(x_0)$ is invertible. We assume*

(i) $\|F'(x_0)^{-1}F(x_0)\| \leq \alpha$,

(ii) $\|F'(x_0)^{-1}(F'(y) - F'(x))\| \leq \omega_0 \|y - x\|$ *for all* $x, y \in D$,

(iii) $h_0 = \alpha\omega_0 \leq \frac{1}{2}$.

(iv) The closed ball $K(x_0, \rho)$ with

$$\rho = \frac{1 - \sqrt{1 - 2h_0}}{\omega_0} \tag{9.38}$$

lies in D.

Then the Newton sequence (x_n),

$$x_{n+1} = x_n + \delta x_n, \quad F'(x_n)\delta x_n = -F(x_n) \tag{9.39}$$

is well-defined, stays in $K(x_0, \rho)$ and converges towards $\bar{x}$ with $F(\bar{x}) = 0$.

If $h_0 < \frac{1}{2}$, then the convergence rate is quadratic.

Remark 9.11. (i) The theorem implies the existence of a zero.

(ii) We obtain quadratic convergence, if the initial value lies in a (typically unknown) neighbourhood of the solution.

(iii) Often one uses Newton-type methods where the search direction is obtained from the equation

$$B_k \delta u = -\nabla F(u_k) \tag{9.40}$$

with a suitable operator B_k.

a) Simplified Newton-methods:

$$B_{k+l} = \nabla^2 F(u_k), \quad l = 0, 1, 2, \ldots, l_{\max}. \tag{9.41}$$

b) Update-methods (BFGS,...)

$$B_{k+1} = B_k + M_k \tag{9.42}$$

with update M_k, so that B_{k+1} satiesfies the condition

$$B_{k+1}\delta u = \nabla F(u_{k+1}) - \nabla F(u_k). \tag{9.43}$$

9.5. Application

For Hilbert spaces Y and U we consider the problem

$$\begin{cases} \min J(u,y), \quad u \in U, \quad y \in Y, \text{ s.t.} \\ a(u,y)(v) = f(v) \quad v \in Y \end{cases} \tag{9.44}$$

with semilinearform $a\colon U \times Y \times Y \to \mathbb{R}$ being linear and bounded in the third argument, and $f \in Y^*$. We further define

$$L(u,y,p) := J(u,y) + a(u,y)(p) - \langle f, p \rangle. \tag{9.45}$$

In the following we assume that all considered derivatives of J, a, and L exist.

In this section we discuss, how the Newton equation

$$\nabla^2 F(u_k)\delta u_k = -\nabla F(u_k) \tag{9.46}$$

for the corresponding reduced cost functional $F\colon U \to \mathbb{R}$ can be solved.

There are the following difficulties:

(i) How can one determine the second derivatives of the reduced cost functional ?

(ii) If U is infinite dimensional, then the dimension $\dim U_h$ after discretization is typically very large. The resulting matrix $\nabla^2 F_h(u_k)$ is typically not sparse.

Remark 9.12. (i) One uses matrix free methods to solve the linear systems.

(ii) In a matrix free method (e.g. the CG-method) for solving the system $Ax = b$ in every step matrix vector multiplications are performed with the matrix A. The matrix A will not be used explicitly.

(iii) Therefore the aim is the efficient evaluation of $\nabla^2 F(u)(\delta u)$ or equivalent of $F''(u)(\delta u, \tau u)$ for a given $u \in U$, a given $\delta u \in U$ and all $\tau u \in U$.

First approach to derive second derivatives

The naive procedure would be to derive from $F(u) = J(u, y[u])$

$$F'(u)(\delta u) = J_y(u,y[u])(\delta y) + J_u(u,y[u])(\delta u) \tag{9.47}$$

with $\delta y = y'[u](\delta u)$ and further

$$\begin{aligned} F(u)''(\delta u, \tau u) = F_{uu}(u,y[u])(\delta u, \tau u) + F_{uy}(u,y[u])(\delta u, \tau y) + F_{yu}(u,y[u])(\delta y, \tau u) \\ + F_{yy}(u,y[u])(\delta y, \tau y) + F''_y(u,y[u]])(\delta\tau y) \end{aligned} \tag{9.48}$$

with $\tau y = y'[u](\tau u)$ and $\delta\tau y = S''(u)(\delta u, \tau u)$.

Remark 9.13. The functions $\delta y \in Y$ and $\tau y \in Y$ satisfy tangent equations and one can also formulate an equation for $\delta\tau y$. For the corresponding evaluation for every pair $(\delta u, \tau u)$ three equations (for δy, τy, and $\delta\tau y$) have to be solved.

Second approach based on the Lagrangian. We saw already that the first derivative can be determined by

$$F'(u)(\delta u) = L_u(u, y, p)(\delta u), \tag{9.49}$$

when $p \in Y$ solves the adjoint equation

$$L_y(u, y, p)(v) = 0 \quad \forall v \in Y. \tag{9.50}$$

Since the state equation is equivalent to

$$L_p(u, y, p)(v) = 0 \quad \forall v \in Y \tag{9.51}$$

the representation of $F'(u)(\delta)$ for given u and δu in U can be written as follows:

$$F'(u)(\delta u) = L_y(u, y, p)(v) + L_u(u, y, p)(\delta u) + L_p(u, y, p)(w) \quad \forall v, w \in Y, \tag{9.52}$$

with $y = y[u]$ and $p = p[u]$. We derive the total derivative w.r.t u in the direction τu using the notation

$$\tau y = y'[u](\tau u), \quad \tau p = p'[u](\tau u). \tag{9.53}$$

Consequently,

$$\begin{aligned} F''(u)(\delta u, \tau u) = {} & L_{uu}(u, y, p)(\delta u, \tau u) + L_{uy}(u, y, p)(\delta u, \tau y) + L_{up}(u, y, p)(\delta u, \tau p) \\ & + L_{yu}(u, y, p)(v, \tau u) + L_{yy}(u, y, p)(v, \tau y) + L_{yp}(u, y, p)(v, \tau p) \\ & + L_{pu}(u, y, p)(w, \tau u) + L_{py}(u, y, p)(w, \tau y) + L_{pp}(u, y, p)(w, \tau p). \end{aligned} \tag{9.54}$$

This representation holds for all $v, w \in Y$. We use $L_{pp} = 0$ and choose $v \in Y$, so that

$$L_{up}(u, y, p)(\delta u, \phi) + L_{yp}(u, y, p)(v, \phi) = 0 \quad \forall \phi \in Y. \tag{9.55}$$

Further we choose $w \in Y$, so that

$$L_{uy}(u, y, p)(\delta u, \psi) + L_{yy}(u, y, p)(v, \psi) + L_{py}(u, y, p)(w, \psi) = 0 \quad \forall \psi \in Y. \tag{9.56}$$

We still have to show existence of v and w. We obtain

$$F''(u)(\delta u, \tau u) = L_{uu}(u, y, p)(\delta u, \tau u) + L_{yu}(u, y, p)(v, \tau u) + L_{pu}(u, y, p)(w, \tau u). \tag{9.57}$$

One can show, that $v \in Y$ solves the tangent equation and hence $v = \delta y = y'[u](\delta u)$. One can check that $w = \delta p = p'[u](\delta u)$.

Theorem 9.14. *Let $F\colon U \times Y \to \mathbb{R}$, the semilinear form a as well as the mappings $U \to Y$, $u \mapsto y[u]$ and $U \to Y$, $u \mapsto p[u]$ twice Fréchet differentiable. Let u and δu in U be given, $y = y[u]$, $p = p[u] \in Y$ and let $\delta y \in Y$ and $\delta p \in Y$ be solutions of*

$$L_{up}(u, y, p)(\delta u, \phi) + L_{yp}(u, y, p)(\delta y, \phi) = 0 \quad \forall \phi \in Y \tag{9.58}$$

and

$$L_{uy}(u, y, p)(\delta u, \psi) + L_{yy}(u, y, p)(\delta y, \phi) + L_{py}(u, y, p)(\delta p, \psi) = 0 \quad \forall \psi \in Y \tag{9.59}$$

be given. Then we have the representation

$$F''(u)(\delta u, \tau u) = L_{uu}(u, y, p)(\delta u, \tau u) + L_{yu}(u, y, p)(\delta y, \tau u) + L_{pu}(u, y, p)(\delta p, \tau u). \tag{9.60}$$

The equation for $\delta y \in Y$ (tangent equation) has the explicit form

$$\delta y \in Y : \quad a_y(u, y)(\delta y, \phi) = -a_u(u, y)(\delta u, \phi) \quad \forall \phi \in Y. \tag{9.61}$$

The equation for $\delta p \in Y$ is called 'Dual-for-Hessian' equation and has the explicit form

$$\begin{aligned} a_y(u, y)(\psi, \delta p) = {} & -J_{yy}(u, y)(\delta y, \psi) + a_{yy}(u, y)(\delta u, \psi, p) \\ & - J_{u,y}(u, y)(\delta u, \psi) + a_{uy}(u, y)(\delta u, \psi, p) \quad \forall \psi \in Y. \end{aligned} \tag{9.62}$$

Thus, the representation for $F''(u)(\delta u, \tau u)$ has the form

$$\begin{aligned} F''(u)(\delta u, \tau u) = {} & J_{uu}(u, y)(\delta u, \tau u) + a_{uu}(u, y)(\delta u, \tau u, p) \\ & + J_{yu}(u, y)(\delta y, \tau u) + a_{yu}(u, y)(\delta y, \tau u, p) \\ & + a_u(u, y)(\tau u, \delta p). \end{aligned} \tag{9.63}$$

If u and δu in U are given and $y = y[u] \in Y$ and $p = p[u] \in Y$ already computed, then, by solving two equations for $\delta y \in Y$ and $\delta p \in Y$ the second derivatives $F''(u)(\delta u, \tau u)$ can be computed for $\tau u \in U$. We have for $\nabla^2 F(u)\delta u$:

$$\nabla^2 F(u)\delta u = \nabla_u(L_u(u, y, p)(\delta u) + L_y(u, y, p)(\delta y) + L_p(u, y, p)(\delta p)). \tag{9.64}$$

Example 9.15. We consider the optimization problem[1]

$$\begin{cases} \min\limits_{(u,y)\in U\times Y} J(u,y) = \frac{1}{2}\left\|y - y_{\mathrm{d}}\right\|^2_{L^2(\Omega)} + \dfrac{\alpha}{2}\left\|u\right\|^2_{L^2(\Omega)}, \quad \alpha > 0, \quad \text{s.t.} \\ \qquad -\Delta y + y^3 = f + u \quad \text{in } \Omega, \\ \qquad\qquad y = 0 \quad \text{on } \partial\Omega, \\ \qquad U := L^2(\Omega),\ Y := H^1_0(\Omega). \end{cases} \tag{9.65}$$

Here we have

$$a(u,y)(v) = (\nabla y, \nabla p) + (y^3, v) - (u, v). \tag{9.66}$$

We have

- Dual equation:

$$p \in Y: \quad (\nabla v, \nabla p)_{L^2(\Omega)} + (3y^2 v, p)_{L^2(\Omega)} = -(y - y_d, v)_{L^2(\Omega)} \quad \forall v \in Y. \tag{9.67}$$

- Tangent equation

$$\delta y \in Y: \quad (\nabla \delta y, \nabla v)_{L^2(\Omega)} + (3y^2 \delta y, p)_{L^2(\Omega)} = (\delta u, v)_{L^2(\Omega)} \quad \forall v \in Y. \tag{9.68}$$

- Dual-for-Hessian-equation

$$\delta p \in Y: \quad (\nabla v, \nabla \delta p)_{L^2(\Omega)} + (3y^2 v, \delta p)_{L^2(\Omega)} = -(\delta y, v)_{L^2(\Omega)} + (6y\delta y, vp)_{L^2(\Omega)} \quad \forall v \in Y. \tag{9.69}$$

- The representation for $F''(u)(\delta u, \tau u)$ is

$$F''(u)(\delta u, \tau u) = \alpha(\delta u, \tau u)_{L^2(\Omega)} - (\tau u, \delta p)_{L^2(\Omega)} \tag{9.70}$$

  and hence

$$\nabla^2 F(u)\delta u = \alpha \delta u - \delta p. \tag{9.71}$$

We obtain the following realization of the Newton method:

(i) Choose initial value $u_0 \in U$ and set $k = 0$.

(ii) Solve the state equation, compute $y_k = y[u_k]$ and $F(u_k) = J(u_k, y_k)$.

(iii) Solve the adjoint equation, determine $p_k \in Y$ by

$$a_y(u_k, y_k)(v, p_k) = -J_y(u_k, y_k)(v) \quad \forall v \in V. \tag{9.72}$$

[1] Existence can be shown similar to the Neumann problem.

(iv) Compute $\nabla F(u_k)$ by

$$\nabla F(u_k) = \nabla_u L(u_k, y_k, p_k). \tag{9.73}$$

(v) Check stopping criterion for u_k:

$$\|\nabla F(u_k)\|_U < \varepsilon \text{ implies STOP.} \tag{9.74}$$

(vi) Solve (approximative) the Newton equation $\nabla^2 F(u_k)\delta u_k = -\nabla F(u_k)$ with the CG-methods.

→ In every CG-step solve the product

$$\nabla^2 F(u_k)\delta u = \nabla_u (L_u(u, y, p)(\delta y) + L_y(u, y, p)(\delta u) + L_p(u, y, p)(\delta p)) \tag{9.75}$$

by solving the tangent equation for $\delta y \in Y$ and the dual-for-Hessian-equation for $\delta p \in Y$.

(vii) Determine σ_k (e.g. by Armijo rule).

(viii) Set $u_{k+1} = u_k + \sigma_k s_k$.

(ix) Set $k = k + 1$ and goto (ii).

Remark 9.16. (i) In step 7 we have to solve the state equation possibly several times to check the Armijo rule.

(ii) In total many linear and nonlinear PDEs have to be solved within this Newton algorithm. This requires efficient algorithms.

(iii) If the dimension of U is small, it may be more efficient to assemble the matrix $\nabla^2 F(u)$. Therefore $(\dim U)$ tangent equations have to be solved.

(iv) It is very important in the algorithm how the different stopping criteria are chosen.

9.6. Semismooth Newton methods

We dicsuss an optimization algorithm to solve optimal control problems for partial differential equations with constraints on the controls. There exists different methods for optimization problems with inequality constraints:

(i) Primal-dual-active-set-strategies (PDAS): an iteration over active/inactive sets using adjoint information.

(ii) Semismooth Newton methods: The optimality conditions are formulated as a semismooth equation and a generalized Newton method is applied. PDAS methods can often be formulated as semismooth Newton methods.

(iii) Penalty methods: The inequality constraints are added with penalty functions.

Here we study semi-smooth Newton methods.

Definition 9.17. *Let X, Z be Banach spaces $D \subset X$ open set and $F\colon D \to Z$ a nonlinear mapping. The mapping F is* Newton differentiable *on the open set $U \subset D$, if there exist a family of mappings $G\colon U \to L(X,Z)$ so that*

$$\lim_{\|h\|_X \to 0} \frac{1}{\|h\|_X} \|F(x+h) - F(x) - G(x+h)h\|_Z = 0 \tag{9.76}$$

for all $x \in U$.

In comparison to the definition of Frechet differentiable mappings here $G(x)h$ is replaced by $G(x+h)h$.

Example 9.18. *Let X be a Hilbert space. Then the norm-functional $F(x) = \|x\|_X$ is Newton differentiable. It can be easily checked that $G(x+h)h = \left(\frac{x+h}{|x+h|}, h\right)_X$ and $G(0)h = (\lambda, h)_X$ for some λ with $\lambda \in X$ is a Newton derivative.*

This definition allows for a Newton method for solving the equation $F(x) = 0$ to prove superlinear convergence.

Theorem 9.19. *Let $(X, \|\cdot\|)$ be Banach space with associated norm $\|\cdot\|$, $D \subset X$ open and $\bar{x} \in D$ a solution of $F(x) = 0$. Let U be an open neighbourhood of $\bar{x}$ such that F Newton differentiable on U and the set*

$$\{\left\|G(x)^{-1}\right\| \;:\; x \in U\} \tag{9.77}$$

is bounded. Then the iteration

$$x_{k+1} = x_k + \delta x_k, \quad G(x_k)\delta x_k = -F(x_k) \tag{9.78}$$

converges against $\bar{x}$ *superlinear if* $\|x_0 - \bar{x}\|$ *small enough.*

Proof. Let $r > 0$ such that $B_r(\bar{x}) \subset U$ and $M > 0$ such that

$$\left\|G(x)^{-1}\right\| \le M \quad \text{for all } x \in B_r(\bar{x}). \tag{9.79}$$

Let $\eta \in (0, 1/2)$ arbitrary. There exists a $\rho \in (0, r)$ such that

$$\|F(\bar{x} + h) - F(\bar{x}) - G(\bar{x} + h)h\| \le \frac{\eta}{M} \|h\| \tag{9.80}$$

for all h with $\|h\| \le \rho$.

Let $x_0 \in B_\rho(\bar{x})$. We show by induction that x_k in $B_\rho(\bar{x})$ for all k. Let $x_k \in B_\rho(\bar{x})$ then there holds

$$\begin{aligned} x_{k+1} - \bar{x} &= x_{k+1} - x_k + x_k - \bar{x} \\ &= -G(x_k)^{-1}F(x_k) + x_k - \bar{x} \\ &= -G(x_k)^{-1}(F(x_k) - F(\bar{x}) - G(x_k)(x_k - \bar{x})). \end{aligned} \tag{9.81}$$

Thus

$$\|x_{k+1} - \bar{x}\| \le M\frac{\eta}{M} \|x_k - \bar{x}\| = \eta \|x_k - \bar{x}\| < \|x_k - \bar{x}\| . \tag{9.82}$$

Hence, $B_\rho(\bar{x}) \ni x_{k+1} \to x$.

Using the definition of Newton differentiability we have

$$\|x_{k+1} - \bar{x}\| = \left\|G(x_k)^{-1}\right\| \|(F(x_k) - F(\bar{x}) - G(x_k)(x_k - \bar{x}))\| \le o(\|x_k - x\|). \tag{9.83}$$

□

For the application of the semismooth Newton method to optimal control problems with inequality constraints we discuss the Newton differentiability of the max-function.

Proposition 9.20. *The function* $f\colon \mathbb{R} \to \mathbb{R}$, $f(x) := \max(0, x)$ *is Newton differentiable on* $\mathbb{R}$ *with Newton derivative*

$$g(x) := \begin{cases} 0, & x < 0 \\ 1, & x > 0 \\ \delta, & x = 0 \end{cases} \tag{9.84}$$

with arbitrary $\delta \in \mathbb{R}$.

Proof. We consider

$$d(x,h) = |\max(0,x+h) - \max(0,x) - g(x+h)h|. \tag{9.85}$$

If $x \neq 0$ and $|h| < |x|$, then there holds $d(x,h) = 0$. If $x = 0$ for all h, then $d(x,h) = 0$. Hence, f is Newton differentiable. □

Remark 9.21. *The generalization to the* max *operator on function spaces*

$$F(u)(x) = \max(0, u(x)), \quad x \in \Omega \tag{9.86}$$

is not trivial. A candidate for the Newton derivative is

$$G(u)v(x) = \begin{cases} 0, & u(x) < 0, \\ v(x), & u(x) > 0, \\ \delta v(x), & u(x) = 0 \end{cases} \tag{9.87}$$

with $\delta \in \mathbb{R}$.

Theorem 9.22. *(i) In general G is not a Newton derivative of the* max *operator $F\colon L^p(\Omega) \to L^p(\Omega)$ for $1 \leq p \leq \infty$.*

(ii) The max *operator $F\colon L^q(\Omega) \to L^p(\Omega)$ with $1 \leq p < q \leq \infty$ is Newton differentiable and G is the Newton derivative.*

For a proof we refer to Hintermüller, Ito, and Kunisch [14].

Proposition 9.23. *Let $H\colon D \subset L^p(\Omega) \to L^q(\Omega)$, $1 \leq p < q < \infty$ Fréchet-differentiable in D and let*

$$\phi\colon L^q(\Omega) \to L^p(\Omega) \text{ be Newton-differentiable} \tag{9.88}$$

with Newton derivative G. Then $F = \varphi(H)\colon D \subset L^p(\Omega) \to L^p(\Omega)$ is Newton differentiable with Newton derivative

$$G(H)H' \in L(L^p(\Omega), L^p(\Omega)). \tag{9.89}$$

Proof. We refer to Ito and Kunisch [16]. □

9.7. Application

We consider the semi-smooth Newton method for the following model problem

$$\begin{cases} \min \quad J(u,y) = \dfrac{1}{2}\left\|y - y_{\mathrm{d}}\right\|^2_{L^2(\Omega)} + \dfrac{\alpha}{2}\left\|u\right\|^2_{L^2(\Omega)}, \quad u \in U_{\mathrm{ad}}, \quad y \in Y, \quad \text{s.t.} \\ \qquad -\Delta y = u \quad \text{in } \Omega, \\ \qquad\quad y = 0 \quad \text{on } \partial\Omega \end{cases} \tag{9.90}$$

with $U = L^2(\Omega)$, $Y = H^1_0(\Omega)$, $y_{\mathrm{d}} \in L^2(\Omega)$, $\alpha > 0$ and

$$U_{\mathrm{ad}} = \{u \in U \ : \ u \le u_M \quad \text{a.e. in } \Omega\}, \quad u_M \in \mathbb{R}. \tag{9.91}$$

Let $S\colon L^2(\Omega) \to L^2(\Omega)$, $S(u) := y[u]$.

We know that the optimality condition given by

$$(\alpha\bar{u} - \bar{p}, \delta u - \bar{u}) \ge 0 \quad \text{for all } \delta u \in U_{\mathrm{ad}}, \tag{9.92}$$

where

$$p \in Y: \quad (\nabla v, \nabla \bar{p})_{L^2(\Omega)} = -(\bar{y} - y_{\mathrm{d}}, v)_{L^2(\Omega)} \qquad \text{for all } v \in Y. \tag{9.93}$$

The variational inequality is equivalent to

$$\bar{u} = P_{U_{\mathrm{ad}}}\left(\frac{1}{\alpha}\bar{p}\right). \tag{9.94}$$

Since the projection $P_{U_{\mathrm{ad}}}$ holds pointwise, the last equality can be refomulated as

$$\alpha(\bar{u} - u_M) + \max(0, \alpha u_M - \bar{p}) = 0. \tag{9.95}$$

For each control $u \in U$ the corresponding adjoint state is given as (exercise[2])

$$p[u] := -S^*(Su - y_{\mathrm{d}}). \tag{9.96}$$

Hence, the optimality condition can be rewritten as an operator equation $F(u) = 0$ with

$$F(u) = \alpha(u - u_M) + \max(0, \alpha u_M - p[u]). \tag{9.97}$$

Lemma 9.24. *The mapping*

$$F\colon L^2(\Omega) \to L^2(\Omega) \tag{9.98}$$

is Newton differentiable with Newton derivative

$$G_F(u)(h) = \alpha h + G_{\max}(\alpha u_M - p[u]))S^*Sh \tag{9.99}$$

with

$$(G_{\max}(v)\phi)(x) = \begin{cases} \phi(x), & v(x) \ge 0, \\ 0, & v(x) < 0. \end{cases} \tag{9.100}$$

[2] Define z as the r.h.s. of the adjoint equation and show $S^*z = p$.

Proof. The statement follows directly from the chain rule in Proposition 9.23, since the mapping

$$B\colon u \mapsto \alpha u_M - p[u] \tag{9.101}$$

is affin-linear and continuous (and hence Fréchet-differentiable) as $B\colon L^2(\Omega) \to H^1(\Omega) \subset L^p(\Omega)$ for $p > 2$ and the later inclusion is compact. □

Theorem 9.25. *The semi-smooth Newton method*

$$u_{k+1} = u_k + \delta u_k, \quad G_F(u_k)\delta u_k = -F(u_k) \tag{9.102}$$

converges superlinear towards $\bar{u}$ for $\|u_0 - \bar{u}\|_{L^2(\Omega)}$ sufficient small.

Proof. To apply Theorem 9.19 we proceed in two steps. First, we show that G_F is invertible, then second, the uniform boundedness.

Let

$$\begin{aligned} I &:= \{x \in \Omega \ : \ p[u](x) \leq \alpha u_M(x) \text{ a.e.}\}, \\ A &:= \Omega \setminus I. \end{aligned} \tag{9.103}$$

We denote the corresponding characteristic functions by χ_I and χ_A and define the extension-by-zero operator $E_I\colon L^2(I) \to L^2(\Omega)$ as its adjoint $E_I^*\colon L^2(\Omega) \to L^2(I)$ as a restriction operator. Accordingly, E_A and E_A^* are defined. Further, we denote the identity map on $L^2(I)$ by id_I and on $L^2(A)$ by id_A. Für $h \in L^2(\Omega)$ there holds

$$G_F(u)(h) = g, \quad g := \begin{cases} \alpha h & \text{on } A, \\ \alpha h + S^*Sh & \text{on } I. \end{cases} \tag{9.104}$$

Thus we have with $D = S^*S$

$$G_F(u)(h) = \begin{pmatrix} \alpha\,\mathrm{id}_I + E_I^* D E_I & E_I^* D E_A \\ 0 & \alpha\,\mathrm{id} \end{pmatrix} \begin{pmatrix} h_I \\ h_A \end{pmatrix} \tag{9.105}$$

with $h_I := E_I^* h$ and $h_A := E_A^* h$. Consequently, we obtain that for given $w \in U$ there exists a unique $h \in U$ such that $w = G_F(u)h$. Since $G_F(u) \in L(L^2(\Omega), L^2(\Omega))$ we obtain from the bounded inverse theorem $G_F(u)^{-1} \in L(L^2(\Omega), L^2(\Omega))$.

