

Applied and Numerical Harmonic Analysis

$$\widehat{f}(\gamma) = \int f(x) e^{-2\pi i x \gamma} dx$$

Kristian Bredies
Dirk Lorenz

Mathematical Image Processing



Birkhäuser



Applied and Numerical Harmonic Analysis

Series Editor

John J. Benedetto

University of Maryland
College Park, MD, USA

Editorial Advisory Board

Akram Aldroubi

Vanderbilt University
Nashville, TN, USA

Douglas Cochran

Arizona State University
Phoenix, AZ, USA

Hans G. Feichtinger

University of Vienna
Vienna, Austria

Christopher Heil

Georgia Institute of Technology
Atlanta, GA, USA

Stéphane Jaffard

University of Paris XII
Paris, France

Jelena Kovačević

Carnegie Mellon University
Pittsburgh, PA, USA

Gitta Kutyniok

Technische Universität Berlin
Berlin, Germany

Mauro Maggioni

Duke University
Durham, NC, USA

Zuowei Shen

National University of Singapore
Singapore, Singapore

Thomas Strohmer

University of California
Davis, CA, USA

Yang Wang

Michigan State University
East Lansing, MI, USA

More information about this series at <http://www.springer.com/series/4968>

Kristian Bredies • Dirk Lorenz

Mathematical Image Processing



Kristian Bredies
Institute for Mathematics and Scientific
University of Graz
Graz, Austria

Dirk Lorenz
Braunschweig, Germany

ISSN 2296-5009 ISSN 2296-5017 (electronic)
Applied and Numerical Harmonic Analysis
ISBN 978-3-030-01457-5 ISBN 978-3-030-01458-2 (eBook)
<https://doi.org/10.1007/978-3-030-01458-2>

Library of Congress Control Number: 2018961010

© Springer Nature Switzerland AG 2018

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This book is published under the imprint Birkhäuser, www.birkhauser-science.com by the registered company Springer Nature Switzerland AG

The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

ANHA Series Preface

The *Applied and Numerical Harmonic Analysis (ANHA)* book series aims to provide the engineering, mathematical, and scientific communities with significant developments in harmonic analysis, ranging from abstract harmonic analysis to basic applications. The title of the series reflects the importance of applications and numerical implementation, but richness and relevance of applications and implementation depend fundamentally on the structure and depth of theoretical underpinnings. Thus, from our point of view, the interleaving of theory and applications and their creative symbiotic evolution is axiomatic.

Harmonic analysis is a wellspring of ideas and applicability that has flourished, developed, and deepened over time within many disciplines and by means of creative cross-fertilization with diverse areas. The intricate and fundamental relationship between harmonic analysis and fields such as signal processing, partial differential equations (PDEs), and image processing is reflected in our state-of-the-art *ANHA* series.

Our vision of modern harmonic analysis includes mathematical areas such as wavelet theory, Banach algebras, classical Fourier analysis, time-frequency analysis, and fractal geometry, as well as the diverse topics that impinge on them.

For example, wavelet theory can be considered an appropriate tool to deal with some basic problems in digital signal processing, speech and image processing, geophysics, pattern recognition, biomedical engineering, and turbulence. These areas implement the latest technology from sampling methods on surfaces to fast algorithms and computer vision methods. The underlying mathematics of wavelet theory depends not only on classical Fourier analysis, but also on ideas from abstract harmonic analysis, including von Neumann algebras and the affine group. This leads to a study of the Heisenberg group and its relationship to Gabor systems, and of the metaplectic group for a meaningful interaction of signal decomposition methods. The unifying influence of wavelet theory in the aforementioned topics illustrates the justification for providing a means for centralizing and disseminating information from the broader, but still focused, area of harmonic analysis. This will be a key role of *ANHA*. We intend to publish with the scope and interaction that such a host of issues demands.

Along with our commitment to publish mathematically significant works at the frontiers of harmonic analysis, we have a comparably strong commitment to publish major advances in the following applicable topics in which harmonic analysis plays a substantial role:

<i>Antenna theory</i>	<i>Prediction theory</i>
<i>Biomedical signal processing</i>	<i>Radar applications</i>
<i>Digital signal processing</i>	<i>Sampling theory</i>
<i>Fast algorithms</i>	<i>Spectral estimation</i>
<i>Gabor theory and applications</i>	<i>Speech processing</i>
<i>Image processing</i>	<i>Time-frequency and time-scale analysis</i>
<i>Numerical partial differential equations</i>	<i>Wavelet theory</i>

The above point of view for the *ANHA* book series is inspired by the history of Fourier analysis itself, whose tentacles reach into so many fields.

In the last two centuries Fourier analysis has had a major impact on the development of mathematics, on the understanding of many engineering and scientific phenomena, and on the solution of some of the most important problems in mathematics and the sciences. Historically, Fourier series were developed in the analysis of some of the classical PDEs of mathematical physics; these series were used to solve such equations. In order to understand Fourier series and the kinds of solutions they could represent, some of the most basic notions of analysis were defined, e.g., the concept of "function." Since the coefficients of Fourier series are integrals, it is no surprise that Riemann integrals were conceived to deal with uniqueness properties of trigonometric series. Cantor's set theory was also developed because of such uniqueness questions.

A basic problem in Fourier analysis is to show how complicated phenomena, such as sound waves, can be described in terms of elementary harmonics. There are two aspects of this problem: first, to find, or even define properly, the harmonics or spectrum of a given phenomenon, e.g., the spectroscopy problem in optics; second, to determine which phenomena can be constructed from given classes of harmonics, as done, for example, by the mechanical synthesizers in tidal analysis.

Fourier analysis is also the natural setting for many other problems in engineering, mathematics, and the sciences. For example, Wiener's Tauberian theorem in Fourier analysis not only characterizes the behavior of the prime numbers, but also provides the proper notion of spectrum for phenomena such as white light; this latter process leads to the Fourier analysis associated with correlation functions in filtering and prediction problems, and these problems, in turn, deal naturally with Hardy spaces in the theory of complex variables.

Nowadays, some of the theory of PDEs has given way to the study of Fourier integral operators. Problems in antenna theory are studied in terms of unimodular trigonometric polynomials. Applications of Fourier analysis abound in signal processing, whether with the fast Fourier transform (FFT), or filter design, or the

adaptive modeling inherent in time-frequency-scale methods such as wavelet theory. The coherent states of mathematical physics are translated and modulated Fourier transforms, and these are used, in conjunction with the uncertainty principle, for dealing with signal reconstruction in communications theory. We are back to the *raison d'être* of the *ANHA* series!

University of Maryland
College Park, MD, USA

John J. Benedetto
Series Editor

Preface

Mathematical imaging is the treatment of mathematical objects that stand for images where an “image” is just what is meant in everyday conversation, i.e., a picture of a real scene, a photograph, or a scan. In this book, we treat images as continuous objects, i.e., as image of a continuous scene or, put differently, as a function of a continuous variable. The closely related field of *digital imaging*, on the other hand, treats discrete images, i.e., images that are described by a finite number of values or pixels. Mathematical imaging is a subfield of *computer vision* where one tries to understand how information is stored in images and how information can be extracted from images in an automatic way. Methods of computer vision usually use underlying mathematical models for images and the information therein. To apply methods for continuous images in practice, i.e., to digital images, one has to *discretize* the methods. Hence, mathematical imaging and digital imaging are closely related and often methods in both fields are developed simultaneously. A method based on a mathematical model is useful only if it can be implemented in an efficient way and the mathematical treatment of a digital method often reveals the underlying assumptions and may explain observed effects.

This book emphasizes the mathematical character of imaging and as such is geared toward students of mathematical subjects. However, students of computer science, engineering, or natural sciences who have a knack for mathematics may also find this book useful. We assume knowledge of introductory courses like linear algebra, calculus, and numerical analysis; some basics of real analysis and functional analysis are advantageous. The book should be suited for students in their third year; however, later chapters of the book use some advanced mathematics. In this book, we give an overview of mathematical imaging; we describe methods and solutions for standard problems in imaging. We will also introduce elementary tools as histograms and linear and morphological filters since they often suffice to solve a given task. A special focus is on methods based on multiscale representations, partial differential equations, and variational methods. In most cases, we illustrate how the methods can be realized practically, i.e., we derive applicable algorithms. This book can serve as the basis for a lecture on mathematical imaging, but is also possible to use parts in lectures on applied mathematics or advanced seminars.

The introduction of the book outlines the mathematical framework and introduces the basic problems of mathematical imaging. Since we will need mathematics from quite different fields, there is a chapter on mathematical basics. Advanced readers may skip this chapter, just use it to brush up their knowledge, or use it as a reference for the terminology used in this book. The chapter on mathematical basics does not cover the basics we will need. Many mathematics facts and concepts are introduced when they are needed for specific methods. Mathematical imaging itself is treated in Chaps. 3–6. We organized the chapters according to the methods, and not according to the problems. Somehow we present a box of tools that shall serve as a reservoir of methods so that the user can pick, combine, and develop tools that seem to be best suited for the problem at hand. These mathematical chapters conclude with exercises which shall help to develop a deeper understanding of the methods and techniques. Some exercises involve programming, and we would like to encourage all readers to try to implement the method in their favorite programming language. As with every book, there are a lot of topics which did not find their way into the book. We would still like to mention some of these topics in the sections called “Further developments.”

Finally, we would like to thank all the people who contributed to this book in one way or another: Matthias Bremer, Jan Hendrik Kobarg, Christian Kruschel, Rainer Löwen, Peter Maaß, Markus Müller (who did a large part of the translation from the German edition), Tobias Preusser, and Nadja Worliczek.

Graz, Austria
Braunschweig, Germany
August 2018

Kristian Bredies
Dirk Lorenz

Contents

1	Introduction	1
1.1	What Are Images?	1
1.2	The Basic Tasks of Imaging	5
2	Mathematical Preliminaries	15
2.1	Fundamentals of Functional Analysis	15
2.1.1	Analysis on Normed Spaces	16
2.1.2	Banach Spaces and Duality	23
2.1.3	Aspects of Hilbert Space Theory	29
2.2	Elements of Measure and Integration Theory	32
2.2.1	Measure and Integral	32
2.2.2	Lebesgue Spaces and Vector Spaces of Measures	38
2.2.3	Operations on Measures	45
2.3	Weak Differentiability and Distributions	49
3	Basic Tools	55
3.1	Continuous and Discrete Images	55
3.1.1	Interpolation	55
3.1.2	Sampling	59
3.1.3	Error Measures	61
3.2	Histograms	62
3.3	Linear Filters	68
3.3.1	Definition and Properties	69
3.3.2	Applications	75
3.3.3	Discretization of Convolutions	81
3.4	Morphological Filters	86
3.4.1	Fundamental Operations: Dilation and Erosion	88
3.4.2	Concatenated Operations	92
3.4.3	Applications	95
3.4.4	Discretization of Morphological Operators	97
3.5	Further Developments	101
3.6	Exercises	105

4 Frequency and Multiscale Methods	109
4.1 The Fourier Transform	109
4.1.1 The Fourier Transform on $L^1(\mathbf{R}^d)$	109
4.1.2 The Fourier Transform on $L^2(\mathbf{R}^d)$	112
4.1.3 The Fourier Transform for Measures and Tempered Distributions	120
4.2 Fourier Series and the Sampling Theorem	125
4.2.1 Fourier Series	125
4.2.2 The Sampling Theorem	126
4.2.3 Aliasing	128
4.3 The Discrete Fourier Transform	135
4.4 The Wavelet Transform	141
4.4.1 The Windowed Fourier Transform	141
4.4.2 The Continuous Wavelet Transform	144
4.4.3 The Discrete Wavelet Transform	149
4.4.4 Fast Wavelet Transforms	156
4.4.5 The Two-Dimensional Discrete Wavelet Transform	161
4.5 Further Developments	165
4.6 Exercises	166
5 Partial Differential Equations in Image Processing	171
5.1 Axiomatic Derivation of Partial Differential Equations	172
5.1.1 Scale Space Axioms	173
5.1.2 Examples of Scale Spaces	176
5.1.3 Existence of an Infinitesimal Generator	186
5.1.4 Viscosity Solutions	191
5.2 Standard Models Based on Partial Differential Equations	196
5.2.1 Linear Scale Spaces: The Heat Equation	196
5.2.2 Morphological Scale Space	199
5.3 Nonlinear Diffusion	206
5.3.1 The Perona-Malik Equation	207
5.3.2 Anisotropic Diffusion	222
5.4 Numerical Solutions of Partial Differential Equations	229
5.4.1 Diffusion Equations	234
5.4.2 Transport Equations	240
5.5 Further Developments	246
5.6 Exercises	247
6 Variational Methods	251
6.1 Introduction and Motivation	251
6.2 Foundations of the Calculus of Variations and Convex Analysis	263
6.2.1 The Direct Method	263
6.2.2 Convex Analysis	270
6.2.3 Subdifferential Calculus	285
6.2.4 Fenchel Duality	301

Contents	xiii
6.3 Minimization in Sobolev Spaces and BV	316
6.3.1 Functionals with Sobolev Penalty	316
6.3.2 Practical Applications	334
6.3.3 The Total Variation Penalty	351
6.3.4 Generalization to Color Images	385
6.4 Numerical Methods	391
6.4.1 Solving a Partial Differential Equation	392
6.4.2 Primal-Dual Methods	396
6.4.3 Application of the Primal-Dual Methods	415
6.5 Further Developments	425
6.6 Exercises	432
References	445
Picture Credits	453
Notation	455
Index	461
Applied and Numerical Harmonic Analysis (81 Volumes)	469

Chapter 1

Introduction



1.1 What Are Images?

We omit the philosophical aspect of the question “What are images?” and aim to answer the question “What kind of images are there?” instead. Images can be produced in many different ways:

Photography: Photography produces two-dimensional images by projecting a scene of the real world through some optics onto a two-dimensional image plane. The optics are focused onto some plane, called the focal plane, and objects appear more blurred the farther they are from the focal plane. Hence, photos usually have both sharp and blurred regions.

At first, photography was based on chemical reactions to map the different values of brightness and color onto photographic film. Then some other chemical reactions were used to develop the film and to produce photoprints. Each of the different chemical reactions happens with some slight uncontrolled variations, and hence the photoprint does not exactly correspond to the real brightness and color values. In particular, photographic film has a certain granularity, which amounts to a certain noise in the picture.

Nowadays, most photos are obtained digitally. Here, the brightness and color are measured digitally at certain places—the pixels, or picture elements. This results in a matrix of brightness or color values. The process of digital picture acquisition also results in some noise in the picture.

Scans: To digitize photos one may use a scanner. The scanner illuminates the photo row by row and measures the brightness or color along the lines. Usually this does not result in some additional blur. However, a scanner operates at some resolution, which results in a reduction of information. Moreover, the scanning process may result in some additional artifacts. Older scans are often pale and may contain some contamination. The correction of such errors is an important problem in image processing.

Microscopy: Digital and analog microscopy is a kind of mixture of photography and scanning. One uses measurement technology similar to photography but since the light cone is very wide in this case, the depth of field is very low. Consequently, objects that are not in the focal plane appear very blurred and are virtually invisible. This results in images that are almost two-dimensional. Confocal microscopy exacerbates this effect in that only the focal plane is illuminated, which suppresses objects not in the focal plane even further.

Indirect imaging: In some scenarios one cannot measure the image data directly. Prominent examples are computerized tomography (CT) and ultrasound imaging. In the case of CT, for example, one acquires X-ray scans of the object from different directions. These scans are then used to reconstruct a three-dimensional image of the density of the object.

Here one needs some mathematics to reconstruct the image in the first place [103]. The reconstruction often generates some artifacts, and due to noise from the measurement process, the reconstructed images are also not noise-free. Other indirect imaging modalities are single-photon emission computerized tomography (SPECT), positron-emission-tomography (PET), seismic tomography, and also holography.

Generalized images: Data different from the above cases can also be treated as images. In industrial engineering, for example, one measures surfaces to check the smoothness of some workpiece. This results in a two-dimensional “elevation profile” that can be treated just like an image. In other cases one may have measured a chemical concentration or the magnetization.

A slightly different example is liquid chromatography-mass spectrometry (LC/MS). Here one measures time-dependent one-dimensional mass spectra. The resulting two-dimensional data has a mass and a time dimension and can be treated by imaging techniques.

Some examples of different kinds of images can be seen in Fig. 1.1.

As we have seen, images do not have to be two-dimensional. Hence, we will work with d -dimensional images in general (which includes volume data and movies). In special cases we will restrict ourselves to one- or two-dimensional data.

Let us return to the question of what images are. We take a down-to-earth viewpoint and formulate an answer mathematically: an image is a function that maps every point in some domain of definition to a certain color value. In other words: an image u is a map from an *image domain* Ω to some *color space* F :

$$u : \Omega \rightarrow F.$$

We distinguish between discrete and continuous image domains:

- discrete d -dimensional images, for example $\Omega = \{1, \dots, N_1\} \times \dots \times \{1, \dots, N_d\}$.
- continuous d -dimensional images, for example $\Omega \subset \mathbf{R}^d$, or specifically $\Omega = [0, a_1] \times \dots \times [0, a_d]$.



$$f = \begin{bmatrix} 3 \cdot \frac{401404}{800} & -2 \cdot \frac{28104}{40} \\ -\frac{401404}{800} & + \frac{28104}{40} \end{bmatrix}$$

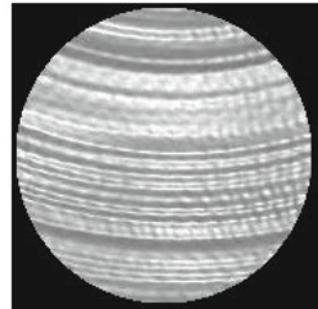
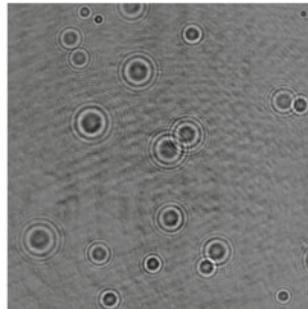
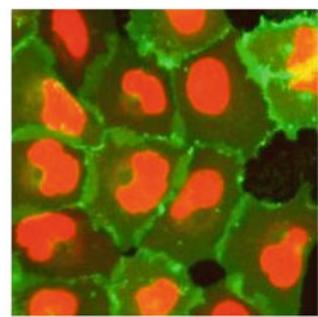


Fig. 1.1 Different types of images. First row: Photos. Second row: A scan and a microscopy image of cells. Third row: An image from indirect measurements (holography image of droplets) and a generalized image (“elevation profile” of a surface produced by a turning process)

Different color spaces are, for example:

- Black-and-white images (also binary images): $F = \{0, 1\}$.
- Grayscale images with discrete color space with k -bit depth: $F = \{0, \dots, 2^k - 1\}$.
- Color images with k -bit depth for each of N color channels: $F = \{0, \dots, 2^k - 1\}^N$.
- Images with continuous gray values: $F = [0, 1]$ or $F = \mathbf{R}$.
- Images with continuous colors: $F = [0, 1]^3$ or $F = \mathbf{R}^3$.

The field of digital image processing treats mostly discrete images, often also with discrete color space. This is reasonable in the sense that images are most often generated in discrete form or have to be transformed to a discrete image before further automatic processing. The methods that are used are often motivated by continuous considerations. In this book we take the viewpoint that our images are continuous objects ($\Omega \subset \mathbf{R}^d$). Hence, we will derive methods for continuous images with continuous color space. Moreover, we will deal mostly with grayscale images ($F = \mathbf{R}$ or $F = [0, 1]$).

The mathematical treatment of color images is a delicate subject. For example, one has to be aware of the question of how to measure distances in the color space: is the distance from red to blue larger than that from red to yellow? Moreover, the perception of color is very complex and also subjective. Colors can be represented in different color spaces and usually they are encoded in different color channels. For example, there are the *RGB space*, where colors are mixed additively from the red, green, and blue channels (as on screens and monitors) and the *CMYK space*, where colors are mixed subtractively from the cyan (C), magenta (M), yellow (Y), and black (K for black) channels (as is common in print). In the RGB space, color values are encoded by triplets $(R, G, B) \in [0, 1]^3$, where the components represent the amount of the respective color; $(0, 0, 0)$ represents the color black, $(1, 1, 1)$ stands for white. This is visualized in the so-called RGB cube; see Fig. 1.2. Also the colors cyan, magenta, and yellow appear as corners of the color cube. To process color images one often uses the so-called *HSV space*: a color is described by the channels *Hue*, *Saturation*, and *Value*. In the HSV space a color is encoded by a triplet $(H, S, V) \in [0, 360[\times [0, 100] \times [0, 100]$. The hue H is interpreted as an angle, the saturation S and the value V as percentages. The HSV space is visualized as a cylinder; see Fig. 1.3. Processing only the V-channel for the value (and leaving the other channels untouched) often leads to fairly good results in practice.

The goal of image processing is to automate or facilitate the evaluation and interpretation of images. One speaks of *high-level* methods if one obtains certain information from the images (e.g., the number of objects, the viewpoint of the

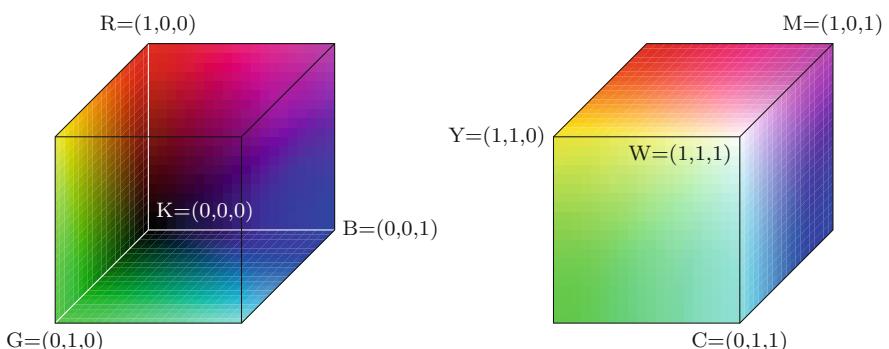
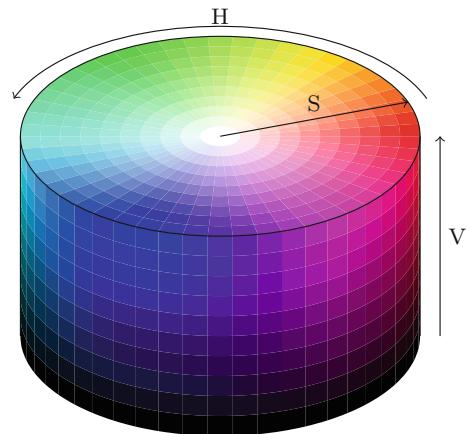


Fig. 1.2 RGB space, visualized as a cube

Fig. 1.3 HSV space, visualized as a cylinder



camera, the size of an object, and even the meaning of the scene). *Low-level* methods are methods that produce new and improved images out of given images. This book treats mainly low-level methods.

For the automatic processing of images one usually focuses on certain properties and structures of interest. These may be, for example:

Edges, corners: An edge describes the boundary between two different structures, e.g., between different objects. However, a region in the shade may also be separated from a lighter area by an edge.

Smooth regions: Objects with uniform color appear as smooth regions. If the object is curved, the illumination creates a smooth transition of the brightness.

Textures: The word “texture” mostly stands for something like a pattern. This refers, for example, to the fabric of a cloth, the structure of wallpapers, or fur.

Periodic structures: Textures may feature some periodic structures. These structures may have different directions and different frequencies and also occur as the superposition of different periodic structures.

Coherent regions: Coherent regions are regions with a similar orientation of objects as, for example, in the structure of wood or hair.

If such structures are to be detected or processed automatically, one needs good models for the structures. Is an edge adequately described by a sudden change of brightness? Does texture occur where the gray values have a high local variance? The choice of the model then influences how methods are derived.

1.2 The Basic Tasks of Imaging

Many problems in imaging can be reduced to only a few basic tasks. This section presents some of the classical basic tasks. The following chapters of this book will introduce several tools that can be used for these basic tasks. For a specific real-

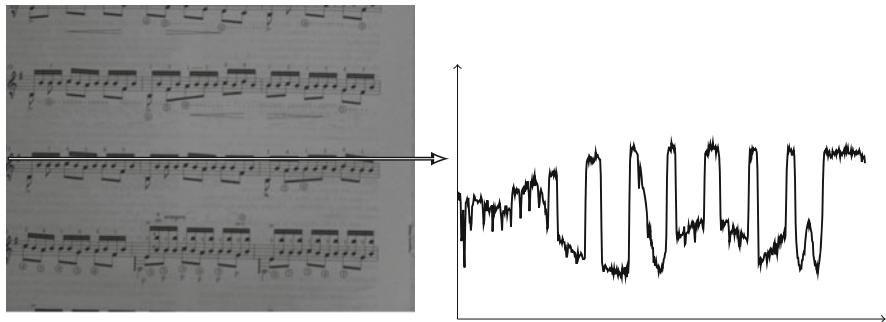


Fig. 1.4 Unfavorable light conditions lead to noisy images. Left: Photo taken in dim light. Right: Gray values along the depicted line

world application one usually deals with several problems and tasks and has to combine and adapt methods or even invent new methods.

Denoising: Digital images contain erroneous information. Modern cameras that can record images with several megapixels still produce noisy images, see Fig. 1.4; in fact, it is usually the case that an increase of resolution also results in a higher noise level. The camera's chip uses the photon count to measure the brightness. Since the emission of photons is fundamentally a random process, the measurement is also a random variable and hence contains some noise. The presence of noise is an inherent problem in imaging. The task of denoising is:

- Identify and remove the noise but at the same time preserve all important information and structure.

Noise does not pose a serious problem for the human eye. We have no problems with images with high noise level, but computers are different. To successfully denoise an image, one needs a good model for the noise and the image. Some reasonable assumptions are, for example:

- The noise is additive.
- The noise is independent of the pixel and comes from some distribution.
- The image consists of piecewise smooth regions that are separated by lines.

In this book we will treat denoising at the following places: □→ Example 3.12, □→ Example 3.25, □→ Sect. 3.4.4, □→ Sect. 3.5, □→ Example 4.19, □→ Example 5.5, □→ Remark 5.21, □→ Example 5.39, □→ Example 5.40, □→ Example 6.1, □→ Application 6.94, and □→ Example 6.124.

Image decomposition: This usually refers to an additive decomposition of an image into different components. The underlying assumption is that an image is

a superposition of different parts, e.g.,

$$\text{image} = \text{cartoon} + \text{texture} + \text{noise}.$$

Here, “cartoon” refers to a rough sketch of the image in which textures and similar components are omitted, and “texture” refers to these textures and other fine structure.

A decomposition of an image can be successful if one has good models for the different components.

In this book we will treat image decomposition at these places: $\square\rightarrow$ Example 4.20, $\square\rightarrow$ Sect. 6.5.

Enhancement, deblurring: Besides noise, there are other errors that may be present in images:

- Blur due to wrong focus: If the focus is not adjusted properly, one point in the image is mapped to an entire region on the film or chip.
- Blur due to camera motion: The object or the camera may move during exposure time. One point of the object is mapped to a line on the film or chip.
- Blur due to turbulence: This occurs, e.g., as “shimmering” of the air above a hot street but also is present in the observation of astronomic objects.
- Blur due to erroneous optics: One of the most famous examples is the Hubble Telescope. Only after the launch of the telescope was it recognized that one mirror had not been made properly. Since a fix in orbit was not considered appropriate at the beginning, elaborate digital methods to correct the resulting errors were developed.

See Fig. 1.5 for illustrations of these defects.

The goal of enhancement is to reduce the blur in images. The more is known about the type of blur, the better. Noise is a severe problem for enhancement and deblurring, since usually, noise is also amplified during deblurring or sharpening. Methods for deblurring are developed at the following places in this book: $\square\rightarrow$ Application 3.24, $\square\rightarrow$ Remark 4.21, $\square\rightarrow$ Example 6.2, $\square\rightarrow$ Application 6.97, and $\square\rightarrow$ Example 6.127.

Edge detection: One key component to the understanding of images is the detection of edges:

- Edges separate different objects or an object from the background.
- Edges help to infer the geometry of a scene.
- Edges describe the shape of an object.

Edges pose different questions:

- How to define an edge mathematically?
- Edges exist at different scales (e.g., fine edges describe the shape of bricks, while coarse edges describe the shape of a house). Which edges are important and should be detected?



Fig. 1.5 Blur in images. Top left: Blur due to wrong focus; top right: motion blur due to camera shake; bottom left: shimmering; bottom right: an image from the Hubble Telescope before the error was corrected

We cover edge detection at the following places: \rightarrow Application 3.23 and \rightarrow Application 5.33.

Segmentation: The goal of segmentation is to decompose an image into different objects. In its simplest form, one object is to be separated from the background. At first glance, this sounds very similar to edge detection. However, in segmentation one focuses on decomposing the whole image into regions, and it may be that different objects are not separated by an edge. However, if all objects are

separated by edges, edge detection is a method for segmentation. In other cases, there are other methods to separate objects without focusing on edges.

There are several problems:

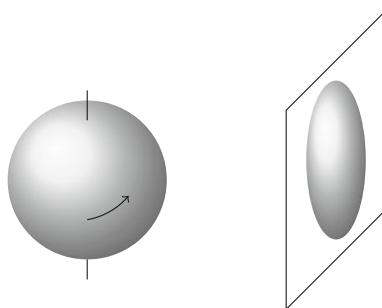
- While object boundaries are easy to see for the human eye, they are not easy to detect automatically if they are not sharp but blurred due to conditions as described above.
- Objects may have different color values, and often these are distorted by noise.
- Edges are distorted by noise, too, and may be rough, even without noise.

Segmentation is treated at the following places in this book: \rightarrow Application 3.11, \rightarrow Application 3.40, and in \rightarrow Sect. 6.5.

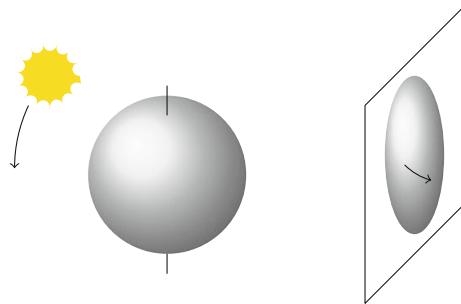
Optical flow computation: Movement is another attribute that can be used to identify and characterize objects and help to understand a scene. If one considers image sequences instead of images one can infer information about movements from digital data.

The movement of an object may result in a change of the gray value of a specific pixel. However, a change of the gray value may also be caused by other means, e.g., a change in illumination. Hence, one distinguishes between the *real field of motion* and the so-called optical flow. The real field of motion of a scene is the projection of the motion in the three-dimensional scene onto the image plane. One aims to extract this information from an image scene. The *optical flow* is the pattern of apparent motion, i.e., the change that can be seen. The real field of motion and the optical flow coincide only in special cases:

- A single-colored ball rotates. The optical flow is zero, but the real field of motion is not.

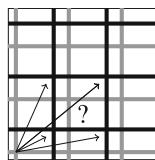


- A ball at rest is illuminated by a moving light source. The real field of motion is zero, but the optical flow is not.

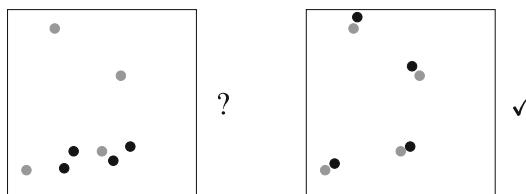


The *correspondence problem* is a consequence of the fact that the optical flow and the real field of motion do not coincide. In some cases different fields of motion may cause the same difference between two images. Also there may be some points in one image that may have moved to more than one place in the other image:

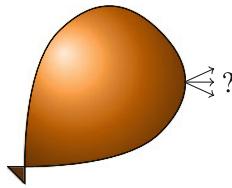
- A regular grid is translated. If we observe only a small part of the grid, we cannot detect a translation that is approximately a multiple of the grid size.



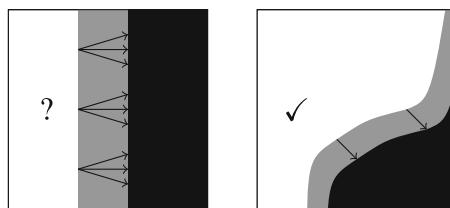
- Different particles move. If the mean distance between particles is larger than their movements, we may find a correspondence; if the movement is too large, we cannot.



- A balloon is inflated. Since the surface of the object increases, one point does not have a single trajectory, but splits up into multiple points.



The *aperture problem* is related to the correspondence problem. Since we see only a part of the whole scene, we may not be able to trace the motion of some object correctly. If a straight edge moves through the picture, we cannot detect any motion along the direction of the edge. Similarly, we are unable to detect whether a circle is rotating. This problem does not occur if the edges of the object have a varying curvature, as illustrated by this picture:



To solve these problems, we make additional assumptions, for example:

- The illumination does not change.
- The objects do change their shape.
- The motion fields are smooth (e.g., differentiable).

In this book we will not treat methods to determine the optical flow. In Chap. 6, however, we will introduce a class of methods that can be adapted for this task, see also \square Sect. 6.5. Moreover, the articles [17, 24, 78] and the book [8] may be consulted.

Registration: In registration one aims to map one image onto another one. This is used in medical contexts, for example: If a patient is examined, e.g., by CT at different times, the images should show the same information (apart from local variations), but the images will not be aligned similarly. A similar problem occurs if the patient is examined with both a CT scan and a scan with magnetic resonance tomography. The images show different information, but may be aligned similarly.

The task in registration is to find a deformation of the image domains such that the content of one image is mapped to the content of the other one. Hence, the

problem is related to that of determining the optical flow. Thus, there are similar problems, but there are some differences:

- Both images may come from different imaging modalities with different properties (e.g., the images may have a different range of gray values or different characteristics of the noise).
- There is no notion of time regularity, since there are only two images.
- In practice, the objects may not be rigid.

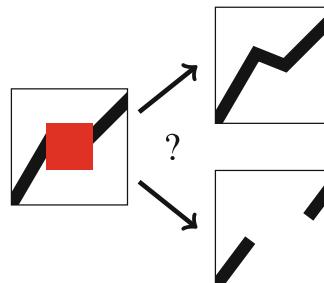
As for optical flow, we will not treat registration in this book. Again, the methods from Chap. 6 can be adapted for this problem, too; see \Rightarrow Sect. 6.5. Moreover, one may consult the book [100].

Restoration (inpainting): Inpainting means the reconstruction of destroyed parts of an image. Reasons for missing parts of an image may be:

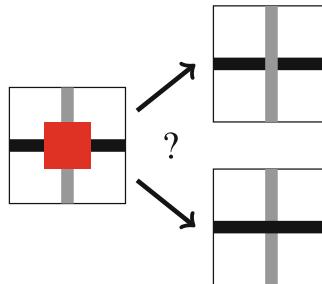
- Scratches in old photos.
- Occlusion of objects by other objects.
- Destroyed artwork.
- Occlusion of an image by text.
- Failure of sensors.
- Errors during transmission of images.

There may be several problems:

- If a line is covered, it is not clear whether there may have been two different, separated, objects.



- If there is an occluded crossing, one cannot tell which line is in front and which is in back.



We are going to treat inpainting in this book at the following places: $\square \rightarrow$ Sect. 5.5, $\square \rightarrow$ Example 6.4, $\square \rightarrow$ Application 6.98, and $\square \rightarrow$ Example 6.128.

Compression: “A picture is worth a thousand words.” However, it needs even more disk space:

$$\begin{array}{lll} 1 \text{ letter} & = & 1 \text{ byte} \\ 1 \text{ word} & \approx & 8 \text{ letters} = 8 \text{ bytes} \\ 1000 \text{ words} & \approx & 8 \text{ KB.} \end{array}$$

$$\begin{array}{lll} 1 \text{ pixel} & = & 3 \text{ bytes} \\ 1 \text{ picture} & \approx & 4,000,000 \text{ pixels} \approx 12 \text{ MB} \end{array}$$

So one picture is worth about 1,500,000 words!

To transmit an uncompressed image with, say, four megapixels via email with an upstream capacity of 128 KB/s, the upload will take about 12 min. However, image data is usually somewhat redundant, and an appropriate compression allows for significant reduction of this time. Several different compression methods have entered our daily lives, e.g. JPEG, PNG, and JPEG2000.

One distinguishes between lossless and lossy compression. Lossless compression allows for a reconstruction of the image that is accurate bit by bit. Lossy compression, on the other hand, allows for a reconstruction of an image that is very similar to the original image. Inaccuracies and artifacts are allowed as long as they are not disturbing to the human observer.

- How to measure the “similarity” of images?
- Compression should work for a large class of images. However, a simple reduction of the color values works well for simple graphics or diagrams, but not for photos.

We will treat compression of images in this book at the following places: $\square \rightarrow$ Sect. 3.1.3, $\square \rightarrow$ Application 4.53, $\square \rightarrow$ Application 4.73, and $\square \rightarrow$ Remark 6.5.

Chapter 2

Mathematical Preliminaries



Abstract Mathematical image processing, as a branch of applied mathematics, is not a self-contained theory of its own, but rather builds on a variety of different fields, such as Fourier analysis, the theory of partial differential equations, and inverse problems. In this chapter, we deal with some of those fundamentals that commonly are beyond the scope of introductory lectures on analysis and linear algebra. In particular, we introduce several notions of functional analysis and briefly touch upon measure theory in order to study classes of Lebesgue spaces. Furthermore, we give an introduction to the theory of weak derivatives as well as Sobolev spaces. The following presentation is of reference character, focusing on the development of key concepts and results, omitting proofs where possible. We also give references for further studies of the respective issues.

Mathematical image processing, as a branch of applied mathematics, is not a self-contained theory of its own, but rather builds on a variety of different fields, such as Fourier analysis, the theory of partial differential equations, and inverse problems. In this chapter, we deal with some of those fundamentals that commonly are beyond the scope of introductory lectures on analysis and linear algebra. In particular, we introduce several notions of functional analysis and briefly touch upon measure theory in order to study classes of Lebesgue spaces. Furthermore, we give an introduction to the theory of weak derivatives as well as Sobolev spaces. The following presentation is of reference character, focusing on the development of key concepts and results, omitting proofs where possible. We also give references for further studies of the respective issues.

2.1 Fundamentals of Functional Analysis

For image processing, mainly those aspects of functional analysis are of interest that deal with function spaces (as mathematically, images are modeled as functions). Later, we shall see that, depending on the space in which an image is contained,

it exhibits different analytical properties. Functional analysis allows us to abstract from concrete spaces and obtain assertions based on these abstractions. For this purpose, the notions of normed spaces, Banach- and Hilbert spaces are essential.

2.1.1 Analysis on Normed Spaces

Let \mathbf{K} denote either the field of real numbers \mathbf{R} or complex numbers \mathbf{C} . For complex numbers, the real part, the imaginary part, the conjugate, and the absolute value are respectively defined by

$$z = a + ib \quad \text{with} \quad a, b \in \mathbf{R} : \quad \operatorname{Re} z = a, \quad \operatorname{Im} z = b, \quad \bar{z} = a - ib, \quad |z| = \sqrt{z\bar{z}}.$$

Definition 2.1 (Normed Space) Let X be a vector space over \mathbf{K} . A function $\|\cdot\| : X \rightarrow [0, \infty[$ is called a *norm* if it exhibits the following properties:

1. $\|\lambda x\| = |\lambda| \|x\|$ for $\lambda \in \mathbf{K}$ and $x \in X$, (positive homogeneity)
2. $\|x + y\| \leq \|x\| + \|y\|$ for $x, y \in X$, (triangle inequality)
3. $\|x\| = 0 \Leftrightarrow x = 0$. (positive definiteness)

The pair $(X, \|\cdot\|)$ is then called a *normed space*.

Two norms $\|\cdot\|_1, \|\cdot\|_2$ on X are called *equivalent* if there exist constants $0 < c < C$ such that

$$c\|x\|_1 \leq \|x\|_2 \leq C\|x\|_1 \quad \text{for all } x \in X.$$

In order to distinguish norms, we may add the name of the underlying vector space to it, for instance, $\|\cdot\|_X$ for the norm on X . It is also common to refer to X itself as the normed space if the norm used is obvious due to the context. Norms on finite-dimensional spaces will be denoted by $|\cdot|$ in many cases. Since in finite-dimensional spaces all norms are equivalent, they play a different role from that in infinite-dimensional spaces.

Example 2.2 (Normed Spaces) Obviously, the pair $(\mathbf{K}, |\cdot|)$ is a normed space. For $N \geq 1$ and $1 \leq p < \infty$,

$$|x|_p = \left(\sum_{i=1}^N |x_i|^p \right)^{1/p} \quad \text{and} \quad |x|_\infty = \max_{i \in \{1, \dots, N\}} |x_i|$$

define equivalent norms on \mathbf{K}^N . The triangle inequality for $|\cdot|_p$ is also known as the *Minkowski inequality*. For $p = 2$, we call $|\cdot|_2$ the *Euclidean vector norm* and normally abbreviate $|\cdot| = |\cdot|_2$.