To verify the estimate we proceed as follows. We have to show that $\|G_F(u)^{-1}\|_{L(L^2(\Omega),L^2(\Omega))}$ is bounded on a neighbourhood of $\bar{u}$. We show that there is a constant $C > 0$ such that

$$\left\|G_F(u)^{-1}(w)\right\|_{L^2(\Omega)} \leq C \left\|w\right\|_{L^2(\Omega)} \quad \text{for all } u, w \in L^2(\Omega). \tag{9.106}$$

Let $u \in U$, χ_I the characteristic function of the set

$$I = \{x \in \Omega \ : \ \alpha u_M - p[u](x) \geq 0\} \tag{9.107}$$

and χ_A the characteristic function of $A = \Omega \setminus I$. Let $h \in L^2(\Omega)$ and

$$w = G_F(u)(h). \tag{9.108}$$

On the set A there holds

$$G_F(u)(h) = \alpha h \tag{9.109}$$

and hence

$$\|h\chi_A\|_{L^2(\Omega)} \leq \frac{1}{\alpha} \|w\chi_A\|_{L^2(\Omega)} . \tag{9.110}$$

On the set I there holds

$$G_F(u)(h) = \alpha h + S^*Sh \tag{9.111}$$

and we obtain

$$\begin{aligned} (w, h\chi_I)_{L^2(\Omega)} &= \alpha \|h\chi_I\|^2_{L^2(\Omega)} + (S^*Sh, h\chi_I)_{L^2(\Omega)} \\ &= \alpha \|h\chi_I\|^2_{L^2(\Omega)} + (S^*Sh\chi_I, h\chi_I)_{L^2(\Omega)} + (S^*Sh\chi_A, h\chi_I)_{L^2(\Omega)} \\ &= \alpha \|h\chi_I\|^2_{L^2(\Omega)} + \|Sh\chi_I\|^2_{L^2(\Omega)} + (S^*Sh\chi_A, h\chi_I)_{L^2(\Omega)} \end{aligned} \tag{9.112}$$

Hence we have

$$\alpha \|h\chi_I\|^2_{L^2(\Omega)} \leq \|w\chi_I\|_{L^2(\Omega)} \|h\chi_I\|_{L^2(\Omega)} + K^2 \|h\chi_A\|_{L^2(\Omega)} \|h\chi_I\|_{L^2(\Omega)} \tag{9.113}$$

with $K = \|S\|$ and hence

$$\begin{aligned} \alpha \|h\chi_I\|_{L^2(\Omega)} &\leq \|w\chi_I\|_{L^2(\Omega)} + K^2 \|h\chi_A\|_{L^2(\Omega)} \\ &\leq \|w\chi_I\|_{L^2(\Omega)} + \frac{K^2}{\alpha} \|w\chi_A\|_{L^2(\Omega)} . \end{aligned} \tag{9.114}$$

Finally we obtain

$$\alpha \|h\|_{L^2(\Omega)} \leq C(\alpha) \|w\|_{L^2(\Omega)} . \tag{9.115}$$

We conclude the statement. □

Remark 9.26. In this case the semismooth Newton method is equivalent to a primal-dual-active-set-strategy.

Primal-dual-active-set-strategy:

1. Choose $u_0 \in U$ and set $y_0 := Su_0$, $p_0 := -S(y_0 - y_d)$.

2. Set $\mu_0 := -\alpha u_0 + p_0$.

3. Check the stopping criterion.

4. For (u_k, y_k) determine

$$\begin{aligned} A_{k+1} &= \{x \in \Omega \mid \mu_k(x)\alpha(u_k - u_M)(x) > 0\}, \\ I_{k+1} &= \Omega \setminus A_{k+1}. \end{aligned} \tag{9.116}$$

5. Determine $(y_{k+1}, u_{k+1}) \in Y \times U$ as the solution of

$$\begin{cases} \min \quad J(u,y), \quad \text{s.t.} \\ \qquad u = u_M \text{ on } A_{k+1} \text{ and} \\ \qquad (\nabla y, \nabla v)_{L^2(\Omega)} = (u,v)_{L^2(\Omega)} \quad \forall v \in Y. \end{cases} \tag{9.117}$$

6. Determine the adjoint state $p_{k+1} = -S^*(y_{k+1} - y_d)$ and set

$$\mu_{k+1} = -\alpha u_{k+1} + p_{k+1}. \tag{9.118}$$

Remark 9.27. (i) One can show: If $A_k = A_{k+1}$, then (u_k, y_k) is the solution of the optimization problem. One can use this condition as a stopping criterion.

(ii) Superlinear convergence follows from Theorem 9.25. Usually one needs less than 10 iterations.

10. Numerical approximation

Contents

10.1. Galerkin discretization and Céa lemma

Let Y be Hilbert space and $a\colon Y \times Y \to \mathbb{R}$ a continuous, positive definite bilinear and coercive form, i.e. there exist $M > 0$ and $\alpha > 0$ such that

$$\begin{aligned} |a(y,z)| &\leq M \left\|y\right\|_Y \left\|z\right\|_Y && \text{for all } y, z \in Y, \\ a(y,y) &\geq \alpha \left\|y\right\|_Y^2 && \text{for all } y \in Y. \end{aligned} \tag{10.1}$$

Furthermore, let $f \in Y^*$. We consider the following equation

$$y \in Y : a(y,v) = \langle f, v\rangle \quad \text{for all } v \in Y. \tag{10.2}$$

We want to illustrate the discretization concept for the Posisson equation whose variational formulation is given as

$$y \in Y := H_0^1(\Omega) : (\nabla y, \nabla v)_{L^2(\Omega)} = \langle f, v\rangle \quad \text{for all } v \in Y. \tag{E}$$

This is an infinite dimensional problem. With a Galerkin discretization one generates a finite dimensional problem (E_h), whose solution should approximate the solution of (E). Let $Y_h \subset Y$ be a finite dimensional subspace with

$$\dim Y_h = N_h \in \mathbb{N}. \tag{10.3}$$

Later, h will denote a discretization parameter. The spaces Y_h become bigger for $h \to 0$. The Galerkin approximation of (E) in the space Y_h is defined as follows:

$$y_h \in Y_h : a(u_h, v_h) = \langle f, v_h \rangle \quad \text{for all } v_h \in Y_h. \tag{E_h}$$

Theorem 10.1. *There exists a unique solution $u_h \in Y_h$ of (E_h) and we have*

$$\|u_h\|_Y \leq \frac{1}{\alpha} \|f\|_{Y^*}. \tag{10.4}$$

Proof. The space Y_h is a finite dimensional subspace of a Hilbert space and hence a Hilbert space by itself. We conclude with Lax-Milgram. □

For the computation of u_h we choose a basis $\{\phi_1, \dots, \phi_{N_h}\} \subset Y_h$ and consider the corresponding representation

$$u_h = \sum_{j=1}^{N_h} u_j \phi_j. \tag{10.5}$$

Theorem 10.2. *Problem (E_h) is equivalent to the system*

$$AY = F \tag{10.6}$$

with the coefficient vector $Y = (y_1, \dots, y_{N_h})^\top \in \mathbb{R}^{N_h}$ with a positive definite matrix $A \in \mathbb{R}^{N_h \times N_h}$ and $F \in \mathbb{R}^{N_h}$ given by

$$A_{ij} = a(\phi_j, \phi_i), \quad F_i = \langle f, \phi_i \rangle. \tag{10.7}$$

Proof. We have that we can write problem (P_h) equivalently as

$$\sum_{j=1}^{N_h} y_j a(\phi_j, \phi_i) = \langle f, \phi_i \rangle, \quad i = 1, \dots, N_h. \tag{10.8}$$

It remains to verify the positive definiteness of A. For a vector $z \in \mathbb{R}^{N_h}$, $z \neq 0$ and $z_h \in Y_h$ with $z_h = \sum_{j=1}^{N_h} z_j \phi_j$ we have

$$z^\top A z = \sum_{i,j=1}^{N_h} z_j A_{ij} z_j = a(z_h, z_h) \geq \alpha \|z_h\|_Y^2 > 0. \tag{10.9}$$

□

Remark 10.3. The matrix A is called stiffness matrix. If the bilinear form $a(\cdot,\cdot)$ is symmetric, then the stiffness matrix is also symmetric.

Remark 10.4. The space Y_h should be chosen on the one hand such that y_h is a good approximation for y and on the other hand such that the system $AY = F$ can be solved as efficiently as possible.

Lemma 10.5 (Céa lemma)**.** *Let $y \in Y$ be the solution of (E) and $y_h \in Y_h$ solution of (E_h). Then we have the error estimate*

$$\|y - y_h\|_Y \le \frac{M}{\alpha} \inf_{v_h \in Y_h} \|y - v_h\|_Y \,. \tag{10.10}$$

Proof. By definition of y and y_h we have

$$\begin{aligned} a(y,v) &= \langle f, v\rangle \quad \text{for all } v \in Y, & (10.11)\\ a_h(y_h, v_h) &= \langle f, v_h\rangle \quad \text{for all } v_h \in Y_h. & (10.12)\end{aligned}$$

Since $Y_h \subset Y$ we have by subtraction

$$a(y - y_h, v_h) = 0 \quad \text{for all } v_h \in Y_h. \tag{10.13}$$

For $v_h \in Y_h$ we have

$$\begin{aligned} \alpha \|y - y_h\|_Y^2 &\le a(y - y_h, y - y_h)\\ &= a(y - y_h, y - v_h) + a(y - y_h, v_h - y_h)\\ &\le M \|y - y_h\|_Y \|y - v_h\|_Y \,. \end{aligned} \tag{10.14}$$

Then we obtain the assertion

$$\|y - y_h\|_Y \le \frac{M}{\alpha} \|y - v_h\|_Y \quad \text{for all } v_h \in Y_h. \tag{10.15}$$

□

Remark 10.6. The relation (10.13) is called Galerkin orthogonality. If the bilinearform $a(\cdot,\cdot)$ is symmetric, then $a(\cdot,\cdot)$ is a scalar product on Y. Hence, (10.13) implies, that the error $e = y - y_h$ is orthogonal to all functions $v_h \in Y_h$.

The Lemma von Céa means that up to a constant M/α y_h is the best approximation of y in Y_h w.r.t. the norm $\|\cdot\|_Y$, i.e. the error in the numerical solution depends mainly on the fact how well the solution y can be approximated by functions in Y_h.

10.2. Finite element ansatz

To set up a finite dimensional discrete state space Y_h we proceed as follows: One considers a triangulation of the domain in finite number of subdomains (elements) of simple structure, e.g. intervals in one dimension, triangles or quadrahedrials in two dimensions, tetrahedrons or hexahedrons in three dimensions. We consider functions, which are polynoms on each element and which additionally satisfy global regularity conditions (e.g. continuity).

Let $\Omega \subset \mathbb{R}^2$ polygonal domain and $\mathcal{T}_h$ a mesh consisting of open cells $K \subset \Omega$ being either triangles or quadrahedrials, we refer to Ern and Guermon [11, Section 1.3 Meshes: Basic concepts] for definitions.

We use the following notation:

$$\begin{cases} \text{cell parameter:} & h_K = \operatorname{diam}(K), \\ \text{radius of inner circle:} & \rho_K = \max\{\rho > 0 \mid B_\rho \subset \bar{K}\}, \\ \text{discretization parameter:} & h\colon \Omega \to \mathbb{R}, \quad h|_K = h_K, \\ \text{maximimum cell parameter:} & h_{\max} = \max\limits_{K\in\mathcal{T}_h} h_K, \\ \text{set of nodes} & \mathcal{N}(\mathcal{T}_h), \\ \text{set of edges} & \mathcal{E}(\mathcal{T}_h). \end{cases} \tag{10.16}$$

Definition 10.7. *A mesh $\mathcal{T}_h$ is called admissible, if*

(i) $\overline{\Omega} = \bigcup_{K\in\mathcal{T}_h} \overline{K}$,

(ii) If $\overline{K}_1 \cap \overline{K}_2$ consists of exact one point $x \in \Omega$, then x is a node of the zell K_1 and K_2,

(iii) If $\overline{K}_1 \cap \overline{K}_2 \neq \emptyset$ and $\overline{K}_1 \cap \overline{K}_2$ not a single point, then we have $\overline{K}_1 \cap \overline{K}_2 = E$, where $E \subset \partial K_1 \cap \partial K_2$ is an edge of the cell K_1 as well as of the cell K_2.

The conditions in Definition 10.7 can be relaxed.

Definition 10.8. *For a sequence (h_i), $h_i > 0$, we call the family of meshes $\{\mathcal{T}_{h_1}, \mathcal{T}_{h_2}, ...\}$ shape-regular if there exists a $\kappa > 0$, such that*

$$\frac{h_K}{\rho_K} \leq \kappa \quad \textit{for all } K \in \bigcup_i \mathcal{T}_{h_i}. \tag{10.17}$$

Definition 10.9. *The space of linear finite elements is defined as*

$$Y_h := \{v_h \in C(\overline{\Omega}) \mid v_h|_K \in \mathcal{P}_1(K) \textit{ for all } K \in \mathcal{T}_h \textit{ and } v_h|_{\partial\Omega} = 0\}. \tag{10.18}$$

The condition $v_h|_{\partial\Omega} = 0$ is chosen for Dirichlet boundary conditions; if one considers, e.g., a equation with homogeneous Neumann boundary condition that means,

$$\begin{cases} -\Delta y + y = f & \text{in } \Omega, \\ \partial_n y = 0 & \text{on } \partial\Omega, \end{cases} \tag{10.19}$$

the condition is removed from the definition of the finite element space.

Definition 10.10. *A finite element space is called conform (Y-conform), if $Y_h \subset Y$.*

Theorem 10.11. *The space Y_h defined in* (10.18) *is a subspace of $Y = H_0^1(\Omega)$.*

Remark 10.12. The basis of Y_h should be chosen such that the stiffness matrix A_h is sparse, i.e. as many as possible entries are zero. Let $A_h \in \mathbb{R}^{N_h \times N_h}$. The aim is that the number of non-zero entries behaves as $\mathcal{O}(N)$. Usually this property is obtained by the condition that the number of non-zero entries per line is bounded uniform in N, i.e.

$$d_i := \#\{1 \le j \le N \mid A_{ij} \neq 0\}, \quad d_i \le C \tag{10.20}$$

with a constant C independent of N. Therefore, usually the basis functions have local support.

Definition 10.13 (Nodal basis)**.** *For each node $x_i \in \mathcal{N}(\mathcal{T}_h)$, $x_i \notin \partial\Omega$, we consider a basisfunction $\phi_h^i \in Y_h$ with*

$$\phi_h^i(x_j) = \delta_{ij} \tag{10.21}$$

with $\delta_{ij} = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases}$.

Theorem 10.14. *The functions $\{\phi_h^i\}$ are a basis of the space Y_h.*

Proof. See, e.g., Ern and Guermond [11, Prop. 1.1]. □

Every function $v_h \in Y_h$ has the representation

$$v_h(x) = \sum_{x_i \in \mathcal{N}(\mathcal{T}_h), x_i \notin \partial\Omega} v_h(x)\phi_h^i(x). \tag{10.22}$$

10.3. Interpolation with finite elements

The aim of the a priori error analysis is to show estimates of type

$$\|\nabla(y - y_h)\|_{L^2(\Omega)} \leq ch. \tag{10.23}$$

By the Céa-lemma we know that

$$\|\nabla(y - y_h)\|_{L^2(\Omega)} \leq C \inf_{v_h \in Y_h} \|\nabla(y - v_h)\|_{L^2(\Omega)}. \tag{10.24}$$

This rise the question how good functions y can be approximated in the space Y_h. We will construct a interpolation $i_h y$ and derive estimates for the interpolation error $y - i_h y$.

Definition 10.15. *The nodal interpolant for the space Y_h is given by*

$$i_h \colon C(\bar{\Omega}) \to Y_h, \quad v \mapsto \sum_{x_i \in \mathcal{N}(\mathcal{T}_h), x_i \notin \partial\Omega} v(x_i) \cdot \phi_h^i. \tag{10.25}$$

Proposition 10.16. *Let $v \in C(\bar{\Omega})$ with $v|_{\partial\Omega} = 0$. Then we have*

$$i_h v(x_i) = v(x_i) \quad \text{for all } x_i \in \mathcal{N}(\mathcal{T}_h) \tag{10.26}$$

and

$$i_h v_h = v_h \quad \text{for all } v_h \in Y_h. \tag{10.27}$$

This follows directly from the definition of the interpolation oeprator.

Theorem 10.17 (Interpolation estimates)**.** *Let $\{\mathcal{T}_h\}_{h>0}$ a family of shape-regular meshes, and let $i_h \colon Y \cap C(\overline{\Omega}) \to Y_h$ the nodal interpolant defined in (10.25) in the space of linear finite elements Y_h. Then we have for all functions $v \in Y \cap H^2(\Omega)$ and all cells $K \in \mathcal{T}_h$: There exists a $c > 0$ such that*

(i) $\|v - i_h v\|_{L^2(K)} \leq ch_K^2 \|\nabla^2 v\|_{L^2(K)}$,

(i) $\|\nabla(v - i_h v)\|_{L^2(K)} \leq ch_K \|\nabla v\|_{L^2(K)}$,

Proof. See, e.g., Ern and Guermond [11, Sec. 1.5]. □

Corollary 10.18. *Let $\{\mathcal{T}_h\}_{h>0}$ a family of shape-regular meshes and let $i_h \colon Y \cap C(\overline{\Omega}) \to Y_h$ the nodal interpolant. Then we have for all $v \in Y \cap H^2(\Omega)$:*

(i) $\|v - i_h v\|_{L^2(\Omega)} \leq ch^2 \|\nabla^2 v\|_{L^2(\Omega)}$,

(ii) $\|\nabla(v - i_h v)\|_{L^2(\Omega)} \leq ch \|\nabla^2 v\|_{L^2(\Omega)}$,

Proof. See, e.g., Ern and Guermond [11, Sec. 1.5]. □

10.4. A priori error analysis

Let $f \in L^2(\Omega)$ and $y \in Y$ be the variational solution of the Poisson equation

$$(\nabla y, \nabla v)_{L^2(\Omega)} = (f, v)_{L^2(\Omega)} \quad \text{for all } v \in Y \tag{10.28}$$

and satisfy the additional regularity condition $y \in H^2(\Omega)$. Let $u_h \in Y_h$ be the finite-element solution on mesh $\mathcal{T}_h$ from a family of shape-regular triangulations $\{\mathcal{T}_h\}$

$$(\nabla y_h, \nabla v_h)_{L^2(\Omega)} = (f, v_h)_{L^2(\Omega)} \quad \text{for all } v_h \in Y_h. \tag{10.29}$$

Theorem 10.19. *There exists a $c > 0$ such that for all $h > 0$ we have*

$$\|\nabla(y - y_h)\|_{L^2(\Omega)} \leq ch \left\|\nabla^2 y\right\|_{L^2(\Omega)}. \tag{10.30}$$

Proof. The constant c in the Céa-lemma 10.5 is in case of the Poisson equation equal to 1. With Corollary 10.18 we have

$$\|\nabla(y - y_h)\|_{L^2(\Omega)} \leq \|\nabla(y - i_h y)\|_{L^2(\Omega)} \leq ch \left\|\nabla^2 y\right\|_{L^2(\Omega)}. \tag{10.31}$$

□

Theorem 10.20 (Aubin-Nitsche)**.** *We have for $c > 0$ that*

$$\|y - y_h\|_{L^2(\Omega)} \leq ch^2 \left\|\nabla^2 y\right\|_{L^2(\Omega)}. \tag{10.32}$$

Proof. Let $e := y - y_h$ with $\|e\|_{L^2(\Omega)} > 0$. Let $z \in Y$ be the solution of

$$\begin{cases} -\Delta z = \dfrac{1}{\|e\|_{L^2(\Omega)}} e & \text{in } \Omega, \\ \quad z = 0 & \text{on } \partial\Omega \end{cases} \tag{10.33}$$

and the corresponding variational formulation

$$(\nabla v, \nabla z)_{L^2(\Omega)} = \frac{1}{\|e\|_{L^2(\Omega)}} (e, v)_{L^2(\Omega)} \quad \forall v \in Y. \tag{10.34}$$

Since Ω is polygonal and convex, we have by Theorem 4.63[1] $z \in Y \cap H^2(\Omega)$ and for some $c > 0$

$$\|z\|_{H^2(\Omega)} \leq c \left\| \frac{1}{\|e\|_{L^2(\Omega)}} e \right\|_{L^2(\Omega)} . \tag{10.35}$$

We obtain

$$\|e\|_{L^2(\Omega)} = \frac{1}{\|e\|_{L^2(\Omega)}} (e, e)_{L^2(\Omega)} = (\nabla e, \nabla z)_{L^2(\Omega)}. \tag{10.36}$$

Using the Galerkin orthogonality we have

$$\|e\|_{L^2(\Omega)} = (\nabla e, \nabla (z - i_h z))_{L^2(\Omega)}. \tag{10.37}$$

With Cauchy-Schwarz inequality, Corollary 10.18, and (10.35) we further have

$$\|e\|_{L^2(\Omega)} = \|\nabla e\|_{L^2(\Omega)} \|\nabla (z - i_h z)\|_{L^2(\Omega)} \leq ch \left\|\nabla^2 y\right\|_{L^2(\Omega)} h \left\|\nabla^2 z\right\|_{L^2(\Omega)} \leq ch^2 \left\|\nabla^2 y\right\|_{L^2(\Omega)} . \tag{10.38}$$

□

[1] The theorem holds also on convex and polygonal dpomains.

10.5. Finite element methods for optimal control problems

Let U Hilbert space, $U_{\mathrm{ad}} \subset U$ a non-empty, convex, and closed set, $B\colon U \to L^2(\Omega)$ a linear, bounded operator, $y_{\mathrm{d}} \in L^2(\Omega)$ and $\alpha > 0$.

$$\begin{cases} \min\limits_{(u,y)\in U_{\mathrm{ad}}\times Y} J(u,y) := \tfrac{1}{2}\left\|y - y_{\mathrm{d}}\right\|^2_{L^2(\Omega)} + \dfrac{\alpha}{2}\left\|u\right\|^2_U , \\ \text{subject to } -\Delta y = Bu \quad \text{in } \Omega, \\ \qquad\qquad\quad\; y = 0 \quad \text{on } \partial\Omega. \end{cases} \tag{10.39}$$

We set $F(u) := J(u, y[u])$ and obtain the reduced optimal control problem

$$\min F(u), \qquad u \in U_{\mathrm{ad}}. \tag{10.40}$$

We know that there exists a unique solution $\bar{u} \in U_{\mathrm{ad}}$ of (10.39), which is charaterized by the optimality codition

$$F'(\bar{u})(\delta u - \bar{u}) \geq 0 \quad \text{for all } \delta u \in U_{\mathrm{ad}}. \tag{10.41}$$

The discrete problem. To discretize the problem (10.39) we consider a finite dimensional space $Y_h \subset Y$ as the space of linear finite elements. Furthermore, we consider a finite-dimensional subspace $U_h \subset U$ and the discrete admissible set $U_{ad,h} := U_{ad} \cap U_h$.

Let $U_{ad,h}$ be a nonempty set. The discrete optimal control problem is given as

$$\begin{cases} \min\limits_{(u_h,y_h)\in U_{ad,h}\times Y_h} J(u_h,y_h) := \tfrac{1}{2}\left\|y_h - y_d\right\|^2_{L^2(\Omega)} + \dfrac{\alpha}{2}\left\|u_h\right\|^2_U , \\ \text{subject to } (\nabla y_h, \nabla v)_{L^2(\Omega)} = (Bu_h, v)_{L^2(\Omega)} \quad \text{for all } v \in Y_h. \end{cases} \tag{10.42}$$

By the Lax-Milgram theorem 4.59 the discrete state equation has a unique solution and there exists a linear and bounded discrete-control-to-state operator $y_h\colon U \to Y_h$. With the reduced discrete cost functional

$$F_h\colon U \to \mathbb{R}, \quad F_h(u) = J(u, y_h[u]) \tag{10.43}$$

the discrete optimal control problem is given by

$$\min F_h(u_h), \quad u_h \in U_{ad,h}. \tag{10.44}$$

The problem has a unique solution $u_h \in U_{ad,h}$ (this can be shown by similar arguments as in Theorem 5.3) characterized by

$$F_h'(\bar{u}_h)(\delta u_h - \bar{u}_h) \geq 0 \quad \text{for all } \delta u_h \in U_{ad,h} \tag{10.45}$$

Derivatives. For the continuous reduced cost function $F\colon U \to Y$ we know the representation of the derivatives (cf. Section 6.5):

$$F'(u)(\delta u) = L_u(u, y, p)(\delta u), \tag{10.46}$$

with $L(u, y, p) := J(u, y) + (Bu, p)_{L^2(\Omega)} - (\nabla y, \nabla p)_{L^2(\Omega)}$, where $p = p[u] \in Y$ is the solution of the adjoint equation

$$(\nabla v, \nabla p)_{L^2(\Omega)} = (y - y_d, v)_{L^2(\Omega)} \quad \text{for all } v \in Y. \tag{10.47}$$

That means we have the representation

$$F'(u)(\delta u) = \alpha(u, \delta u)_{L^2(\Omega)} + (B\delta u, p)_{L^2(\Omega)} = (\alpha u + B^* p, \delta u)_{L^2(\Omega)}. \tag{10.48}$$

With $y_h[u] \in Y_h$ solving the discrete state equation and $p_h := p_h[u] \in Y_h$ solving the discrete adjoint equation

$$(\nabla v_h, \nabla p_h)_{L^2(\Omega)} = (y_h[u] - y_d, v_h)_{L^2(\Omega)} \quad \text{for all } v_h \in Y_h \tag{10.49}$$

we obtain analogously

$$F_h'(u)(\delta u) = (\alpha u + B^* p_h, \delta u)_{L^2(\Omega)}. \tag{10.50}$$

10.6. A priori error estimates (no control constraints)

Lemma 10.21. *The second derivatives $F''(u)$ and $F_h''(u)$ do not depend on u and are positiv definit, i.e. we have*

$$F''(u)(v,v) \geq \alpha \|v\|_U^2 \qquad \textit{for all } u \in U, \tag{10.51}$$

$$F_h''(u)(v,v) \geq \alpha \|v\|_U^2 \qquad \textit{for all } u \in U. \tag{10.52}$$

Proof. The functions F and F_h are quadratic in u. Consequently the second derivatives do not depend on u. We have

$$F'(u)(\delta u) = (Su - y_d, S\delta u)_{L^2(\Omega)} + \alpha(u, \delta u)_U \tag{10.53}$$

and

$$F''(u)(\delta u, \tau u) = (S\tau u, S\delta u)_{L^2(\Omega)} + \alpha(\tau u, \delta u)_U \tag{10.54}$$

Hence we obtain

$$F''(u)(\delta u, \delta u) = \|S\delta u\|_{L^2(\Omega)}^2 + \alpha \|\delta u\|_U^2 \geq \alpha \|\delta u\|_U^2, \tag{10.55}$$

and analogously

$$F_h''(u)(\delta u, \delta u) = \|S_h\delta u\|_{L^2(\Omega)}^2 + \alpha \|\delta u\|_U^2 \geq \alpha \|\delta u\|_U^2. \tag{10.56}$$

□

Lemma 10.22. *There exists a constant $L > 0$ independent of h such that*

$$|F_h'(u)(\delta u) - F_h'(v)(\delta u)| \leq L \|u - v\|_U \|\delta u\|_U \quad \textit{for all } u, v, \delta u \in U. \tag{10.57}$$

Proof. We have for $w \in U$

$$F_h'(w)(\delta w) = (\alpha w - B^* p_h[w], \delta w)_{L^2(\Omega)}. \tag{10.58}$$

We obtain for $u, v \in U$

$$\begin{aligned} |F_h'(u)(\delta u) - F_h'(v)(\delta u)| &= |(\alpha(u - v, \delta u) + (B^*(p_h[u] - p_h[v]), \delta u)| \\ &\leq \alpha \|u - v\|_U \|\delta u\|_U + \|B^*\| \|p_h[u] - p_h[v]\|_{L^2(\Omega)} \|\delta u\|_U. \end{aligned} \tag{10.59}$$

It remains to estimate the difference $\|p_h[u] - p_h[v]\|_{L^2(\Omega)}$. Let $w_h = y_h[u] - y_h[v]$, i.e.

$$(\nabla w_h, \nabla v_h)_{L^2(\Omega)} = (B(u - v), v_h)_{L^2(\Omega)} \quad \text{for all } v_h \in Y_h. \tag{10.60}$$

With $v_h = w_h$ we obtain

$$\|\nabla w_h\|_{L^2(\Omega)}^2 = (B(u-v), w_h) \leq \|B\| \|u - v\|_U \|w_h\|_{L^2(\Omega)} \leq c_p \|B\| \|u - v\|_U \|\nabla w_h\|_{L^2(\Omega)} \tag{10.61}$$

with the Poincare constante c_P. Hence we have

$$\|w_h\|_{L^2(\Omega)} \leq c_p \|\nabla w_h\|_{L^2(\Omega)} \leq c_p^2 \|B\| \|u-v\|_U . \tag{10.62}$$

For $q_h := p_h[u] - p_h[v]$ we obtain using

$$\|\nabla q_h\|_{L^2(\Omega)}^2 \leq c_p \|w_h\|_{L^2(\Omega)} \|\nabla q_h\|_{L^2(\Omega)} \tag{10.63}$$

that

$$\|q_h\|_{L^2(\Omega)} \leq c_p^2 \|w_h\|_{L^2(\Omega)} \leq c_p^4 \|B\| \|u-v\|_U . \tag{10.64}$$

□

Lemma 10.23. *Let Ω polygonal and convex*[2] *Then we have*

$$|F'(u)(\delta u) - F_h'(u)(\delta u)| \leq ch^2(\|u\|_U + \|y_d\|_{L^2(\Omega)}) \|\delta u\|_U \tag{10.65}$$

with $c > 0$ independent of h.