The norm on a vector space directly implies a topology on X , the *norm topology* or “strong” topology: for $x \in X$ and $r > 0$, define the *open r-ball around x* as

the set

$$B_r(x) = \{y \in X \mid \|x - y\| < r\}.$$

A subset $U \subset X$ is

- *open* if it consists of *interior points* only, i.e., for every $x \in U$, there exists an $\varepsilon > 0$ such that $B_\varepsilon(x) \subset U$,
- a *neighborhood* of $x \in X$ if x is an interior point of U ,
- *closed* if it consists of *limit points* only, i.e., for every $x \in X$ for which for every $\varepsilon > 0$, the sets $B_\varepsilon(x)$ intersect the set U , one has also $x \in U$,
- *compact* if every covering of U by a family of open sets has a finite subcover, i.e.,

$$V_i \text{ open, } i \in I \text{ with } U \subset \bigcup_{i \in I} V_i \quad \Rightarrow \quad \exists J \subset I, J \text{ finite with } U \subset \bigcup_{j \in J} V_j.$$

The fact that the set U is a compact subset of X we denote by $U \subset\subset X$.

Furthermore, let the *interior* of U , abbreviated by $\text{int}(U)$, be the set of all interior points, and the *closure* of U , denoted by \overline{U} , the set of all limit points. The set difference $\partial U = \overline{U} \setminus \text{int}(U)$ is called the *boundary*. For open r -balls, one has

$$\overline{B_r(x)} = \{x \in X \mid \|x - y\| \leq r\},$$

which is why the latter is also referred to as a *closed r -ball*. We say that a subset $U \subset X$ is *dense* in X if $\overline{U} = X$. In particular, X is called *separable* if it possesses a countable and dense subset.

Normed spaces are first countable (i.e., each point has a countable neighborhood base, cf. [122]). That is why we can also describe the terms closed and compact also by means of sequences and their convergence properties:

- We say that a sequence $(x_n) : \mathbb{N} \rightarrow X$ converges to $x \in X$ if $(\|x_n - x\|)$ is a null sequence. This is also denoted by $x_n \rightarrow x$ for $n \rightarrow \infty$ or $x = \lim_{n \rightarrow \infty} x_n$.
- The subset U is closed if and only if for every sequence (x_n) in U with $x_n \rightarrow x$, the limit x lies in U as well (*sequential closedness*).
- The subset U is compact if and only if every sequence (x_n) in U has a convergent subsequence (*sequential compactness*).

For nonempty subsets $V \subset X$, we naturally obtain a topology on V through restriction, which is referred to as the *relative topology*. The notions introduced above result simply through substituting X by the subset V in the respective definitions.

Example 2.3 (Construction of Normed Spaces)

- For finitely many normed spaces $(X_i, \|\cdot\|_{X_i})$, $i = 1, \dots, N$, and a norm $\|\cdot\|$ on \mathbf{R}^N , the product space

$$Y = X_1 \times \dots \times X_N, \quad \|(x_1, \dots, x_N)\|_Y = \|(\|x_1\|_{X_1}, \dots, \|x_N\|_{X_N})\|$$

is a normed space.

- For a subspace U of a normed space $(X, \|\cdot\|_X)$, the pair $(U, \|\cdot\|_X)$ is a normed space again. Its topology corresponds to the relative topology on U .
- If $(X, \|\cdot\|_X)$ is a normed space and $Y \subset X$ is a closed subspace, the quotient vector space

$$X/Y = \{[x] \mid x_1 \sim x_2 \text{ if and only if } x_1 - x_2 \in Y\}$$

can be constructed with the following norm:

$$\|[x]\|_{X/Y} = \inf\{\|x + y\|_X \mid y \in Y\}.$$

These topological constructions lead to a notion of continuity:

Definition 2.4 (Bounded, Continuous, and Closed Functions) A mapping $F : X \supset U \rightarrow Y$ between the normed spaces $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ is *bounded* if there exists $C > 0$ such that $\|F(x)\|_Y \leq C$ for all $x \in U$. The domain U of F is denoted by $\text{dom}(F)$.

The mapping is *continuous in $x \in U$* if for every $\varepsilon > 0$, there exists $\delta > 0$ such that the implication

$$\|x - y\|_X < \delta \quad \Rightarrow \quad \|F(x) - F(y)\|_Y < \varepsilon$$

holds. If F is continuous at every point $x \in U$, we call F simply *continuous*. If, furthermore, δ does not depend on x , then F is *uniformly continuous*.

Finally, we call F *closed* if the graph

$$\text{graph}(F) = \{(x, y) \in X \times Y \mid y = F(x)\} \subset X \times Y$$

is closed.

Continuity in $x \in U$ can equivalently be expressed through *sequential continuity*, i.e., for every sequence (x_n) in U , one has $x_n \rightarrow x \Rightarrow F(x_n) \rightarrow F(x)$ for $n \rightarrow \infty$. The weaker property that F is closed is expressed with sequences as follows: for every sequence (x_n) in U with $x_n \rightarrow x$ such that $F(x_n) \rightarrow y$ for some $y \in Y$, we always have $x \in \text{dom } F$ and $y = F(x)$.

On normed spaces, a stronger notion of continuity is of importance as well:

Definition 2.5 (Lipschitz Continuity) A mapping $F : X \supset U \rightarrow Y$ between the normed spaces $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ is called *Lipschitz continuous* if there

exists a constant $C > 0$ such that for all $x, y \in U$, one has

$$\|F(x) - F(y)\|_Y \leq C\|x - y\|_X.$$

The infimum over all these constants C is called the *Lipschitz constant*.

Sets of uniformly continuous mappings can be endowed with a norm structure:

Definition 2.6 (Spaces of Continuous Mappings) Let $U \subset X$ be a non-empty subset of the normed space $(X, \|\cdot\|_X)$, endowed with the relative topology, and let $(Y, \|\cdot\|_Y)$ be a normed space.

The vector space of continuous mappings we denote by:

$$\mathcal{C}(U, Y) = \{F : U \rightarrow Y \mid F \text{ continuous}\}.$$

The set and the norm

$$\mathcal{C}(\overline{U}, Y) = \{F : U \rightarrow Y \mid F \text{ bounded and uniformly continuous}\}, \quad \|F\|_\infty = \sup_{x \in U} \|F(x)\|_Y$$

form the normed space of the (uniformly) *continuous mappings* on U .

If U and Y are separable, then $\mathcal{C}(\overline{U}, Y)$ is separable as well.

Uniformly continuous mappings on U can always be continuously extended onto \overline{U} , which is indicated by the notation $\mathcal{C}(\overline{U}, Y)$. This notation is quite common in the literature, but it is misused easily: For unbounded U , it is possible that $U = \overline{U}$, but also that $\mathcal{C}(U, Y) \neq \mathcal{C}(\overline{U}, Y)$.

Another important case is that of continuous mappings between two normed spaces that are linear. These mappings are characterized by continuity at the origin, or equivalently, through boundedness on the unit sphere, and they form a normed space themselves. Of course, linear mappings can also be discontinuous; and in this case mappings that are defined on a dense subspace are most interesting.

Definition 2.7 (Linear and Continuous Mappings) Let $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ be normed spaces.

A linear mapping $F : \text{dom } F \rightarrow Y$ that is defined on a subspace $\text{dom } F \subset X$ with $\overline{\text{dom } F} = X$ is called *densely defined*; we call the set $\text{dom } F$ the domain of F .

The notation $\mathcal{L}(X, Y) = \{F : X \rightarrow Y \mid F \text{ linear and continuous}\}$ denotes the set of linear and continuous mappings, which, together with the norm

$$\|F\| = \inf \{M \geq 0 \mid \|Fx\|_Y \leq M\|x\|_X \text{ for all } x \in X\},$$

forms the *space of linear and continuous mappings* between X and Y . We also refer to the norm given on $\mathcal{L}(X, Y)$ as *operator norm*.

Linear and continuous mappings are often also called *bounded* linear mappings. Note that densely defined and continuous mappings can be extended onto the whole of X , which is why densely defined linear mappings are often also called *unbounded*.

Depending on the situation, we will also use the following characterizations of the operator norm:

$$\|F\| = \sup_{\|x\|_X \leq 1} \|Fx\|_Y = \sup_{x \neq 0} \frac{\|Fx\|_Y}{\|x\|_X}.$$

In this context, let us note the following:

- The set $\ker(F) = \{x \in X \mid Fx = 0\}$ denotes the *kernel* of F .
- The *range* of F is given by $\text{rg}(F) = \{Fx \in Y \mid x \in X\}$.
- If F is bijective, the inverse F^{-1} is continuous if and only if there exists $c > 0$ such that

$$c\|x\|_X \leq \|Fx\|_Y \quad \text{for all } x \in X.$$

In this case, we have $\|F^{-1}\| \leq c^{-1}$.

- If F and F^{-1} are continuous, we call F a *linear isomorphism*. If in particular, $\|F\| = \|F^{-1}\| = 1$, then F is an *isometric isomorphism*.

Definition 2.8 (Embeddings) We say that a normed space X is *continuously embedded* into another normed space Y , denoted by $X \hookrightarrow Y$, if:

- $X \subset Y$ (or X can be identified with a subset of Y)
- the identity map $\text{id} : X \rightarrow Y$ is continuous.

In other words

$$X \hookrightarrow Y \iff \exists C > 0 \forall x \in X : \|x\|_Y \leq C\|x\|_X.$$

Normed spaces are also the starting point for the definition of the differentiability of a mapping. Apart from the classical definition of Fréchet differentiability, we will also introduce the weaker notion of Gâteaux differentiability.

Definition 2.9 (Fréchet Differentiability) Let $F : U \rightarrow Y$ be a mapping defined on the open subset $U \subset X$ between the normed spaces $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$. Then, F is *Fréchet differentiable* (or differentiable) at $x \in U$ if there exists $DF(x) \in \mathcal{L}(X, Y)$ such that for every $\varepsilon > 0$, there exists $\delta > 0$ such that

$$0 < \|h\|_X < \delta \Rightarrow x + h \in U \quad \text{and} \quad \frac{\|F(x + h) - F(x) - DF(x)h\|_Y}{\|h\|_X} < \varepsilon.$$

The linear and continuous mapping $DF(x)$ is also called the *(Fréchet) derivative* at the point x .

If F is differentiable at every point $x \in U$, then F is called *(Fréchet) differentiable*, and $DF : U \rightarrow \mathcal{L}(X, Y)$, given by $DF : x \mapsto DF(x)$, denotes the *(Fréchet) derivative*. If DF is continuous, we call F *continuously differentiable*.

Since $\mathcal{L}(X, Y)$ is also a normed space, we can consider the differentiability of DF as well: If the second derivative exists at a point x , we denote it by $D^2F(x)$ and note that it lies in the space $\mathcal{L}(X, \mathcal{L}(X, Y))$, which exactly corresponds to the continuous bilinear mappings $G : X \times X \rightarrow Y$ (cf. [51]). Together with the notion of k -linearity or multilinearity, we can formulate the analogue for $k \geq 1$:

$$\underbrace{\mathcal{L}(X, \dots, \mathcal{L}(X, Y) \dots)}_{k\text{-times}} \sim \mathcal{L}^k(X, Y) = \{G : \underbrace{X \times \dots \times X}_{k\text{-times}} \rightarrow Y \mid G \text{ } k\text{-linear and continuous}\},$$

$$\|G\| = \inf \left\{ M > 0 \mid \|G(x_1, \dots, x_k)\|_Y \leq M \prod_{i=1}^k \|x_i\|_X \text{ for all } (x_1, \dots, x_k) \in X^k \right\},$$

where the latter norm coincides with the respective operator norm. Usually, we regard the k th derivative as an element in $\mathcal{L}^k(X, Y)$ (the space of k -linear continuous mappings), or equivalently, as a mapping $D^k F : U \rightarrow \mathcal{L}^k(X, Y)$. If such a derivative exists in $x \in U$ and is continuous at this point, then $D^k F(x)$ is symmetric.

Example 2.10 (Differentiable Mappings)

- A linear and continuous mapping $F \in \mathcal{L}(X, Y)$ is infinitely differentiable with itself as the first derivative and 0 as every higher derivative.
- On \mathbf{K}^N , every polynomial is infinitely differentiable.
- Functions on $U \subset \mathbf{K}^N$ that possess continuous partial derivatives are also continuously differentiable.

In the case of functions, i.e., $X = \mathbf{R}^N$ and $Y = \mathbf{K}$, the following notations are common:

$$\nabla F = \left(\frac{\partial F}{\partial x_1} \cdots \frac{\partial F}{\partial x_N} \right), \quad \nabla^2 F = \begin{pmatrix} \frac{\partial^2 F}{\partial x_1^2} & \cdots & \frac{\partial^2 F}{\partial x_1 \partial x_N} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 F}{\partial x_N \partial x_1} & \cdots & \frac{\partial^2 F}{\partial x_N^2} \end{pmatrix}.$$

By means of that notation and under the assumption that the respective partial derivatives are continuous, the Fréchet derivatives can be represented by matrix vector products

$$DF(x)y = \nabla F(x)y, \quad D^2F(x)(y, z) = z^T \nabla^2 F(x)y.$$

For higher derivatives, there are similar summation notations. In the former situation, the vector field ∇F is called the *gradient* of F , while the matrix-valued mapping $\nabla^2 F$ is, slightly abusively, referred to as the *Hessian matrix*. In the case of the gradient, it has become common in the literature not to distinguish between a row vector and a column vector—in the sense that one can multiply the gradient at a point by a matrix from the left as well. We will also make use of this fact as long as ambiguities are impossible, but we will point this out again at suitable locations.

Finally, for $U \subset \mathbf{R}^N$ and $F : U \rightarrow \mathbf{K}^M$, we introduce the notion of the *Jacobian matrix*, for which, in the case of continuity of the partial derivatives, one has:

$$\nabla F = \begin{pmatrix} \frac{\partial F_1}{\partial x_1} & \cdots & \frac{\partial F_1}{\partial x_N} \\ \vdots & \ddots & \vdots \\ \frac{\partial F_M}{\partial x_1} & \cdots & \frac{\partial F_M}{\partial x_N} \end{pmatrix}, \quad DF(x)y = \nabla F(x)y.$$

Apart from that, let us introduce two specific and frequently used differential operators: For a vector field $F : U \rightarrow \mathbf{K}^N$ with $U \subset \mathbf{R}^N$ a nonempty open subset, the function

$$\operatorname{div} F = \operatorname{trace} \nabla F = \sum_{i=1}^N \frac{\partial F_i}{\partial x_i}$$

is called the *divergence* and the associated operator div the *divergence operator*. For functions $F : U \rightarrow \mathbf{K}$, the operator Δ with

$$\Delta F = \operatorname{trace} \nabla^2 F = \sum_{i=1}^N \frac{\partial^2 F}{\partial x_i^2}$$

denotes the *Laplace operator*.

In order to keep track of higher-order partial derivatives of a function, one often uses the so-called multi-index notation. A multi-index is given by $\alpha \in \mathbf{N}^d$ and for $\alpha = (\alpha_1, \dots, \alpha_d)$, we write

$$\frac{\partial^\alpha}{\partial x^\alpha} = \frac{\partial^{\alpha_1}}{\partial x^{\alpha_1}} \cdots \frac{\partial^{\alpha_d}}{\partial x^{\alpha_d}}.$$

We will also use the notation

$$\partial^\alpha = \frac{\partial^\alpha}{\partial x^\alpha}.$$

By $|\alpha| = \sum_{k=1}^d \alpha_k$, we denote the *order* of the multi-index. By means of multi-indices, we can, for instance, formulate the *Leibniz rule* for the higher-order derivatives of a product in a compact fashion: with $\alpha! = \prod_{k=1}^d \alpha_k!$ and $\binom{\alpha}{\beta} = \frac{\alpha!}{\beta!(\alpha-\beta)!}$ for $\beta \leq \alpha$ (i.e., $\beta_k \leq \alpha_k$ for $1 \leq k \leq d$), one has

$$\partial^\alpha (f g) = \sum_{\beta \leq \alpha} \binom{\alpha}{\beta} (\partial^{\alpha-\beta} f)(\partial^\beta g).$$

Let us finally introduce a weaker variant of differentiability than the Fréchet differentiability:

Definition 2.11 (Gâteaux Differentiability) A mapping $F : X \supset U \rightarrow Y$ of a non-empty subset U between the normed spaces $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ is *Gâteaux differentiable* at $x \in U$ if there exists $DF(x) \in \mathcal{L}(X, Y)$ such that for every $y \in X$, the mapping $F_{x,y} : \lambda \mapsto F(x + \lambda y)$, defined on a neighborhood of $0 \in \mathbf{K}$, is differentiable at 0 and the (*Gâteaux*) derivative satisfies $DF_{x,y} = DF(x)y$. In particular, the mapping F is called Gâteaux differentiable if it is Gâteaux differentiable at every $x \in U$.

The essential difference between Gâteaux and Fréchet differentiability is the quality of approximation of the respective derivative: In the case of Fréchet differentiability, the linearization $F(x) + DF(x)(\cdot - x)$ approximates better than $\varepsilon \|\cdot - x\|_X$ for every $\varepsilon > 0$ and uniformly in a neighborhood of x . For Gâteaux differentiability, this holds only for every direction.

2.1.2 Banach Spaces and Duality

Apart from the concept of normed spaces, the notion of *completeness*, i.e., the existence of limits of Cauchy sequences, is essential for various fundamental results.

Definition 2.12 (Banach Space) A normed space $(X, \|\cdot\|_X)$ is *complete* if every Cauchy sequence converges in X , i.e., for every sequence (x_n) in X with the property

for all $\varepsilon > 0$, there exists $n_0 \in \mathbf{N}$ such that for all $n, m \geq n_0$, one has: $\|x_n - x_m\|_X < \varepsilon$,

there exists some $x \in X$ with $x_n \rightarrow x$ for $n \rightarrow \infty$.

A complete normed space is called a *Banach space*.

Example 2.13 (Banach Spaces) The field $(\mathbf{K}, |\cdot|)$ is a Banach space, and due to that fact, all $(\mathbf{K}^N, |\cdot|_p)$ of Example 2.2 are Banach spaces as well. Furthermore, $\mathcal{L}(X, Y)$ is a Banach space if $(Y, \|\cdot\|_Y)$ is one.

In Banach spaces, several fundamental assertions hold that link pointwise properties to global ones (cf. [22, 122]).

Theorem 2.14 (Baire Category Theorem) Let X be a Banach space and (A_n) a sequence of closed sets such that $\bigcup_n A_n = X$. Then there exists n_0 such that the set A_{n_0} contains an open ball.

Theorem 2.15 (Banach Steinhaus or Uniform Boundedness Principle) *Let X be a Banach space, Y a normed space and $(F_i) \subset \mathcal{L}(X, Y)$, $i \in I \neq \emptyset$, a family of linear and continuous mappings. Then*

$$\sup_{i \in I} \|F_i x\|_Y < \infty \quad \text{for all } x \in X \quad \Rightarrow \quad \sup_{i \in I} \|F_i\| < \infty.$$

Theorem 2.16 (Open/Inverse Mapping Theorem) *A linear and continuous mapping $F \in \mathcal{L}(X, Y)$ between the Banach spaces X and Y is open if and only if it is surjective.*

In particular, a bijective, linear and continuous mapping F between Banach spaces always possesses a continuous inverse F^{-1} .

We now turn to a construction that is important in functional analysis, the *dual space* $X^* = \mathcal{L}(X, \mathbf{K})$ associated to X . According to Example 2.13, X^* is always a Banach space, since \mathbf{K} is a Banach space. The norm on X^* is due to Definition 2.7 or to one of the characterizations as

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

The subspaces of X and X^* can be related in the following way: For a subspace $U \subset X$, the *annihilator* is defined as the set

$$U^\perp = \{x^* \in X^* \mid x^*(x) = 0 \text{ for all } x \in U\} \quad \text{in } X^*,$$

and for the subspace $V \subset X^*$, the definition reads

$$V^\perp = \{x \in X \mid x^*(x) = 0 \text{ for all } x^* \in V\} \quad \text{in } X.$$

The sets U^\perp and V^\perp are always closed subspaces. Annihilators are used for the characterizations of dual spaces of subspaces.

Example 2.17 (Dual Spaces)

1. The dual space X^* of an N -dimensional normed space X is again N -dimensional, i.e., equivalent to itself. In particular, $(\mathbf{K}^N, \|\cdot\|)^* = (\mathbf{K}^N, \|\cdot\|_*)$, where $\|\cdot\|_*$ is the norm dual to $\|\cdot\|$.
2. The dual space of $Y = X_1 \times \cdots \times X_N$ of Example 2.3 can be regarded as

$$Y^* = X_1^* \times \cdots \times X_N^*, \quad \|(x_1^*, \dots, x_N^*)\|_{Y^*} = \|(\|x_1^*\|_{X_1^*}, \dots, \|x_N^*\|_{X_N^*})\|_*$$

with the norm $\|\cdot\|_*$ dual to $\|\cdot\|$.

3. The dual space of a subspace $U \subset X$ is given by the quotient space $U^* = X^*/U^\perp$ endowed with the quotient norm; cf. Example 2.3.

The consideration of the dual space is the basis for the notion of weak convergence, for instance, and in many cases, X^* reflects important properties of the *predual space* X . It is common to regard the application of elements in X^* to elements in X as a bilinear mapping called *duality pairing*:

$$\langle \cdot, \cdot \rangle_{X^* \times X} : X^* \times X \rightarrow \mathbf{K}, \quad \langle x^*, x \rangle_{X^* \times X} = x^*(x).$$

The subscript is often omitted if the spaces are evident due to the context. Of course, one can iterate the generation of the dual space, the next space being the *bidual space* X^{**} . This space naturally contains X , and the *canonical injection* is given by

$$J : X \rightarrow X^{**}, \quad \langle J(x), x^* \rangle_{X^{**} \times X^*} = \langle x^*, x \rangle_{X^* \times X}, \quad \|J(x)\|_{X^{**}} = \|x\|_X.$$

The latter equality of the norms is a consequence of the Hahn-Banach extension theorem, which in a rephrased statement, reads

$$\begin{aligned} \|x\|_X &= \inf \{L \geq 0 \mid |\langle x^*, x \rangle| \leq L \|x^*\|_{X^*} \forall x^* \in X^*\} \\ &= \sup_{\|x^*\|_{X^*} \leq 1} |\langle x^*, x \rangle| = \sup_{x^* \neq 0} \frac{|\langle x^*, x \rangle|}{\|x^*\|_{X^*}}. \end{aligned}$$

Hence, the bidual space is always at least as large as the original space. We can now consider the closure of $J(X)$ in X^{**} and naturally obtain a Banach space that contains X in a certain sense.

Theorem 2.18 (Completion of Normed Spaces) *For every normed space $(X, \|\cdot\|_X)$, there exists a Banach space $(\tilde{X}, \|\cdot\|_{\tilde{X}})$ and a norm-preserving mapping $J : X \rightarrow \tilde{X}$ such that $J(X)$ is dense in \tilde{X} .*

Often, one identifies $X \subset \tilde{X}$ and considers the completed space \tilde{X} instead of X . A completion can also be constructed by taking equivalence classes of Cauchy sequences; according to the inverse mapping theorem, this procedure yields an equivalent Banach space.

The notion of the bidual space is also essential in another context: If the injection $J : X \rightarrow X^{**}$ is surjective, X is called *reflexive*. Reflexive spaces play a particular role in the context of weak convergence:

Definition 2.19 (Weak Convergence, Weak*-Convergence) A sequence (x_n) in a normed space $(X, \|\cdot\|_X)$ converges *weakly* to some $x \in X$ if for every $x^* \in X^*$, one has

$$\lim_{n \rightarrow \infty} \langle x^*, x_n \rangle_{X^* \times X} = \langle x^*, x \rangle_{X^* \times X}.$$

In this case, we write $x_n \rightharpoonup x$ for $n \rightarrow \infty$.

Analogously, a sequence (x_n^*) in X^* converges in the *weak** sense to some $x^* \in X^*$ if

$$\lim_{n \rightarrow \infty} \langle x_n^*, x \rangle_{X^* \times X} = \langle x^*, x \rangle_{X^* \times X}$$

for every $x \in X$, which we also denote by $x_n^* \xrightarrow{*} x^*$ for $n \rightarrow \infty$.

Note that the definition coincides with the notion of the convergence of sequences in the weak or weak* topology, respectively. However, we will not go into further detail here. While the convergence in the norm sense implies convergence in the weak sense, the converse holds only in finite-dimensional spaces. Also, in the dual space X^* , weak* convergence is in general a weaker property than weak convergence; for reflexive spaces, however, the two notions coincide. According to Theorem 2.15 (Banach Steinhaus), weakly or weak*-convergent sequences, respectively, are at least still bounded, i.e., $x_n \rightharpoonup x$ for $n \rightarrow \infty$ implies $\sup_n \|x_n\|_X < \infty$ and, the analogue holds for weak* convergence.

Of course, notions such as continuity and closedness of mappings can be generalized for these types of convergence.

Definition 2.20 (Weak/Weak* Continuity, Closedness) Let X, Y be normed spaces and $U \subset X$ a nonempty subset.

- A mapping $F : U \rightarrow Y$ is called *{strongly, weakly}-{strongly, weakly continuous}* if the {strong, weak} convergence of a sequence (x_n) to some $x \in U$ implies the {strong, weak} convergence of $(F(x_n))$ to $F(x)$.
- The mapping is *{strongly, weakly}-{strongly, weakly} closed* if the {strong, weak} convergence of (x_n) to some $x \in X$ and the {strong, weak} convergence of $(F(x_n))$ to some $y \in Y$ imply: $x \in U$ and $y = F(x)$.

In the case that X or Y is a dual space, the corresponding weak* terms are defined analogously with weak* convergence.

One of the main reasons to study these types of convergences is compactness results, whose assertions are similar to the Heine Borel theorem in the finite-dimensional case.

- A subset $U \subset X$ is *weakly sequentially compact* if every sequence in U possesses a weakly convergent subsequence with limit in U .
- We say that a subset $U \subset X^*$ is *weak*-sequentially compact* if the analogue holds for weak* convergence.

Theorem 2.21 (Banach-Alaoglu for Separable Spaces) *Every closed ball in the dual space of a separable normed space is weak*-sequentially compact.*

Theorem 2.22 (Eberlein Šmulyan) *A normed space is reflexive if and only if every closed ball is weakly sequentially compact.*

Dual spaces and weak and weak* convergence can naturally be used in connection with linear and continuous mappings as well. Corresponding examples are the adjoint as well as the notions of weak and weak* sequential continuity.

Definition 2.23 (Adjoint Mapping) For $F \in \mathcal{L}(X, Y)$,

$$\langle F^*y^*, x \rangle_{X^*\times X} = \langle y^*, Fx \rangle_{Y^*\times Y}, \quad \text{for all } x \in X, y^* \in Y^*,$$

defines the *adjoint mapping* $F^* \in \mathcal{L}(Y^*, X^*)$.

Remark 2.24

- If F is linear and continuous, then F^* is linear and continuous. Furthermore, $\|F^*\| = \|F\|$; i.e., taking the adjoint is a linear, continuous, and norm-preserving mapping.
- If $\text{rg}(F)$ is dense in Y , then the mapping F^* is injective. Conversely, if F is injective, then $\text{rg}(F^*)$ is dense.
- Every mapping $F \in \mathcal{L}(X, Y)$ is also *weakly sequentially continuous*, since for (x_n) in X with $x_n \rightharpoonup x$ and arbitrary $y^* \in Y^*$, one has

$$\langle y^*, Fx_n \rangle_{Y^*\times Y} = \langle F^*y^*, x_n \rangle_{X^*\times X} \rightarrow \langle F^*y^*, x \rangle_{X^*\times X} = \langle y^*, Fx \rangle_{Y^*\times Y},$$

i.e., $Fx_n \rightharpoonup Fx$ in Y . Analogously, we infer that an adjoint mapping $F^* \in \mathcal{L}(Y^*, X^*)$ is *weak*-sequentially continuous*, i.e., $y_n^* \xrightarrow{*} y^*$ implies $F^*y_n^* \xrightarrow{*} F^*y^*$.

The adjoint can also be defined for densely defined, unbounded mappings:

Definition 2.25 (Adjoint of Unbounded Mappings) Let $\text{dom } F \subset X$ be a dense subspace of a normed space $(X, \|\cdot\|_X)$ and $F : \text{dom } F \rightarrow Y$ a linear mapping in a normed space $(Y, \|\cdot\|_Y)$. Then, the mapping $F^* : \text{dom } F^* \rightarrow X^*$, defined on

$$\text{dom } F^* = \{y^* \in Y^* \mid x \mapsto \langle y^*, Fx \rangle_{Y^*\times Y} \text{ is continuous on } \text{dom } F\} \subset Y^*$$

and satisfying that F^*y^* is the extension of $x \mapsto \langle y^*, Fx \rangle_{Y^*\times Y}$ onto the whole of X , is also called the *adjoint*.

For reflexive Banach spaces, the adjoint is densely defined and closed. For a densely defined and closed mapping F between X and Y , the following relations between the kernel and the range hold:

$$\ker(F) = (\text{rg}(F^*))^\perp \text{ in } X \quad \text{and} \quad \ker(F^*) = (\text{rg}(F))^\perp \text{ in } Y^*.$$

The circumstances under which these identities hold for interchanged annihilators are given in the following theorem on linear mappings with a closed range.

Theorem 2.26 (Closed Range Theorem) *Let X, Y be Banach spaces and $F : X \supset \text{dom } F \rightarrow Y$ a densely defined, closed mapping. Then the following assertions are equivalent:*

1. $\text{rg}(F)$ is closed in Y ,
2. $\text{rg}(F^*)$ is closed in X^* ,
3. $\ker(F)^\perp = \text{rg}(F^*)$ in X^* ,
4. $\ker(F^*)^\perp = \text{rg}(F)$ in Y .

An important class of operators, which in the linear and continuous case relates weak and strong convergence, is that of compact mappings:

Definition 2.27 (Compact Mappings) A mapping $F : X \rightarrow Y$ between two normed spaces X and Y is called *compact* if $F(U)$ is relatively compact for every bounded $U \subset X$.

The set of linear, continuous, and compact mappings is denoted by $\mathcal{K}(X, Y)$.

In the context of embeddings of Banach spaces X and Y , we also say that X is compactly embedded into Y if for the identity map $\text{id} : X \rightarrow Y$ of Definition 2.8, one has $\text{id} \in \mathcal{K}(X, Y)$.

In functional analysis, there exists an extensive theory on linear, continuous, and compact mappings, since results on linear mappings between finite-dimensional spaces that do not hold true for general linear and continuous mappings can be transferred in this case. For this subject, we refer to [22, 122] and only mention some elementary properties:

- The space $\mathcal{K}(X, Y)$ is a closed subspace of $\mathcal{L}(X, Y)$. Every linear mapping with finite-dimensional range is compact.
- Elements in $\mathcal{K}(X, Y)$ are weakly-strongly continuous or *completely continuous*. If X is reflexive, a linear and continuous mapping is compact if and only if it is weakly-strongly continuous.
- An analogous statement can be found for weak* convergence, dual spaces X^* and Y^* with X separable, and adjoints of linear and continuous mappings, i.e., for F^* with $F \in \mathcal{L}(Y, X)$.

Finally, let us touch upon the role of dual spaces in the separation of convex sets.

Definition 2.28 A subset $A \subset X$ of a normed space X is called *convex* if for every $x, y \in A$ and $\lambda \in [0, 1]$, one has $\lambda x + (1 - \lambda)y \in A$.

Theorem 2.29 (Hahn-Banach Separation Theorems) *Let $A, B \subset X$ be nonempty, disjoint, convex subsets of the normed space $(X, \|\cdot\|_X)$.*

1. *If A is open, then there exist $x^* \in X^*$, $x^* \neq 0$, and $\lambda \in \mathbf{R}$ such that*

$$\operatorname{Re} \langle x^*, x \rangle \leq \lambda \quad \text{for all } x \in A \quad \text{and} \quad \operatorname{Re} \langle x^*, x \rangle \geq \lambda \quad \text{for all } x \in B.$$

2. If A is closed and B is compact, then there exist $x^* \in X^*$, $\lambda \in \mathbf{R}$, and $\varepsilon > 0$ such that

$$\operatorname{Re} \langle x^*, x \rangle \leq \lambda - \varepsilon \quad \text{for all } x \in A \quad \text{and} \quad \operatorname{Re} \langle x^*, x \rangle \geq \lambda + \varepsilon \quad \text{for all } x \in B.$$

In the case above, the set $\{x \in X \mid \operatorname{Re} \langle x^*, x \rangle = \lambda\}$ is interpreted as a closed hyperplane that separates the sets A and B . Hence, dual spaces contain enough elements to separate two different points, for instance, but also points from closed sets, etc.

2.1.3 Aspects of Hilbert Space Theory

The concept of a Banach space is very general. Thus, it is not surprising that many desired properties (such as reflexivity, for instance) do not hold in general and have to be required separately. In the case of a Hilbert space, however, one has additional structure at hand due to the inner product, which naturally yields several properties. Let us give a brief summary of the most important of these properties.

Definition 2.30 (Inner Product) Let X be a \mathbf{K} -vector space. A mapping $(\cdot, \cdot)_X : X \times X \rightarrow \mathbf{K}$ is called an *inner product* if it satisfies:

1. $(\lambda_1 x_1 + \lambda_2 x_2, y)_X = \lambda_1 (x_1, y)_X + \lambda_2 (x_2, y)_X$ for $x_1, x_2, y \in X$ and $\lambda_1, \lambda_2 \in \mathbf{K}$,
(linearity)
2. $(x, y)_X = \overline{(y, x)_X}$ for $x, y \in X$,
(Hermitian symmetry)
3. $(x, x)_X \geq 0$ and $(x, x)_X = 0 \Leftrightarrow x = 0$.
(positive definiteness)

An inner product on X induces a norm by $\|x\|_X = \sqrt{(x, x)_X}$, which satisfies the *Cauchy-Schwarz inequality*:

$$|(x, y)_X| \leq \|x\|_X \|y\|_X \quad \text{for all } x, y \in X.$$

Based on that, one can easily deduce the continuity of the inner product as well.

For a \mathbf{K} -vector space with inner product and the associated normed space $(X, \|\cdot\|_X)$, the terms *inner product space* and *pre-Hilbert space* are common. We also denote this space by $(X, (\cdot, \cdot)_X)$. Together with the notion of completeness, we are lead to the following definition.

Definition 2.31 (Hilbert Space) A *Hilbert space* is a complete pre-Hilbert space $(X, (\cdot, \cdot)_X)$. Depending on $\mathbf{K} = \mathbf{R}$ or $\mathbf{K} = \mathbf{C}$, it is called a *real* or a *complex* Hilbert space, respectively.

Example 2.32 For $N \geq 1$, the set \mathbf{K}^N is a finite-dimensional Hilbert space if the inner product is chosen as

$$(x, y)_2 = \sum_{i=1}^N x_i \overline{y_i}.$$

We call this inner product the *Euclidean inner product* and also write $x \cdot y = (x, y)_2$.

Analogously, the set $\ell^2 = \{x : \mathbf{N} \rightarrow \mathbf{K} \mid \sum_{i=1}^{\infty} |x_i|^2 < \infty\}$, endowed with the inner product

$$(x, y)_2 = \sum_{i=1}^{\infty} x_i \overline{y_i},$$

yields an infinite-dimensional separable Hilbert space. (Note that the inner product is well defined as a consequence of the Cauchy-Schwarz inequality.)

For (pre-)Hilbert spaces, the notion of *orthogonality* is characteristic:

Definition 2.33 (Orthogonality) Let $(X, (\cdot, \cdot)_X)$ be a pre-Hilbert space.

- Two elements $x, y \in X$ are called *orthogonal* if $(x, y)_X = 0$, which we also denote by $x \perp y$. A set $U \subset X$ whose elements are mutually orthogonal is called an *orthogonality system*.
- The subspaces $U, V \subset X$ are *orthogonal*, denoted by $U \perp V$, if orthogonality holds for every pair $(x, y) \in U \times V$. For a subspace $W \subset X$ with $W = U + V$, we also write $W = U \perp V$.
- For a subspace $U \subset X$, the subspace of all vectors

$$U^\perp = \{y \in X \mid x \perp y \text{ for all } x \in U\}$$

is called the *orthogonal complement* of U .

If $x, y \in X$ are orthogonal, then the norm satisfies the *Pythagorean theorem*

$$\|x + y\|_X^2 = \|x\|_X^2 + \|y\|_X^2.$$

The analogous assertion on the square of norms of sums remains true for finite orthogonal systems as well as countable orthogonal systems whose series converge in X .

The orthogonal complement U^\perp is closed and for closed subspaces U in the Hilbert space X , one has $X = U \perp U^\perp$. That implies the existence of the *orthogonal projection* on U : if $(X, (\cdot, \cdot)_X)$ is a Hilbert space and $U \subset X$ is a closed subspace, then there exists a unique $P \in \mathcal{L}(X, X)$ with

$$P^2 = P, \quad \operatorname{rg}(P) = U, \quad \text{and} \quad \ker(P) = U^\perp.$$

Obviously, $Q = I - P$ represents the orthogonal projection onto U^\perp .

The notion of orthogonality is the foundation for orthonormal systems as well as orthonormal bases.

Definition 2.34 (Orthonormal System)

- A subset U of a pre-Hilbert space $(X, (\cdot, \cdot)_X)$ is called an *orthonormal system* if $(x, y)_X = \delta_{x,y}$ for all $x, y \in U$.
- An orthonormal system is called *complete* if there is no orthonormal system that contains U as a strict subset.
- An at most countable and complete orthonormal system is also referred to as an *orthonormal basis*.

For orthonormal systems $U \subset X$, *Bessel's inequality* holds:

$$x \in X : \quad \sum_{y \in U} |(x, y)_X|^2 \leq \|x\|_X^2,$$

where $(x, y) \neq 0$ is true for at most countably many $y \in U$, i.e., the sum is to be interpreted as a convergent series.

If U is complete, then equality holds as long as X is a Hilbert space. This relationship is called *Parseval's identity*:

$$x \in X : \quad \|x\|_X^2 = \sum_{y \in U} |(x, y)_X|^2.$$

The latter can also be interpreted as a special case of the *Parseval identity*:

$$x_1, x_2 \in X : \quad (x_1, x_2)_X = \sum_{y \in U} (x_1, y)_X \overline{(x_2, y)_X}.$$

One can show that in every Hilbert space, there exists a complete orthonormal system (cf. [22]). Furthermore, a Hilbert space X is separable if and only if there exists an orthonormal basis in X . Due to the Parseval identity, every separable Hilbert space is thus isometrically isomorphic to either ℓ^2 or \mathbf{K}^N for some $N \geq 0$. In particular, the Parseval relation implies that every sequence of orthonormal vectors (x^n) weakly converges to zero.

Let us finally consider the dual spaces of Hilbert spaces. Since the inner product is continuous, for every $y \in X$ we obtain an element $J_X y \in X^*$ by means of $\langle J_X y, x \rangle_{X^* \times X} = (x, y)_X$. The mapping $J_X : X \rightarrow X^*$ is *semi linear*, i.e., we have

$$J_X(\lambda_1 x_1 + \lambda_2 x_2) = \overline{\lambda_1} J_X x_1 + \overline{\lambda_2} J_X x_2 \quad \text{for all } x_1, x_2 \in X \text{ and } \lambda_1, \lambda_2 \in \mathbf{K}.$$

The remarkable property of a Hilbert space is now that the range of J_X is the whole of the dual space:

Theorem 2.35 (Riesz Representation Theorem) *Let $(X, (\cdot, \cdot)_X)$ be a Hilbert space. For every $x^* \in X^*$, there exists $y \in X$ with $\|y\|_X = \|x^*\|_{X^*}$ such that*

$$\langle x^*, x \rangle_{X^* \times X} = (x, y)_X \quad \text{for all } x \in X.$$

The mapping $J_X^{-1} : X^* \rightarrow X$ that assigns $y \in X$ to the $x^* \in X^*$ above is also called *Riesz mapping*. It is semi linear and norm-preserving, which yields that X and X^* are isometrically isomorphic in this sense. Immediately, this implies that a Hilbert space is reflexive.

If X and Y are Hilbert spaces, then for every $F \in \mathcal{L}(X, Y)$, we can additionally construct the so-called *Hilbert space adjoint* $F^* = J_X^{-1}F^*J_Y \in \mathcal{L}(Y, X)$ by means of the Riesz mapping. Obviously, this mapping can equivalently be defined by

$$(Fx, y)_Y = (x, F^*y)_X \quad \text{for all } x \in X, y \in Y.$$

If $F^* = F$, then the mapping F is called *self-adjoint*.

We will often use the identification of a Hilbert space with its dual without further notice and simply write $F^* = F^* \in \mathcal{L}(Y, X)$, for instance, as long as this is evident from the context.

2.2 Elements of Measure and Integration Theory

For our purposes, it is mainly the functional-analytic aspects of measure and integration theory that are of interest: On the one hand, the function spaces associated to the Lebesgue integral exhibit several “good” properties. On the other hand, they contain objects that are of interest for image processing: classical notions, such as the continuity of functions, are too restrictive for images, since jumps in the function values may appear. Additionally, noise is generally regarded as a discontinuous perturbation. The Lebesgue spaces comply with such assumptions and at the same time, they provide a suitable analytical structure.

2.2.1 Measure and Integral

The notion of the Lebesgue integral is based on measuring the contents of sets by means of a so-called measure.