Proof. We have

$$|F'(u)(\delta u) - F_h'(u)\delta u)| = |(B^*(p[u] - p_h[u]), \delta u)| \leq \|B^*\| \|p[u] - p_h[u]\|_{L^2(\Omega)} \|\delta u\|_U . \tag{10.66}$$

Let

$$d_y := y[u] - y_h[u], \quad d_p := p[u] - p_h[u]. \tag{10.67}$$

By Theorem 10.20 we have

$$\|d_y\|_{L^2(\Omega)} \leq ch^2 \|\nabla y[u]\|_{L^2(\Omega)} . \tag{10.68}$$

With the H^2-regularity we have

$$\|d_y\|_{L^2(\Omega)} \leq ch^2 \|B\| \|u\|_U . \tag{10.69}$$

The dual solution $p[u] \in Y$ and $p_h[u] \in Y_h$ are given by

$$\begin{aligned} (\nabla v, \nabla p[u]) &= (y[u] - y_d, v) \qquad \text{for all } v \in Y, \\ (\nabla v_h, \nabla p_h[u]) &= (y_h[u] - y_d, v_h) \qquad \text{for all } v_h \in Y_h. \end{aligned} \tag{10.70}$$

We further consider the Galerkin projection $\tilde{p}_h[u] \in Y_h$ defined by

$$(\nabla v_h, \nabla \tilde{p}_h[u]) = (y[u] - y_d, v_h) \quad \text{for all } v_h \in Y_h. \tag{10.71}$$

[2] Theorem 4.63 holds also on polygonal and convex domains.

For the error $p[u] - \tilde{p}_h[u]$ we have by Theorem 10.20

$$\|p[u] - \tilde{p}_h[u]\|_{L^2(\Omega)} \leq ch^2 \left\|\nabla^2 p[u]\right\|_{L^2(\Omega)} \leq ch^2 \|y[u] - y_d\|_{L^2(\Omega)} \leq ch^2(\|u\|_U + \|y_d\|_{L^2(\Omega)}). \tag{10.72}$$

For the difference $\tilde{d}_p = \tilde{p}_h[u] - p_h[u] \in Y_h$ we have

$$(\nabla \tilde{d}_p, \nabla v_h) = (d_y, v_h) \quad \text{for all } v_h \in Y_h. \tag{10.73}$$

With $v_h = \tilde{d}_p$ and obtain as in the proof of Lemma 10.22

$$\left\|\tilde{d}_p\right\|_{L^2(\Omega)} \leq c \|d_y\|_{L^2(\Omega)}. \tag{10.74}$$

We obtain

$$\|d_p\|_{L^2(\Omega)} \leq ch^2(\|u\|_U + \|y_d\|_{L^2(\Omega)}). \tag{10.75}$$

□

Theorem 10.24. *Let $U = U_h$ be finite dimensional and $\Omega \subset \mathbb{R}^n$, $n = 2, 3$, polygonal and convex. Let $\bar{u} \in U_{ad}$ be the solution of the continuous problem (10.39) and $\bar{u}_h \in U_{ad,h} = U_{ad}$ the solution of the discrete problem (10.42). Then there exists a $c > 0$ such that*

$$\|\bar{u} - \bar{u}_h\|_U \leq ch^2(\|\bar{u}\|_U + \|y_d\|_{L^2(\Omega)}). \tag{10.76}$$

Proof of Theorem 10.24. By Lemma 10.21 we have for all $v \in U$

$$\alpha \|\bar{u} - \bar{u}_h\|_U^2 \leq F_h''(v)(\bar{u} - \bar{u}_h, \bar{u} - \bar{u}_h). \tag{10.77}$$

Since $F_h''(v)$ does not depend on v, we have

$$\alpha \|\bar{u} - \bar{u}_h\|_U^2 \leq F_h''(v)(\bar{u} - \bar{u}_h, \bar{u} - \bar{u}_h) = F_h'(\bar{u})(\bar{u} - \bar{u}_h) - F_h'(\bar{u}_h)(\bar{u} - \bar{u}_h). \tag{10.78}$$

Using the optimality conditions for $\bar{u}$ and $\bar{u}_h$ we have

$$-F_h'(\bar{u}_h)(\bar{u} - \bar{u}_h) \leq 0 \leq -F'(\bar{u})(\bar{u} - \bar{u}_h). \tag{10.79}$$

We obtain with Lemma 10.23

$$\alpha \|\bar{u} - \bar{u}_h\|_U^2 \leq F_h'(\bar{u})(\bar{u} - \bar{u}_h) - F'(\bar{u})(\bar{u} - \bar{u}_h) \leq ch^2(\|\bar{u}\|_U + \|y_d\|_{L^2(\Omega)}) \|\bar{u} - \bar{u}_h\|_U. \tag{10.80}$$

□

Theorem 10.25. *Let $U = L^2(\Omega)$ and $U_{ad} = U$. Furthermore, let $B = \mathrm{id}$, $U_h = Y_h$ be the space of continuous linear finite elements and let $\Omega \subset \mathbb{R}^2$ be polygonal and convex. Let $\bar{u} \in U_{ad}$ the solution of the continuous problem (10.39) and $\bar{u}_h \in U_h$ the solution of the discrete problem (10.42). Then we have*

$$\|\bar{u} - \bar{u}_h\|_{L^2(\Omega)} \leq ch^2(\|\bar{u}\|_{H^2(\Omega)} + \|y_d\|_{L^2(\Omega)}). \tag{10.81}$$

Proof. From the optimality system from (10.39) we have

$$\alpha\bar{u} + \bar{p} = 0. \tag{10.82}$$

By the H^2-regularity, see Theorem 4.63, we have $\bar{p} \in H_0^1(\Omega) \cap H^2(\Omega)$. For the interpolation error we have

$$\|\bar{u} - i_h\bar{u}\|_{L^2(\Omega)} \leq ch^2 \left\|\nabla^2\bar{u}\right\|_{L^2(\Omega)}. \tag{10.83}$$

Following the proof of Theorem 10.24 we have

$$\alpha \|i_h\bar{u} - \bar{u}_h\|_{L^2(\Omega)} \leq F_h''(v)(i_h\bar{u}-\bar{u}_h, i_h\bar{u}-\bar{u}_h) = F_h'(i_h\bar{u})(i_h\bar{u}-\bar{u}_h) - F_h'(\bar{u}_h)(i_h\bar{u}-\bar{u}_h). \tag{10.84}$$

With optimality conditions for $\bar{u}$ and $\bar{u}_h$ we have

$$F_h'(\bar{u}_h)(i_h\bar{u} - \bar{u}_h) = 0 = F'(\bar{u})(i_h\bar{u} - \bar{u}_h). \tag{10.85}$$

We have

$$\begin{aligned}\alpha \|i_h\bar{u} - \bar{u}_h\|_{L^2(\Omega)} &\leq F_h'(i_h\bar{u}_h)(i_h\bar{u} - \bar{u}_h) - F_h'(\bar{u})(i_h\bar{u} - \bar{u}_h)\\ &= F_h'(i_h\bar{u})(i_h\bar{u} - \bar{u}_h) - F_h'(\bar{u})(i_h\bar{u} - \bar{u}_h)\\ &\quad + F_h'(\bar{u})(i_h\bar{u} - \bar{u}_h) - F'(\bar{u})(i_h\bar{u} - \bar{u}_h)\end{aligned} \tag{10.86}$$

We estimate the first term by Lemma 10.23:

$$F_h'(i_h\bar{u})(i_h\bar{u} - \bar{u}_h) - F_h'(\bar{u})(i_h\bar{u} - \bar{u}_h) \leq L \|i_h\bar{u} - \bar{u}\|_{L^2(\Omega)} \|i_h\bar{u} - \bar{u}_h\|_{L^2(\Omega)} \tag{10.87}$$

For the second term we use Lemma 10.22

$$F_h'(\bar{u})(i_h\bar{u} - \bar{u}_h) - F'(\bar{u})(i_h\bar{u} - \bar{u}_h) \leq ch^2(\|\bar{u}\|_{L^2(\Omega)} + \|y_d\|_{L^2(\Omega)}) \|i_h\bar{u} - \bar{u}_h\|_{L^2(\Omega)}. \tag{10.88}$$

Consequently, we have

$$\alpha \|i_h\bar{u} - \bar{u}_h\|_{L^2(\Omega)} \leq L \|i_h\bar{u} - \bar{u}\|_{L^2(\Omega)} + ch^2(\|\bar{u}\|_{L^2(\Omega)} + \|y_d\|_{L^2(\Omega)}. \tag{10.89}$$

Together with the interpolation estimate (10.18) we conclude. □

10.7. A priori error estimates (with control constraints)

Finally we consider the case that inequality constraints imposed on the controls, i.e. we have $U = L^2(\Omega)$, $B = \mathrm{id}$ with

$$U_{\mathrm{ad}} = \{u \in U \mid u_m \leq u(x) \leq u_M \text{ a.a. in } \Omega\} \tag{10.90}$$

with $u_m, u_M \in \mathbb{R}$, $u_m < u_M$. For the optimal control $\bar{u}$ we have

$$\bar{u} = P_{[u_m,u_M]}\left(-\frac{1}{\alpha}\bar{p}\right) \tag{10.91}$$

with the pointwise projection $P_{[u_m,u_M]}$, see Theorem 7.6(iv).

For the discretization of the control space U we consider on the triangulation $\mathcal{T}_h$ cellwise constant functions; we denote the space by $U_h = U_h^0$.

The L^2-projection $\pi\colon U \to U_h$ is defined by

$$(\pi_h u, v_h)_{L^2(\Omega)} = (u, v_h)_{L^2(\Omega)} \quad \forall v_h \in U_h. \tag{10.92}$$

By direct calculation one can show that for U_h that

$$\pi_h u|_K = \frac{1}{|K|}\int_K u(x)\mathrm{d}x \qquad \text{for all } K \in \mathcal{T}_h. \tag{10.93}$$

Lemma 10.26. *Let $u \in H^1(\Omega)$. Then we have*

$$\|u - \pi_h u\|_{L^2(\Omega)} \leq ch\,\|\nabla u\|_{L^2(\Omega)}\,. \tag{10.94}$$

Proof. The proof uses standard arguments based on the Bramble-Hilbert lemma; see Ern and Guermond [11, Prop. 1.135]. □

Theorem 10.27. *Let $U = L^2(\Omega)$, $B = \mathrm{id}$,*

$$U_{ad} = \{u \in U \mid u_m \leq u(x) \leq u_M, \quad \text{a.e. in } \Omega\}, \tag{10.95}$$

with $U_h = U_h^0$ and $\Omega \in \mathbb{R}^2$ a polygonal and convex. Let $\bar{u} \in U_{ad}$ the solution of the continuous problem (10.39) and $\bar{u}_h \in U_{ad,h}$ the solution of the discrete solution. Then we have

$$\|\bar{u} - \bar{u}_h\|_{L^2(\Omega)} \leq ch(\|\bar{u}\|_{H^1(\Omega)} + \|y_d\|_{L^2(\Omega)}). \tag{10.96}$$

Proof. From the representation (10.93) we have for every $u \in U$, that $\pi_h u \in U_{ad,h}$. By Lemma 10.26 it is sufficient to prove an estimate for $\pi_h \bar{u} - \bar{u}_h$. As in the proof of Theorem 10.25 we have

$$\begin{aligned} \alpha \left\| \pi_h \bar{u} - \bar{u}_h \right\|_{L^2(\Omega)}^2 &\leq F_h''(v)(\pi_h \bar{u} - \bar{u}_h, \pi_h \bar{u} - \bar{u}_h) \\ &= F_h'(\pi_h \bar{u})(\pi_h \bar{u} - \bar{u}_h) - F_h'(\bar{u}_h)(\pi_h \bar{u} - \bar{u}_h). \end{aligned} \tag{10.97}$$

With the optimality conditions for $\bar{u}$ and $\bar{u}_h$ we have

$$-F_h'(\bar{u}_h)(\pi_h \bar{u} - \bar{u}_h) \leq 0 \leq -F_h'(\bar{u})(\bar{u} - \bar{u}_h) \leq -F'(\bar{u})(\pi_h \bar{u} - \bar{u}_h) - F'(\bar{u})(\bar{u} - \pi_h \bar{u}_h). \tag{10.98}$$

We obtain

$$\begin{aligned} \alpha \left\| \pi \bar{u} - \bar{u}_h \right\|_{L^2(\Omega)}^2 &\leq F_h'(\pi_h \bar{u})(\pi_h \bar{u} - \bar{u}_h) - F'(\bar{u})(\pi_h \bar{u} - \bar{u}_h) - F'(\bar{u})(\bar{u} - \pi_h \bar{u}) \\ &= (F'(\pi_h \bar{u})(\pi_h \bar{u} - \bar{u}_h) - F_h'(\bar{u})(\pi_h \bar{u} - \bar{u}_h)) \\ &\quad + (F_h'(\bar{u})(\pi_h \bar{u} - \bar{u}_h) - F'(\bar{u})(\pi_h \bar{u} - \bar{u}_h)) - F'(\bar{u})(\bar{u} - \pi_h \bar{u}). \end{aligned} \tag{10.99}$$

For the last term we have

$$\begin{aligned} -F'(\bar{u})(\bar{u} - \pi_h \bar{u}) &= -(\alpha \bar{u} - \bar{p}, \bar{u} - \pi_h \bar{u})_{L^2(\Omega)} \\ &= -((\alpha \bar{u} - \bar{p})_{L^2(\Omega)} - \pi_h(\alpha \bar{u} - \bar{p}), \bar{u} - \pi_h \bar{u})_{L^2(\Omega)} \\ &\leq ch^2 \left\| \nabla(\alpha \bar{u} - \bar{p}) \right\|_{L^2(\Omega)} \left\| \nabla \bar{u} \right\|_{L^2(\Omega)}. \end{aligned} \tag{10.100}$$

The first two terms are estimated as in the Theorem 10.24 by Lemma 10.23 and Lemma 10.22. We have

$$F_h'(\pi_h \bar{u})(\pi_h \bar{u} - \bar{u}_h) - F'(\bar{u})(\pi_h \bar{u} - \bar{u}_h) \leq L \left\| \pi_h \bar{u} - \bar{u} \right\|_{L^2(\Omega)} \left\| \pi_h \bar{u} - \bar{u}_h \right\|_{L^2(\Omega)} \tag{10.101}$$

and

$$F_h'(\pi_h \bar{u})(\pi_h \bar{u} - \bar{u}_h) - F'(\bar{u})(\pi_h \bar{u} - \bar{u}_h) \leq ch^2(\left\| \bar{u} \right\|_{L^2(\Omega)} + \left\| y_{\mathrm{d}} \right\|_{L^2(\Omega)}) \left\| \pi_h \bar{u} - \bar{u}_h \right\|_{L^2(\Omega)}. \tag{10.102}$$

Summarizing we have

$$\begin{aligned} \alpha \left\| \pi_h \bar{u} - \bar{u}_h \right\|_{L^2(\Omega)}^2 &\leq L \left\| \pi_h \bar{u} - \bar{u} \right\|_{L^2(\Omega)} \left\| \pi_h \bar{u} - \bar{u}_h \right\|_{L^2(\Omega)} \\ &\quad + ch^2(\left\| \bar{u} \right\|_{L^2(\Omega)} + \left\| y_{\mathrm{d}} \right\|_{L^2(\Omega)}) \left\| \pi_h \bar{u} - \bar{u}_h \right\|_{L^2(\Omega)} \\ &\quad + ch^2 \left\| \nabla(\alpha \bar{u} - \bar{p}) \right\|_{L^2(\Omega)} \left\| \nabla \bar{u} \right\|_{L^2(\Omega)}. \end{aligned} \tag{10.103}$$

With Young's inequality we have for the first two terms

$$L \left\| \pi_h \bar{u} - \bar{u} \right\|_{L^2(\Omega)} \left\| \pi_h \bar{u} - \bar{u}_h \right\|_{L^2(\Omega)} \leq c \left\| \pi_h \bar{u} - \bar{u} \right\|_{L^2(\Omega)}^2 + \frac{\alpha}{4} \left\| \pi_h \bar{u} - \bar{u}_h \right\|_{L^2(\Omega)}^2 \tag{10.104}$$

and

$$ch^2(\|\bar{u}\|_{L^2(\Omega)}+\|y_\mathrm{d}\|_{L^2(\Omega)})\,\|\pi_h\bar{u}-\bar{u}_h\|_{L^2(\Omega)} \le ch^4(\|\bar{u}\|_{L^2(\Omega)}+\|y_\mathrm{d}\|_{L^2(\Omega)})^2+\frac{\alpha}{4}\,\|\pi_h\bar{u}-\bar{u}_h\|^2_{L^2(\Omega)}\,. \tag{10.105}$$

Hence

$$\begin{aligned}\frac{\alpha}{2}\,\|\pi_h\bar{u}-\bar{u}_h\|^2_{L^2(\Omega)} &\le c\,\|\pi_h\bar{u}-\bar{u}\|^2_{L^2(\Omega)}+ch^4(\|\bar{u}\|_{L^2(\Omega)}+\|y_\mathrm{d}\|_{L^2(\Omega)})^2\\ &\quad+ch^2\,\|\nabla(\alpha\bar{u}-\bar{p})\|_{L^2(\Omega)}\,\|\nabla\bar{u}\|_{L^2(\Omega)}\,.\end{aligned} \tag{10.106}$$

Furthermore, we have

$$\begin{aligned}\|\nabla(\alpha\bar{u}-\bar{p})\|_{L^2(\Omega)} &\le C\,\|\nabla\bar{u}\|_{L^2(\Omega)}+\|\nabla(\bar{y}-y_d)\|_{L^2(\Omega)}\\ &\le C\,\|\nabla\bar{u}\|_{L^2(\Omega)}+\|\nabla\bar{y}\|_{L^2(\Omega)}+\|\nabla y_d\|_{L^2(\Omega)}\\ &\le C\,\|\nabla\bar{u}\|_{L^2(\Omega)}+\|\bar{u}\|_{L^2(\Omega)}+\|\nabla y_d\|_{L^2(\Omega)}\end{aligned} \tag{10.107}$$

This implies the statement. □

Part III.

Theory II

11. Optimal control of parabolic equations

Contents

Throughout this section assume that $\Omega \subset \mathbb{R}^n$ is an open and bounded subset, we set $Q := (0,T) \times \Omega$ and $\Sigma := (0,T) \times \Omega$, $T > 0$. We will analyze problems as the following one

$$\begin{cases} \min \quad J(u,y) := \dfrac{1}{2} \|y(T) - y_{\mathrm{d}}\|^2_{L^2(\Omega)} + \dfrac{\alpha}{2} \|u\|^2_{L^2(Q)}, \quad \alpha > 0 \\ \text{subject to} \\ y_t - \Delta y = u \quad \text{in } Q, \quad y = 0 \quad \text{on } \Sigma, \\ \quad y(0,\cdot) = y_0 \quad \text{on } \Omega, \quad u_m \le u \le u_M \quad \text{a.e. in } Q \end{cases} \tag{11.1}$$

with $u_m, u_M \in L^2(Q)$, $u_m < u_M$, $y_0, y_d \in L^2(\Omega)$. At first we will study the underlying parabolic equations.

11.1. Bochner spaces

Let X be separable Banach space. We consider mappings

$$t \in [0,T] \mapsto y(t) \in X. \tag{11.2}$$

We extend the notion of measurability, integrability, and weak differentiability.

Definition 11.1.

(i) *A function* $s\colon [0,T] \to X$ *is called* simple *if it has the form*

$$s(t) = \sum_{i=1}^{m} \boldsymbol{1}_{E_i}(t) y_i,$$

with Lebesgue measurable sets $E_i \subset [0,T]$ *and* $y_i \in X$.

(ii) *A function* $f\colon [0,T] \to X$ *is called* strongly measurable *if there exists simple functions* $s_k\colon [0,T] \to X$ *such that*

$$s_k(t) \to f(t) \quad \textit{for almost all } t \in [0,T]. \tag{11.3}$$

Definition 11.2 (Bochner integral)**.**

(i) *For a simple function* $s(t) = \sum_{i=1}^{m} \mathbf{1}_{E_i}(t) y_i$ *we define the integral*

$$\int_0^T s(t)\mathrm{d}t = \sum_{i=1}^{m} y_i \mu(E_i). \tag{11.4}$$

(ii) *We say that* $f\colon [0,T] \to X$ *is Bochner-integrable if there exists a subsequence* (s_k) *of simple functions such that* $s_k(t) \to f(t)$ *a.e. and*

$$\int_0^T \|s_k(t) - f(t)\|_X \,\mathrm{d}t \to 0 \quad \textit{as } k \to \infty. \tag{11.5}$$

(iii) *If* f *is Bochner integrable we define*

$$\int_0^T f(t)\mathrm{d}t := \lim_{k\to 0} \int_0^T s_k(t)\mathrm{d}t. \tag{11.6}$$

Theorem 11.3. *A strongly measurable function* $f\colon [0,T] \to X$ *is Bochner integrable if and only if* $t \mapsto \|f(t)\|_X$ *is Lebesgue integrable. In this case*

$$\left\| \int_0^T f(t)\mathrm{d}t \right\|_X \le \int_0^T \|f(t)\|_X \,\mathrm{d}t \tag{11.7}$$

and for all $u^* \in X^*$ *the function* $t \mapsto \langle u^*, f(t)\rangle_X$ *is integrable with*

$$\left\langle u^*, \int_0^T f(t)\mathrm{d}t \right\rangle_X = \int_0^T \langle u^*, f(t)\rangle_X \mathrm{d}t. \tag{11.8}$$

Definition 11.4. *Let X be a separable Banach space. We define the norms*

$$\|y\|_{L^p(0,T;X)} := \left(\int_0^T \|y(t)\|_X^p \,\mathrm{d}t\right)^{1/p} \qquad \textit{for } 1 \leq p < \infty, \tag{11.9}$$

$$\|y\|_{L^\infty(0,T;X)} := \textit{ess sup}_{t\in[0,T]} \|y(t)\|_X \quad \textit{for } p = \infty, \tag{11.10}$$

and for $1 \leq p \leq \infty$ the space

$$L^P(0,T;X) := \left\{ y\colon [0,T] \to X \textit{ strongly measurable } \;:\; \|y\|_{L^p(0,T;X)} < \infty \right\}. \tag{11.11}$$

Definition 11.5 (Weak time derivative)**.** *Let $y \in L^1(0,T;X)$. We say that $v \in L^1(0,T;X)$ is the weak derivative of y, written $y_t = v$, if*

$$\int_0^T \varphi'(t)y(t)\mathrm{d}t = -\int_0^T \varphi(t)v(t)\mathrm{d}t \quad \forall \varphi \in C_c^\infty(0,T). \tag{11.12}$$

Lemma 11.6. *For any $y \in L^p(0,T;X)$, $1 \leq p < \infty$, there is a sequence (s_k) of simple functions with $s_k \to y$ a.e. and $s_k \to y \in L^p(0,T;X)$. Moreover functions of the form*

$$\sum_{i=1}^m \varphi_i(t)y_i, \quad \varphi_i \in C_c^\infty(0,T),\; y_i \in X \tag{11.13}$$

are dense in $L^p(0,T;X)$ for $1 \leq p < \infty$. In particular, $C_c^\infty(0,T;X)$ as well as $C^k([0,T];X)$ are dense in $L^p(0,T;X)$ for $1 \leq p < \infty$, $k \in \mathbb{N}_0$.

Theorem 11.7. *Let X be separable Banach space. Then for $1 \leq p \leq \infty$ the spaces $L^p(0,T;X)$ are Banach spaces.*

For $1 \leq p < \infty$ the dual space of $L^p(0,T;X)$ can isometrically be identified with $L^q(0,T;X^*)$, $\frac{1}{p} + \frac{1}{q} = 1$, by means of the pairing

$$\langle v, y\rangle_{L^q(0,T;X^*),L^p(0,T;X)} = \int_0^T \langle v(t), y(t)\rangle_X \mathrm{d}t. \tag{11.14}$$

If H is a separable Hilbert space then $L^2(0,T;H)$ is a Hilbert space with inner product

$$(y,v)_{L^2(0,T;H)} := \int_0^T (y(t), v(t))_H \mathrm{d}t. \tag{11.15}$$

11.2. Gelfand triple

Let H be a Hilbert space identified by Riesz

$$R^{-1}\colon H \to H^* \tag{11.16}$$

with its dual, we write then $H^* \equiv H$, whose scalar product is denoted by $(\cdot,\cdot)_H$. Let the Hilbert space V, with duality product denoted by $\langle \cdot,\cdot\rangle_V$, be densely and continuously embedded in H: we write $V \overset{d}{\hookrightarrow} H$ and denote by J the (continuous) injection; there exists $c_V > 0$ such that

$$\|Ju\|_H \leq c_V \|u\|_V, \quad \forall v \in V. \tag{11.17}$$

The adjoint mapping $J^*\colon H \to V^*$, since $H^* \equiv H$, is defined by

$$\langle J^* h, v\rangle_V = \langle h, Jv\rangle_H = (h, Jv)_H, \quad \forall h \in H,\ v \in V \ \ (= (Rh, Jv)_H). \tag{11.18}$$

It may be interpreted as the restriction of elements of H (seen as linear forms over H) to V.

Since V is a dense subset of H, J^* is injective (cf. Example 5.8) and we may therefore interpret it as an inclusion operator from H into V^*. The orthogonal of the range of J^* is the set of $v \in V$ such that

$$0 = \langle J^* h, v\rangle_V = (h, Jv)_H, \quad \forall h \in H. \tag{11.19}$$

Taking $h = Jv$ we deduce that $v = 0$, i.e., the inclusion $H \subset V^*$ is dense. We obtain the Gelfand triple

$$V \overset{d}{\hookrightarrow} H \equiv H^* \overset{d}{\hookrightarrow} V^*. \tag{11.20}$$

By (11.18), $J^*J\colon V \to V^*$ (which can be viewed as the canonical injection of V into V^*) satisfies

$$\langle J^* Jv, v'\rangle_{V^*,V} = (Jv, Jv')_H, \quad \forall v, v' \in V. \tag{11.21}$$

(not omitting the Riesz projection, $\langle J^* RJv, v\rangle_{V^*,V} = \langle RJv, Jv\rangle_{H^*,H} = (Jv, Jv)_H$, for all $v, v \in V$). We recall that $\langle \cdot,\cdot\rangle_V$ denotes the duality product between V^* and V. In practice the injections J and J^* are often understated, and so (11.21) reads

$$\langle v, v' \rangle_V = (v, v')_H \quad \forall v, v' \in V. \tag{11.22}$$

Remark 11.8. (i) The reader should pay attention to the fact that, in general, $(v, v')_V$ (scalar product in V) is different from $\langle v, v' \rangle_V = (v, v')_H$.

(ii) Alternatively one can think of identifying V with V^* and looking at

$$H^* \subset V^* \equiv V \subset H. \tag{11.23}$$

Then we have

$$\langle u, v \rangle_H = (u, v)_V \quad \forall u, v \in V. \tag{11.24}$$

However, this is not favourable, since then, e.g., the integration by parts formula has to be done w.r.t. the $(\cdot, \cdot)_V$ product and not $(\cdot, \cdot)_H$.

Remark 11.9. It is possible to introduce a Gelfand triple also for pairs (V, H) with V Banach and H Hilberts space.

Example 11.10. Take for instance $V = H^1(\Omega)$, densely embedded in $H = L^2(\Omega)$. Let v and v' belong to V, with derivatives denoted by Dv and Dv'. Then

$$(v, v')_V = \int_\Omega v(x)v(x)\mathrm{d}x + \int_\Omega Dv(x)Dv(x)\mathrm{d}x; \quad \langle v, v' \rangle_V = (v, v)_H = \int_\Omega v(x)v(x)\mathrm{d}x. \tag{11.25}$$

11.3. A priori estimates

We study the backwards heat equation

$$\begin{cases} \dot{y}(t,x) + \Delta y(t,x) = f(t,x), & (t,x) \in (0,T) \times \Omega, \\ u(t,x) = 0, & \text{in } (0,T) \times \partial\Omega, \\ y(T,x) = g(x) & \text{in } \Omega. \end{cases} \tag{11.26}$$

with

$$f \in L^2(Q), \quad Q := [0,T] \times \Omega; \quad g \in L^2(\Omega). \tag{11.27}$$

We can state a variational formulation using test functions depending on the space only:

$$\begin{cases} \text{We have that } y(T,\cdot) = g, \text{ and for a.a. } t \in (0,T): \\ \displaystyle\int_\Omega (\dot{y}(t,x)v(x) - \nabla y(t,x) \cdot \nabla v(x) - f(t,x)v(x))\mathrm{d}x = 0, \\ \qquad\qquad\qquad\qquad \text{for all } v \in H_0^1(\Omega). \end{cases} \tag{11.28}$$

We will elaborate more on this in the next chapter, and turn now to the a priori estimates

First a priori estimate. This is an estimate of the solution in the space

$$L^\infty(0,T;L^2(\Omega)) \cap L^2(0,T;H_0^1(\Omega)), \tag{11.29}$$

obtained by multiplying the equation by the solution and integrating in space and time. For $\tau \in [0,T)$ we set

$$Q_\tau := [\tau,T] \times \Omega \tag{11.30}$$

Lemma 11.11.

$$\|y\|^2_{L^\infty(0,T;L^2(\Omega))} \leq e^T \left(\|g\|_2^2 + \|f\|^2_{L^2(Q)} \right), \tag{11.31}$$

$$\int_Q \|\nabla y\|_2^2 \, \mathrm{d}x\mathrm{d}t \leq \tfrac{1}{2}(Te^T + 1) \left(\|g\|_2^2 + \|f\|^2_{L^2(Q)} \right) \tag{11.32}$$

Proof. Taking $v = y(t,\cdot)$ in the variational formulation (11.28), integrate over Q_τ and observe that

$$2\int_{Q_\tau} \dot{y}(t,x)y(t,x)\mathrm{d}x\mathrm{d}t = \int_{Q_\tau} \frac{\mathrm{d}}{\mathrm{d}t}y(t,x)^2\mathrm{d}x\mathrm{d}t = \int_\tau^T \frac{\mathrm{d}}{\mathrm{d}t}\|y(t,\cdot)\|_2^2\,\mathrm{d}t = \|g\|_2^2 - \|y(\tau,\cdot)\|_2^2 . \tag{11.33}$$

Majorizing the contribution of f as usual we get

$$\tfrac{1}{2}\|y(\tau,\cdot)\|_2^2 + \int_\tau^T \|\nabla y(t,\cdot)\|_2^2\,\mathrm{d}x\mathrm{d}t \le \tfrac{1}{2}\|g\|_2^2 + \tfrac{1}{2}\int_\tau^T (\|f(t,\cdot)\|_2^2 + \|y(t,\cdot)\|_2^2)\mathrm{d}t. \tag{11.34}$$

Set $\beta(t) := \|y(t,\cdot)\|_2^2$. By the above inequalities

$$\beta(\tau) \le \beta(T) + \int_\tau^T (\|f(t,\cdot)\|_2^2 + \beta(t))\mathrm{d}t. \tag{11.35}$$

We then deduce (11.31) from the Gronwall lemma below, with parameters

$$a := \beta(T) + \|f\|_{L^2(Q)}^2 ; \quad b := 1 \tag{11.36}$$

and get the other estimate with (11.34). □

Lemma 11.12 (Simplified Gronwall lemma)**.** *Let $a \ge 0$, $b > 0$ and $\gamma(t)$ satisfy*

$$\gamma(t) \le a + b\int_t^T \gamma(s)\mathrm{d}s. \tag{11.37}$$

Then

$$\gamma(t) \le a e^{b(T-t)}. \tag{11.38}$$

Proof. We have that $\theta(t) := e^{bt}\int_t^T \gamma(s)\mathrm{d}s$ satisfies

$$\dot{\theta}(t) = b\theta(t) - e^{bt}\gamma(t) \ge b\theta(t) - ae^{bt} - b\theta(t) = -ae^{bt}. \tag{11.39}$$

Then $\theta(T) = 0$ implies $\theta(t) = -\int_t^T \dot{\theta}(s)\mathrm{d}s \le a\int_t^T e^{bs}\mathrm{d}s = \frac{a}{b}(e^{bT} - e^{bt})$, and finally $\gamma(t) \le a + be^{-bt}\theta(t) = ae^{b(T-t)}$, as was to be shown. □

Second a priori estimates. This is an estimate of

$$u \in L^\infty(0,T;H_0^1(\Omega)); \text{ and } \dot{u} \in L^2(0,T;L^2(\Omega)), \tag{11.40}$$

obtained by multiplying the equation by the time derivative of the solution. More precisely:

Lemma 11.13 (Second parabolic estimate)**.** *If $f \in L^2(Q)$ and $g \in H^1_0(\Omega)$, then $\dot{y} \in L^2(Q)$, $\nabla y \in L^\infty(0,T;L^2(\Omega))$ and:*

$$\max\left(\|\dot{y}\|^2_{L^2(Q)}, \|\nabla y\|_{L^\infty(0,T;L^2(\Omega))^n} \right) \leq \|\nabla g\|^2_{L^2(Q)} + \|f\|^2_{L^2(Q)}. \tag{11.41}$$

Proof. We multiply the heat equation by $\dot{y}$ and integrate over space. Using

$$-2\int_\Omega \Delta y(t,x)\dot{y}(t,x) = 2\int_\Omega \nabla y(t,x)\cdot\nabla\dot{y}(t,x)\mathrm{d}x = \int_\Omega \frac{d}{\mathrm{d}t}|\nabla y(t,x)|^2\mathrm{d}x = \frac{\mathrm{d}}{\mathrm{d}t}\|\nabla y\|^2_2, \tag{11.42}$$

we obtain

$$\int_\Omega |\dot{y}(t,x)|^2\mathrm{d}x + \tfrac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\nabla y(t,\cdot)\|^2_{L^2(\Omega)^n} = \int_\Omega f(t,x)\dot{y}(t,x)\mathrm{d}x. \tag{11.43}$$

For $\tau \in [0,T)$, integrating over $t \in [\tau,T]$ we get

$$\int_\tau^T \|\dot{y}(t,\cdot)\|^2_{L^2(\Omega)}\mathrm{d}t + \tfrac{1}{2}\|\nabla y(\tau,\cdot)\|^2_{L^2(\Omega)^n} \leq \tfrac{1}{2}\|\nabla g\|^2_{L^2(\Omega)^n} + \int_{Q_\tau} f(t,x)\dot{y}(t,x)\mathrm{d}x\mathrm{d}t. \tag{11.44}$$

Using Young's inequality and multiplying by 2, we deduce that

$$\int_\tau^T \|\dot{y}(t,\cdot)\|^2_{L^2(\Omega)}\mathrm{d}t + \|\nabla y(\tau,\cdot)\|^2_{L^2(\Omega)^n} \leq \|\nabla g\|^2_{L^2(\Omega)^n} + \|f\|^2_{L^2(Q)}, \tag{11.45}$$

from which the conclusion easily follows. □

11.4. Existence and uniqueness

The uniqueness follows from the first parabolic estimate. The existence can be obtained by the Galerkin approach, i.e., solving the variational formulation where the space V is approximated by a finite dimensional subspace.

Heat equation: Dirichlet boundary conditions. We consider the backwards heat equation on the open set $\Omega \subset \mathbb{R}^n$, with homogeneous Dirichlet boundary conditions:

$$\begin{cases} \dot{y}(t,x) + \Delta y(t,x) = f(t,x), & (t,x) \in (0,T) \times \Omega, \\ y(t,x) = 0, & (t,x) \in (0,T) \times \partial\Omega, \\ y(x,T) = g(x), & x \in \Omega. \end{cases} \tag{11.46}$$

We can associate with the heat equation the following variational formulation (compare to (4.3)):

$$\begin{cases} y(T,\cdot) = g, \text{ and for a.a. } t \in (0,T), \text{ for all } v \in H_0^1(\Omega): \\ \displaystyle\int_\Omega (\dot{y}(t,x)v(t,x) - \nabla y(t,x) \cdot \nabla v(t,x) - f(t,x)v(t,x))\mathrm{d}x = 0. \end{cases} \tag{11.47}$$

Here we consider the Lions-Magenes variational setting for parabolic equations. The Gelfand triple setting is as follows. Given two Hilbert spaces such that $V \subset H$ with continuous inclusion, identifying H with its dual we have that $H \subset V^*$ with continuous inclusion. Now define

$$W(0,T) := \{y \in L^2(0,T;V); \quad \dot{y} \in L^2(0,T;V^*)\}. \tag{11.48}$$

We have that

$$W(0,T) \subset C(0,T;H) \tag{11.49}$$

with continuous inclusion, and the following integration by parts formula holds for all $0 \le t' \le t'' \le T$:

$$(y(t''), z(t''))_H - (y(t'), z(t'))_H = \int_{t'}^{t''} (\langle \dot{y}(t), z(t)\rangle_V + \langle \dot{z}(t), y(t)\rangle_V)\mathrm{d}t. \tag{11.50}$$

Let the bilinear form $a(t;y,z)$ over $V \times V$ be uniformly (in time) continuous and semicoercive, i.e. for some $\alpha > 0$ and $\lambda \ge 0$:

$$a(t;y,y) \ge \alpha \|y\|_V^2 - \lambda \|y\|_H^2, \quad \text{for all } y \in V. \tag{11.51}$$

Let $f \in L^2(0,T;V^*)$. Consider the abstract parabolic equation

$$\langle \dot{y}, z\rangle_V + a(t;y,z) = \langle f(t), z\rangle_V, \qquad \text{for all } z \in V, \text{for a.a. } t \in (0,T); \quad y(0) = y_0. \tag{11.52}$$

Theorem 11.14. *Let $y_0 \in H$ and $f \in L^2(0,T;V^*)$. Then (11.52) has a unique solution in $W(0,T)$. We have a stronger result under the* semi-symmetry hypothesis

$$\begin{cases} a(t;y,z) = a_0(t;y,z) + a_1(t;y,z); \\ a_0 \in C^1, \text{ symmetric, coercive}; a_1 \text{ unif. continuous on } V \subset H. \end{cases} \tag{11.53}$$

Assume that $y_0 \in V$, $f \in L^2(0,T;H)$, and the semi symmetry hypothesis (11.53) holds. Then the solution of (11.52) satisfies

$$y \in L^\infty(0,T;V); \quad \dot{y} \in L^2(0,T;H). \tag{11.54}$$

The first statement follows from a Galerkin argument and the first parabolic estimate. Thereby, the regularity $\dot{y} \in L^2(0,T;V^*)$ results directly from the equation. The second statement uses the second parabolic estimate.

Remark 11.15. With $a(t;y,z)$, bilinear continuous over $V \times V$ is associated $A(t) \in L(V,V^*)$ such that

$$a(t;y,z) = \langle A(t)y, z\rangle_V. \tag{11.55}$$

So we may as well write the parabolic equation (11.53) as

$$\dot{y}(t) + A(t)y(t) = f(t) \text{ in } L^2(0,T;V^*); \quad y(0) = y_0 \text{ in } H. \tag{11.56}$$

11.5. Control of parabolic equations

Linear-quadratic setting. Previous setting with $f(t) = Bu(t)$, $B \in L(U, V^*)$, U Hilbert space, $u \in L^2(0,T;U)$. Cost function

$$J(u,y) := \tfrac{1}{2}\int_0^T (\|y(t) - y_d(t)\|_H^2 + \|u(t)\|_H^2)\mathrm{d}t + \tfrac{1}{2}\|y(T) - y_{dT}\|_H^2 . \tag{11.57}$$

Lagrange multiplier

$$(p,q) \in L^2(0,T;V^*) \times H. \tag{11.58}$$

Lagrangian function

$$L := J(u,y) + \int_0^T (\langle Bu(t) - A(t)y(t) - \dot{y}(t), p(t)\rangle_V)\mathrm{d}t + (q, y_0 - y(0))_H. \tag{11.59}$$

Costate equation

$$\begin{aligned} 0 = L_y z = \int_0^T &((y(t) - y_d(t), z(t))_H + \langle -A(t)z(t) - \dot{z}(t), p(t)\rangle_V)\mathrm{d}t \\ &+ (y(T) - y_{dT}(T), z(T))_H - (q, z(0))_H. \end{aligned} \tag{11.60}$$

Choosing $z \in D(0,T;V)$ we deduce that (taking $\dot{p}$ in a weak sense)

$$-\dot{p}(t) + A^*(t)p(t) = y(t) - y_d(t), \tag{11.61}$$

so that $p \in W(0,T)$. Integrating by parts we deduce

$$p(T) = y(T) - y_d(T); \quad p(0) = q. \tag{11.62}$$

The derivative of reduced cost $F(u) := J(u, y[u])$ is

$$F'(u)v = L_u v = \int_0^T (B^*p(t) + u(t), v(t))_H \mathrm{d}t. \tag{11.63}$$

11.6. Boundary control of parabolic equations (setting of Section 5.1)

We consider the following problem:

$$\begin{cases} \min \quad J(u,y) := \frac{1}{2}\left\|y(T)-y_{\mathrm{d}}\right\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\left\|u\right\|^2_{L^2((0,T)\times\partial\Omega)}, \quad \alpha > 0, \\ \quad \text{subject to} \\ \qquad y_t - \Delta y = 0 \quad \text{in } Q, \quad \partial_n y + y = u \quad \text{on } \Sigma, \\ \qquad y(0,\cdot) = y_0 \quad \text{in } \Omega, \quad u_m \leq u \leq u_M \quad \text{on } \Sigma, \end{cases} \tag{11.64}$$

with $u_m, u_M \in L^2(\Sigma)$, $u_m < u_M$, $y_0, y_d \in L^2(\Omega)$. Let $V := H^1(\Omega)$, $H := L^2(\Omega)$.

Let for $y \in W(0,T)$ and $v \in V$

$$a(y(t),v) := \int_\Omega \nabla y(t)\cdot\nabla v \mathrm{d}x + \int_{\partial\Omega} y(t)v \mathrm{d}s, \tag{11.65}$$

$$\langle u(t), v\rangle_V := (u(t),v)_{L^2(\partial\Omega)}. \tag{11.66}$$

the weak formulation of the state equation given by (16.14)–(16.15). Let

$$U := L^2(Q), \quad Y := W(0,T), \quad Z := L^2(0,T;H^1(\Omega)^*)\times L^2(\Omega). \tag{11.67}$$

Then it is easy to check that a is uniformly continuous and coercive. The weak formulation defines a bounded affine linear operator

$$e\colon (u,y)\in U\times Y \mapsto \mathcal{A}y + \begin{pmatrix} Bu \\ -y_0 \end{pmatrix} \in Z. \tag{11.68}$$

By Theorem 11.14 and the a priori estimates the equation $e(u,y)=0$ has a unique bounded affine linear solution operator $u \mapsto y[u]$ and $e_y(u,y)\in L(Y,Z)$,

$$e_y(u,y)v = \begin{pmatrix} v_t + Av \\ v(0) \end{pmatrix}, \tag{11.69}$$

has a bounded inverse. Moreover, by using the imbedding $Y \hookrightarrow C([0,T];L^2(\Omega))$, the objective function $J\colon U\times Y \to \mathbb{R}$ is obviously continously F-differentiable.

Hence, Assumption 7.5 is satisfied. Let $(\bar{u},\bar{y})\in U\times Y$ be local solution of (11.64), which is a global solution, since the problem is convex. Then Corollary 7.8 yields necessary optimality conditions, where the Lagrangian is given by

$$\begin{aligned} L(u,y,p,q) &:= J(u,y) + \langle (p,q), e(u,y)\rangle_Z \\ &= J(u,y) + \int_0^T \big(c(y(t),p(t)) - (u(t),p(t))_{L^2(\partial\Omega)}\big)\,\mathrm{d}t + (y(0)-y_0,q)_{L^2(\Omega)}, \\ c(y(t),p(t)) &:= \langle y_t(t),p(t)\rangle_V + a(y(t),p(t)) \end{aligned} \tag{11.70}$$

with $(p,q) \in L^2(0,T;V) \times L^2(\Omega)$. Hence, the optimality system is given in the form

$$\int_0^T (c(\bar{y}(t),p(t)) - (\bar{u}(t),p(t))_{L^2(\partial\Omega)})\mathrm{d}t = 0 \quad \forall v \in L^2(0,T;V), \tag{11.71}$$

$$\int_0^T (c(v(t),\bar{p}(t))\mathrm{d}t + (\bar{y}(T)-y_d, v(T))_{L^2(\Omega)} + (v(0),\bar{q})_{L^2(\Omega)} = 0 \quad \forall v \in Y, \tag{11.72}$$

$$u_m \le \bar{u} \le u_M, \quad (\alpha\bar{u} - \bar{p}, u - \bar{u})_{L^2(\Sigma)} \ge 0 \quad \forall u \in U, \text{ with } u_m \le u \le u_M. \tag{11.73}$$

Since $e_y(\bar{y},\bar{u}) \in L(Y,Z)$ has a bounded inverse, there exists a unique adjoint state $(\bar{p},\bar{q}) \in Z^* = L^2(0,T;V) \times L^2(\Omega)$.

To identify the adjoint equation, assume $\bar{p} \in W(0,T)$ (justified later). Then integration by parts in the term $\langle v_t(t),\bar{p}(t)\rangle_V$, using the integration by parts formula the adjoint equation is equivalent to

$$\begin{aligned} &\int_0^T (-\langle \bar{p}_t(t), v(t)\rangle_V + a(v(t),\bar{p}(t)))\mathrm{d}t + (\bar{y}(T) - y_d + \bar{p}(T), v(T))_{L^2(\Omega)} \\ &+ (v(0), \bar{q} - \bar{p}(0))_{L^2(\Omega)} = 0 \quad \forall v \in Y. \end{aligned} \tag{11.74}$$

Using the fact that $C_c^\infty(0,T;V) \subset Y$ is dense in $L^2(0,T;V)$, we conclude that for $\bar{p} \in Y$ the adjoint equation is equivalent to

$$\begin{cases} \displaystyle\int_0^T (-\langle \bar{p}_t(t), v(t)\rangle_V + a(v(t),\bar{p}(t)))\,\mathrm{d}t = 0 \quad \forall v \in L^2(0,T;V), \\ \bar{p}(T) = -(\bar{y}(T) - y_d), \quad \bar{q} = \bar{p}(0). \end{cases} \tag{11.75}$$

But this variational equation is the weak formulation of

$$-\bar{p}_t - \Delta\bar{p} = 0, \quad \bar{p}(T) = -(\bar{y}(T) - y_d), \quad (\partial_n\bar{p} + \bar{p})|_{(0,T)\times\partial\Omega} = 0 \tag{11.76}$$

and has unique solution $\bar{p} \in Y$, which is together with $\bar{q} = \bar{p}(0)$ the unique adjoint state. By applying Theorem 7.7 we obtain

Theorem 11.16. *For optimal solution* $(\bar{y}, \bar{u})$ *of* (11.64) *there exists* $\bar{p} \in Y$, $\lambda_{u_m}, \lambda_{u_M} \in L^2(\Sigma)$ *such that the following optimality system holds in a weak sense:*

$$\begin{cases} \bar{y}_t - \Delta \bar{y} = 0, & \partial_n \bar{y}|_\Sigma + \bar{y} = \bar{u}, & \bar{y}(0) = y_0, \\ -\bar{p}_t - \Delta \bar{p} = 0, & \partial_n \bar{p}|_\Sigma + \bar{p} = \bar{0}, & \bar{p}(T) = -(\bar{y}(T) - y_d), \\ \alpha \bar{u} - \bar{p} + \lambda_{u_M} - \lambda_{u_m} = 0, & & \\ \bar{u} \geq u_m, & \lambda_{u_m} \geq 0, & \lambda_{u_m}(\bar{u} - u_m) = 0, \\ \bar{u} \leq u_M, & \lambda_{u_M} \geq 0, & \lambda_{u_M}(u_M - \bar{u}) = 0. \end{cases} \tag{11.77}$$

12. State constraints

In this chapter we consider semilinear optimal control problems with constraints on the state.

Contents

12.1. Functions of bounded variation

Here we briefly recall some useful properties of the space of bounded variation functions and its relation with measure spaces. For details we refer to [13] and [8]. Let $T > 0$. Consider the set $\mathcal{S}^n(\tau, \tau')$ of finite increasing sequences σ_n in $[\tau, \tau']$, of the form

$$\tau = t_0 < t_1 < \cdots < t_n = \tau'. \tag{12.1}$$

The variation of a function $\mu\colon [0, T] \to \mathbb{R}$ on $[\tau, \tau']$, where $0 \le \tau < \tau' \le T$ is

$$\operatorname{var}_{[\tau,\tau']}(\mu) := \sup_n \left\{ \sum_{k=0}^{n-1} |\mu_{t_{k+1}} - \mu_{t_k}|; \quad t \in S_n(\tau, \tau') \right\}. \tag{12.2}$$

(i) $BV(0, T)$ denotes the space of functions of bounded variations.

(ii) $BV(0, T)$ is the space of differences of nondecreasing functions.

(iii) A function $\mu \in BV(0, T)$ has, for all $\tau \in [0, T]$, right and left limits denoted $\mu_{\tau\pm}$ at time τ (its value at time 0 (resp. T) is understood as its left (resp. right) limit at that point).

(iv) Its jump at time $\tau \in [0, T]$ is defined as $[\mu_\tau] := \mu_{\tau+} - \mu_{\tau-}$. Since the variation of μ is finite, there are finitely many jumps of absolute value greater than $1/n$, for all positive integer n, so that the set $D(\mu)$ of discontinuity times is countable.

(v) The set $C([0, T])$ of continuous functions over $[0, T]$, endowed with the norm $\|z\| := \max\{|z_t|;\ t \in [0, T]\}$ is a Banach space. With each $g \in C([0, T])$ and $\mu \in BV(0, T)$ we can associate the *Stieltjes integral*

$$\int_0^T g_t \mathrm{d}\mu_t := \lim_n \frac{1}{n} \sum_{k=0}^{n-1} g_{\hat{t}_k} (\mu_{t_{k+1}} - \mu_{t_k}), \tag{12.3}$$

where $t := k\frac{T}{n}$, for $k = 0$ to n, and $\hat{t}$ is an arbitrary element of $[t_k, t_{k+1}]$. For $\mu = x$ we obtain the Riemann integral.

(vi) The topological dual of $C([0, T])$ denoted by $M([0, T])$ is isomorphic to the quotient space $BV(0, T)/\mathbb{R}$ (Riesz).

(vii) Denote by 'd' the operator that with $\mu \in BV(0, T)$ associates the linear form $g \mapsto \int_0^T g_t \mathrm{d}\mu_t$.