Definition 2.36 (Measurable Space) Let Ω be a nonempty set. A family of subsets \mathfrak{F} is a σ -algebra if

1. $\emptyset \in \mathfrak{F}$,
2. for all $A \in \mathfrak{F}$, one has $\Omega \setminus A \in \mathfrak{F}$,
3. $A_i \in \mathfrak{F}, i \in \mathbb{N}$, implies that $\bigcup_{i \in \mathbb{N}} A_i \in \mathfrak{F}$.

The pair (Ω, \mathfrak{F}) is called a *measurable space*, and the sets in \mathfrak{F} are called *measurable*. The smallest σ -algebra that contains a family of subsets \mathfrak{G} is the σ -algebra induced by \mathfrak{G} . For a topological space Ω , the σ -algebra induced by the open sets is called a *Borel algebra*, denoted by $\mathfrak{B}(\Omega)$.

Roughly speaking, in a σ -algebra, taking the complement as well as (countable) intersections and unions of sets are allowed. In particular, a Borel algebra contains all open, all closed, and all compact sets. We consider σ -algebras, since they are a suitable class of systems of sets whose elements can be measured or integrated.

Definition 2.37 (Positive Measure) A *measure* on a measurable space (Ω, \mathfrak{F}) is a function $\mu : \mathfrak{F} \rightarrow [0, \infty]$ with the following properties:

1. $\mu(\emptyset) = 0$,
2. $A_i \in \mathfrak{F}, i \in \mathbf{N}$, mutually disjoint implies that

$$\mu\left(\bigcup_{i \in \mathbf{N}} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i).$$

If $\mu(\Omega) < \infty$, then μ is a *finite measure*; in the special case that $\mu(\Omega) = 1$, it is called a *probability measure*. If there exists a sequence (A_i) in \mathfrak{F} for which $\mu(A_i) < \infty$ for all $i \in \mathbf{N}$ as well as $\Omega = \bigcup_{i \in \mathbf{N}} A_i$, then the measure is called σ -*finite*.

The triple $(\Omega, \mathfrak{F}, \mu)$ is called a *measure space*.

Often, the concrete measurable space to which a measure is associated to, is of interest: in the case $\mathfrak{F} = \mathcal{B}(\Omega)$, we speak of a *Borel measure*; if additionally one has $\mu(K) < \infty$ for all compact sets K , we call μ a *positive Radon measure*.

Example 2.38 (Measures)

1. On every nonempty set Ω , a measure is defined on the power set $\mathfrak{F} = \mathfrak{P}(\Omega)$ in an obvious manner:

$$\mu(A) = \begin{cases} \text{card}(A) & \text{if } A \text{ is finite,} \\ \infty & \text{otherwise.} \end{cases}$$

This measure is called the *counting measure* on Ω . One can restrict it to a Borel measure, but in general, it is not a positive Radon measure: in the important special case of the standard topology on \mathbf{R}^d , there are compact sets with infinite measure.

2. The following example is of a similar nature: For $\Omega \subset \mathbf{R}^d$, $x \in \Omega$, and $A \in \mathfrak{F} = \mathcal{B}(\Omega)$,

$$\delta_x(A) = \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{otherwise,} \end{cases}$$

defines a positive Radon measure, the *Dirac measure in x* .

3. For half-open cuboids $[a, b[= \{x \in \mathbf{R}^d \mid a_i \leq x_i < b_i\} \in \mathfrak{B}(\mathbf{R}^d)$ with $a, b \in \mathbf{R}^d$, we define

$$\mathcal{L}^d([a, b[) = \prod_{i=1}^d (b_i - a_i).$$

One can show that this function possesses a unique extension to a Radon measure on $\mathfrak{B}(\mathbf{R}^d)$ (cf. [123]). This measure, denoted by \mathcal{L}^d as well, is the *d-dimensional Lebesgue measure*. It corresponds to the intuitive idea of the “volume” of a *d*-dimensional set, and we also write $|\Omega| = \mathcal{L}^d(\Omega)$.

4. A different approach to the Lebesgue measure assumes that the volume of *k*-dimensional unit balls is known: For an integer $k \geq 0$, this volume is given by

$$\omega_k = \frac{\pi^{k/2}}{\Gamma(1+k/2)}, \quad \Gamma(k) = \int_0^\infty t^{k-1} e^{-t} dt,$$

where Γ is known as the *gamma function*. For $k \in [0, \infty[,$ a volume can be defined even for “fractional dimensions.” For an arbitrary bounded set $A \subset \mathbf{R}^d$, one now expects that the *k*-dimensional volume is at most $\omega_k \operatorname{diam}(A)^k / 2^k$ with

$$\operatorname{diam}(A) = \sup \{|x - y| \mid x, y \in A\}, \quad \operatorname{diam}(\emptyset) = 0.$$

This suffices to define for $A \in \mathfrak{B}(\mathbf{R}^d)$,

$$\mathfrak{H}^k(A) = \lim_{\delta \rightarrow 0} \frac{\omega_k}{2^k} \inf \left\{ \sum_{i=1}^{\infty} \operatorname{diam}(A_i)^k \mid A \subset \bigcup_{i \in \mathbb{N}} A_i, \operatorname{diam}(A_i) < \delta \right\}.$$

The latter constitutes a Borel measure, the *k-dimensional Hausdorff measure*, which plays a major role in geometric measure theory. A fundamental result in this theory is that this measure coincides with the Lebesgue measure in the case $k = d$. Furthermore, the surface area of a *k*-dimensional surface can be expressed through \mathfrak{H}^k (cf. [62]).

In order to integrate with respect to these measures on subsets as well, one first needs the notion of the restriction of a measure.

Definition 2.39 Let $(\Omega, \mathfrak{F}, \mu)$ be a measure space and Ω' a measurable set.

1. Then

$$A \in \mathfrak{F} : \quad (\mu \llcorner \Omega')(A) = \mu(A \cap \Omega')$$

defines another measure on Ω , the *restriction of μ to Ω'* .

2. Together with the σ -algebra

$$\mathfrak{F}|_{\Omega'} = \{A \cap \Omega' \mid A \in \mathfrak{F}\},$$

the triple $(\Omega', \mathfrak{F}|_{\Omega'}, \mu|_{\Omega'})$ defines the *measure space restricted to Ω'* .

Example 2.40

- For $(\mathbf{R}^d, \mathcal{B}(\mathbf{R}^d), \mathcal{L}^d)$ and $\Omega \subset \mathbf{R}^d$ nonempty and open, we obtain the restricted measure space $(\Omega, \mathcal{B}(\Omega), \mathcal{L}^d|_{\Omega})$. This space leads to the standard Lebesgue integration on Ω .
- The Dirac comb $\mu = \sum_{n=0}^{\infty} \delta_n$ can also be interpreted as a restriction: $\mu = \mathfrak{H}^0|_{\mathbf{N}}$. Note that in contrast to \mathfrak{H}^0 , the measure μ is σ -finite.
- For an \mathfrak{H}^k -integrable Ω' , the term $\mu = \mathfrak{H}^k|_{\Omega'}$ is a finite measure, while \mathfrak{H}^k is not σ -finite for $k < d$. This is of importance in the context of theorems that assume (σ -)finiteness.

If a specific measure is given on a σ -algebra, it is possible to extend the measure and the σ -algebra in such a way that in some sense, as many sets as possible can be measured.

Definition 2.41 (Null Sets, Almost Everywhere, μ -Measurability) Let $(\Omega, \mathfrak{F}, \mu)$ be a measure space.

- If for $N \subset \Omega$, there exists some $A \in \mathfrak{F}$ with $N \subset A$ and $\mu(A) = 0$, the set N is called a *μ -null set*.
- If a statement $P(x)$ is true for all $x \in A$ and $\Omega \setminus A$ is a null set, we say that $P(x)$ holds *μ -almost everywhere in Ω* .
- The σ -algebra \mathfrak{F}_{μ} , given by

$$A \in \mathfrak{F}_{\mu} \iff A = B \cup N, \quad B \in \mathfrak{F}, \quad N \text{ null set},$$

is the *completion of \mathfrak{F} with respect to μ* . Its elements are the *μ -measurable sets*.

- For $A \in \mathfrak{F}_{\mu}$, we extend $\mu(A) = \mu(B)$ using $B \in \mathfrak{F}$ above.

The extension of μ to \mathfrak{F}_{μ} results in a measure again, which we tacitly denote by μ as well. For the Lebesgue measure, the construction $\mathcal{B}(\mathbf{R}^d)|_{\mathcal{L}^d}$ yields the *Lebesgue measurable sets*. The measure space

$$(\Omega, \mathcal{B}(\Omega)|_{\mathcal{L}^d}, \mathcal{L}^d|_{\Omega})$$

associated to $\Omega \in \mathcal{B}(\mathbf{R}^d)|_{\mathcal{L}^d}$ presents the basis for the standard Lebesgue integration on Ω . If nothing else is mentioned explicitly, the notions of measure and integration theory refer to this measure space.

We can now establish the integral for nonnegative functions:

Definition 2.42 (Measurable Nonnegative Functions, Integral) Let $(\Omega, \mathfrak{F}, \mu)$ be a measure space. A function $f : \Omega \rightarrow [0, \infty]$ is called *measurable* if

$$f^{-1}([a, b]) \in \mathfrak{F} \quad \text{for all } 0 \leq a \leq b.$$

For measurable functions with finite range, called *step functions*, the *integral* is defined by

$$\int_{\Omega} f(t) d\mu(t) = \int_{\Omega} f d\mu = \sum_{a \in f(\Omega)} \mu(f^{-1}(\{a\}))a \in [0, \infty].$$

The integral for general measurable nonnegative functions is defined by means of

$$\int_{\Omega} f(t) d\mu(t) = \int_{\Omega} f d\mu = \sup \left\{ \int_{\Omega} g d\mu \mid g \text{ step function, } g \leq f \right\};$$

if this supremum is finite, we call f *integrable*.

The value of the integral is in particular invariant with respect to changes in the integrand on a measurable null set. Thus, it is possible to define measurability and integrability for functions that are defined only almost everywhere on Ω (extending the function with 0, for instance). That is why, despite the sloppiness, we use the notation “ $f : \Omega \rightarrow [0, \infty]$ measurable” also in the case that the domain of f is not the whole of Ω .

The notion of the integral for nonnegative functions presents the basis for the Lebesgue integral. Generally, measurable functions with values in a Banach space are defined to be integrable if their pointwise norm is integrable.

Definition 2.43 (Measurability/Integrability for Vector-Valued Mappings) Let $(\Omega, \mathfrak{F}, \mu)$ be a measure space and $(X, \|\cdot\|_X)$ a Banach space.

A step function $f : \Omega \rightarrow X$, i.e., $f(\Omega)$ is finite, is called *measurable* if $f^{-1}(\{x\}) \in \mathfrak{F}$ for all $x \in X$. A general mapping $f : \Omega \rightarrow X$ is called μ -*measurable* or measurable if there exists a sequence (f_n) of measurable step functions that converges pointwise to f almost everywhere.

If f is a step function, the *integral* is given by

$$\int_{\Omega} f(t) d\mu(t) = \int_{\Omega} f d\mu = \sum_{x \in f(\Omega)} \mu(f^{-1}(\{x\}))x.$$

Otherwise, we set

$$\int_{\Omega} f(t) d\mu(t) = \int_{\Omega} f d\mu = \lim_{n \rightarrow \infty} \int_{\Omega} f_n d\mu$$

where (f_n) is a sequence of step functions that converges pointwise to f almost everywhere. If $\int_{\Omega} \|f(t)\|_X d\mu(t) < \infty$, then f is called *integrable*.

Note that the latter form of the integral is the more interesting construction. One can show that this construction indeed makes sense, i.e., that for every integrable f , there exists a sequence of step functions that converges pointwise almost everywhere to f (in the sense of norm convergence) and that the limit in the definition exists (for which the completeness of the space is needed). Furthermore, the integral is independent of the choice of the sequence. The linearity of the integral with respect to the integrand also follows immediately from this definition as well.

- If X is the space of real numbers, the measurability of $f : \Omega \rightarrow \mathbf{R}$ is equivalent to the measurability of $f_+ = \max(0, f)$ and $f_- = -\min(0, f)$. Also, f is integrable if and only if f_+ are f_- are integrable; additionally,

$$\int_{\Omega} f d\mu = \int_{\Omega} f_+ d\mu - \int_{\Omega} f_- d\mu.$$

The integral is *monotonic*, i.e., for $f, g : \Omega \rightarrow \mathbf{R}$ integrable or f, g nonnegative and measurable, we have

$$f \leq g \quad \text{almost everywhere} \quad \Rightarrow \quad \int_{\Omega} f d\mu \leq \int_{\Omega} g d\mu.$$

- For $f : \Omega \rightarrow \mathbf{C}$, the measurability of f is equivalent to the measurability of $\operatorname{Re} f$ and $\operatorname{Im} f$. Integrability is also the case if and only if the real and imaginary parts are integrable. For the integral, one has

$$\int_{\Omega} f d\mu = \int_{\Omega} \operatorname{Re} f d\mu + i \int_{\Omega} \operatorname{Im} f d\mu.$$

The analogue holds for N -dimensional \mathbf{K} -vector spaces X .

- If X is a general Banach space, the integral introduced in Definition 2.43 is also called the *Bochner integral* (cf. [147]). Mappings that are integrable in this sense exhibit, apart from the image of a null set, a separable range in X . This is particularly important for infinite-dimensional Banach spaces.
- An important relationship between the integral of vector-valued mappings and nonnegative functions is given by the following fundamental estimate:

$$\left\| \int_{\Omega} f(t) d\mu(t) \right\|_X \leq \int_{\Omega} \|f(t)\|_X d\mu(t).$$

- In connection with the integral sign, several special notations are common: if A is a μ -measurable subset of Ω , then the integral of a measurable/integrable f over A is defined by

$$\int_A f \, d\mu = \int_{\Omega} \bar{f} \, d\mu, \quad \bar{f}(t) = \begin{cases} f(t) & \text{if } t \in A, \\ 0 & \text{otherwise.} \end{cases}$$

This corresponds to the integral over the restriction of μ to A .

In the case of the Lebesgue measure \mathcal{L}^d and a Lebesgue measurable subset $\Omega \subset \mathbf{R}^d$, one often abbreviates

$$\int_{\Omega} f(t) \, d\mathcal{L}^d(t) = \int_{\Omega} f(t) \, dt = \int_{\Omega} f \, dt,$$

where the latter notation can lead to misunderstandings and is used only if it is evident that f is a function of t .

The notions μ -measurability and μ -integrability for non-negative functions as well as vector-valued mappings present the respective analogues for the completed measure space $(\Omega, \mathfrak{F}_{\mu}, \mu)$. They constitute the basis for the *Lebesgue spaces*, which we will introduce now as spaces of equivalence classes of measurable functions.

2.2.2 Lebesgue Spaces and Vector Spaces of Measures

Definition 2.44 (Lebesgue Spaces) Let $(\Omega, \mathfrak{F}_{\mu}, \mu)$ be a complete measure space and $(X, \|\cdot\|_X)$ a Banach space. For a μ -measurable $f : \Omega \rightarrow X$, we denote by $[f]$ the equivalence class associated to f under the relation

$$f \sim g \iff f = g \text{ almost everywhere.}$$

Note that $[f]$ is measurable or integrable if there exists a measurable or integrable representative, respectively.

Let $p \in [1, \infty[$. The *Lebesgue space of p -integrable functions* is the set

$$L_{\mu}^p(\Omega, X) = \left\{ [f] \mid f : \Omega \rightarrow X \text{ } \mu\text{-measurable, } \int_{\Omega} \|f(t)\|_X^p \, d\mu(t) < \infty \right\}$$

endowed with the norm

$$\|[f]\|_p = \left(\int_{\Omega} \|f(t)\|_X^p \, d\mu(t) \right)^{1/p}.$$

The space of the *essentially bounded mappings* is defined as the set

$$L_\mu^\infty(\Omega, X) = \{[f] \mid f : \Omega \rightarrow X \text{ } \mu\text{-measurable, } \|f(\cdot)\|_X : \Omega \rightarrow \mathbf{R} \text{ bounded}\}$$

together with the norm

$$\|[f]\|_\infty = \inf \left\{ \sup_{t \in \Omega} \|g(t)\|_X \mid g : \Omega \rightarrow X \text{ measurable, } f \sim g \right\}.$$

For real-valued functions $f : \Omega \rightarrow [-\infty, \infty]$, we define the *essential supremum* and the *essential infimum* by, respectively,

$$\text{ess sup } f = \inf \left\{ \sup_{x \in \Omega} g(x) \mid g : \Omega \rightarrow [-\infty, \infty] \text{ measurable, } f \sim g \right\} \text{ and}$$

$$\text{ess inf } f = \sup \left\{ \inf_{x \in \Omega} g(x) \mid g : \Omega \rightarrow [-\infty, \infty] \text{ measurable, } f \sim g \right\}.$$

Example 2.45 (Standard Lebesgue Spaces) In the case that $\Omega \subset \mathbf{R}^d$ is non-empty and measurable, we denote the spaces associated to $(\Omega, \mathcal{B}(\Omega), \mathcal{L}^d \llcorner \Omega)$ simply by $L^p(\Omega, X)$. For $X = \mathbf{K}$, we also write $L^p(\Omega)$. These spaces are the *standard Lebesgue spaces*.

Regarding the transition to equivalence classes and the definition of the Lebesgue spaces, let us remark the following:

- The norms used in Definition 2.44 are well-defined mappings on the equivalence classes, i.e., they do not depend on the respective representative. As is common in the literature, we will tacitly select a representative for $f \in L^p(\Omega, X)$ and then denote it by f as well, as long as all that follows is independent of the representative.
- $\|\cdot\|_p$ is indeed a norm on a vector space: The positive homogeneity can be deduced from the definition directly. For integrals, one also has the *Minkowski inequality*

$$\left(\int_\Omega \|f + g\|_X^p d\mu \right)^{1/p} \leq \left(\int_\Omega \|f\|_X^p d\mu \right)^{1/p} + \left(\int_\Omega \|g\|_X^p d\mu \right)^{1/p},$$

which yields the triangle inequality for $p \in [1, \infty[$; the case $p = \infty$ can be proved in a direct way. Finally, the requirement of the positive definiteness reflects the fact that a transition to equivalence classes is necessary, since the integral of a nonnegative, measurable function vanishes if and only if it is zero almost everywhere.

For sequences of measurable or integrable functions in the sense of Lebesgue, several convergence results hold. The most important of these findings are as follows. Proofs can be found in [50, 53], for instance.

Lemma 2.46 (Fatou's Lemma) *Let $(\Omega, \mathfrak{F}_\mu, \mu)$ be a measure space and (f_n) a sequence of nonnegative measurable functions $f_n : \Omega \rightarrow [0, \infty]$, $n \in \mathbf{N}$. Then,*

$$\int_{\Omega} \liminf_{n \rightarrow \infty} f_n(t) d\mu(t) \leq \liminf_{n \rightarrow \infty} \int_{\Omega} f_n(t) d\mu(t).$$

Theorem 2.47 (Lebesgue's Dominated Convergence Theorem) *Let $(\Omega, \mathfrak{F}_\mu, \mu)$ be a measure space, X a Banach space, $1 \leq p < \infty$ and (f_n) a sequence in $L_\mu^p(\Omega, X)$ that converges to $f : \Omega \rightarrow X$ pointwise μ -almost everywhere. Assume that there exists $g \in L^p(\Omega)$ that satisfies $\|f_n(t)\|_X \leq g(t)$ μ -almost everywhere for all $n \in \mathbf{N}$. Then, $f \in L_\mu^p(\Omega, X)$ as well as*

$$\lim_{n \rightarrow \infty} \int_{\Omega} \|f_n(t) - f(t)\|_X^p d\mu(t) = 0.$$

Let us now turn to the functional-analytical aspects, first to the completeness of the Lebesgue spaces, which is based on the following result:

Theorem 2.48 (Fischer-Riesz) *Let $1 \leq p < \infty$ and $L_\mu^p(\Omega, X)$ a Lebesgue space. For every Cauchy sequence (f_n) in $L_\mu^p(\Omega, X)$, there exists a subsequence (f_{n_k}) that converges to $f \in L_\mu^p(\Omega, X)$ pointwise μ -almost everywhere.*

In particular, that $f_n \rightarrow f$ in $L_\mu^p(\Omega, X)$ for $n \rightarrow \infty$.

This assertion holds also for $p = \infty$, but the restriction to subsequences is not necessary.