12.2. Existence of a multiplier in an abstract setting

Given a Banach space X, a closed convex subset $S \subseteq X$ and a point $\bar{s} \in S$, the *normal cone* to S at $\bar{s}$ is defined as

$$N_S(\bar{s}) := \{x^* \in X^*;\ \langle x^*, s - \bar{s}\rangle \leq 0,\ \text{ for all } s \in S\}. \tag{12.4}$$

Consider the *abstract* problem

$$\min_{x \in K} f(x); \quad g(x) \in K_Y. \tag{P}$$

Here X and Y are Banach spaces, $f\colon X \to \mathbb{R}$ and $g\colon X \to Y$ are C^1, K is a nonempty closed convex subset of X, and K_Y is a nonempty closed convex subset of Y, with nonempty interior. The generalized Lagrangian of this problem is $L\colon X \times \mathbb{R} \times Y^* \to \mathbb{R}$ defined by

$$L(x, \alpha, \lambda) := \alpha f(x) + \langle \lambda, g(x)\rangle_Y. \tag{12.5}$$

Theorem 12.1. *Let $\bar{x}$ be a local solution of (P). Then there exists $\alpha \geq 0$ and $\lambda \in N_{K_Y}(g(\bar{x}))$, not both equal to 0, such that*

$$\alpha f'(\bar{x}) + g'(\bar{x})^* \lambda + N_K(\bar{x}) \ni 0. \tag{12.6}$$

Proof. (a) We claim that there exists no $h \in \mathcal{R}_K(\bar{x})$[1] such that

$$f'(\bar{x})h < 0, \quad g(\bar{x}) + g'(\bar{x})h \in \operatorname{int} K_Y. \tag{12.7}$$

Indeed, otherwise, setting $x_t := \bar{x} + th$ for $t \geq 0$, we would have, for $t > 0$ small enough, $x_t \in K$, and $f(x_t) < f(\bar{x})$. Also, since $y_h := g(\bar{x}) + g'(\bar{x})h$ is such that $B(y_h, \lambda)$ belongs to int K_Y for some $\lambda > 0$, we have that K_Y contains the convex hull say K_0 of $g(\bar{x})$ and $B(y, \lambda)$. For $t \in (0, 1)$ small enough,

$$g(x_t) = (1-t)g(\bar{x}) + ty_h + o(t) \in (1-t)g(\bar{x}) + tB(y_h, \lambda) \subset K_0 \subset K, \tag{12.8}$$

so that x_t is feasible and $f(x_t) < f(\bar{x})$, contradicting the local optimality of $\bar{x}$.

[1] $\mathcal{R}_K(x) := \{h \in X\ :\ \exists t > 0;\ x + th \in K\}$.

(b) Assuming w.l.o.g. that $g(\bar{x}) = 0$, it follows that the two following convex sets have empty intersection:

$$K_1 := (f'(\bar{x}), g'(\bar{x}))\mathcal{R}(\bar{x}), \quad K_2 := (-\infty, 0) \times \operatorname{int} K_Y. \tag{12.9}$$

Since K_2 has a nonempty interior, by the Hahn Banach theorem, K_1 and K_2 can be separated, i.e., there exists a nonzero $(\alpha, \lambda) \in \mathbb{R} \times Y^*$ such that

$$\alpha\beta_1 + \langle \lambda, y_1 \rangle_Y \geq \alpha\beta_2 + \langle \lambda, y_2 \rangle_Y, \quad \text{for all } (\beta_1, y_1) \in K_1, \quad (\beta_2, y_2) \in K_2. \tag{12.10}$$

Taking $\beta_2 \downarrow -\infty$ we deduce that $\alpha \geq 0$. Taking next $\beta_2 \uparrow 0$ we deduce that

$$\alpha f'(\bar{x})(x - \bar{x}) + \langle \lambda, g'(\bar{x})(x - \bar{x}) \rangle_Y \geq \langle \lambda, y \rangle_Y \quad \text{for all } y \in \operatorname{int} K_Y. \tag{12.11}$$

Taking $y \in \operatorname{int} K_Y$, $y \to 0$, we get (12.6). Taking $x = \bar{x}$, observing that the above inequality also holds for all $y \in K_Y$, we obtain that $\lambda \in N_{K_Y}(g(\bar{x}))$. The conclusion follows. □

12.3. Control of a semilinear elliptic equation

We consider a problem close to the one considered above, but with a distributed (and not boundary) control, and state constraint. The problem is given as

$$\begin{cases} J(u,y) := \frac{1}{2}\displaystyle\int_\Omega ((y(x)-y_d(x))^2 + u(x)^2)\mathrm{d}x, \\ y(x) - \Delta y(x) + y(x)^3 = f(x) + u(x); \quad x \in \Omega; \quad y(x) = 0 \quad \text{on } \partial\Omega, \\ \textbf{Control and state constraints:} \\ \qquad\qquad u \in K_U; \quad y(x) \le 1, \quad \text{for all } x \in \Omega. \end{cases} \tag{12.12}$$

The state space is $Y := H^2(\Omega) \cap H_0^1(\Omega)$. We assume $n \le 3$ so that the state will be continuous, indeed by the Sobolev inclusions for some $\beta \in (0,1)$, with $C_0^{0,\beta}(\Omega)$ the space of Hölder functions on $\bar\Omega$ with Hölder exponent β, and zero value at the boundary:

$$\text{For } n \le 3 : Y \subset C_0^{0,\beta}(\Omega). \tag{12.13}$$

The reduced cost and state constraints are

$$F(u) := J(u, y[u]); \quad G(u) \in K_Y \tag{12.14}$$

with $U := L^2(\Omega), \quad G : U \to C_0(\Omega),$

$$K_Y := \{y \in C_0(\Omega);\ y(x) \le 1 \text{ for all } x \in \Omega\}. \tag{12.15}$$

Here $C_0(\Omega)$ space of continuous functions over $\bar\Omega$ with zero value at the boundary. Dual space $M_0(\bar\Omega)$ of bounded Borel measures vanishing at the boundary. We consider the costate equation

$$\bar p - \Delta \bar p + 3\bar y^2 \bar p = \bar y - y_d + \mathrm{d}\mu \tag{12.16}$$

This must be understood as

$$\int_\Omega \bar p(x)(\Delta z(x) - (1+3\bar y^2(x))z(x))\mathrm{d}x + \int_\Omega (\bar y(x) - y_d(x))z(x)\mathrm{d}x + \int_\Omega z(x)\mathrm{d}\bar\mu(x) = 0,$$
$$\text{for all } z \in Y. \tag{12.17}$$

Equivalently, defining for $v \in L^2(\Omega)$, $z[v] \in Y$ as the unique solution of the linearized state equation

$$z - \Delta z + 3\bar y^2 z = v \tag{12.18}$$

the costate equation may be written as

$$\int_\Omega \bar{p}(x)v(x)\mathrm{d}x = \int_\Omega (\bar{y}(x) - y_d(x))z[v](x)\mathrm{d}x + \int_\Omega z[v](x)d\bar{\mu}(x), \quad \text{for all } v \in U. \tag{12.19}$$

Since the r.h.s. is a linear continuous form on U, by the Riesz theorem, the costate equation has a unique solution $\bar{p} \in U$.

Theorem 12.2. *Let $(\bar{u}, \bar{y})$ be a solution of the optimal control problem. Then there exists a nonzero pair $(\alpha, \bar{\mu}) \in \mathbb{R}^+ \times M_0(\bar{\Omega})$, $\bar{\mu}$ nonnegative with support on the contact set*

$$\Omega_0 := \{x \in \Omega; y_b(x) = 1\} \tag{12.20}$$

with $\bar{p} \in L^2(\Omega)$, solution of the costate equation and

$$\int_\Omega (\bar{p}(x) + \bar{u}(x))(v(x) - \bar{u}(x))\mathrm{d}x \geq 0, \quad \textit{for all } v \in K_U. \tag{12.21}$$

Proof. Observe that K_Y has a nonempty interior and that $F(u)$ and $G(u)$ are of class C^1. By Theorem 12.1, there exists a nonzero pair $(\alpha, \bar{\mu}) \in \mathbb{R}_+ \times M_0(\bar{\Omega})$, such that $\bar{\mu} \in N_{K_Y}(y_b)$ and

$$\alpha F'(\bar{u}) + G'(\bar{u})^* \bar{\mu} + N_{K_U}(\bar{u}) \ni 0. \tag{12.22}$$

That $\bar{\mu} \in N_{K_Y}(y_b)$ means

$$\int_\Omega (z(x) - y_b(x))\mathrm{d}\bar{\mu}(x) \leq 0, \quad \text{for all } z \in K_Y. \tag{12.23}$$

Taking $z = y_b - z'$, with $z' \geq 0$ arbitrary, deduce that $\mathrm{d}\bar{\mu} \geq 0$. If $x_0 \in \Omega$ and $y_b(x_0) < 1$, then for small enough ε, $B(x_0, \varepsilon) \subset \Omega$, and $y_b(x) < 1 - \varepsilon$ for all $x \in B(x_0, \varepsilon)$. Taking $z = y_b + z'$, with $z'(x) \in (0, \varepsilon)$ for all $x \in \Omega$, $z'(x)$ positive over $B(x_0, \varepsilon)$, get $\int_{B(x_0,\varepsilon)} z'(x)\mathrm{d}\bar{\mu}(x) \leq 0$. Since $\mathrm{d}\bar{\mu} \geq 0$, this means that the support of $\mathrm{d}\bar{\mu}$ does not contain x_0. Now (12.22) means, there exists $q \in -N_{K_U}(\bar{u})$ such that for all $v \in U$:

$$\int_\Omega qv = \int_\Omega \alpha(y_b - y_d)z + \int_\Omega z\mathrm{d}\bar{\mu} + \alpha \int_\Omega \bar{u}v, \tag{12.24}$$

proving that $q = \bar{p} + \bar{u}$. Since $q \in -N_{K_U}(\bar{u})$ implies

$$\langle q, v - \bar{u} \rangle_{C_0(\Omega)} \geq 0, \tag{12.25}$$

the conclusion follows. □

12.4. Control of a parabolic equations and alternative costates

State equation

$$\dot{y} - \Delta y = u, \quad y(0) = y_0 \tag{12.26}$$

with homogeneous boundary condition, where $u \in U := L^2(Q)$ and initial condition $y_0 \in H_0^1(\Omega)$. The cost function is

$$J(u, y) := \tfrac{1}{2} \int_Q (u(x,t)^2 + (y(x,t) - y_d(x,t))^2) \mathrm{d}x\mathrm{d}t + \tfrac{1}{2} \int_\Omega (y(x,T) - y_{dT}(x))^2 \mathrm{d}x. \tag{12.27}$$

The state constraint is

$$\int_\Omega c(x) y(x,t) \leq 0, \quad \text{for all } t \in [0,T], \tag{12.28}$$

where

$$c \in H^2(\Omega) \cap H_0^1(\Omega). \tag{12.29}$$

The state space and space of the state constraint are

$$Y := L^2(0,T; H^2(\Omega) \cap H_0^1(\Omega)) \cap H^1(0,T; L^2(\Omega)); \quad Z := C([0,T]). \tag{12.30}$$

Lagrangian

$$\beta J(u,y) + (p, u + \Delta y - \dot{y})_{L^2(Q)} + \int_\Omega p_0(y_0 - y(0)) + \int_0^T \int_\Omega c(x) y(x,t) \mathrm{d}\mu(t). \tag{12.31}$$

Here $\mu \in BV_0(0,T)$ space of BV (bounded variation) functions over $[0,T]$, with zero value at time T. It is well-known that given a bounded variation function $\mu \in BV(0,T)$ its distributional derivative $\mathrm{d}\mu$ is in the *space* $\mathcal{M}(0,T)$ *of finite Radon measures.* And, conversely, any element $\mathrm{d}\mu \in \mathcal{M}(0,T)$ can be identified with a function μ of bounded variation that vanishes at time T.

Costate equation: for any $z \in Y$:

$$\begin{aligned} 0 = L_y z = \int_Q (\beta(y - y_d) z + p\Delta z - p\dot{z}) + \beta(y(T) - y_{dT}, z(T))_{L^2(\Omega)} - \int_\Omega p_0 z(0) \\ + \int_0^T \int_\Omega c(x) z(x,t) \mathrm{d}\mu(t). \end{aligned} \tag{12.32}$$

Easy case: if $\mu \in W^{1,1}(0,T)$, searching for p in Y, integrating by parts in space and time, we recover a classical costate equation

$$-\dot{p} - \Delta p = \beta(y - y_d) + c\dot{\mu}; \tag{12.33}$$

with initial and final conditions

$$p(T) = \beta(y(T) - y_{dT}); \quad p(0) = p_0. \tag{12.34}$$

General case (Alternative costates): observe that the above expression suggests that the

$$\textit{alternative costate } p_1 := p + \mu c \text{ is smooth.} \tag{12.35}$$

Eliminating $p = p_1 - \mu c$ from (12.32), we get

$$\begin{aligned} 0 = \beta \int_\Omega ((y - y_d)z + (p_1 - \mu c)\Delta z - p_1 \dot{z}) + \beta(y(T) - y_{dT}, z(T))_{L^2(\Omega)} \\ - \int_\Omega p_0 z(0) + \int_0^T \int_\Omega c(x)(z(x,t)\mathrm{d}\mu(t) + \mu(t)\dot{z}(x,t)\mathrm{d}t). \end{aligned} \tag{12.36}$$

Now if φ and ψ belong to $BV([0,T])$, one of them being continuous, we have the integration by parts formula, see e.g. [7]:

$$\int_0^T (\varphi(t)\mathrm{d}\Psi(t) + \Psi(t)\mathrm{d}\varphi(t)) = \varphi(T)\Psi(T) - \varphi(0)\Psi(0). \tag{12.37}$$

Apply this with $\varphi(t) := \int_\Omega c(x)z(x,t)$ and $\Psi(t) = \mu(t)$. Note that $\varphi \in W^{1,1}(0,T)$, so that $\varphi \in BV([0,T])$. So (12.36) implies, since $\mu(T) = 0$:

$$\begin{aligned} 0 = \beta \int_\Omega ((y - y_d)z + (p_1 - \mu c)\Delta z - p_1 \dot{z}) - \int_\Omega p_0(x)z(x,0) \\ + \beta(y(T) - y_{dT}, z(T))_{L^2(\Omega)} - \int_\Omega c(x)z(x,0)\mu(0). \end{aligned} \tag{12.38}$$

We obtain that

$$-\dot{p}_1 - \Delta p_1 = \beta(y - y_d) - \mu \Delta c, \tag{12.39}$$

with initial and final conditions

$$p_1(T) = \beta(y(T) - y_{dT}); \quad p_1(0) = p_0 + \mu(0)c. \tag{12.40}$$

We deduce that $p_1 \in Y$. Since

$$\mu \in BV([0,T]), \quad y \in H^1(0,T;L^2(\Omega)) \subset BV(0,T;L^2(\Omega)) \tag{12.41}$$

we have that $p \in BV(0,T;L^2(\Omega))$. So its value at time 0 and T are well-defined and by the definition of p_1:

$$p(T) = p_1(T) = \beta(y(T) - y_{dT}); \quad p(0) = p_1(0) - \mu(0)c = p_0. \tag{12.42}$$

Note also that $p \in L^\infty(0,T;H^1_0(\Omega))$.

12.5. Time optimal control

Consider the problem of minimizing the horizon $T > 0$, such that the solution of the parabolic equation

$$\dot{y} - \Delta y = u, \quad y(0) = y_0 \tag{12.43}$$

with $u(t) \in K_U$ for a.a. t, satisfies $y(T) \in K$, where K is a closed convex subset of $H := L^2(\Omega)$, with nonempty interior, and K_U is a closed convex subset of H. We assume that $y_0 \in H$, and $y_0 \notin K$. The normalized time state equation is

$$\dot{y}(\tau) = T(\Delta y(\tau) + u(\tau)) \tag{12.44}$$

with $\tau \in (0,1)$. The normalized minimum time problem is

$$\begin{cases} \min T; \quad \text{s.t. } (12.44) \\ y(1) \in K, \text{ and } u(\tau) \in K_U \text{ for a.a. } \tau. \end{cases} \tag{12.45}$$

Asssume that (T, u, y) is solution of this problem, with $T > 0$. We apply Theorem (12.1). This amounts to consider the Lagrangian function

$$\beta T + \int_0^1 \langle p(t), T(u + \Delta y - \dot{y}\rangle_H \mathrm{d}t + \langle p_0, y_0 - y(0)\rangle + \langle \lambda, y(1)\rangle_H. \tag{12.46}$$

Costate equation

$$0 = L_y z = \int_0^1 \langle p(t), T\Delta z - \dot{z}\rangle_H \mathrm{d}t - \langle p_0, z(0)\rangle_H + \langle \lambda, z(1)\rangle_H. \tag{12.47}$$

Corresponding PDE

$$-\dot{p}(\tau) - T\Delta p(\tau) = 0, \quad \tau \in (0,1), \tag{12.48}$$

with boundary conditions

$$p(1) = \lambda \in N_K(y(1)); \quad p(0) = p_0. \tag{12.49}$$

Optimality condition: w.r.t. the control

$$p(\tau) + N_{K_U}(u(\tau)) \ni 0, \quad \tau \in (0,1), \tag{12.50}$$

but also the horizon

$$0 = L_T = \beta + \int_0^1 \langle p(\tau), (u(\tau) + \Delta y(\tau)\rangle_H \mathrm{d}\tau \tag{12.51}$$

Remark 12.3. We may define the pre-Hamiltonian

$$H(\tau) = \beta + \langle p(\tau), u(\tau) + \Delta y(\tau) \rangle_H. \tag{12.52}$$

As in the ODE setting, we see that the optimality condition w.r.t. T is that the average value of the pre-Hamiltonian is zero.

Remark 12.4. For $t \in [0, T]$, we may define $\bar{u}(t) := u(t/T)$ and $y_b(t) := y(t/T)$, and express the results using $(\bar{u}, y_b)$ rather than (u, y).

12.6. Numerical treatment of state constraints: Moreau-Yosida regularization

We consider

$$\begin{cases} \min J(u,y), \quad \text{s.t.:} \\ -\Delta y = u, \quad y_{\partial\Omega} = 0, \\ y(x) \leq y_b \text{ a.e. in } \Omega \end{cases} \tag{12.53}$$

with

$$J(u,y) := \tfrac{1}{2} \|y - y_d\|^2_{L^2(\Omega)} + \frac{\alpha}{2} \|u\|^2_{L^2(\Omega)}, \alpha > 0. \tag{12.54}$$

Due to the poor regularity of the multipliers involved in (12.53), a Moreau-Yosida regularization is frequently used for the numerical solution of the optimization problem. This approach consists in penalizing the pointwise state constraints by means of the C^1-function

$$\max(0, \bar{\lambda} + \gamma(y - y_b))^2, \tag{12.55}$$

with some fixed $\bar{\lambda} \in L^2(\Omega)$, yielding the following problem:

$$\begin{cases} \min J(u,y) + \dfrac{1}{2\gamma} \displaystyle\int_\Omega \max(0, \bar{\lambda} + \gamma(y - y_d))^2 \mathrm{d}x, \\ \qquad\qquad\qquad\qquad \text{s.t. : } -\Delta y = u, \quad y_{\partial\Omega} = 0. \end{cases} \tag{12.56}$$

Existence of a solution can be argued in a similar manner as for the unconstrained problem. Moreover, a first-order optimality system may be derived

Theorem 12.5. *Let $(\bar{u}, \bar{y})$ be a local optimal solution to (12.56) and let $\bar{\lambda} = 0$. Then there exists an adjoint state $p \in L^2(\Omega)$ such that (with the associated operator A) we have*

$$\begin{cases} A\bar{y} = \bar{u}, \\ A^*p = \bar{y} - y_d + \mu, \\ p + \alpha\bar{u} = 0, \\ \mu = \max(0, \gamma(\bar{y} - y_b)). \end{cases} \tag{12.57}$$

Proof. The existence of an adjoint state is obtained by following the lines of the proof of Theorem 3.3. In what follows let us introduce the variable

$$\mu := \max(0, \gamma(\bar{y} - y_b)) \in L^2(\Omega). \tag{12.58}$$

By computing the derivative of the reduced cost functional we obtain:

$$F'[\bar{u}]h = \langle J_y(\bar{u}, y[u]), y'[\bar{u}]h\rangle_Y + \gamma \int_\Omega \max(0, \bar{y} - y_b)y'[\bar{u}]h\mathrm{d}x + J_u(\bar{u}, y[\bar{u}])h \tag{12.59}$$

which, using the adjoint equation, implies that

$$F'[\bar{u}]h = \langle A^*p, y'[\bar{u}]h\rangle_Y + J_u(\bar{u}, y[\bar{u}])h = (p, Ay'[\bar{u}]h)_{L^2(\Omega)} + J_u(\bar{u}, y[\bar{u}])h \tag{12.60}$$

Considering the linearized equation $Ay'(\bar{u})h = h$, we finally get that

$$F'(\bar{u})h = (p, h)_{L^2(\Omega)} + J_u(\bar{u}, y[u])h \tag{12.61}$$

and therefore,

$$p + \alpha\bar{u} = 0 \quad \text{in } L^2(\Omega). \tag{12.62}$$

□

The solutions so obtained yield a sequence $\{(y_\gamma, u_\gamma)\}_{\gamma>0}$ that approximates the solution to (12.53) in the following sense.

Theorem 12.6. *The sequence $\{(y_\gamma, u_\gamma)\}_\gamma > 0$ of solutions to (12.56) contains a subsequence which converges strongly in $L^2(\Omega) \times Y$ to an optimal solution $(\bar{u}, \bar{y})$ of (12.53).*

Proof. Let $(\bar{u}, \bar{y}) \in U \times Y$ be a solution to (12.53). From the properties of the regularized cost functional we know that

$$J_\gamma(u_\gamma, y_\gamma) \leq J_\gamma(\bar{u}, \bar{y}) = J(\bar{u}, \bar{y}). \tag{12.63}$$

Consequently, since $\alpha > 0$, the sequence $\{u_\gamma\}_\gamma > 0$ is uniformly bounded in $L^2(\Omega)$, which implies that $\{y_\gamma\}_\gamma > 0$ is uniformly bounded in Y. Therefore, there exists a subsequence, denoted the same, such that $y_\gamma \rightharpoonup \hat{y}$ weakly in Y and $u_\gamma \rightharpoonup \hat{u}$ weakly in $L^2(\Omega)$. Additionally, from (12.63) the term

$$\frac{1}{2\gamma} \|\max(0, \gamma(y_\gamma - y_b))\|^2_{L^2(\Omega)} \tag{12.64}$$

is uniformly bounded with respect to γ. Hence,

$$\lim_{\gamma\to\infty} \|\max(0, y_\gamma - y_b)\|_{L^2(\Omega)} = 0. \tag{12.65}$$

Applying Fatous Lemma to the previous term we get that $\bar{y} \leq y_b$. Considering additionally that

$$J(\hat{u}, \hat{y}) \leq \liminf J(u_\gamma, y_\gamma) \leq \limsup J_\gamma(u_\gamma, y_\gamma) \leq J(\bar{u}, \bar{y}), \tag{12.66}$$

we get that $(\hat{u}, \hat{y})$ is solution of (12.53). Subsequently, we denote the optimal pair by $(\bar{u}, \bar{y})$. To verify strong convergence, let us first note that, due to (12.66)

$$\lim_{\gamma \to 0} \|y - y_d\|^2_{L^2(\Omega)} + \alpha \|u_\gamma\|^2_{L^2} = \|\bar{y} - z\|^2_{L^2(\Omega)} + \alpha \|\bar{u}\|^2_{L^2(\Omega)} \tag{12.67}$$

and, hence, $u_\gamma \to \bar{u}$ strongly in $L^2(\Omega)$. From the state equations it can be verified that the difference $y_\gamma - y$ satisfies the equation $A(y_\gamma - y) = u_\gamma - \bar{u}$, which thanks to the bounded- ness of A^{-1} implies that $y_\gamma \to \bar{y}$ strongly in Y. □

For the numerical solution of the optimality system the difficulty arises from the last nonsmooth equation. Using again the generalized derivative of the max function, the semismooth Newton step is given by

$$\delta\mu - \gamma \chi_{y > y_b} \delta_y = -\mu + \max(0, \gamma(y - y_b)). \tag{12.68}$$

13. Variational Inequality Constraints

Another type of nonsmooth optimization problems occurs when the constraints are given by so-called partial variational inequalities. An elliptic variational inequality problem has the following form:

Let Y be Hilbert space and $K \subset Y$ a non-mepty, closed, convex subset. Find $y \in K$ such that

$$a(y, v-y) \geq \langle f, v-y \rangle_Y, \qquad \text{for all } v \in Y, \tag{13.1}$$

where Y, U are Hilbert function spaces, $a(\cdot,\cdot)$ is Lipschitz continuous and coercive bilinear form and $f \in Y^*$.

Theorem 13.1 (Lions-Stampacchia)**.** *The variational inequality* (13.1) *has a unique solution. The mapping $Y^* \to K \subset Y$, $f \mapsto y$ is Lipschitz continuous with Lipschitz constant $1/\kappa$, where $\kappa > 0$ is the coercivity constant of the operator A.*

Furthermore, we have the following result. Let $j\colon Y \to \mathbb{R}$ be convex and lower semicontinuous. We consider the problem: Find $y \in Y$ such that

$$a(y, v-y) + j(v) - j(y) \geq \langle f, v-y \rangle_Y, \qquad \text{for all } v \in Y. \tag{13.2}$$

Theorem 13.2 (Lions-Stampacchia)**.** *The variational inequality* (13.2) *has a unique solution.*

Inequalities of this type arise in contact mechanics, elastoplasticity, viscoplastic fluid flow, among others.

The optimization of variational inequalities is closely related to the field of mathematical programming with equilibrium constraints (MPEC), which has received increasing interest in the past years, both in finite-dimensions and in function spaces. Due to the nondifferentiable structure of the constraints, the characterization of solutions via optimality conditions becomes challenging, and the same extends to the numerical solution of such problems.

A tracking type distributed optimization problem can be formulated as follows: Let $Y := H_0^1(\Omega)$ and $U = L^2(\Omega)$.

$$\begin{cases} \min\limits_{u\in U, y\in Y} J(u,y) = \frac{1}{2}\|y-y_d\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u\|^2_U\,, \\ \text{s.t.}: a(y,v-y)+j(v)-j(y) \geq \langle u, v-y\rangle_Y, \quad \text{for all } v \in Y. \end{cases} \tag{13.3}$$

Then

$$Y \mapsto L^2(\Omega) \mapsto Y^* \tag{13.4}$$

with compact and continuous embeddings. Existence of an optimal solution to (13.3) is shown in the following result.

Theorem 13.3. *There exists an optimal solution for problem* (13.3)*.*

Proof. Since the cost functional is bounded from below, there exists a minimizing sequence $\{(u_n, y_n)\}$, i.e.,

$$J(u_n, y_n) \to \inf_{u \in U} J(u,y), \tag{13.5}$$

where y_n stands for the unique solution to

$$a(y_n, v-y_n) + j(v) - j(y_n) \geq \langle u_n, v-y_n\rangle_Y \quad \text{for all } v \in Y. \tag{13.6}$$

From the structure of the cost functional it also follows that $\{u_n\}$ is bounded in U. Additionally, it follows from (13.6) that $\{y_n\}$ is bounded in Y, since using continuity and coercivity of a

$$c\|y_n\|_Y\|v\|_Y + j(v) \geq a(y_n,v)+j(v) \geq c(\|y_n\|^2_Y + j(y_n) - \|u_n\|_U\|v\|_Y - \|u_n\|_U\|y_n\|_Y). \tag{13.7}$$

Therefore, there exists a subsequence (denoted in the same way) such that $u_n \rightharpoonup \bar{u}$ weakly in U and $y_n \rightharpoonup \bar{y}$ weakly in Y. Due to the compact embedding $L^2(\Omega) \mapsto Y^*$ it then follows that $u_n \mapsto \bar{u}$ strongly in Y^*. From (13.6) we directly obtain that

$$a(y_n, y_n) - a(y_n, v) + j(v) - j(y_n) - \langle u_n, y_n - v\rangle_Y \leq 0, \quad \text{for all } v \in Y. \tag{13.8}$$

Thanks to the convexity and continuity of $a(\cdot,\cdot)$ and $j(\cdot)$ we may take the limit inferior in the previous inequality and obtain that

$$a(\bar{y}, \bar{y}) - a(\bar{y}, v) + j(v) - j(\bar{y}) - \langle \bar{u}, \bar{y} - v\rangle_Y \leq 0, \quad \text{for all } v \in Y, \tag{13.9}$$

which implies that $\bar{y}$ solves the VI with $\bar{u}$ on the right hand side. Thanks to the weakly lower semicontinuity of the cost functional we finally obtain that which implies the result. □

14. Sufficient optimality conditions

Contents

In this section we study problems of the form

$$\min f(x); \quad x \in K, \tag{14.1}$$

with K nonempty, closed and convex subset of the Banach space X, and $f \colon X \to \mathbb{R}$ has, at each $\bar{x} \in K$, a second-order Taylor expansion:

$$f(\bar{x}+h) = f(\bar{x}) + Df(\bar{x})h + \tfrac{1}{2}D^2 f(\bar{x})(h,h) + o(\|h\|_2), \tag{14.2}$$

where $Df(\bar{x}) \in X^*$ and

$$h \mapsto D^2 f(\bar{x})(h,h) \tag{14.3}$$

is a continuous quadratic form. By the above extension, $Df(\bar{x})$ is the Fréchet derivative of f at $\bar{x}$. But $D^2 f(\bar{x})(h,h)$ need not be the derivative of $Df(x)$ at $\bar{x}$. The polyhedricity hypothesis on K, stated later, allows to develop an abstract theory, concerning second-order optimality conditions, as well as the sensitivity analysis. This abstract material will be applied to optimal control problems with bound constraints on the controls

14.1. Second-order optimality conditions

Definition 14.1. *Let $\bar{x}$ belong to the convex set K. The radial cone to K at $\bar{x}$ is*

$$\mathcal{R}_K(\bar{x}) := \mathbb{R}_+(K-\bar{x}) = \{\alpha(x-\bar{x}), \alpha > 0, x \in K\}. \tag{14.4}$$

The tangent cone to K at $\bar{x}$ is the closure of the radial cone, or equivalently

$$T_K(\bar{x}) := \{h \in X; \quad \operatorname{dist}(\bar{x}+th) = o(t), \quad t \geq 0\}. \tag{14.5}$$

The normal cone to K at $\bar{x}$ is

$$N_K(\bar{x}) := \{x^* \in X^*; \langle x^*, x - \bar{x}\rangle_X \leq 0, \text{ for all } x \in K\}. \tag{14.6}$$

The critical cone at $\bar{x}$, for problem (14.1), is

$$C(\bar{x}) := \{h \in T_K(\bar{x}); \quad Df(\bar{x})h \leq 0\}. \tag{14.7}$$

The critical cone $C(x)$ represents those directions for which the linearization does not provide any information about optimality of x.

Proposition 14.2 (Second-order necessary optimality conditions)**.** *Let $\bar{x}$ be a local solution of (14.1). Then it satisfies the first-order necessary optimality condition*

$$Df(\bar{x})h = 0, \qquad \text{for all } h \in C(\bar{x}), \tag{14.8}$$

and the second-order necessary optimality condition

$$D^2 f(\bar{x})(h,h) \geq 0, \qquad \text{for all } h \in \overline{\mathcal{R}_K(\bar{x}) \cap Df(\bar{x})^\perp}. \tag{14.9}$$

Proof. Relation (14.8) follows from the definition of $C(\bar{x})$ and the classical first order necessary optimality condition

$$Df(\bar{x})h \geq 0, \text{ for all } h \in T_K(\bar{x}). \tag{14.10}$$

If in addition $h \in \mathcal{R}_K(\bar{x}) \cap Df(\bar{x})^\top$, then $\bar{x} + th \in K$ for $t > 0$ small enough, so that

$$0 \leq \lim_{t \downarrow 0} \frac{f(\bar{x} + th) - f(\bar{x})}{\frac{1}{2}t^2} = D^2 f(\bar{x})(h,h). \tag{14.11}$$

Since $h \mapsto D^2 f(\bar{x})(h,h)$ is continuous, (14.9) follows. □

Definition 14.3. *Let Y be a Hilbert space. We say that the quadratic form $Q : Y \to \mathbb{R}$ is a Legendre form if (i) Q is weakly sequentially l.s.c., and (ii) if $y_k \rightharpoonup y$ in Y and $Q(y_k) \to Q(y)$, then $y_k \to y$ strongly.*

Example 14.4. Let $q_0 \colon Y \to \mathbb{R}$, $q_0(y) := \|y\|_Y^2$. Then q is a Legendre form. Let $q_1(y) := \|Ay\|_Z^2$, where Z is a Hilbert space, and $A \in L(Y,Z)$ is compact. Then $q(y) := q_0(y) + q_1(y)$ is a Legendre form.

Proposition 14.5 (Second-order sufficient optimality condition)**.** *Let $\bar{x} \in K$ satisfy the first-order necessary optimality condition (14.8). Assume that X is a Hilbert space, $h \mapsto D^2 f(\bar{x})(h,h)$ is a Legendre form, and*

$$D^2 f(\bar{x})(h,h) > 0, \ \text{for all } h \in C(\bar{x}), \quad h \neq 0. \tag{14.12}$$

Then $\bar{x}$ is a local solution of (14.1), satisfying a quadratic growth condition: there exists $\alpha > 0$ such that

$$f(x) \geq f(\bar{x}) + \alpha \|x - \bar{x}\|_X^2, \quad \text{for all } x \in K, \quad \text{close enough to } \bar{x}. \tag{14.13}$$

Proof. If the conclusion does not hold, there exists a sequence x_k in K, $x_k \neq x$ for all k, converging to $\bar{x}$, such that

$$f(x_k) \leq f(\bar{x}) + o(\|x - \bar{x}\|^2). \tag{14.14}$$

Set $t_k := \|x_k - \bar{x}\|$, $h_k := t_k^{-1}(x_k - \bar{x})$, $x_k = \bar{x} + t_k h_k$, and $\|h_k\| = 1$. By the Taylor expansion, we have that

$$f(x_k) = f(\bar{x}) + t_k Df(\bar{x})h_k + \tfrac{1}{2}t_k^2 D^2 f(\bar{x})(h_k, h_k) + o(t_k^2). \tag{14.15}$$

Combining with (14.13), we obtain

$$Df(\bar{x})h_k + \tfrac{1}{2}t_k D^2 f(\bar{x})(h_k, h_k) \leq o(t_k). \tag{14.16}$$

Extracting if necessary a subsequence, since X is a Hilbert space, we may assume that $h_k \rightharpoonup \bar{h}$, whence $Df(\bar{x})h_k \rightharpoonup \bar{h}$. Passing to the limit in (14.16), get $Df(\bar{x})\bar{h} \leq 0$. Also, $\bar{h} \in T_K(\bar{x})$ (since $h_k \in T_K(\bar{x})$, and $T_K(\bar{x})$ is closed convex, hence weakly sequentially closed). Therefore, $\bar{h} \in C(\bar{x})$. By the first-order necessary optimality condition, $Df(\bar{x})h_k \geq 0$. Combining with (14.16), get

$$D^2 f(\bar{x})(h_k, h_k) \leq o(1). \tag{14.17}$$

Since $D^2 f(\bar{x})$ is weakly sequentially,

$$D^2 f(\bar{x})(\bar{h}, \bar{h}) \leq \limsup D^2 f(\bar{x})(h_k, h_k) \leq 0. \tag{14.18}$$

Since $\bar{h}$ is a critical direction, (14.12) implies $\bar{h} = 0$, and therefore, the inequalities in (14.18) are equalities. Since $D^2 f(\bar{x})$ is a Legendre form,

$$h_k \rightharpoonup \bar{h}, \ \text{and so,} \ \|\bar{h}\| = 1. \tag{14.19}$$

But then (14.12) contradicts (14.18). □

14.2. Polyhedricity theory

Polyhedric sets. It seems that there is a large gap between the necessary and sufficient second-order optimality conditions of the previous section, since they apply resp. to directions in the sets

$$\overline{\mathcal{R}_K(\bar{x}) \cap Df(\bar{x})^\perp}; \quad C(\bar{x}). \tag{14.20}$$

These sets are in general very different, as the following example shows.

Example 14.6. Take $X := \mathbb{R}^2$, endowed with the Euclidean norm, K its closed unit ball, $f(x) := x_2$. The solution is $\bar{x} = (0, -1)^\top$, and $C(\bar{x}) = \mathbb{R} \times \{0\}$, whereas

$$\overline{\mathcal{R}_K(\bar{x}) \cap Df(\bar{x})^\perp} = \{(0,0)\}. \tag{14.21}$$

Yet the two sets in (14.20) happen to be equal for a significant class of problems. Note that the first-order necessary optimality condition writes

$$-Df(\bar{x}) \in N_K(\bar{x}). \tag{14.22}$$

Definition 14.7. *Let $x \in K$ and $q \in N_K(x)$. We say that K is polyhedric at x for the normal q, if*

$$T_K(x) \cap q^\perp = \overline{\mathcal{R}_K(x) \cap q^\perp}. \tag{14.23}$$

If this holds for any $x \in K$ and $q \in N_K(x)$, we say that K is polyhedric.

Proposition 14.8. *Let X be a Hilbert space, K be polyhedric, and $\bar{x} \in K$ be such that $D^2 f(\bar{x})$ is a Legendre form. Then $\bar{x}$ is a local solution of (14.1) satisfying a quadratic growth condition iff it satisfies (14.8) and (14.12).*

Proof. By Proposition 14.5, if (14.8) and (14.12) hold, then $\bar{x}$ is a local solution satisfying the quadratic growth condition. Conversely, let the quadratic growth condition holds. Then $\bar{x}$ satisfies the first-order necessary optimality condition (14.8), and is for $\alpha > 0$ small enough a local minimum of the problem

$$\min f(x) - \tfrac{1}{2}\alpha \|x - \bar{x}\|^2; \quad x \in K. \tag{14.24}$$

By Proposition 14.2, this implies the second-order necessary condition

$$D^2 f(\bar{x})(h,h) - \alpha \|h\|^2 \geq 0, \qquad \text{for all } h \in \overline{\mathcal{R}_K(\bar{x}) \cap Df(\bar{x})^\perp}, \tag{14.25}$$

which implies (14.12). □

Stability of solutions. Consider now a family of optimization problems of the form

$$\min f(x,u); \quad x \in K \tag{P_u}$$

with X Hilbert space, U Banach space, K nonempty closed and convex subset of X, and $f : X \times U \to \mathbb{R}$ of class C^2. Let $\bar{x} \in X$, $\bar{u} \in U$. We assume that $D^2_{xx} f(\bar{x}, \bar{u})$ is a Legendre form, and that $\bar{x}$ is a local solution of (P_u) satisfying the second-order sufficient condition

$$D_x f(\bar{x},\bar{u})h = 0 \quad \text{and } D^2_{xx} f(\bar{x},\bar{u})(h,h) > 0, \quad \text{for all } h \in C(\bar{x},\bar{u}),\ h \neq 0. \tag{14.26}$$

Here $C(\bar{x}, \bar{u})$ denotes the critical cone

$$C(\bar{x},\bar{u}) := \{h \in T_K(\bar{x}); D_x f(\bar{x},\bar{u})h \le 0\}. \tag{14.27}$$

By Proposition 14.5, the quadratic growth condition holds. More precisely, define the local problem

$$\min f(x,u); \quad x \in K, \quad \|x - \bar{x}\| \le \theta, \tag{$P_{u,\theta}$}$$

with $\theta > 0$. Assuming $\theta > 0$ small enough, $\bar{x}$ is the unique solution of ($P_{\bar{u},\theta}$), and there exists $\alpha > 0$ such that

$$f(x,\bar{u}) \ge f(\bar{x},\bar{u}) + \alpha \|x - \bar{x}\|^2, \quad \text{for all } x \in K, \quad \|x - \bar{x}\| \le \theta. \tag{14.28}$$

We next establish the stability of a solution of the local problem, w.r.t. a perturbation.

Proposition 14.9. *Assume that f is weakly sequentially l.s.c., $Df(x,u)$ is Lipschitz over bounded sets, and (14.28) holds. Then, for all $u \in U$, the local problem ($P_{u,\theta}$) has at least a solution, and if $x_u \in S(P_{u,\theta})$, we have that*

$$\|x_u - \bar{x}\| = O(\|u - \bar{u}\|). \tag{14.29}$$

Proof. Being bounded, a minimizing sequence for ($P_{u,\theta}$) has a weak limit point x_u. Since K is weakly sequentially closed, $x_u \in K$. As f is weakly sequentially l.s.c., it follows that $x_u \in S((P_{u,\theta}))$. Combining the relations

$$f(x_u,\bar{u}) = f(x_u,u) + \int_0^1 D_u f(x_u, u + \sigma(\bar{u} - u))(\bar{u} - u)\mathrm{d}\sigma, \tag{14.30}$$

$$f(\bar{x},\bar{u}) = f(\bar{x},u) + \int_0^1 D_u f(\bar{x}, u + \sigma(\bar{u} - u))(\bar{u} - u)\mathrm{d}\sigma, \tag{14.31}$$

with the quadratic growth condition (14.28), the inequality $f(\bar{x}, u) - f(x_u, u) \geq 0$, and the fact that $Df(\bar{x}, \bar{u})$ is Lipschitz over bounded sets, we get

$$\begin{aligned}
\alpha \|x_u - \bar{x}\|^2 &\leq f(x_u, \bar{u}) - f(\bar{x}, \bar{u}) && (14.32)\\
&\leq f(x_u, \bar{u}) - f(x_u, u) + f(\bar{x}, u) - f(\bar{x}, \bar{u}) && (14.33)\\
&= \int_0^1 [D_u f(x_u, u + (\bar{u} - u)) - D_u f(\bar{x}, u + (\bar{u} - u))](\bar{u} - u) \mathrm{d}\sigma \\
& && (14.34)\\
&= O(\|x_u - \bar{x}\| \, \|u - \bar{u}\|), && (14.35)
\end{aligned}$$

and this implies (14.29). □

Sensitivity analysis. In the framework of the previous section, consider a mapping $\mathbb{R}^+ \to U$, $t \mapsto u(t)$, such that, for some $d \in U$:

$$u(t) = \bar{u} + td + r(t); \quad \|r(t)\|_U = o(t). \tag{14.36}$$

Set $v(t) := \mathrm{val}(P_{u(t),\theta})$, where $\theta > 0$ is such that (14.28) holds. Define the subproblem

$$\min_{h \in C(\bar{x})} D^2 f(\bar{x}, \bar{u})((h, d), (h, d)). \tag{SP}$$

Theorem 14.10. *Assume that f is weakly sequentially l.s.c., $Df(\bar{x}, \bar{u})$ is Lipschitz over bounded sets, $D^2 f(\bar{x})$ is a Legendre form, and the second-order sufficient condition (14.26) holds. Then we have the following expansion for the value function:*

$$v(t) = v(0) + D_u f(\bar{x}, \bar{u})(u(t) - \bar{u}) + \tfrac{1}{2} t^2 \, \mathrm{val}\,(\mathrm{SP}) + o(t^2), \quad t \geq 0. \tag{14.37}$$

Also, any weak limit $\bar{h}$ of an extracted sequence of $(x_t - \bar{x})/t$, $t \geq 0$, is a strong limit-point, and satisfies $\bar{h} \in S(SP)$.

Proof. a) **Upper estimate**. Let $\varepsilon > 0$. Since K is polyhedric, there exists $h \in \mathcal{R}_K(\bar{x}) \cap Df(\bar{x})^\perp$ such that

$$D^2 f(\bar{x}, \bar{u})((h, d), (h, d)) \leq \mathrm{val}(SP) + \varepsilon. \tag{14.38}$$

We have the Taylor expansion (using (14.26))

$$f(\bar{x} + th, u(t)) = f(\bar{x}, \bar{u}) + D_u f(\bar{x}, \bar{u})(u(t) - \bar{u}) + \tfrac{1}{2} t^2 D^2 f(\bar{x}, \bar{u})((h, d), (h, d)) + o(t^2). \tag{14.39}$$

Since $\bar{x}+th \in K$ for small enough $t>0$, we have that

$$v(t) \le f(\bar{x}+th, u(t)) \le f(\bar{x},\bar{u})+D_u f(\bar{x},\bar{u})(u(t)\bar{u})+\tfrac{1}{2}t^2(\mathrm{val}(SP)+\varepsilon)+o(t^2). \tag{14.40}$$

This being true for any $\varepsilon > 0$, we obtain

$$v(t) \le f(\bar{x},\bar{u}) + D_u f(\bar{x},\bar{u})(u(t)\bar{u}) + \tfrac{1}{2}t^2\,\mathrm{val}(SP) + o(t^2). \tag{14.41}$$

b) **Lower estimate**. Let $x_t \in S(P_{u(t),\theta})$. By Proposition 14.9, we have that

$$\|x_t - \bar{x}\| = O(\|u(t)-\bar{u}\|) = O(t), \tag{14.42}$$

so that $h_t := (x_t - \bar{x})/t$ is bounded. Let $\bar{h}$ be a weak limit for a subsequence. We have that

$$\begin{aligned} f(x_t, u(t)) &= f(\bar{x}+th_t, u(t)) \\ &= f(\bar{x},\bar{u}) + Df(\bar{x},\bar{u})(x_t-\bar{x}, u(t)-\bar{u}) \\ &\quad + \tfrac{1}{2}t^2 D^2 f(\bar{x},\bar{u})((h_t,d),(h_t,d)) + o(t^2). \end{aligned} \tag{14.43}$$

Comparing to (14.41), we obtain after division by $\frac{1}{2}t^2$:

$$2t^{-1}D_x f(\bar{x},\bar{u})h_t + D^2 f(\bar{x},\bar{u})((h_t,d),(h_t,d)) \le \mathrm{val}(SP) + o(1). \tag{14.44}$$

This implies $D_x f(\bar{x},\bar{u})h_t \le O(t)$, whence $D_x f(\bar{x},\bar{u})\bar{h} \le 0$. Since $h_t \in \mathcal{R}_K(\bar{x})$, we have that $\bar{h} \in T_K(\bar{x})$, and therefore

$$\bar{h} \text{ is a critical direction.} \tag{14.45}$$

On the other hand, $h_t \in \mathcal{R}_K(\bar{x})$ so that

$$D_x f(\bar{x},\bar{u})h_t \ge 0. \tag{14.46}$$

Combining with (14.44) and the weak sequentially l.s.c. of $D^2 f(\bar{x},\bar{u})$, we get that

$$D^2 f(\bar{x},\bar{u})((\bar{h},d),(\bar{h},d)) \le \liminf_{t\downarrow 0} D^2 f(\bar{x},\bar{u})((h_t,d),(h_t,d)) \le \mathrm{val}(SP). \tag{14.47}$$

Since $\bar{h} \in C(\bar{x})$, this implies $\bar{h} \in S(SP)$ and therefore

$$D^2 f(\bar{x},\bar{u})((h_t,d),(h_t,d)) \to D^2 f(\bar{x},\bar{u})((\bar{h},d),(\bar{h},d)). \tag{14.48}$$

Since $\bar{h}$ is a weak limit of a sequence extracted from h_t, it follows that

$$D^2_{xx} f(\bar{x},\bar{u})(h_t,h_t) \to D^2_{xx} f(\bar{x},\bar{u})(\bar{h},\bar{h}). \tag{14.49}$$

Thus, we have

$$v(t) \ge v(0) + D_u f(\bar{x},\bar{u})(u(t)-\bar{u}) + \tfrac{1}{2}t^2\,\mathrm{val}(SP) + o(t^2). \tag{14.50}$$

As $D^2_{xx} f(\bar{x},\bar{u})$ is a Legendre form, we deduce that $\bar{h}$ is a strong limit point of h_t. In particular, if $(SP) = \{\bar{h}\}$, then $h_t \to \bar{h}$. □

We refer to Bonnans and Shapiro [9] for further results.

Bound constraints in spaces of square summable functions. Let $X := L^2(\Omega)$, where $\Omega \subset \mathbb{R}^n$ open. Set

$$K := \{x \in L^2(\Omega); \quad \check{x} \leq x \leq \hat{x} \text{ a.e.}\}. \tag{14.51}$$

Here $\check{x}$ and $\hat{x}$ are measurable functions over Ω, $\check{x}$ with values in $\{-\infty\} \times \mathbb{R}$, $\hat{x}$ with values in $\mathbb{R} \times \{\infty\}$, such that K is nonempty.

Set $K(\omega) := [\check{x}(\omega), \bar{x}(\omega)]$. Let $x \in X$. It is easily established that its projection $y = P_K(x)$ onto K is characterized by

$$y(\omega) = P_{K(\omega)}x(\omega) = \max(\check{x}(\omega), \min(x(\omega), \hat{x}(\omega))) \text{ a.e..} \tag{14.52}$$

Given $x \in K$, we set

$$I(x) := \{\omega \subset \Omega; x(\omega) = \check{x}(\omega)\}; \tag{14.53}$$
$$J(x) := \{\omega \subset \Omega; x(\omega) = \hat{x}(\omega)\}; \tag{14.54}$$
$$L(x) := \Omega \setminus (I(x) \cup J(x)). \tag{14.55}$$

Lemma 14.11. *(i) The set K is closed and convex. (ii) Let $x \in K$. Then*

$$T_K(x) = \{h \in X; h \geq 0 \text{ a.e. on } I(x); h \leq 0 \text{ a.e. on } J(x)\}. \tag{14.56}$$
$$N_K(x) = \{h \in X; h \leq 0 \text{ a.e. on } I(x); h \geq 0 \text{ a.e. on } J(x); h = 0 \text{ a.e. on } L(x)\}. \tag{14.57}$$

In addition, let $q \in N_K(x)$. Then

$$T_K(x) \cap q^\perp = \{h \in T_K(x); \quad h(\omega)q(\omega) = 0 \text{ a.e.}\}. \tag{14.58}$$

(iii) The set K is polyhedric.

Proof. (i) The convexity of K given in (14.51) is obvious. Let x_k be a sequence in K converging to $\bar{x} \in X$. Extracting if necessary a subsequence, we may assume that $x_k \to \bar{x}$ a.e. It follows that $\bar{x} \in K$.

(ii) We just give the idea of the proof. One may check first, by direct arguments, the expression of $N_K(\bar{x})$, and then deduce the one of $T_K(\bar{x})$, using that for all $g \in N_K(\bar{x})$:

$$h \in T_K(\bar{x}) \quad \text{iff} \quad (g, h)_X \leq 0. \tag{14.59}$$

Namely, assume, $h < 0$ on a nonzero subset $\omega \subset I(\bar{x})$. Then choose, $g \in N_k(\bar{x})$ such that

$$g(x) = \begin{cases} h(x) & \text{for } x \in \omega, \\ 0 & \text{for else.} \end{cases} \tag{14.60}$$

We obtain $(g, h)_X = (g, h)_{L^2(\omega)} > 0$ and hence a contradiction.

Now let $h \in T_K(x) \cap q^\perp$, $q \in N_K(\bar{x})$. By (14.56) and (14.57), $h(\omega)q(\omega)$ is a.e. nonpositive, hence

$$h \perp q \quad \text{iff} \quad h(\omega)q(\omega) = 0 \text{ a.e.} \tag{14.61}$$

(iii) Let

$$h \in T_K(x) \cap q^\perp. \tag{14.62}$$

For $\varepsilon > 0$, set

$$h_\varepsilon := ((P_K(x + \varepsilon h) - x)/\varepsilon. \tag{14.63}$$

Clearly,

$$h_\varepsilon \in \mathcal{R}_K(x) \tag{14.64}$$

since

$$x + \varepsilon h_\varepsilon = P_K(x + \varepsilon h) \in K. \tag{14.65}$$

Since projections are nonexpansive, we have in view of the punctual characterization of P_K, $|h_\varepsilon(\omega)| \leq |h(\omega)|$ a.e., since

$$|h_\varepsilon(\omega)| = \frac{1}{\varepsilon}|P_{K(\omega)}(x(\omega) + \varepsilon h(\omega)) - P_{(\omega)}K(x)| \leq \frac{1}{\varepsilon}|x(\omega) + \varepsilon h(\omega) - x(\omega)| = |h(\omega)|. \tag{14.66}$$

In view of the expression of $P_K(x)$, $h_\varepsilon \to h$ a.e, since $x(\omega) + \varepsilon h(\omega) \in K(\omega)$ for $\varepsilon > 0$ sufficiently small. By the dominated convergence theorem,

$$h_\varepsilon \to h \quad \text{ in } X. \tag{14.67}$$

Also $h(\omega)q(\omega) = 0$ and $|h_\varepsilon(\omega)| \leq |h(\omega)|$ imply that

$$h_\varepsilon(\omega)q(\omega) = 0 \tag{14.68}$$

a.e., since

$$|h_\varepsilon(\omega)q(\omega)| = |h_\varepsilon(\omega)||q(\omega)| \leq |h(\omega)||q(\omega)| = |h(\omega)q(\omega)| = 0. \tag{14.69}$$

So, by (14.67), (14.64), and (14.69) we have found a sequence $h_\varepsilon \in \mathcal{R}_K(\bar{x}) \cap q^\perp$ with $h_\varepsilon \to h$ in X. This shows

$$T_K(x) \cap q^\perp \subset \overline{\mathcal{R}_K(\bar{x}) \cap q^\perp}. \tag{14.70}$$

The inclusion $' \supset'$ is obvious. □

Example 14.12. Consider the problem

$$\min_{x \in L^2(\Omega)^+} f(x), \tag{14.71}$$

with f of class $C^2 : L^2(\Omega) \to \mathbb{R}$. Set

$$I(\bar{x}) := \{\omega \in \Omega;\ \bar{x}(\omega) = 0\}. \tag{14.72}$$

The sufficient condition for quadratic growth is that, assuming $D^2 f(\bar{x})$ to be Legendre:

$$\begin{cases} Df(\bar{x})(\omega)h(\omega) = 0, \ \text{a.e.} \\ D^2 f(\bar{x})(h,h) > 0 \end{cases} \quad \text{whenever} \quad \begin{cases} \ h \geq 0 \text{ over } I(\bar{x}),\ h \neq 0, \text{ with} \\ Df(\bar{x})(\omega)h(\omega) = 0 \text{ a.e.} \end{cases} \tag{14.73}$$

Proof. We have

$$h \in T_K(\bar{x}) \cap Df(\bar{x})^{\perp} \quad \text{iff} \quad Df(\bar{x})h = 0 \text{ a.e. and } h \in T_K(\bar{x}). \tag{14.74}$$

If the necessary optimality condition of first-order holds

$$Df(\bar{x})h = 0 \quad \text{for all } h \in C(\bar{x}) \tag{14.75}$$

then

$$T_K(\bar{x}) \cap Df(\bar{x})^{\perp} = C(\bar{x}). \tag{14.76}$$

Since K is polyhedric, we have

$$T_K(\bar{x}) \cap Df(\bar{x})^{\perp} = \overline{\mathcal{R}_K(\bar{x}) \cap Df(\bar{x})^{\perp}}. \tag{14.77}$$

That means we obtain no-gap second order conditons, and using (14.58) we obtain the result. □

We say that a **no-gap condition** holds, when the only change between necessary or sufficient second-order optimality conditions is between a strict and non strict inequality. In that case it is usually possible to obtain a characterization of the second-order growth condition. No-gap conditions are obtained in the **polyhedric framework**, in the case when the Hessian of Lagrangian is a Legendre form, originating in the works by Haraux and Mignot and applied to optimal control problems by, e.g., Sokolowski and Bonnans. For further results on polyhedricity see also Wachsmuth.