Corollary 2.49 *The Lebesgue spaces $L_\mu^p(\Omega, X)$ are Banach spaces. If X is a Hilbert space, then $L_\mu^2(\Omega, X)$ is also a Hilbert space with the inner product*

$$(f, g)_2 = \int_{\Omega} (f(t), g(t))_X d\mu(t).$$

Example 2.50 (Sequence Spaces) Using the counting measure μ on \mathbf{N} yields Lebesgue spaces that contain sequences in X (the empty set \emptyset being the only null set). They are accordingly called *sequence spaces*. We use the notation $\ell^p(X)$ for general Banach spaces X and write ℓ^p in the case of $X = \mathbf{K}$. The norm of a sequence $f : \mathbf{N} \rightarrow X$ can be simply expressed by

$$\|f\|_p = \left(\sum_{n=1}^{\infty} \|f_n\|_X^p \right)^{1/p}, \quad \|f\|_{\infty} = \sup_{n \in \mathbf{N}} \|f_n\|_X.$$

Due to the above results, Lebesgue spaces become part of Banach and Hilbert space theory. This also motivates further topological considerations, such as the characterization of the dual spaces, for instance: if $L_\mu^p(\Omega, X)$ is a Hilbert space, i.e., $p = 2$ and X is a Hilbert space, then the Riesz representation theorem

(Theorem 2.35) immediately yields $L_\mu^2(\Omega, X) = (L_\mu^2(\Omega, X))^*$ in the sense of the Hilbert space isometry

$$J : L_\mu^2(\Omega, X) \rightarrow (L_\mu^2(\Omega, X))^*, \quad (Jf^*)f = \int_{\Omega} (f(t), f^*(t)) d\mu(t).$$

The situation is similar for a general p : With the exception of $p = \infty$, the corresponding dual spaces are again Lebesgue spaces with an analogous isometry. However, there are restrictions if the range X is of infinite dimension (cf. [50]).

Theorem 2.51 (Dual Space of a Lebesgue Space) *Let X be a reflexive Banach space, $1 \leq p < \infty$, and $L_\mu^p(\Omega, X)$ the Lebesgue space with respect to the σ -finite measure space $(\Omega, \mathfrak{F}_\mu, \mu)$. Denote by p^* the dual exponent, i.e., the solution of $1/p + 1/p^* = 1$ (with $p^* = \infty$ if $p = 1$). Then,*

$$J : L_\mu^{p^*}(\Omega, X^*) \rightarrow (L_\mu^p(\Omega, X))^*, \quad (Jf^*)f = \int_{\Omega} \langle f^*(t), f(t) \rangle_{X^* \times X} d\mu(t)$$

defines a Banach space isometry.

Remark 2.52

- The fact that the above J is a continuous mapping with

$$\|Jf^*\|_{L_\mu^p(\Omega, X)^*} \leq \|f^*\|_{p^*}$$

leads to *Hölder's inequality*:

$$\int_{\Omega} |\langle f^*(t), f(t) \rangle_{X^* \times X}| d\mu(t) \leq \left(\int_{\Omega} \|f(t)\|_X^p d\mu(t) \right)^{\frac{1}{p}} \left(\int_{\Omega} \|f^*(t)\|_{X^*}^{p^*} d\mu(t) \right)^{\frac{1}{p^*}}.$$

Equality is usually obtained by choosing a normed sequence (f^n) that satisfies $\langle Jf^*, f^n \rangle \rightarrow \|f^*\|_{p^*}$. The surjectivity of J , however, is a deeper result of measure theory: it is based on the *Radon-Nikodym theorem*, an assertion about when a measure can be expressed as an integral with respect to another measure (cf. [71]).

- As mentioned before, every finite-dimensional space is reflexive. Thus, the condition that X has to be reflexive is relevant only in the infinite dimensional case. In particular, there are examples of nonreflexive spaces X for which the above characterization of the dual space does not hold.

One can easily calculate that the “double dual exponent” p^{**} satisfies $p^{**} = p$ for finite p . Therefore, we have the following corollary.

Corollary 2.53 *The spaces $L_\mu^p(\Omega, X)$ are reflexive if $p \in]1, \infty[$ and X is reflexive.*

In particular, according to the Eberlein-Šmulyan theorem (Theorem 2.22), every bounded sequence in $L_\mu^p(\Omega, X)$, $1 < p < \infty$, possesses a weakly convergent subsequence.

Functions in $L^p(\Omega, X)$ can be approximated by continuous functions, in particular by those with compact support:

Definition 2.54 Let $\Omega \subset \mathbf{R}^d$ be a nonempty, open subset endowed with the relative topology on \mathbf{R}^d .

A function $f \in \mathcal{C}(\Omega, X)$ has a *compact support* in Ω if the set

$$\text{supp } f = \overline{\{t \in \Omega \mid f(t) \neq 0\}}$$

is compact in Ω .

We denote the subspace of continuous functions with compact support by

$$\mathcal{C}_c(\Omega, X) = \{f \in \mathcal{C}(\Omega, X) \mid f \text{ has compact support}\} \subset \mathcal{C}(\Omega, X).$$

Theorem 2.55 For $1 \leq p < \infty$, the set $\mathcal{C}_c(\Omega, X)$ is dense in $L^p(\Omega, X)$, i.e., for every $f \in L^p(\Omega, X)$ and every $\varepsilon > 0$, there exists some $g \in \mathcal{C}_c(\Omega, X)$ such that $\|f - g\|_p \leq \varepsilon$.

Apart from Lebesgue spaces, which contain classes of measurable functions, we also consider Banach spaces of measures. In particular, we are interested in the spaces of signed or vector-valued Radon measures. These spaces, in some sense, contain functions, but additionally also measures that cannot be interpreted as functions or equivalence classes of functions. Thus, they already present a set of generalized functions, i.e., of objects that in general can no longer be evaluated pointwise. In the following, we give a summary of the most important corresponding results, following the presentation in [5, 61].

Definition 2.56 (Vector-Valued Measure) A function $\mu : \mathfrak{F} \rightarrow X$ on a measurable space (Ω, \mathfrak{F}) into a finite-dimensional Banach space X is called a *vector-valued measure* if

1. $\mu(\emptyset) = 0$ and
2. for $A_i \in \mathfrak{F}$, $i \in \mathbf{N}$, with A_i mutually disjoint, one has

$$\mu\left(\bigcup_{i \in \mathbf{N}} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i).$$

For $X = \mathbf{R}$, the function μ is also called a *signed measure*.

In the special case of $\mathfrak{F} = \mathfrak{B}(\Omega)$, we speak of a *vector-valued finite Radon measure*.

An essential point that distinguishes vector-valued measures from positive measures is the fact that the “value” ∞ is not allowed for the former, i.e., every

measurable set exhibits a “finite measure.” Furthermore, there is the following fundamental relationship between vector-valued measures and positive measures in the sense of Definition 2.37.

Definition 2.57 (Total Variation Measure) Let μ be a vector-valued measure on a measurable space (Ω, \mathfrak{F}) with values in X . Then the measure $|\mu|$ given by

$$A \in \mathfrak{F} : |\mu|(A) = \sup \left\{ \sum_{i=1}^{\infty} \|\mu(A_i)\|_X \mid A = \bigcup_{i \in \mathbb{N}} A_i, A_i \in \mathfrak{F} \text{ mutually disjoint} \right\}$$

denotes the *total variation measure* or *variation measure* associated to μ .

One can show that $|\mu|$ always yields a positive finite measure. If μ is a vector-valued finite Radon measure, then $|\mu|$ is a positive finite Radon measure. In fact, one can interpret the total variation measure as a kind of absolute value; in particular, we have the following.

Theorem 2.58 (Polar Decomposition of Measures) For every vector-valued measure μ on Ω , there exists an element $\sigma \in L_{|\mu|}^{\infty}(\Omega, X)$ that satisfies $\|\sigma(t)\|_X = 1$ $|\mu|$ -almost everywhere as well as

$$\mu(A) = \int_A \sigma(t) d|\mu|(t)$$

for all $A \in \mathfrak{F}$. The pair $(|\mu|, \sigma)$ is called the *polar decomposition* of μ .

Viewed from the other side, every positive measure yields a vector-valued measure by means of the above representation (together with a function that is measurable with respect to that measure and exhibits norm 1 almost everywhere). In this sense, the notion of a vector-valued measure does not yield anything new at first sight. However, we gain a Banach space structure:

Theorem 2.59 (Space of Radon Measures) The set

$$\mathfrak{M}(\Omega, X) = \{\mu : \mathfrak{B}(\Omega) \rightarrow X \mid \mu \text{ vector-valued finite Radon measure}\},$$

endowed with the norm $\|\mu\|_{\mathfrak{M}} = |\mu|(\Omega)$, is a Banach space.

Example 2.60 (Vector-Valued Finite Radon Measures) Let $\Omega \subset \mathbf{R}^d$, $d \geq 1$, be a nonempty open subset. To every $k \in [0, d]$ and every element $f \in L_{\mathfrak{H}^k}^1(\Omega, X)$, we can associate a measure $\mu_f \in \mathfrak{M}(\Omega, X)$ by means of

$$\mu_f(A) = \int_A f(t) d\mathfrak{H}^k(t).$$

Since $\|\mu_f\|_{\mathfrak{M}} = \|f\|_1$, this mapping is injective and the range is a closed subspace.

For $k = d$, one can interpret $L^1(\Omega, X)$ as a subspace of $\mathfrak{M}(\Omega, X)$ since $\mathfrak{H}^d = \mathcal{L}^d$. For $k < d$, due to the requirement that f be integrable, μ_f can be nontrivial only on “thin” sets such as suitable k -dimensional surfaces. For $k = 0$, the measure μ_f can be expressed through an at most countable sequence of Dirac measures: $\mu_f = \sum_{i=1}^{\infty} x_i \delta_{t_i}$ with $x \in \ell^1(X)$ and $t_i \in \Omega$ mutually disjoint.

A characterization of $\mathfrak{M}(\Omega, X)$ that allows for far-reaching functional-analytical implications is the identification of the space of Radon measures with a suitable Banach space, namely the dual space of the continuous functions that, roughly speaking, vanish on the boundary of Ω .

Definition 2.61 The closure of $\mathcal{C}_c(\Omega, X)$ in $\mathcal{C}(\Omega, X)$ is denoted by $\mathcal{C}_0(\Omega, X)$.

Theorem 2.62 (Riesz-Markov Representation Theorem) *Let $\Omega \subset \mathbf{R}^d$, $d \geq 1$, be a nonempty open subset and $F \in \mathcal{C}_0(\Omega, X^*)^*$. Then there exists a unique $\mu \in \mathfrak{M}(\Omega, X)$ (with the polar decomposition $|\mu|$ and σ) such that*

$$\langle F, f \rangle = \int_{\Omega} f \, d\mu = \int_{\Omega} \langle f(t), \sigma(t) \rangle \, d|\mu|(t) \quad \text{for all } f \in \mathcal{C}_0(\Omega, X^*).$$

Furthermore, one has $\|F\| = \|\mu\|_{\mathfrak{M}}$.

Example 2.63 For $\Omega = [a, b]$, the Riemann integral for $f \in \mathcal{C}_0(\Omega)$ defines an element in $\mathcal{C}_0(\Omega)^*$ by means of

$$f \mapsto \int_a^b f(t) \, dt,$$

and therefore, it yields a finite, even positive, Radon measure that coincides with the restriction of the one-dimensional Lebesgue measure to $[a, b]$.

By means of the characterization of $\mathfrak{M}(\Omega, X)$ as a dual space, we immediately obtain the notion of weak* convergence in the sense of Definition 2.19: in fact, $\mu_n \xrightarrow{*} \mu$ for a sequence (μ_n) in $\mathfrak{M}(\Omega, X)$ and $\mu \in \mathfrak{M}(\Omega, X)$ if and only if

$$\int_{\Omega} f \, d\mu_n \rightarrow \int_{\Omega} f \, d\mu \quad \text{for all } f \in \mathcal{C}_0(\Omega, X^*).$$

In the case that Ω is a subset of \mathbf{R}^d , the Riesz-Markov representation theorem, together with the Banach-Alaoglu theorem (Theorem 2.21), yields the following compactness result, which is similar to the weak sequential compactness in $L_{\mu}^p(\Omega, X)$.

Theorem 2.64 *Let $\Omega \subset \mathbf{R}^d$ be a nonempty, open subset. Then every bounded sequence (μ_n) in $\mathfrak{M}(\Omega, X)$ possesses a weak*-convergent subsequence.*

2.2.3 Operations on Measures

For two measure spaces given on Ω_1 and Ω_2 , respectively, one can easily construct a measure on the Cartesian product $\Omega_1 \times \Omega_2$. This, for example, is helpful for integration on $\mathbf{R}^{d_1+d_2} = \mathbf{R}^{d_1} \times \mathbf{R}^{d_2}$.

Definition 2.65 (Product Measure) For measure spaces $(\Omega_1, \mathfrak{F}_1, \mu_1)$ and $(\Omega_2, \mathfrak{F}_2, \mu_2)$, the product $\mathfrak{F}_1 \otimes \mathfrak{F}_2$ denotes the σ -algebra induced by the sets $A \times B$, $A \in \mathfrak{F}_1$ and $B \in \mathfrak{F}_2$.

A product measure $\mu_1 \otimes \mu_2$ is a measure given on $\mathfrak{F}_1 \otimes \mathfrak{F}_2$ that satisfies

$$(\mu_1 \otimes \mu_2)(A) = \mu_1(A) \mu_2(B) \quad \text{for all } A \in \mathfrak{F}_1, B \in \mathfrak{F}_2.$$

The triple $(\Omega_1 \times \Omega_2, \mathfrak{F}_1 \otimes \mathfrak{F}_2, \mu_1 \otimes \mu_2)$ denotes a *product measure space*.

Based on that, we can define measurability and integrability for product measures, of course. The key question, whether a product measure is uniquely defined and whether in the case of its existence, the integral can be calculated by means of a double integral, is answered by Fubini's theorem. This result often justifies changing the order of integration as well.

Theorem 2.66 (Fubini) Let $(\Omega_1, \mathfrak{F}_1, \mu_1)$ and $(\Omega_2, \mathfrak{F}_2, \mu_2)$ be σ -finite measure spaces and X a Banach space.

Then there exists a unique product measure $\mu_1 \otimes \mu_2$, the corresponding product measure space $(\Omega_1 \times \Omega_2, \mathfrak{F}_1 \otimes \mathfrak{F}_2, \mu_1 \otimes \mu_2)$ is σ -finite and one has the following:

1. The measure of every measurable set $A \subset \Omega_1 \times \Omega_2$ can be expressed through

$$(\mu_1 \otimes \mu_2)(A) = \int_{\Omega_2} \mu_1(A_t) d\mu_2(t) = \int_{\Omega_1} \mu_2(A_s) d\mu_1(s)$$

by means of the sets $A_t = \{s \in \Omega_1 \mid (s, t) \in A\}$ and $A_s = \{t \in \Omega_2 \mid (s, t) \in A\}$. In particular, A_t is μ_2 -almost everywhere μ_1 -measurable, whereas A_s is μ_1 -almost everywhere μ_2 -measurable and the functions $t \mapsto \mu_1(A_t)$ and $s \mapsto \mu_2(A_s)$ are μ_2 - and μ_1 -measurable, respectively.

2. A $(\mu_1 \otimes \mu_2)$ -measurable mapping $f : \Omega_1 \times \Omega_2 \rightarrow X$ is integrable if and only if $t \mapsto \int_{\Omega_1} \|f(s, t)\|_X d\mu_1(s)$ is μ_2 -integrable or $s \mapsto \int_{\Omega_2} \|f(s, t)\|_X d\mu_2(t)$ is μ_1 -integrable. In particular, these functions are always measurable, and in the case of integrability, one has

$$\begin{aligned} \int_{\Omega_1 \times \Omega_2} f(s, t) d(\mu_1 \otimes \mu_2)(s, t) &= \int_{\Omega_2} \left(\int_{\Omega_1} f(s, t) d\mu_1(s) \right) d\mu_2(t) \\ &= \int_{\Omega_1} \left(\int_{\Omega_2} f(s, t) d\mu_2(t) \right) d\mu_1(s). \end{aligned}$$

Remark 2.67 The respective assertions hold for the completed measure space $(\Omega_1 \times \Omega_2, (\mathfrak{F}_1 \otimes \mathfrak{F}_2)_{\mu_1 \otimes \mu_2}, \mu_1 \otimes \mu_2)$ as well.

Example 2.68 One can show that the product of Lebesgue measures is again a Lebesgue measure: $\mathcal{L}^{d-n} \otimes \mathcal{L}^n = \mathcal{L}^d$ for integers $1 \leq n < d$ (cf. [146]). This fact facilitates integration on \mathbf{R}^d :

$$\int_{\mathbf{R}^d} f(t) dt = \int_{\mathbf{R}^n} \int_{\mathbf{R}^{d-n}} f(t_1, t_2) dt_1 dt_2, \quad t = (t_1, t_2)$$

for $f \in L^1(\mathbf{R}^d, X)$. According to Fubini's theorem, the value of the integral does not depend on the order of integration.

For a measure space $(\Omega_1, \mathfrak{F}_1, \mu_1)$, a measurable space $(\Omega_2, \mathfrak{F}_2)$, and a measurable mapping $\varphi : \Omega_1 \rightarrow \Omega_2$, one can abstractly define the *pushforward measure*:

$$\mu_2(A) = \mu_1(\varphi^{-1}(A)), \quad A \in \mathfrak{F}_2.$$

By means of that, $(\Omega, \mathfrak{F}_2, \mu_2)$ becomes a measure space, and integration on Ω_2 with respect to the pushforward measure can be defined, and one has

$$\int_{\Omega_2} f d\mu_2 = \int_{\Omega_1} f \circ \varphi d\mu_1.$$

In applications, we wish to integrate on Ω_2 with respect to the Lebesgue measure and to “pull back” the integral to Ω_1 by means of the coordinate transformation φ . The following theorem shows that \mathcal{L}^d is then the pushforward of a particular measure.

Theorem 2.69 (Change of Variables Formula for the Lebesgue Measure) *Let $\Omega_1, \Omega_2 \subset \mathbf{R}^d$ be nonempty, open and $\varphi : \Omega_1 \rightarrow \Omega_2$ a diffeomorphism, i.e., φ is invertible and φ as well as φ^{-1} are continuously differentiable.*

Then for all Lebesgue measurable subsets $A \subset \Omega_2$, one has

$$\mathcal{L}^d(A) = \int_{\varphi^{-1}(A)} |\det \nabla \varphi| d\mathcal{L}^d.$$

If $f : \Omega_2 \rightarrow [0, \infty]$ is Lebesgue measurable or respectively $f \in L^1(\Omega_2, X)$ for a Banach space X , we have

$$\int_{\Omega_2} f d\mathcal{L}^d = \int_{\Omega_1} |\det \nabla \varphi|(f \circ \varphi) d\mathcal{L}^d;$$

in particular, $|\det \nabla \varphi|(f \circ \varphi)$ is Lebesgue measurable or lies in $L^1(\Omega_1, X)$, respectively.

We are also interested in integration on surfaces and parts of surfaces. Here, we confine ourselves to integration with respect to the \mathfrak{H}^{d-1} -measure on the boundary $\partial\Omega'$ of a so-called domain Ω . In order to be able to calculate such integrals, we perform a transformation by means of a suitable parameterization. For this purpose, we need the following notion of regularity for sets and their boundaries.

Definition 2.70 (Domain/Bounded Lipschitz Domain) A nonempty, open, and connected set $\Omega \subset \mathbf{R}^d$ is called a *domain*. If additionally, Ω is bounded, we speak of a *bounded domain*.

A bounded domain Ω is a bounded *Lipschitz domain* or possesses the *Lipschitz property* if there exist a finite open cover U_1, \dots, U_n of the boundary $\partial\Omega$, open subsets V_1, \dots, V_n , as well as for every $j = 1, \dots, n$, a Lipschitz continuous mapping $\psi_j : V_j \rightarrow U_j$ with a Lipschitz continuous inverse such that for all $x \in U_j$, one has $x \in \Omega$ if and only if $(\psi_j^{-1}(x))_d < 0$ (i.e., the d -th component of the preimage of x under ψ_j is negative).

This condition is equivalent to the commonly used requirement that locally and after a change of coordinates, the boundary of Ω can be represented as the graph of a Lipschitz function (cf. [137]). By means of the mappings ψ_j , the boundary of Ω can be “locally flattened”; see Fig. 2.1.

Example 2.71

1. Every open cuboid $\Omega =]a_1, b_1[\times]a_2, b_2[\times \cdots \times]a_d, b_d[$ with $a_i < b_i$, for $i = 1, \dots, d$, is a bounded Lipschitz domain.
2. More generally, every convex bounded domain possesses the Lipschitz property.
3. Let be $\Omega = \{x \in \mathbf{R}^d \mid \phi(x) < 0\}$ for a continuously differentiable function $\phi : \mathbf{R}^d \rightarrow \mathbf{R}$. Then an application of the inverse function theorem implies that Ω possesses the Lipschitz property if it is bounded and $\nabla\phi(x) \neq 0$ for all $x \in \partial\Omega$.