15. Optimal control in a semigroup setting

Contents

15.1. Uniform operators

Let $A \in L(H)$, where H is a Banach space. Consider the ODE

$$\dot{x}(t) = Ax(t), \quad t > 0; \quad x(0) = x_0. \tag{15.1}$$

Here $x_0 \in H$. This ODE has a unique solution $x(t) = S(t)x_0$, where $S(t) \in L(H)$ satisfies

$$S(t) = e^{tA}, \tag{15.2}$$

where $e^{tA} := \sum_{k=0}^{\infty} \frac{t^k}{k!} A^k$. We have $\|S(t)\| \leq e^{\|A\|}$ and $S(t)$ has the group property

$$S(t)S_{t'} = S_{t+t'}, \qquad \text{for all } t, t' \in \mathbb{R}. \tag{15.3}$$

Also $t \mapsto S(t)$ is of class C^{∞}, $\mathbb{R} \to L(H)$, and

$$\frac{\mathrm{d}}{\mathrm{d}t} S(t) = AS(t) = S(t)A. \tag{15.4}$$

Next define the *resolvent set*

$$\rho(A) := \{\lambda \in \mathbb{R} \,;\; (\lambda I - A) \text{ is surjective and has a bounded inverse}\}. \tag{15.5}$$

We remind that, if $A \in L(H)$ is bijective, then its inverse is continuous. This is a consequence of the open mapping theorem.

15.2. Strongly continuous semigroups

Definition 15.1. *Let H be a Banach space, S a family of mappings $S(t) \in L(H)$ with $t \in [0, \infty)$. S is called* semigroup, *if*

$$S(t) \circ S(s) = S(t+s) \quad \forall t, s \geq 0, \tag{15.6}$$

$$S(0) = id. \tag{15.7}$$

Additionally, we assume that the semigroup is strongly continuous, i.e.

$$[0, \infty) \ni t \mapsto S(t)x \in H \qquad \text{is continuous for all } x \in H. \tag{15.8}$$

Definition 15.2. *Let H_1 be a subspace of H and*

$$A \colon H_1 \to H \quad \text{linear.} \tag{15.9}$$

We say that A is an unbounded operator on H with domain

$$D(A) = H_1, \tag{15.10}$$

and that

$$A \text{ is closed if its graph is closed.} \tag{15.11}$$

By the definition, the graph of A is closed if, when x_k belongs to $D(A)$ and $(x_k, Ax_k) \to (\bar{x}, \bar{y})$ in $H \times H$, then $\bar{x} \in D(A)$ and $\bar{y} = Ax$.

Example 15.3. *Let $\Omega \subset \mathbb{R}^n$ open with smooth boundary, $H = L^2(\Omega)$, $D(A) = H^2(\Omega) \cap H_0^1(\Omega)$, $Ay = -\Delta y$. If $(y_k, Ay_k) \to (\bar{y}, \bar{z})$ in $L^2(\mathbb{R}^n)$. Then, since*

$$\|y\|_{H^2} \leq c \, \|\Delta y\|_{L^2(\Omega)} \,, \tag{15.12}$$

we have $y_k \to \bar{y}$ in $H^2(\mathbb{R}^n)$, so that A is closed.

Definition 15.4 (Generator)**.** *The generator of a semigroup S auf H is*

$$A\colon x \mapsto \lim_{h\to 0} \frac{S(h)x - x}{h} \tag{15.13}$$

defined on $D(A) \subset H$,

$$D(A) := \left\{ x \in H \mid \lim_{h\to 0} \frac{S(h)x - x}{h} \right\}. \tag{15.14}$$

Example 15.5 (Exponential function)**.** For $H = \mathbb{R}^n$ and a matrix A we consider the equation

$$\partial_t y = Ay, \quad y(0) = y_0. \tag{15.15}$$

In the case $n = 1$ and $A = a$ the solution is given by $y(t) = y_0 e^{at}$. The semigroup $S(t)\colon \mathbb{R} \to \mathbb{R}$ ist gegeben durch $S(t) = e^{at}$. For general n (15.15) is also solvable (cf. Analysis II). The solution operator is

$$S(t) = e^{At}. \tag{15.16}$$

This operator can be defined as $S(t)\colon y_0 \mapsto y(t)$, where y is solution to the equation. A different possibility is to define e^{At} by a series. By elementar argument one can show that the series gives a solution to the equation.

We ask for the generator of the semigroup. The mapping $t \mapsto e^{At}$ ist differentiable and we find by

$$\partial_t e^{At} = Ae^{At} \quad \text{in } t = 0 \tag{15.17}$$

that

$$\lim_{h\to 0} \frac{S(h)x - x}{h} = (\partial_t e^{At} x)_{t=0} = (Ae^{At}x)_{t=0} = Ax. \tag{15.18}$$

Therefore, we call

$$A\colon \mathbb{R}^n \to \mathbb{R}^n \tag{15.19}$$

the generator of the semigroup e^{At}.

Example 15.6. For $H = L^p(\mathbb{R}, \mathbb{R})$ define the right shift $S(t)$

$$(S(t)y_0)(x) = y_0(x - t). \tag{15.20}$$

The generator of the semigroup is $A = -\partial_x\colon W^{1,p}(\mathbb{R}, \mathbb{R}) \to L^p(\mathbb{R}, \mathbb{R})$.

Proof. For $y \in W^{1,p}$ we have

$$\lim_{h\to 0} \frac{S(h)y - y}{h} = \lim_{h\to 0} \frac{y(\cdot - h) - y(\cdot)}{h} = -\partial_x. \tag{15.21}$$

Conversely, we have $-\partial_x y \in L^p$ only for functions $y \in W^{1,p}$. Therefore, there exists at most one solution. □

Example 15.7. On $L^p(\mathbb{R}^+;\mathbb{R})$ $p > 1$, we define $S(t)$ by

$$(S(t)y_0)(x) = \begin{cases} y_0(x-t) & \text{for } x - t > 0, \\ 0, & \text{else.} \end{cases} \tag{15.22}$$

The generator is formally again $A = -\partial_x$, but this time with a different domain

$$A = -\partial_x \colon H \supset D(A) = \{y \in W^{1,p}(\mathbb{R}^+,\mathbb{R}) \mid y(0) = 0\} \to L^p(\mathbb{R}^+,\mathbb{R}). \tag{15.23}$$

Proof. We consider $y \in W^{1,p}(\mathbb{R}^+,\mathbb{R})$ mit $y(0) \neq 0$. Then $(S(h)y - y)/h$ on $(0,h)$ is of order $1/h$. Such a sequence is unbounded in L^p für $p > 1$. The limit does not exist. Conversely, for $y \in W^{1,p}(\mathbb{R}^+)$ with $y(0) = 0$ the trivial extension is in $W^{1,p}(\mathbb{R})$. Thus, the limit exist for such y. □

We see:

1. $D(A) \neq \{x \in H \mid Ax \text{ can be defined}\}$
2. $D(A)$ contains important information, e.g. boundary values
3. The initial condition needs not to be in $D(A)$.

Lemma 15.8. *There exists $M > 0$ and $\omega \in \mathbb{R}$ such that*

$$\|S(t)\| \leq Me^{\omega t}. \tag{15.24}$$

Proof. For $\tau > 0$, let

$$\mathcal{S}(\tau) := \{S(t)\ ;\ t \in [0,\tau]\}. \tag{15.25}$$

We claim that, for $\tau > 0$ small enough, $\mathcal{S}(\tau)$ is bounded. Indeed, otherwise there would exist a positive sequence $t_k \downarrow 0$ such that $S(t_k)$ is unbounded. Then, by the Banach-Steinhaus theorem, see Brezis, there exists $x \in H$ such that $S(t_k)x$ is unbounded, which contradicts the hypotheses that the semigroup is strongly continuous.

So, there exists $M > 0$ such that $\|S(t)\| \leq M$, for all $t \in [0, \tau]$. Using the semigroup property we see that we may take $\tau = 1$. Since $S_0 = I$, $M \geq 1$. Set $\omega := \log M$. So if $t \in [n-1, n)$ with $n \in \mathbb{N}$, $n \geq 1$, then

$$\|S(t)\| = \|S^n(t/n)\| \leq M^n \leq M^{t+1} = Me^{\omega t}. \tag{15.26}$$

□

Corollary 15.9. *For a semigroup $S(t)$ on a Banach space H, the following assertions are equivalent.*

(i) S is strongly continuous,

(ii) $\lim_{h\downarrow 0} S(h)x = x$ for all $x \in H$.

For any C^0 semigroup $S(t)$ and $x \in H$, for all $h \geq 0$ and $t > 0$, when $t \downarrow 0$, we have that

$$\frac{1}{t}\int_h^{h+t} S(s)x\mathrm{d}s \to S(h)x. \tag{15.27}$$

Lemma 15.10. *Given $x \in H$ and $t > 0$, set $y := \int_0^t S(s)x\mathrm{d}s$. Then $y \in D(A)$ and*

$$Ay = S(t)x - x, \tag{15.28}$$

and if $x \in D(A)$ and $0 \leq s < t$, then $S(t)x \in D(A)$. Further, for all $t \geq 0$, $S(t)x$ is differentiable and

$$\frac{\mathrm{d}}{\mathrm{d}t}S(t)x = AS(t)x = S(t)Ax; \quad S(t)x - S(s)x = \int_s^t S(\tau)Ax\mathrm{d}\tau = \int_s^t AS(\tau)x\mathrm{d}\tau. \tag{15.29}$$

Proof. Observe that using the semigroup property we have

$$\frac{S(h) - \mathrm{id}}{h}y = \frac{1}{h}\int_h^{t+h} S(s)x\mathrm{d}s - \frac{1}{h}\int_0^t S(s)x\mathrm{d}s = \frac{1}{h}\int_t^{t+h} S(s)x\mathrm{d}s - \frac{1}{h}\int_0^h S(s)x\mathrm{d}s. \tag{15.30}$$

When $h \downarrow 0$, the r.h.s. converges to $S(t)x - x$. Relation (15.28) follows. Next, let $x \in D(A)$ and $0 \leq s < t$. Then, when $h \downarrow 0$:

$$\frac{S(h) - \mathrm{id}}{h}S(t)x = S(t)\frac{S(h) - \mathrm{id}}{h}x \to S(t)Ax. \tag{15.31}$$

For the first equality we use again the semigroup property. Therefore $S(t)x \in D(A)$ and $AS(t)x = S(t)Ax$. In addition

$$\frac{S(t)x - S(t-h)x}{h} - S(t)Ax = S(t-h)\left(\frac{S(h) - \mathrm{id}}{h}x - Ax\right) + (S(t-h)Ax - S(t)Ax). \tag{15.32}$$

When $h \downarrow 0$, the terms inside the parentheses converge to 0, and since $\|S(t)\| \leq Me^{t\omega}$, the l.h.s. of (15.32) also converges to 0, proving that $S(t)x$ has derivative $AS(t)x = S(t)Ax$. Integrating between s and t we recover (15.29). □

Lemma 15.11. *Any generator has dense domain and closed graph.*

Proof. Lemma 15.10 and (15.27) implies that $D(A)$ is a dense subset of H. Now let x_k in $D(A)$, and $(x_k, Ax_k) \to (\bar{x}, \bar{y})$ in $H \times H$. Then (15.29) implies that

$$S(t)x_k - x_k = \int_0^t S(\tau)Ax_k d\tau \tag{15.33}$$

Passing to the limit, we deduce that $S(t)\bar{x} - \bar{x} = \int_0^t S(\tau)\bar{y}\mathrm{d}\tau$; dividing by $t \downarrow 0$, we obtain that $A\bar{x} = \bar{y}$. So, A is a dense operator. □

15.3. Mild solutions

Consider the equation

$$\dot{y}(t) = Ay(t) + f(t); \quad t \in [0, T]; \quad y(0) = y_0. \tag{15.34}$$

Here A is the generator of a C^0 semigroup $S(t)$, $T > 0$, $y_0 \in H$, and $f \in L^1(0, T; H)$. When A is bounded, $S(t)$ is a continuous function of t, and (15.34) has a unique solution given by the variation of constants formula:

$$y(t) = S(t)y_0 + \int_0^t S(t-s)f(s)\mathrm{d}s. \tag{15.35}$$

We come back to the case when A is the generator of a C^0 semigroup $S(t)$. Since $\|S(t)\| \leq Me^{\omega t}$, the above integral is still well-defined (the measurability of $S(t-s)f(s)$ follows from the approximation of f by a piecewise constant function).

Lemma 15.12. *The mild solution operator, that with $(f, y_0) \in L^1(0, T; H) \times H$ associates $y \in C(0, T; H)$ solution of (15.35), is linear and continuous.*

Proof. Set $M_T := M \max(1, e^{\omega T})$. Then

$$\|y(t)\|_H \leq M_T(\|y_0\|_H + \|f\|_{L^1(0,T;H)}), \tag{15.36}$$

for all $t \in [0, T]$. Also, for $0 \leq t \leq \tau \leq T$:

$$y(\tau) - y(t) = (S(\tau) - S(t))y_0 + \int_t^\tau S(\tau - s)f(s)\mathrm{d}s \\ + \int_0^t (S(\tau - s) - (S(t-s))f(t)\mathrm{d}s.$$

When $\tau \downarrow t$, since S_t is a C^0 semigroup,

$$(S(\tau) - S(t))y_0 \to 0, \quad (S(\tau - s) - S(t-s))f(t) \to 0 \tag{15.37}$$

for a.a. $s \in (0, t)$. By the dominated convergence theorem,

$$\int_0^t (S(\tau - s) - S(t-s))f(t)\mathrm{d}s \to 0. \tag{15.38}$$

Finally

$$\left\| \int_t^\tau S(\tau - s)f(s)\mathrm{d}s \right\|_H \leq M_T \int_t^\tau \|f(s)\|_H \,\mathrm{d}s \tag{15.39}$$

converges to 0 when $\tau \downarrow t$. The conclusion follows. □

15.4. Link to weak solutions

Definition 15.13. *Let A be an unbounded operator over H, with dense domain. Its dual operator A^* is the unbounded operator in H^*, with domain*

$$D(A^*) := \{x^* \in H^*; \exists y^* \in H^*; \langle y^*, x\rangle_H = \langle x^*, Ax\rangle_H, \textit{for all } x \in D(A)\}, \tag{15.40}$$

and for $x^* \in D(A^*)$, we set $A^*x^* := y^*$.

Definition 15.14. *We say that $y \in C(0,T;H)$ is a weak solution to (15.34) if for all $x^* \in D(A^*)$, $\langle x^*, y(t)\rangle_H$ is absolutely continuous, and*

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle x^*, y(t)\rangle_H = \langle A^*x^*, y(t)\rangle_H + \langle x^*, f(t)\rangle_H, \quad \textit{for a.a. } t \in (0,T)\}; \quad y(0) = y_0. \tag{15.41}$$

Lemma 15.15. *Let A be a closed unbounded operator over H, with dense domain. Then (15.34) has a weak solution for each $y_0 \in H$, iff A is the generator of a semigroup. If this is the case, this solution is unique and coincides with the mild solution given by (15.35).*

Proof. See Ball [5]. □

Definition 15.16. *We call resolvent set of the unbounded operator A over H, the set*

$$\rho(A) := \{\lambda \in \mathbb{C}; (\lambda \,\mathrm{id} - A) \textit{ is surjective and has a bounded inverse}\}. \tag{15.42}$$

The corresponding resolvent operator is, for $\lambda \in \rho(A)$, defined by

$$R(\lambda, A) := (\lambda I - A)^{-1}. \tag{15.43}$$

Remark 15.17. For $\lambda \in \rho(A)$, the equation

$$\lambda x - Ax = y \tag{15.44}$$

has for any $y \in H$, a unique solution $x \in D(A)$ such that $\|x\| \leq c_\lambda \|y\|$ where $c_\lambda := \|R(\lambda, A)\|$. Now, we perturb λ, let $\mu \in \mathbb{C}$. Given $y \in H$, we want to find $x \in D(A)$ such that

$$y = \mu x - Ax = (\mu - \lambda)x + \lambda x - Ax, \tag{15.45}$$

so that

$$x = R(\lambda, A)(y - (\mu - \lambda)x). \tag{15.46}$$

This is a fixed point problem whose operator $Tx := R(\lambda, A)(y - (\mu - \lambda)x)$ satisfies

$$\|Tx_0 - Tx\| \leq c_\lambda |\mu - \lambda| \, \|x\| \, . \tag{15.47}$$

So, if $c_\lambda |\mu - \lambda| \, \|x\| < 1$, and if A is closed, then the fixed-point operator has a unique solution; this proves that, if A is closed, the resolvent set is open.

15.5. Hille-Yosida

Definition 15.18. *A strongly continous semigroup $S\colon H \to H$ is called* contractive, *if $\|S(t)\| \leq 1$ for all $t \geq 0$.*

Theorem 15.19. *Let A be an unbounded operator over H. Then A is the generator of a contraction C^0 semigroup iff*

(i) The operator A is closed and has a dense domain

(ii) $(0, \infty) \subset \rho(A)$ and, for each $\lambda > 0$:

$$\|R(\lambda, A)\| \leq \frac{1}{\lambda}. \tag{15.48}$$

Proof. Let A the generator of a contraction C^0 semigroup $S(t)$. By Lemma 15.11, A is closed with dense domain. Set

$$R(\lambda) := \int_0^\infty e^{-\lambda t} S(t) \mathrm{d}t. \tag{15.49}$$

The above Riemann integral is convergent, and

$$\|R(\lambda)\| \leq \int_0^\infty e^{-\lambda t} \mathrm{d}t = \frac{1}{\lambda}. \tag{15.50}$$

In addition, for $h > 0$:

$$\begin{aligned} \frac{S(h) - \mathrm{id}}{h} R(\lambda) &= \frac{1}{h} \int_0^\infty e^{-\lambda t} (S(t+h) - S(t)) \mathrm{d}t \\ &= \frac{e^{\lambda h} - 1}{h} \int_0^\infty e^{-\lambda t} S(t) \mathrm{d}t - \frac{e^{\lambda h}}{h} \int_0^h e^{-\lambda t} S(t) \mathrm{d}t. \end{aligned} \tag{15.51}$$

When $h \downarrow 0$, the r.h.s. converges to $\lambda R(\lambda) - \mathrm{id}$. So, for any $x \in H$, $R(\lambda)x \in D(A)$ and $AR(\lambda)x = \lambda R(\lambda)x - x$, so that $(\lambda \,\mathrm{id} - A)R(\lambda) = \mathrm{id}$, and therefore

$$R(\lambda) = (\lambda \,\mathrm{id} - A)^{-1} \tag{15.52}$$

is equal to the resolvent $R(\lambda, A)$. Therefore, conditions (i) and (ii) are necessary for A to be the generator of a contraction C^0 semigroup. For the proof of sufficiency, see Pazy, p.8. □

15.6. Optimal control in a semigroup setting

Integral equations framework. Consider the state equation

$$y(t) = S(0,t)y_0 + \int_0^t S(s,t)Bu(s)\mathrm{d}s, \quad t \in [0,T], \tag{15.53}$$

e.g.,

$$y(t) = e^{(t-0)\Delta}y_0 + \int_0^t e^{(t-s)\Delta}Bu(s)\mathrm{d}s, \quad t \in [0,T], \tag{15.54}$$

Here U and H are Banach spaces, $B \in L(U,H)$, $u \in L^1(0,T;U)$, and $y_0 \in H$. We assume that for $0 \leq s \leq t \leq T$, $S(s,t) \in L(H)$ satisfies

$$\|S(s)\|_{L(H)} \leq M, \tag{15.55}$$

and $(s,t) \mapsto S(s,t)x$ is continuous, for each $x \in H$. Then the integral in (15.53) is well-defined (again, approximating u by a piecewise constant function). We take as control and state spaces

$$\mathcal{U} := L^1(0,T;U); \quad \mathcal{Y} := L^\infty(0,T;H). \tag{15.56}$$

The cost function is

$$J(u,y) := \int_0^T \ell(u(t),y(t))\mathrm{d}t + \varphi(y(T)), \tag{15.57}$$

where $\ell : U \subset H \to R$, $\varphi \colon H \to \mathbb{R}$ are of class C^1, with derivatives bounded over bounded sets. Denoting by $y[u] \in Y$ the state associated with $u \in U$, and by $F(u) := J(u,y[u])$ the reduced cost, the optimal control problem is

$$\min_{u\in K} F(u). \tag{15.58}$$

The Lagrangian of the problem is

$$L(u,y,q) := J(u,y) + \int_0^T \left\langle q(t), S(0,t)y_0 + \int_0^t S(s,t)Bu(s)\mathrm{d}s - y(t)\right\rangle_H \mathrm{d}t$$
$$+ \left\langle q_T, S(0,T)y_0 + \int_0^T S(s,T)Bu(s)\mathrm{d}s - y(T)\right\rangle_H, \tag{15.59}$$

where q is the multiplier associated with the state equation.

A natural space for the latter is Y, and so, we need to take q in Y^*. We choose to take it in the smaller space

$$Q := L^\infty(0,T;H^*) + Q_T \tag{15.60}$$

with Q_T Dirac measures at time T with values in H^*. Then the Lagrangian function is well-defined. The condition of stationarity of the Lagrangian w.r.t. the state gives

$$q(t) = \ell_y(u(t),y(t)), \quad t \in (0,T); \quad q_T = \varphi'(y(T)) \tag{15.61}$$

Next, the derivative of the Lagrangian w.r.t. u in direction $v \in U$ gives, exchanging the roles of s and t:

$$\begin{aligned} L_u v &= J_u v + \int_0^T \left\langle q(t), \int_0^t S(s,t)Bv(s)\mathrm{d}s \right\rangle_H \mathrm{d}t + \left\langle q_T, \int_0^T S(s,T)Bv(s)\mathrm{d}s \right\rangle_H \\ &= J_u v + \int_0^T \left\langle \int_t^T S(t,s)^* q(s)\mathrm{d}s + S(t,T)^* q_T, Bv(t) \right\rangle_U \mathrm{d}t. \end{aligned} \tag{15.62}$$

So, introduce the costate $p \in L^\infty(0,T;H^*)$ defined by

$$p(t) := S(t,T)^* q_T + \int_t^T S(t,s)^* q(s)\mathrm{d}s. \tag{15.63}$$

Then

$$L_u v = \int_0^T \langle l_u(u(t),y(t)) + B^* p(t), v(t)\rangle_H \mathrm{d}t. \tag{15.64}$$

Note that the costate equation can be written as

$$p(t) := S(t,T)^* \varphi'(y(T)) + \int_t^T S(t,s)^* \ell_y(u(s),y(s))\mathrm{d}s. \tag{15.65}$$

We obtain that

Theorem 15.20. *The state equation* (15.53) *has a unique solution denoted by* $y[u]$*, and the derivative of the reduced cost* $F(u) := J(u,y[u])$ *in direction* $v \in U$ *satisfies*

$$F'(u)v = \int_0^T \langle \ell_u(u(t),y(t)) + B^* p(t), v(t)\rangle_H \mathrm{d}t. \tag{15.66}$$

15.7. Adjoint semigroup

When $S(t,T)$ can be expressed as $S(T-t)$, where S is the generator of a C^0 semigroup, then $S(t)^*$ is a semigroup, called the transpose semigroup. We may wonder if the transpose semigroup is C^0, so that it has a generator.

Theorem 15.21. *If H is reflexive, the adjoint semigroup is of class C^0 and its generator is A^*.*

Proof. See p. 38 in Pazy. □

By the previous Theorem, if H is reflexive, then costate equation (15.65) can be interpreted as a mild formulation for the backward differential equation

$$-\dot{p}(t) = A^* p(t) + l_y(u(t), y(t)), \quad t \in (0,T); \quad p(T) = \varphi'(y(T)). \tag{15.67}$$

15.8. Application

Let $\Omega \subset \mathbb{R}^n$ be bounded with smooth boundary and $\partial\Omega \in C^2$. We consider the unbouded operator $A := -\Delta$ on H,

$$A\colon D(A) = H^2(\Omega) \cap H_0^1(\Omega) \subset H \to H = L^2(\Omega), \tag{15.68}$$

Then A is a strongly continuous semigroup. We want to verify the hypothesis of the **Hille-Yosida theorem**:

- A is dense defined.
- $(0, \infty) \subset \rho(A)$: Show that for all $\lambda \in (0, \infty)$ we have $R(\lambda, A) = H$; that mean the problem

  $$-\Delta y + \lambda y = f \quad \text{in } \Omega, \quad u = 0 \text{ on } \partial\Omega \tag{15.69}$$

  has a unique solution for every $f \in L^2(\Omega)$ in $H_0^1(\Omega)$ by Lax-Milgram, an by H^2-regularity theory then $y \in D(A)$.
- We have for all $\lambda > 0$ that $\|R(\lambda, A)\| \leq \lambda^{-1}$:

  Let $\lambda > 0$. Then

  $$\|\nabla y\|^2 + \lambda \|y\|^2 = (f, y) \leq \|f\| \|y\| . \tag{15.70}$$

- A is closed: Let $x_n \in D(A)$, $x_n \to x$ in H and $Ax_n \to y$ in H, $n \to \infty$. We show that $x \in D(A)$ and $Ax = y$. There exists $z \in D(A)$ such that $(\lambda \operatorname{id} - A)z = \lambda x - y$. Let $w \in H^2(\Omega) \cap H_0^1(\Omega)$ with

  $$(\lambda - A)(x_n - z) = w. \tag{15.71}$$

  Then,

  $$\begin{aligned} \|x_n - z\| &\leq \lambda^{-1} \|(\lambda - A)(x_n - z)\| \\ &= \lambda^{-1} \|\lambda x_n - Ax_n - (\lambda x - y)\| \\ &\leq \|x_n - x\| + \lambda^{-1} \|Ax_n - y\| \to 0 \end{aligned} \tag{15.72}$$

  implying $x = z$, $Ax = y$.

Remark 15.22. For a formulation of a corresponding optimal control problem and the derivation of the optimality system we refer to the previous section. Existence can be shown by classical arguments using the fact that the control to state mapping is linear continuous and therefore weakly sequentially continuous.