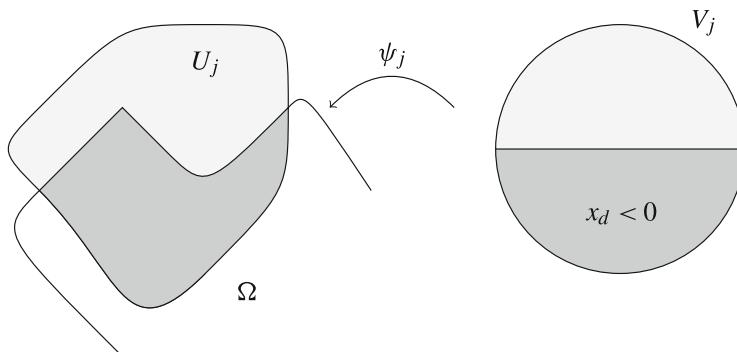


Fig. 2.1 The mappings ψ_j pull the boundary of a Lipschitz domain “locally straight”

In dealing with Lipschitz domains, the possibility of localization is important, in particular for parts of the boundary $\partial\Omega$. For this purpose, we quote the following technical result:

Lemma 2.72 *For every bounded Lipschitz domain Ω and the corresponding sets U_j , there exist an open set $U_0 \subset \Omega$ as well as functions $\varphi_0, \dots, \varphi_n \in \mathcal{D}(\mathbf{R}^d)$ such that $\overline{\Omega} \subset \bigcup_{j=0}^n U_j$ and for $j = 0, \dots, n$, we have*

$$\text{supp } \varphi_j \subset\subset U_j, \quad \varphi_j(x) \in [0, 1] \quad \forall x \in \mathbf{R}^d, \quad \sum_{j=0}^n \varphi_j(x) = 1 \quad \forall x \in \overline{\Omega}.$$

The functions φ_j are called a partition of unity on $\overline{\Omega}$ subordinated to U_0, \dots, U_n .

By means of the previous lemma, an integration on $\partial\Omega$ can be expressed as an integration with respect to \mathfrak{L}^{d-1} (cf. [62]).

Theorem 2.73 (Integration on the Boundary)

1. For a bounded Lipschitz domain Ω and $p \in [1, \infty[$, we have $f \in L_{\mathfrak{H}^{d-1}}^p(\partial\Omega)$ if and only if for a subordinated partition of unity of Lemma 2.72, one has

$$((\varphi_j |f|^p) \circ \psi_j)(\cdot, 0) \in L_{\mathfrak{L}^{d-1}}^p(V_j \cap \mathbf{R}^{d-1}) \quad \text{for } j = 1, \dots, n.$$

2. For $f \in L_{\mathfrak{H}^{d-1}}^1(\partial\Omega)$, the following integral identity holds:

$$\int_{\partial\Omega} f(x) d\mathfrak{H}^{d-1}(x) = \sum_{j=1}^n \int_{V_j \cap \mathbf{R}^{d-1}} J_j(x) ((\varphi_j f) \circ \psi_j)(x, 0) d\mathfrak{L}^{d-1}(x),$$

where $J_j(x) = \sqrt{\det(D_j(x)^T D_j(x))}$ with $D_j(x)$ corresponding to the first $d - 1$ columns of the Jacobian matrix $\nabla \psi_j(x, 0)$ and $D_j(x)^T$ corresponding to its transpose. In particular, the function J_j can be defined \mathfrak{L}^{d-1} -almost everywhere in $V_j \cap \mathbf{R}^{d-1}$ and is measurable and essentially bounded there.

3. There exists an \mathfrak{H}^{d-1} -measurable mapping $v : \partial\Omega \rightarrow \mathbf{R}^d$, called the outer normal, with $|v(x)| = 1$ \mathfrak{H}^{d-1} -almost everywhere such that for all vector fields $f \in L^1(\partial\Omega, \mathbf{R}^d)$,

$$\int_{\partial\Omega} (f \cdot v)(x) d\mathfrak{H}^{d-1}(x) = \sum_{j=1}^n \int_{V_j \cap \mathbf{R}^{d-1}} J_j(x) ((\varphi_j f) \circ \psi_j) \cdot E_j(x, 0) d\mathfrak{L}^{d-1}(x),$$

where E_j is given by $E_j(x) = (\nabla \psi_j(x)^{-T} e_d) / |\nabla \psi_j(x)^{-T} e_d|$ and $\nabla \psi_j(x)^{-T}$ denotes the inverse of the transposed Jacobian matrix $\nabla \psi_j(x)$.

An important application of integration on the boundary is the generalization of the formula for integration by parts to higher dimensions, usually called Gauss's theorem or the divergence theorem.

Theorem 2.74 (Gauss's Theorem) *For $\Omega \subset \mathbf{R}^d$ a bounded Lipschitz domain and $f : \mathbf{R}^d \rightarrow \mathbf{K}^d$ a continuously differentiable vector field, one has*

$$\int_{\partial\Omega} f \cdot v \, d\mathfrak{H}^{d-1} = \int_{\Omega} \operatorname{div} f \, dx,$$

where v denotes the outer normal introduced in Theorem 2.73. In particular, we obtain for continuously differentiable vector fields $f : \mathbf{R}^d \rightarrow \mathbf{K}^d$ and continuously differentiable functions $g : \mathbf{R}^d \rightarrow \mathbf{K}$,

$$\int_{\partial\Omega} f \bar{g} \cdot v \, d\mathfrak{H}^{d-1} = \int_{\Omega} \bar{g} \operatorname{div} f + f \cdot \nabla g \, dx.$$

A proof can be found in [62] again, for instance.

2.3 Weak Differentiability and Distributions

Differentiable functions can form function spaces as well: for a domain $\Omega \subset \mathbf{R}^d$, we define

$$\mathcal{C}^k(\Omega) = \{f : \Omega \rightarrow \mathbf{K} \mid \frac{\partial^\alpha}{\partial x^\alpha} f \in \mathcal{C}(\Omega) \text{ for } |\alpha| \leq k\}$$

and analogously $\mathcal{C}^k(\overline{\Omega})$, endowed with the norm $\|f\|_{k,\infty} = \max_{|\alpha| \leq k} \|\frac{\partial^\alpha}{\partial x^\alpha} f\|_\infty$. The space of functions that are infinitely differentiable is given by

$$\mathcal{C}^\infty(\Omega) = \{f : \Omega \rightarrow \mathbf{K} \mid \frac{\partial^\alpha}{\partial x^\alpha} f \in \mathcal{C}(\Omega) \text{ for all } \alpha \in \mathbf{N}^d\},$$

$\mathcal{C}^\infty(\overline{\Omega})$ defined analogously again. Furthermore, the following notion is common:

$$\mathcal{D}(\Omega) = \{f \in \mathcal{C}^\infty(\Omega) \mid \operatorname{supp} f \text{ is compact in } \Omega\}.$$

For some applications, the classical notion of differentiability is too restrictive: In the solution theory for partial differential equations or in the calculus of variations, for instance, a weaker notion of a derivative is needed. For this purpose, it is fundamental to observe that a large class of functions is characterized through integrals of products with smooth functions. By $L^1_{\text{loc}}(\Omega)$ we denote the space of functions whose absolute value is integrable over every compact subset of Ω .

Lemma 2.75 (Fundamental Lemma of the Calculus of Variations) *Let $\Omega \subset \mathbf{R}^d$ be nonempty and open and $f \in L^1_{\text{loc}}(\Omega)$. Then, $f = 0$ almost everywhere if and only if for every $\phi \in \mathcal{D}(\Omega)$, we have*

$$\int_{\Omega} f\phi \, dx = 0.$$

In reference to the integrals $\int_{\Omega} f\phi \, dx$, we also say that f is “tested” with ϕ . That is, the fundamental lemma claims that a function $f \in L^1_{\text{loc}}(\Omega)$ is determined almost everywhere by testing f with all functions $\phi \in \mathcal{D}(\Omega)$. Therefore, the space $\mathcal{D}(\Omega)$ is also called the *space of test functions*. It is endowed with the following notion of convergence:

Definition 2.76 (Convergence in $\mathcal{D}(\Omega)$) A sequence (ϕ_n) in $\mathcal{D}(\Omega)$ converges to $\phi \in \mathcal{D}(\Omega)$ if

1. there exists a compact set $K \subset\subset \Omega$ such that $\text{supp } \phi_n \subset K$ for all n and
2. for all multi-indices $\alpha \in \mathbf{N}^d$,

$$\partial^\alpha \phi_n \rightarrow \partial^\alpha \phi \quad \text{uniformly in } \Omega.$$

By means of this notion of convergence, we can consider continuous linear functionals on $\mathcal{D}(\Omega)$:

Definition 2.77 (Distributions) Analogously to the dual space of a normed space, let $\mathcal{D}(\Omega)^*$ denote the set of linear and continuous functionals $T : \mathcal{D}(\Omega) \rightarrow \mathbf{K}$. Note that T is continuous if $\phi_n \rightarrow \phi$ in $\mathcal{D}(\Omega)$ implies $T(\phi_n) \rightarrow T(\phi)$. The elements of $\mathcal{D}(\Omega)^*$ are called *distributions*. A sequence (T_n) of distributions converges to T if for all $\phi \in \mathcal{D}(\Omega)$, one has $T_n(\phi) \rightarrow T(\phi)$.

Every function $f \in L^1_{\text{loc}}(\Omega)$ induces a distribution T_f by

$$T_f(\phi) = \int_{\Omega} f(x)\phi(x) \, dx,$$

which is why distributions are also called “generalized functions.” If a distribution T is induced by a function, we call it *regular*. Examples of nonregular distributions are Radon measures $\mu \in \mathfrak{M}(\Omega, \mathbf{K})$, which also introduce distributions through

$$T_\mu(\phi) = \int_{\Omega} \phi(x) \, d\mu(x).$$

The Dirac measures of Example 2.38 are called Dirac distributions or delta distributions in this context.

Test functions also induce distributions, and for arbitrary multi-indices α , according to the rule of integration by parts, one has for $f, \phi \in \mathcal{D}(\Omega)$ that

$$T_{\partial^\alpha f}(\phi) = \int_{\Omega} \frac{\partial^\alpha}{\partial x^\alpha} f(x) \phi(x) dx = (-1)^{|\alpha|} \int_{\Omega} f(x) \frac{\partial^\alpha}{\partial x^\alpha} \phi(x) dx = (-1)^{|\alpha|} T_f(\partial^\alpha \phi).$$

This identity gives rise to the definition of the derivative of a distribution:

$$\partial^\alpha T(\phi) = (-1)^{|\alpha|} T(\partial^\alpha \phi).$$

The derivative of a distribution induced by a function is also called *distributional derivative* of that function. Note that in general, distributional derivatives of functions are not functions, but if they are, they are called weak derivatives:

Definition 2.78 (Weak Derivative) Let $\Omega \subset \mathbf{R}^d$ be a domain, $f \in L^1_{\text{loc}}(\Omega)$, and α a multi-index. If there exists a function $g \in L^1_{\text{loc}}(\Omega)$ such that for all $\phi \in \mathcal{D}(\Omega)$,

$$\int_{\Omega} g(x) \phi(x) dx = (-1)^{|\alpha|} \int_{\Omega} f(x) \partial^\alpha \phi(x) dx,$$

we say that the *weak derivative* $\partial^\alpha f = g$ exists. If for an integer $m \geq 1$ and all multi-indices α with $|\alpha| \leq m$, the weak derivatives $\partial^\alpha f$ exist, then f is m -times *weakly differentiable*.

Note that we denote the classical as well as the weak derivative by the same symbol, which normally will not lead to ambiguities. If necessary, we will explicitly state which derivative is meant. According to Lemma 2.75 (fundamental lemma of the calculus of variations), the weak derivative is uniquely determined almost everywhere.

Definition 2.79 (Sobolev Spaces) Let be $1 \leq p \leq \infty$ and $m \in \mathbf{N}$. The *Sobolev space* $H^{m,p}(\Omega)$ is the set

$$H^{m,p}(\Omega) = \{f \in L^p(\Omega) \mid \partial^\alpha f \in L^p(\Omega) \text{ for } |\alpha| \leq m\}$$

endowed with the so-called *Sobolev norm*

$$\|f\|_{m,p} = \left(\sum_{|\alpha| \leq m} \|\partial^\alpha f\|_p^p \right)^{1/p} \quad \text{if} \quad p < \infty$$

and $\|f\|_{m,\infty} = \max_{|\alpha| \leq m} \|\partial^\alpha f\|_\infty$, otherwise. Furthermore, we define

$$H_0^{m,p}(\Omega) = \overline{\mathcal{D}(\Omega)} \subset H^{m,p}(\Omega),$$

where the closure is taken with respect to the Sobolev norm.

The Sobolev spaces are Banach spaces; for $1 < p < \infty$, they are reflexive, and for $p = 2$, together with the inner product

$$(f, g)_{H^m} = \sum_{|\alpha| \leq m} (\partial^\alpha f, \partial^\alpha g)_2,$$

they form Hilbert spaces; we also write $H^{m,2}(\Omega) = H^m(\Omega)$.

The theory of Sobolev spaces is an extensive mathematical field of study. Since their definition and the notion of the weak derivative are based on the Lebesgue integral, the techniques used in this context are closely connected to integration theory. For instance, rules for derivatives such as the chain rule and the product rule also hold for Sobolev functions; however, their justification is considerably more complex. For this purpose, density results are essential. These results are based on the technique of so-called *mollifiers*, which will be introduced in Chap. 3. For most parts of this book, the elementary properties of Sobolev functions will be sufficient, i.e., those properties that can be proven without getting deeper into the theory. In Sect. 6.3, however, some deeper results will be presented in greater detail, since the direct connection to their application is important there. Anyhow, a comprehensive treatment of Sobolev spaces will be beyond the scope of this book and we refer to the corresponding literature (see [2] or [149], for instance).

For assertions on the properties of Sobolev functions, the nature of the domain Ω is crucial in many cases. In particular, the behavior of the functions close to the boundary $\partial\Omega$ is important, which is why assumptions on the boundary are necessary in many cases. For our purposes, however, the notion of a Lipschitz domain introduced in Sect. 2.2.3 will be sufficient.

Let us remark on the sense in which Sobolev functions can attain values on the boundary: Obviously, every $f \in C(\overline{\Omega}) \cap H^{1,p}(\Omega)$ is continuous on the boundary, and thus the restriction lies in every $L_{\mathfrak{H}^{d-1}}^p(\partial\Omega)$. Being equivalence classes of functions, general Sobolev functions are usually not uniformly continuous; however, they possess a so-called trace on $\partial\Omega$:

Theorem 2.80 (Traces of Sobolev Functions) *Let Ω be a bounded Lipschitz domain and $p \in [1, \infty[$. Then there exists a unique linear and continuous mapping $T : H^{1,p}(\Omega) \rightarrow L_{\mathfrak{H}^{d-1}}^p(\partial\Omega)$ such that for all $u \in H^{1,p}(\Omega) \cap C(\overline{\Omega})$, one has $Tu = u|_{\partial\Omega}$. The mapping T is called the trace operator, and the image $Tu \in L_{\mathfrak{H}^{d-1}}^p(\partial\Omega)$ of $u \in H^{1,p}(\Omega)$ is called the trace of the Sobolev function u .*

A proof can be found in [104], for instance. Roughly speaking, it is based on defining T on $H^{1,p}(\Omega) \cap C(\overline{\Omega})$ and then continuously extending it. An important

property of the trace is the fact that Gauss's theorem holds for Sobolev functions on Lipschitz domains:

Theorem 2.81 (Gauss's Theorem, Weak Form) *If Ω is a bounded Lipschitz domain and $f \in H^{1,1}(\Omega, \mathbf{K}^d) = (H^{1,1}(\Omega))^d$ is a Sobolev vector field. Then*

$$\int_{\partial\Omega} f \cdot v \, d\mathfrak{H}^{d-1} = \int_{\Omega} \operatorname{div} f \, dx$$

where $f \in L^1_{\mathfrak{H}^{d-1}}(\partial\Omega)$ is the Sobolev trace of Theorem 2.80 and v is the outer normal introduced in Theorem 2.73. In particular, we have for $f \in H^{1,p}(\Omega, \mathbf{K}^d)$ and $g \in H^{1,p^*}(\Omega)$,

$$\int_{\partial\Omega} f \bar{g} \cdot v \, d\mathfrak{H}^{d-1} = \int_{\Omega} \bar{g} \operatorname{div} f + f \cdot \nabla g \, dx.$$

The proof is again based on the fact that the assertion holds for smooth functions and vector fields. Then density arguments transfer the result to the general case. For the second claim, we additionally use that the product satisfies $fg \in H^{1,1}(\Omega, \mathbf{K}^d)$ and $\operatorname{div}(fg) = \bar{g} \operatorname{div} f + f \cdot \nabla g$.

Chapter 3

Basic Tools



In this book we regard, admittedly slightly arbitrarily, as basic tools histograms and linear and morphological filters. These tools belong to the oldest methods in mathematical image processing and are discussed in early books on digital image processing as well (cf. [67, 114, 119]).

3.1 Continuous and Discrete Images

In Sect. 1.1 we considered images with continuous and discrete image domains. In this book, we essentially work with continuous image domains. However, there are good reasons to deal with discrete images and in particular with the connection of discrete and continuous images:

- In practice, images are given in discrete form.
- In order to apply continuous methods to discrete images, the method has to be discretized. This can, for instance, be achieved by interpolating the discrete image to a continuous one and then employing the continuous method.
- Also images that are given in discrete form often stem from “continuous brightness distributions.” For this purpose, the continuous scene was sampled. What does this sampled image have to do with the real image?

Let us first deal with the interpolation of images.

3.1.1 Interpolation

We consider a discrete image $U : \{1, \dots, N\} \times \{1, \dots, M\} \rightarrow \mathbf{R}$. If we want to rotate this image, for instance, we can determine the value of a pixel of the

rotated image by calculating where this pixel was located before the rotation. This point, however, will in general not be a pixel of the original image, i.e., we have to evaluate the image at intermediate points. Of course, the same happens for other geometric transformations such as stretching, shrinking, shearing, and shifting, for instance. Let us define the following geometric operations on images, which we will encounter frequently in this book:

Definition 3.1 For $y \in \mathbf{R}^d$ and $A \in \mathbf{R}^{d \times d}$, we define

$$\begin{aligned} t_y : \mathbf{R}^d &\rightarrow \mathbf{R}^d, & d_A : \mathbf{R}^d &\rightarrow \mathbf{R}^d, \\ t_y(x) &= x + y, & d_A(x) &= Ax. \end{aligned}$$

By means of that, we define the linear operators for *translation* (shifting) and *linear coordinate transformation* (scaling) on $\mathcal{C}(\mathbf{R}^d)$ by

$$\begin{aligned} T_y : \mathcal{C}(\mathbf{R}^d) &\rightarrow \mathcal{C}(\mathbf{R}^d), & D_A : \mathcal{C}(\mathbf{R}^d) &\rightarrow \mathcal{C}(\mathbf{R}^d), \\ T_y u &= u \circ t_y, & D_A u &= u \circ d_A. \end{aligned}$$

Remark 3.2 The operators T_y and D_A act “from the right,” i.e., they are applied before the use of the function u . For concatenation, one has

$$\begin{aligned} (T_y D_A u)(x) &= (T_y(u \circ d_A))(x) = (u \circ d_A \circ t_y)(x) = u(A(x + y)), \\ (D_A T_y u)(x) &= (D_A(u \circ t_y))(x) = (u \circ t_y \circ d_A)(x) = u(Ax + y). \end{aligned}$$

See also Exercises 3.1 and 3.2.

A good interpolation scheme is *separable*, i.e., a multidimensional interpolation can be induced by a tensor product of one-dimensional interpolations. Therefore, in the following we consider a one-dimensional image $U : \{1, \dots, N\} \rightarrow \mathbf{R}$. In the case of discrete pictures, we mostly use indices instead of the notation with an argument, i.e., $U_j = U(j)$.