For more details, bilinear control problems in a semigroup setting, applications to heat, wave, and Schrödinger equations, see Aronna, Bonnans, K. [3, 2, 4].

16. Appendix

Contents

16.1. Integration

Basics

- Monotone convergence
- Dominated convergence
- Fatou Lemma

Weak and a.e. convergence

Lemma 16.1. *Let Ω be a measurable subset of $\mathbb{R}^n$, and for $1 < q < 1$, a bounded sequence g_k in $L^q(\Omega)$, converging a.e. to some g. Then $g \in L^q(\Omega)$ and g_k weakly converges to g in $L^q(\Omega)$.*

Lemma 16.2 (Brézis-Lieb)**.** *Let f_k be a bounded sequence in $L^p(\Omega)$, $p \in (1, \infty)$, converging a.e. to some f. Then $f \in L^p(\Omega)$, and*

$$\|f\|_p = \lim_k (\|f_k\|_p - \|f - f_k\|_p). \tag{16.1}$$

And so $f_k \to f$ strongly iff $\|f_k\|_p \to \|f\|_p$.

Compactness in L^p

Lemma 16.3. *Let Ω have finite measure, and $1 \leq p < q < 1$. Let f_k be a bounded sequence in $L^q(\Omega)$, converging a.e. to f. Then $f_k \to f$ in $L^p(\Omega)$.*

16.2. Hölder spaces

Let $S \subset \mathbb{R}^n$ and let Y be a Banach. For $0 < \alpha \leq 1$ and $f : S \to Y$ we call

$$\text{Höl}_\alpha(f, S) := \sup \left\{ \frac{\|f(x) - f(y)\|_Y}{\|x - y\|_Y^\alpha} \mid x, y \in S, \quad x \neq y \right\} \in [0, \infty] \tag{16.2}$$

the Hölder constant of f on S to the exponent α, and in the special case $\alpha = 1$ we call $\text{Lip}(f, S) := \text{Höl}_1(f, S)$ the Lipschitz constant. If $\Omega \subset \mathbb{R}^n$ is open and bounded and $m \geq 0$, then the corresponding Hölder spaces are defined by

$$C^{m,\alpha}(\bar{\Omega}; Y) := \{ f \in C^m(\bar{\Omega}; Y) \mid \text{Höl}_\alpha(\partial^s f, \bar{\Omega}) < \infty \text{ for } |s| = m \}. \tag{16.3}$$

These are Banach spaces with the norm

$$\|f\|_{C^{m,\alpha}(\bar{\Omega})} := \sum_{|s| \leq m} \|\partial^s f\|_{C^0(\bar{\Omega})} + \sum_{|s| \leq m} \text{Höl}_\alpha(f, S). \tag{16.4}$$

Functions in $C^{0,\alpha}(\bar{\Omega}; Y)$ are called Hölder continuous on $\bar{\Omega}$, and Lipschitz continuous in the special case $\alpha = 1$.

Similarly one can also define the metric spaces $C^{m,\alpha}(\Omega; Y)$.

16.3. Mollification

Let ψ be $C^\infty(\mathbb{R}^n, \mathbb{R})$, $\int_{\mathbb{R}^n} \psi(x) \mathrm{d}x = 1$ have support on the unit ball B. The associated family of mollifiers, for $\varepsilon > 0$ and $x \in \mathbb{R}^n$ is:

$$\psi_\varepsilon(x) := \varepsilon^{-n} \psi(x/\varepsilon). \tag{16.5}$$

It has support on εB and unit integral. The mollification of $f \in L^p(\mathbb{R}^n)$, $1 \leq p < \infty$, is ($*$: convolution product)

$$f_\varepsilon(x) := f * \psi_\varepsilon(x) = \int_{\mathbb{R}^n} f(x - y) \psi_\varepsilon(y) \mathrm{d}y \in C^\infty(\mathbb{R}^n, \mathbb{R}). \tag{16.6}$$

Regularizing kernel: convergence

Lemma 16.4. *Let* $f \in L^p(\mathbb{R}^n)$*;* $1 \leq p < 1$*. Then* $f_\varepsilon \to f$ *in* $L^p(\mathbb{R}^n)$*.*

16.4. Existence and uniqueness of weak solutions for parabolic equations

We study the initial-boundary problem

$$\begin{cases} y_t + Ly = f & \text{in } Q, \\ y = 0 & \text{on } [0,T] \times \partial\Omega, \\ y(0,\cdot) = y_0 & \text{in } \Omega; \end{cases} \tag{16.7}$$

we will assume

$$a_{ij}, b_i, c_0 \in L^\infty(Q) \tag{16.8}$$

and for the source term and initial data

$$f \in L^2(0,T;H^{-1}(\Omega)), \quad y_0 \in L^2(\Omega), \tag{16.9}$$

L denotes for each time t a second order partial differential operator

$$Ly := -\sum_{i,j=1}^n \partial_j(a_{ij}(t,x)\partial_i y) + \sum_{i=1}^n b_i(t,x)\partial_i y + c_0(t,x)y \tag{16.10}$$

in divergence.

Definition 16.5. *The partial differential oeprator* $\frac{\partial}{\partial t}+L$ *is called* uniformly parabolic *if there is a constant* $\theta > 0$ *such that*

$$\sum_{i,j=1}^n a_{ij}(t,x)\xi_I\xi_j \geq \theta \|\xi\|^2 \quad \text{for a.a. } (t,x) \in Q \text{ and all } \xi \in \mathbb{R}^n. \tag{16.11}$$

We consider solutions of (16.7) as a Banach space valued function

$$t \in [0,T] \mapsto y(t) \in H_0^1(\Omega). \tag{16.12}$$

We analyze existence and uniqueness in an abstract setting. For Gelfand triple (V,H,V^*), given $f \in L^2(0,T;V^*)$ and $y_0 \in H$ we look for solutions

$$y \in W(0,T) \tag{16.13}$$

of the abstract parabolic evolution problem

$$\langle y_t(t), v\rangle_V + a(t;y(t),v) = \langle f(t), v\rangle_V \quad \forall v \in V \text{ and a.a. } t \in [0,T] \tag{16.14}$$

with the initial condition

$$y(0) = y_0. \tag{16.15}$$

Definition 16.6 (Weak solutions of parabolic PDEs)**.** *Let $\Omega \subset \mathbb{R}^n$ be open and bounded and the coefficients as in (16.8). We consider the Gelfand triple given by $(H_0^1(\Omega), L^2(\Omega), H^{-1}(\Omega))$. Then for $f \in L^2(0,T;H^{-1}(\Omega))$, $y_0 \in L^2(\Omega)$ a function*

$$y \in W(0,T) \tag{16.16}$$

is a weak solution of the initial-boundary value problem (16.7) if y satisfies the variational equation

$$\langle y_t(t), v\rangle_{H_0^1(\Omega)} + a(t;y(t),v) \;=\; \langle f(t), v\rangle_{H_0^1(\Omega)} \quad \forall v \;\in\; H_0^1(\Omega) \text{ and a.a. } t \;\in\; [0,T] \tag{16.17}$$

and the initial condition

$$y(0) = y_0 \tag{16.18}$$

with bilinear form a as given in (16.28).

Definition 16.7 (Weak solution of parabolic PDE, equivalent formulation)**.** *With the same assumption and notation as in Definition 16.6 the following is equivalent. For $f \in L^2(0,T;H^{-1}(\Omega))$, $y_0 \in L^2(\Omega)$ a function*

$$y \in W(0,T) \tag{16.19}$$

is a weak solution of the initial boundary problem (11.1) if y satisfies the variational equation

$$\int_0^T \langle y_t(t), v(t)\rangle_{H_0^1(\Omega)} \mathrm{d}t + \int_0^T a(t;y(t),v(t))\mathrm{d}t = \int_0^T \langle f(t), v(t)\rangle_{H_0^1(\Omega)} \mathrm{d}t \qquad \forall v \in L^2(0,T;H_0^1(\Omega)) \tag{16.20}$$

with $a(\cdot,\cdot;t)$ in (16.28) and the initial condition

$$y(0) = y_0. \tag{16.21}$$

Theorem 16.8. *Definition 16.6 and Definition 16.7 are equivalent.*

Proof. Let $y \in W(0,T)$ be a weak solution according to Definition 16.6. This implies

$$\langle y_t(t), v(t)\rangle_V + a(t;y(t),v(t)) = \langle f(t), v(t)\rangle_V \qquad \forall v \in L^2(0,T;V) \quad \text{and a.a. } t \in (0,T). \tag{16.22}$$

In fact, since $y \in W(0,T)$ and $f \in L^2(0,T;V^*)$, we have to check that both sides in (16.22) are in $L^1(0,T)$. For a simple function $v(t) = \sum_{i=1}^m \mathbf{1}_{E_i}(t)v_i$, $v_i \in V$, it is obvious, since then

$$\begin{aligned}
&\langle y_t(t), v(t)\rangle_V + a(t;y(t),v(t)) - \langle f(t), v(t)\rangle_V \\
&\quad = \sum_{i=1}^m \mathbf{1}_{E_i}(t)(\langle y_t(t), v_i\rangle_V + a(t;y(t),v_i) - \langle f(t), v_i\rangle_{V^*.V}) = 0 \quad \text{for a.a. } t.
\end{aligned} \tag{16.23}$$

For general $v \in L^2(0,T;V)$ choose a sequence v_k of simple functions with $v_k(t) \to v(t)$ almost everywhere. Then (16.22) holds for all v_k outside a set of measure zero (the countable union of the exceptional sets for v_k). Since $v_k(t) \to v(t)$ in V almost everywhere, we conclude that by continuity (16.22) holds also for the limit v. Integrating (16.22) with respect to t shows that (16.20) holds.

Let now $y \in W(0,T)$ be a weak solution according to Definition 16.7. Then (16.17) must hold. Otherwise we find $w \in V$ and a set E of nonzero measure such that for $v = w$ the difference of the left and right hand side of (16.17) is positive on E. But then (16.20) would not hold for $v(t) = \mathbf{1}_E(t)w$. Hence, (16.20) implies (16.17). □

Remark 16.9. *The mapping $v \mapsto a(u,v;t)$ is linear and continuous, that is, there exists $A(t) \in L(V,V^*)$ (formally corresponding to the differential operator L) such that*

$$\langle A(t)y, v\rangle_V = a(t;y,v) \quad \textit{for all } y \textit{ and } v \textit{ in } V. \tag{16.24}$$

Assumption 16.10. *(i) (V,H,V^*) is a separable Gelfand triple.*

(ii) $a(\cdot,\cdot,t)\colon V\times V \to \mathbb{R}$ is for almost all $t \in (0,T)$ a bilinear form and there are $\alpha,\beta > 0$ and $\gamma \geq 0$ with

$$|a(t;v,w)| \leq \alpha \|v\|_V \|w\|_V \quad \forall v,w \in V \textit{ and a.a. } t \in (0,T), \tag{16.25}$$

$$a(t;v,v) + \gamma \|v\|_H^2 \geq \beta \|v\|_V^2 \quad \forall v \in V \textit{ and a.a. } t \in (0,T). \tag{16.26}$$

The mappings $t \mapsto a(t;v,w) \in \mathbb{R}$ are measurable for all $v,w \in V$.

(iii) $y_0 \in H$, $f \in L^2(0,T;V^*)$.

Example 16.11. *Assumption 16.10 is obviously satisfied for the uniformly parabolic initial boundary value problem (11.1) with $H = L^2(\Omega)$, $V = H_0^1(\Omega)$ and*

$$a_{ij}, b_i, c_0 \in L^\infty(Q), \quad f \in L^2(0,T;H^{-1}(\Omega)). \tag{16.27}$$

One easily verifies that the associated bilinear form

$$a(t;y,v) := \int_\Omega \left(\sum_{i,j=1}^n a_{ij}(t)\partial_i y \partial_j v + \sum_{i=1}^n b_i(t)\partial_i y v + c_0 y v \right) \mathrm{d}x, \quad y,v \in H_0^1(\Omega). \tag{16.28}$$

satisfies (16.25) and (16.26).

One can easily verify that under Assumption 16.10 for any $y, v \in L^2(0,T;V)$ the functions $t \mapsto a(t;y(t),v(t))$ is in $L^1(0,T)$. As above (16.14) implies

$$\begin{gathered} \langle y_t(t), v(t)\rangle_V + a(t;y(t),v(t)) = \langle f(t), v(t)\rangle_V \\ \forall v \in L^2(0,T;V) \quad \text{and a.a. } t \in (0,T). \end{gathered} \tag{16.29}$$

and (16.14) and (16.15) are equivalent to the abstract parabolic problem:

$$\begin{cases} \text{Find } y \in W(0,T) \text{ such that} \\ \displaystyle\int_0^T \langle y_t(t), v(t)\rangle_V \mathrm{d}t + \int_0^T a(t;y(t),v(t))\mathrm{d}t = \int_0^T \langle f(t), v(t)\rangle_V \mathrm{d}t \\ \forall v \in L^2(0,T;V) \\ \text{with initial condition } y(0) = y_0. \end{cases} \tag{16.30}$$

Theorem 16.12 (Energy estimate and uniqueness result)**.** *Let Assumption 16.10 hold. Then the abstract parabolic evolution problem has at most one solution $y \in W(0,T)$ and it satisfies the energy estimate*

$$\|y(t)\|_H^2 + \|y\|_{L^2(0,t;V)}^2 + \|y_t\|_{L^2(0,t;V^*)}^2 \le C(\|y_0\|_H^2 + \|f\|_{L^2(0,t;V^*)}^2) \quad \forall t \in (0,t], \tag{16.31}$$

where $C > 0$ depends only on β and γ in Assumption 16.10.

Proof. The proof is obtained by using $v = y$ in (16.29), using integration by parts formula (in time) and applying the Gronwall lemma, cf. [12]. □

16.4.1. Galerkin approximation.

Since V is separable there exists a countable set

$$\{v_k\colon k \in \mathbb{N}\} \subset V \tag{16.32}$$

of linearly independent elements v_k of V, such that the linear span $\{v_k \ : \ k \in \mathbb{N}\}$ is dense in V (take first a countable dense subset and drop elements that lie in the span of previous elements). Moreover, let

$$V_k := \operatorname{span}\{v_1, ..., v_k\}. \tag{16.33}$$

Then $V_k \subset V$ are Hilbert spaces and $\bigcup V_k$ is dense in V. Since V is dense in H, we find

$$y_{0,k} = \sum_{i=1}^{k} \alpha_{ik} v_i \in V_k \quad \text{with } y_{0,k} \to y_0 \in H. \tag{16.34}$$

Now fix $k \in \mathbb{N}$. We look for a function

$$y_k(t) := \sum_{i=1}^{k} \varphi_{ik}(t) v_i, \quad \varphi_{ik} \in H^1(0,T), \tag{16.35}$$

satisfying the finite dimensional *Galerkin approximation* of (16.14), (16.15)

$$\langle (y_k)_t(t), v\rangle_V + a(t; y_k(t), v) = \langle f(t), v\rangle_V \quad \forall v \in V_k \text{ and a.a. } t \in [0,T], \tag{16.36}$$

$$y(0) = y_{0k}. \tag{16.37}$$

One can (easily) show that functions y_k of type (16.35) are in $W(0,T)$ with weak derivative

$$(y_k)_t(t) = \sum_{i=1}^{k} \varphi'_{ik}(t) v_i \in L^2(0,T;V), \tag{16.38}$$

where $\varphi_i \in L^2(0,T)$ are the weak derivatives of $\varphi_i \in H^1(0,T)$. Since it is sufficient to test with the basis $\{v_1, \dots, v_k\}$ in (16.36), we conclude that (16.36) and (16.37) is equivalent to the system of ODEs for $\varphi_1, \dots, \varphi_{kk}$

$$\begin{cases} \displaystyle\sum_{i=1}^{k} (v_i, v_j)_H \varphi'_{ik}(t) + \sum_{i=1}^{k} a(t; v_j, v_j)\varphi_{ik}(t) = \langle f(t), v_j\rangle_V, \\ 1 \le j \le k, \quad \text{a.a. } t \in [0,T], \\ \varphi_{ik}(0) = \alpha_{ik}, \quad 1 \le i \le k. \end{cases} \tag{16.39}$$

Here we have used that $V \overset{d}{\hookrightarrow} H \equiv H^*$ yields $\langle v_i, v_j\rangle_V = (v_i, v_j)_H$.

Theorem 16.13. *Let Assumption 16.10 hold. Then the Galerkin approcimations have unique solutions $y_k \in W(0,T)$ and satisfies the energy estimate (16.31).*

Proof. This follows from standard theory for ODEs with measurable coefficients from (16.39). □

Theorem 16.14. *Let Assumption 16.10 hold. Then the abstract parabolic evolution problem (16.14) and (16.15) has a unique solution $y \in W(0,T)$.*

Corollary 16.15. *Let $\Omega \subset \mathbb{R}^n$ be open and bounded and $\partial_t + L$ with L as in (16.10) be uniformly parabolic, where $a_{ij}, b_i, c_0 \in L^\infty(Q)$. Then for any $f \in L^2(0,T;H^{-1}(\Omega))$ and $y_0 \in L^2(\Omega)$ the initial boundary problem (11.1) has a unique weak solution $y \in W(0,T)$ and satisfies the energy estimate (16.31) with $H = L^2(\Omega)$, $V = H_0^1(\Omega)$, $V^* = H^{-1}(\Omega)$.*

Proof of Theorem 16.14. Since $\|y_{0,k}\|_H \to \|y_0\|_H$ the energy estimate (16.14) yields a constant $C > 0$ such that

$$\|y_k\|_{L^2(0,T;V)} < C, \quad \|(y_k)_t\|_{L^2(0,T;V^*)} < C. \tag{16.40}$$

Now $L^2(0,T;V)$, $L^2(0,T;V^*)$ are Hilbert spaces and thus reflexive. Hence, we find a subsequence (y_{k_i}) with

$$y_{k_i} \rightharpoonup y \text{ in } L^2(0,T;V), \quad (y_{k_i})_t \rightharpoonup w \text{ in } L^2(0,T;V^*). \tag{16.41}$$

It is not difficult to show that this implies $w = y_t$. Now (16.14) implies

$$\begin{aligned}&\int_0^T (\langle (y_k)_t, v\rangle_V + a(t; y_k(t), v) - \langle f(t), v\rangle_V)\varphi(t)\mathrm{d}t = 0\\ &\forall v \in V_k;\ \varphi \in C_c^\infty(0,T)\end{aligned} \tag{16.42}$$

and the first two terms are bounded linear functionals w.r.t. $(y_k)_t$ and y_k, respectively. Limit transition gives

$$\int_0^T (\langle y_t, v\rangle_V + a(t; y(t), v) - \langle f(t), v\rangle_V)\,\varphi(t)\mathrm{d}t = 0 \quad \forall v \in \bigcup_k V_k;\ \varphi \in C_c^\infty(0,T). \tag{16.43}$$

This shows (16.14), where we use that $\bigcup_k V_K$ is dense in V. Finally, also the initial condition (16.15) holds. In fact, let

$$\varphi \in C^\infty([0,T]), \quad \varphi(0) = 1, \quad \varphi(T) = 0. \tag{16.44}$$

Then $t \mapsto w(t) = \varphi(t)v \in W(0,T)$ for all $v \in V$ and $w(0) = v$, $w(T) = 0$ yields by the integration by parts formula

$$\int_0^T (-\langle\varphi'(t)v, y\rangle_V + a(t; y(t), \varphi(t)v) - \langle f(t), \varphi(t)v\rangle_V) = (y(0), v)_H \quad \forall v \in V. \tag{16.45}$$

Similarly, we have by (16.36) and the integration by parts formula

$$\int_0^T (-\langle\varphi'(t)v, y_{k_i}(t)\rangle_V + a(t; y_{k_i}(t), \varphi(t)v) - \langle f(t), \varphi(t)v\rangle_V)\,\mathrm{d}t = (y_{0,k}, v)_H \qquad \forall v \in V_{k_i} \tag{16.46}$$

and the left hand side tends to the left hand side of the previous equation by the weak convergence of y_{k_i}. This gives $(y(0), v)_H = \lim_{k\to\infty}(y_{0,k}, v)_H = (y_0, v)_H$ for all $v \in \bigcup_k V_k$ and hence $y(0) = \lim_{k\to\infty} y_{0,k} = y_0$, since $\bigcup_k V_k$ is dense in V. □

Operator formulation. From (16.22) the weak formulation means that $y_t + Ly = f$ holds in $L^2(0,T;V^*)$. It defines a bounded linear operator

$$\mathcal{A} \colon W(0,T) \to L^2(0,T;H^{-1}(\Omega)) \times L^2(\Omega), \quad y \mapsto \begin{pmatrix} y_t + Ly \\ y(0,\cdot) \end{pmatrix} \tag{16.47}$$

in the sense that for all $(f, y_0) \in L^2(0,T;H^{-1}(\Omega)) \times L^2(\Omega)$

$$\begin{pmatrix} y_t + Ly \\ y(0,\cdot) \end{pmatrix} = \begin{pmatrix} f \\ y_0 \end{pmatrix} \iff \text{(16.17) and (16.18).} \tag{16.48}$$

16.4.2. Higher regularity

We make now the following additional assumption.

Assumption 16.16. *In addition to Assumption 16.10 we assume that*

$$a(v,w;\cdot) \in C^1([0,T]), \qquad a_t(v,w;t) \leq \alpha \|v\|_V \|w\|_V \quad \forall v,w \in V, \tag{16.49}$$

$$y_0 \in \{w \in V \ : \ a(w,\cdot,0) \in H^*\}, \tag{16.50}$$

$$f \in W(0,T). \tag{16.51}$$

Theorem 16.17. *Let Assumption 16.16 hold. Then the solution of (16.14) satisfies in addition $y_t \in W(0,T)$ and satisfies the equation*

$$\begin{aligned} &\langle y_{tt}(t), w\rangle_V + a(y_t(t), w; t) = \langle f_t(t), w\rangle_V - a_t(y(t), w; t), \\ &\langle y_t(0), w\rangle_V = (f(0), w)_H - a(y_), w; 0) \quad \forall w \in V. \end{aligned} \tag{16.52}$$

Proof. See [12]. □

From the temporal regularity we can deduce spatial regularity, if L is for example a uniformly elliptic operator. In fact, we have $y_t, f \in W(0,T) \hookrightarrow C([0,T];H)$ and thus

$$\|y_t(t)\|_H + \|f(t)\|_H \leq C \quad \text{for a.a. } t \in [0,T], \tag{16.53}$$

where $H = L^2(\Omega)$, $V = H_0^1(\Omega)$. This yields

$$a(y(t), w; t) = -\langle y_t(t), w\rangle_{H_0^1(\Omega)} + (f(t), w)_{L^2(\Omega)} \quad \forall w \in H_0^1(\Omega). \tag{16.54}$$

Now using regularity results for uniformly elliptic operators one can obtain for $\partial\Omega \in C^2$, $a_{ij} \in C^1(\Omega)$, $c_0 \in L^\infty(\Omega)$, L uniformly elliptic that

$$\|y(t)\|_{H^2(\Omega)} \leq C(\|y_t\|_{L^\infty(0,T;L^2(\Omega)} + \|f\|_{L^\infty(0,T;L^2(\Omega))} + \|y\|_{L^\infty(0,T;H^1(\Omega)}), \tag{16.55}$$

see [15].

Bibliography

[1] H.-W. Alt. *Linear functional analysis.* Universitext. Springer-Verlag London, Ltd., London, 2016. An application-oriented introduction, Translated from the German edition by Robert Nürnberg.

[2] M. Soledad Aronna, J. Frédéric Bonnans, and Axel Kröner. Correction to: Optimal control of infinite dimensional bilinear systems: application to the heat and wave equations [MR3767765]. *Math. Program.*, 170(2, Ser. A):569–570, 2018.

[3] M. Soledad Aronna, J. Frédéric Bonnans, and Axel Kröner. Optimal control of infinite dimensional bilinear systems: application to the heat and wave equations. *Math. Program.*, 168(1-2, Ser. B):717–757, 2018.

[4] M. Soledad Aronna, Joseph Frédéric Bonnans, and Axel Kröner. Optimal control of PDEs in a complex space setting: application to the Schrödinger equation. *SIAM J. Control Optim.*, 57(2):1390–1412, 2019.

[5] J. M. Ball. Strongly continuous semigroups, weak solutions, and the variation of constants formula. *Proc. Amer. Math. Soc.*, 63(2):370–373, 1977.

[6] F. Bonnans. Lecture notes: Optimal control of pdes. Technical report, INRIA Saclay, 2019.

[7] J.F. Bonnans. Numerical analysis of partial differential equations arising in finance and stochastic control.

[8] J.F. Bonnans. Optimal control problems with state constraints. *preprint*, 2009.

[9] J.F. Bonnans and A. Shapiro. Optimization problems with perturbations: a guided tour. *SIAM Rev.*, 40(2):228–264, 1998.

[10] H. Brézis. *Functional analysis, Sobolev spaces and partial differential equations.* Universitext. Springer, New York, 2011.

[11] A. Ern and J.-L. Guermond. *Theory and practice of finite elements*, volume 159 of *Applied Mathematical Sciences.* Springer-Verlag, New York, 2004.

[12] L.C. Evans. *Partial differential equations.* Amer. Math Soc., Providence, RI, 1998. Graduate Studies in Mathematics 19.

[13] N. Grady. Functions of bounded variation. *preprint.*

[14] M. Hintermüller, K. Ito, and K. Kunisch. The primal-dual active set strategy as a semismooth Newton method. *SIAM J. Optim.*, 13(3):865–888, 2003.

[15] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. *Optimization with PDE constraints*, volume 23 of *Mathematical Modelling: Theory and Applications.* Springer, New York, 2009.

[16] K. Ito and K. Kunisch. Semi-smooth Newton methods for state-constrained optimal control problems. *Systems and Control Lett.*, 50:221–228, 2003.

[17] P. Malliavin. *Integration and probability*, volume 157 of *Graduate Texts in Mathematics.* Springer-Verlag, New York, 1995. French edition: Masson, Paris, 1982.

[18] M. K. V. Murthy and G. Stampacchia. A variational inequality with mixed boundary conditions. *Israel J. Math.*, 13:188–224 (1973), 1972.

[19] J.P. Raymond. Optimal control of partial differential equations. page pp. 121.

[20] F. Tröltzsch. *Optimal control of partial differential equations*, volume 112 of *Graduate Studies in Mathematics.* American Mathematical Society, Providence, RI, 2010. Theory, methods and applications, Translated from the 2005 German original by Jürgen Sprekels.

[21] B. Vexler. Optimal control of partial differential equations. page 120, 2008.

[22] J. Wloka. *Partial differential equations.* Cambridge University Press, Cambridge, 1987. Translated from the German by C. B. Thomas and M. J. Thomas.

[23] Eberhard Zeidler. *Nonlinear functional analysis and its applications. I.* Springer-Verlag, New York, 1986. Fixed-point theorems, Translated from the German by Peter R. Wadsack.