Piecewise Constant Interpolation Based on U , we construct the continuous image $u : [1, N] \rightarrow \mathbf{R}$ through

$$\begin{aligned} u(x) &= U_j \quad \text{if } j - \frac{1}{2} \leq x < j + \frac{1}{2} \\ &= U_{\lfloor x + \frac{1}{2} \rfloor} \end{aligned}$$

where $\lfloor y \rfloor$ denotes the greatest integer that is less than y . This interpolation is also called *nearest-neighbor* interpolation and can be interpreted as a

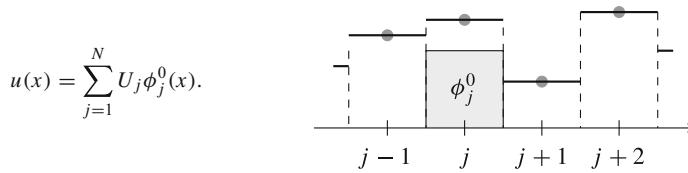
spline-interpolation of zeroth order. For this purpose, we define the vector space

$$V^0 = \left\{ u : [\frac{1}{2}, N + \frac{1}{2}] \rightarrow \mathbf{R} \mid u|_{[j-\frac{1}{2}, j+\frac{1}{2}]} \text{ is constant for } j = 1, \dots, N \right\}.$$

In this vector space, the *plateau functions* form a basis. These functions are given by $\phi_j^0(x) = T_{-j}\phi^0(x) = \phi^0(x - j)$, i.e., translations of

$$\phi^0(x) = \chi_{[-\frac{1}{2}, \frac{1}{2}]}(x) = \begin{cases} 1 & \text{if } x \in [-\frac{1}{2}, \frac{1}{2}], \\ 0 & \text{else.} \end{cases}$$

By means of this, the piecewise constant interpolation of U can be written as



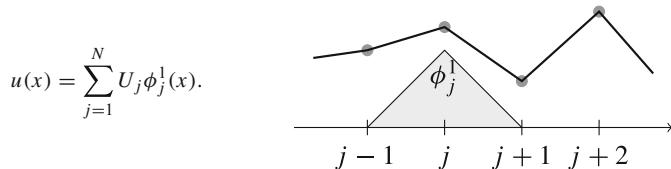
Piecewise Linear Interpolation Piecewise linear interpolation is the classical interpolation ansatz. Between the grid points, the image data is extended linearly, which corresponds to the interpolation with splines of first order. We define the vector space

$$V^1 = \{u : [1, N] \rightarrow \mathbf{R} \mid u \text{ is continuous, } u|_{[j, j+1]} \text{ is linear for } j = 1, \dots, N - 1\}.$$

Obviously, a basis of this vector space is given by the *hat functions* $\phi_j^1(x) = T_{-j}\phi^1(x) = \phi^1(x - j)$ with

$$\phi^1(x) = \begin{cases} x + 1 & \text{if } x \in [-1, 0], \\ 1 - x & \text{if } x \in [0, 1], \\ 0 & \text{otherwise.} \end{cases}$$

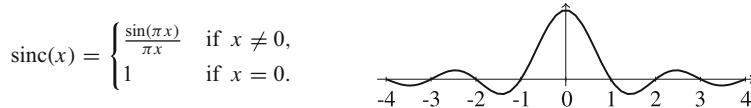
By means of this, the piecewise linear interpolation of U can be written as



Further Interpolation Functions Technically, one can use any interpolating function ϕ for interpolation. We call a function $\phi : \mathbf{R} \rightarrow \mathbf{R}$ an *interpolation function* if

$$\phi(x) = \begin{cases} 1 & \text{if } x = 0 \\ 0 & \text{if } x \in \mathbf{Z} \setminus \{0\} \\ \text{arbitrary} & \text{else.} \end{cases}$$

A function that will play a major role in Sect. 4.2.2 is the sinc function:



Analogously to nearest-neighbor and bilinear interpolation, we obtain the interpolation rule

$$u(x) = \sum_{j=1}^N U_j T_{-j} \text{sinc}(x).$$

Tensor Product Interpolation The methods introduced in the previous paragraphs can easily be generalized to multidimensional images. Here, we state only the formulas for two-dimensional images $U \in \mathbf{R}^{N \times M}$:

Nearest neighbor:

$$u(x, y) = U_{\lfloor x + \frac{1}{2} \rfloor, \lfloor y + \frac{1}{2} \rfloor} = \sum_{i=1}^N \sum_{j=1}^M U_{i,j} \phi_i^0(x) \phi_j^0(y).$$

Piecewise bilinear interpolation:

$$u(x, y) = \sum_{i=1}^N \sum_{j=1}^M U_{i,j} \phi_i^1(x) \phi_j^1(y).$$

General interpolating function:

$$u(x, y) = \sum_{i=1}^N \sum_{j=1}^M U_{i,j} T_{-i} \phi(x) T_{-j} \phi(y).$$

3.1.2 Sampling

In order to digitize a continuous image $u : \Omega \rightarrow \mathbf{R}$, it is usually sampled. This can be achieved in different ways. Let us first describe the so-called *point sampling*, i.e., one defines sampling points x_i in the image domain Ω , and the values $(U_i) = (u(x_i))$ are stored. Usually, images are sampled on regular grids. Let us assume that our image is given on a rectangle or, to make it even simpler, on a square $\Omega = [0, 1]^2$, so a natural choice for the grid is given by $x_{i,j} = (i/N, j/N)$, $i, j = 1, \dots, N$. We can then regard the discrete image $(U_{i,j}) = (u(x_{i,j}))$ as an element $U \in \mathbf{R}^{N \times N}$ as well.

Remark 3.3 (Discrete Images as Delta Comb) We can also apply the continuous approach to sampled images. For this purpose, in the case of point sampling, the following observation is helpful: For a countable set I , let $(U_i)_{i \in I}$ be a discrete image such that $\sum |U_i| < \infty$. On the set I , we define the counting measure μ :

$$\mu(J) = \begin{cases} \text{number of elements in } J & \text{if } J \text{ finite,} \\ \infty & \text{if } J \text{ infinite.} \end{cases}$$

By means of that, we can regard U as an element in $L^1_\mu(I)$, since

$$\|U\|_{L^1_\mu(I)} = \sum_{i \in I} |U_i|.$$

Using the Dirac measure of Example 2.38, we can view U as a delta comb

$$U = \sum U_i \delta_{x_i}$$

and therefore, as a measure on Ω .

Another sampling method is the sampling of mean values. In this case, the domain Ω is partitioned into subsets Ω_i , and the average values of $u : \Omega \rightarrow \mathbf{R}$ on each subset are stored:

$$U_i = \frac{1}{|\Omega_i|} \int_{\Omega_i} u(x) dx.$$

This approach is somewhat closer to what happens inside a digital camera: the chip collects photons over a small area. Mathematically, one can argue that an image u cannot be evaluated pointwise, since it actually corresponds to the “distribution” of brightness. Slightly more generally, we can define the mean sampling also by

means of a test function $\phi \in \mathcal{D}(\mathbf{R}^d)$ with $\phi \geq 0$ and $\int \phi \, dx = 1$. For this purpose, let $x_i \in \mathbf{R}^d$ be the sampling points and

$$U_i = \int_{\mathbf{R}^d} \phi(x - x_i) u(x) \, dx.$$

In this case, the mean sampling is justified for distributions as well, since it corresponds to the application of the distribution to a test function in the sense of Sect. 2.3.

When an image is sampled, some information is obviously lost. However, it can even happen that “wrong” or undesired information creeps in; cf. Fig. 3.1. With both point sampling and mean sampling, errors occur in this way, yet the errors introduced by mean sampling are less obvious. The sampling of continuous images (or signals, respectively) is a particular mathematical theory, which we will cover in Sect. 4.2.2. We can then explain in Sect. 4.2.3 the so-called “alias effect” shown in Fig. 3.1.

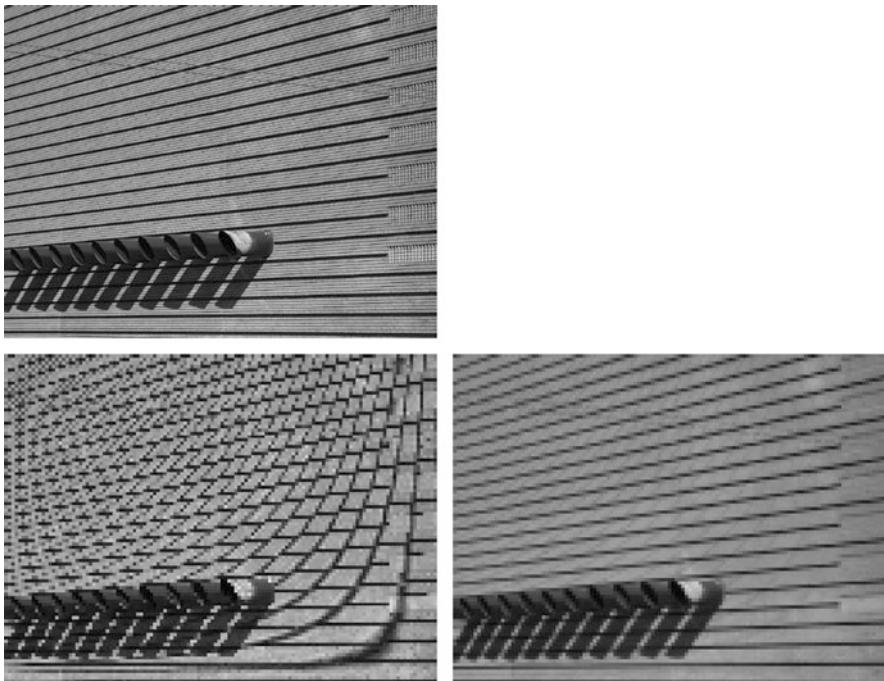


Fig. 3.1 Error due to wrong sampling. Top: Original image in high resolution. Lower left: The same image after an eight times point subsampling, i.e., in the horizontal and vertical directions, every eighth value was taken. Lower right: The same image after an eight times mean subsampling, i.e., the mean was taken of squares of eight by eight pixels, respectively. In order to ease the comparison, the images were scaled to the same size

3.1.3 Error Measures

Definition 3.4 (Mean Squared Error (MSE)) For two continuous images $u, v \in L^2(\Omega)$, the *mean squared error* is given by

$$\text{MSE}(u, v) = \frac{1}{|\Omega|} \|u - v\|_2^2 = \frac{1}{|\Omega|} \int_{\Omega} (u(x) - v(x))^2 dx.$$

If $U, V \in \mathbf{R}^{N \times M}$ are discrete images, we have

$$\text{MSE}(U, V) = \frac{1}{NM} \sum_{i,j} (U_{i,j} - V_{i,j})^2.$$

The mean squared error can be used to evaluate the difference of images, for instance to compare the result of a denoising method with the original image.

In the context of image compression, one compares the uncompressed image with the compressed one. For this purpose, the “peak signal-to-noise ratio,” PSNR, is common. Essentially, the PSNR is a scaled version of the MSE; it measures the ratio of the maximal possible energy of the signal and the energy of the existing noise. The PSNR is usually given logarithmically (more precisely, in decibels):

Definition 3.5 (Peak Signal-to-Noise Ratio (PSNR)) For two continuous images $u, v \in L^2(\Omega)$ with $u, v : \Omega \rightarrow [0, 1]$, the PSNR is given by

$$\text{PSNR}(u, v) = 10 \log_{10} \left(\frac{1}{\text{MSE}(u, v)} \right) \text{db.}$$

If $U, V \in \mathbf{R}^{N \times M}$ are discrete images with $U_{i,j}, V_{i,j} \in [0, 255]$, then we have

$$\text{PSNR}(U, V) = 10 \log_{10} \left(\frac{255^2}{\text{MSE}(u, v)} \right) \text{db.}$$

Note that a higher PSNR value implies a better image quality. We set $\text{PSNR}(u, u) = \infty$; a PSNR value of over 40db typically means that the difference between the images cannot be perceived; cf. Fig. 3.2. The PSNR is designed to measure *noise* or *compression artifacts*. It is not suitable for specifying a distance between two general images that in some sense reflects the “similarity” of these images.

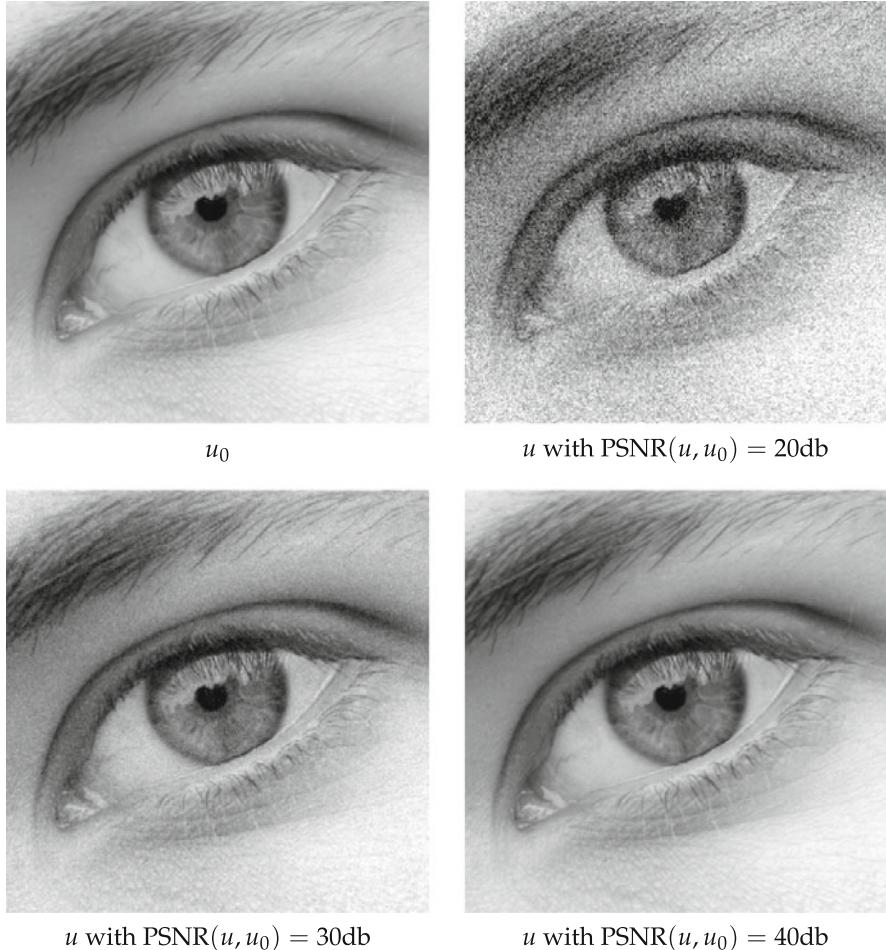


Fig. 3.2 The PSNR for some noisy images

3.2 Histograms

A histogram contains some important properties of the image and is very helpful for several basic applications. Roughly speaking, a histogram specifies how often the different gray values appear in the image. Before we introduce the histogram for continuous images, we consider a basic example:

Example 3.6 (Histogram of a Discrete Image) We consider a discrete image $u : \Omega \rightarrow F$ with $\Omega = \{1, \dots, N\} \times \{1, \dots, M\}$ and $F = \{0, \dots, n\}$. The histogram H_u of u states how often the respective gray values appear in the image:

$$H_u(k) = \#\{(i, j) \in \Omega \mid u_{i,j} = k\}.$$

By means of the Kronecker delta, we can represent the histogram in a different way:

$$H_u(k) = \sum_{i=1}^N \sum_{j=1}^M \delta_{k,u_i,j}.$$

For a continuous image domain Ω , the example can easily be generalized if Ω is endowed with a measure μ . In this case, one sets $H_u(k) = \mu(\{x \in \Omega \mid u(x) = k\})$.

For images with a continuous color space, the example cannot be readily adapted: a particular gray value is attained on a subset of Ω , which can possibly have the measure 0 for every gray value. Therefore, the histogram cannot be defined pointwise for every gray value. Rather, the histogram is a measure itself, which is made precise in the following definition:

Definition 3.7 (Histogram) Let $(\Omega, \mathfrak{F}, \mu)$ be a measure space and $u : \Omega \rightarrow [0, 1]$ a measurable image. Then the *histogram* H_u of u is a measure on $[0, 1]$ defined by

$$H_u(E) = \mu(\{x \in \Omega \mid u(x) \in E\}).$$

We see immediately that $H_u([0, 1]) = \mu(\Omega)$. By means of the *distribution function* $G_u : \mathbf{R} \rightarrow \mathbf{R}$ of the gray values of the image u ,

$$G_u(s) = \mu(\{x \in \Omega \mid u(x) \leq s\}),$$

we obtain an alternative representation of the histogram as the distributional derivative of the distribution function:

Theorem 3.8 Let $(\Omega, \mathfrak{F}, \mu)$ be a σ -finite measure space and $u : \Omega \rightarrow [0, 1]$ a measurable image. Then for the distribution function G_u of u , one has the following

1. $G_u : \mathbf{R} \rightarrow [0, \mu(\Omega)]$ is monotonically increasing with $G_u(s) = 0$ for $s < \text{ess inf } u$ and $G_u(s) = \mu(\Omega)$ for $s > \text{ess sup } u$.
2. $G'_u = H_u$ in the distributional sense.

Proof Obviously, G_u is monotonically increasing. For $s < \text{ess inf } u$, the set $\{x \in \Omega \mid u(x) \leq s\}$ is either empty or a null set. In each case, its measure is zero. For $s > \text{ess sup } u$, the set $\{x \in \Omega \mid u(x) \leq s\}$ comprises the whole of Ω apart from, at most, a set of measure 0, i.e., assertion 1 holds. For assertion 2, we remark that for $\phi = \chi_{]a,b]}$, we have the identity

$$\int_0^1 \phi \, dH_u(t) = G_u(b) - G_u(a) = \int_{\Omega} \phi \circ u \, d\mu,$$

which together with the definition of the integral yields

$$\int_0^1 \phi \, dH_u(t) = \int_{\Omega} \phi \circ u \, d\mu \quad \text{for all } \phi \in \mathcal{D}(]0, 1[).$$

If we now test with $-\phi'$ for $\phi \in \mathcal{D}(]0, 1[)$, we obtain

$$\begin{aligned} - \int_0^1 G_u(t) \phi'(t) \, dt &= - \int_0^1 \int_{\{u(x) \leq t\}} 1 \, d\mu(x) \phi'(t) \, dt = - \int_{\Omega} \int_u^1 \phi'(t) \, dt \, d\mu(x) \\ &= \int_{\Omega} \phi(u(x)) \, d\mu(x) = \int_0^1 \phi \, dH_u(t), \end{aligned}$$

which implies $G'_u = H_u$ in the distributional sense. \square

In Definition 3.7, we have generalized the notion of the histogram of Example 3.6:

Example 3.9 We consider a domain Ω that is split into three disjoint sets Ω_1 , Ω_2 , and Ω_3 . Let the image $u : \Omega \rightarrow [0, \infty[$ be given by

$$u(x) = \begin{cases} s_1 & \text{if } x \in \Omega_1, \\ s_2 & \text{if } x \in \Omega_2, \\ s_3 & \text{if } x \in \Omega_3, \end{cases}$$

with $0 < s_1 < s_2 < s_3$, i.e., the color space is essentially discrete. Then the distribution function reads

$$G_u(s) = \begin{cases} 0 & \text{if } s < s_1, \\ \mu(\Omega_1) & \text{if } s_1 \leq s < s_2, \\ \mu(\Omega_1) + \mu(\Omega_2) & \text{if } s_2 \leq s < s_3, \\ \mu(\Omega_1) + \mu(\Omega_2) + \mu(\Omega_3) = \mu(\Omega) & \text{if } s_3 \leq s. \end{cases}$$

Therefore, the histogram is given by

$$H_u = G'_u = \mu(\Omega_1)\delta_{s_1} + \mu(\Omega_2)\delta_{s_2} + \mu(\Omega_3)\delta_{s_3},$$

where δ_{s_i} denotes the Dirac measure of Example 2.38.

In two small applications, we demonstrate for what purposes the histogram can be utilized.

Application 3.10 (Contrast Improvement Through Histogram Equalization)
Due to bad conditions while shooting a photo or wrongly adjusted optical settings,

the resulting image may exhibit a low contrast. By means of the histogram, a reasonable contrast improvement method can be motivated.

An image with high contrast usually has gray values in all the range available. If we assume this range to be the interval $[0, 1]$, the linear gray value spread

$$s \mapsto \frac{s - \text{ess inf } u}{\text{ess sup } u - \text{ess inf } u}$$

leads to a full coverage of the range. However, this does not necessarily suffice to increase the contrast sufficiently in all parts of the image, i.e., in particular areas, the contrast can still be improvable. One possibility is to distribute the gray values equally over the range of gray values as much as possible. For this purpose, we look for a monotonic function $\Phi : \mathbf{R} \rightarrow [0, 1]$ such that

$$H_{\Phi \circ u}([a, b]) = |b - a| \mu(\Omega).$$

Using the distribution function, this reads

$$G_{\Phi \circ u}(s) = s \mu(\Omega).$$

If we assume that Φ is invertible, we obtain

$$\begin{aligned} s \mu(\Omega) &= \mu(\{x \in \Omega \mid \Phi(u(x)) \leq s\}) \\ &= \mu(\{x \in \Omega \mid u(x) \leq \Phi^{-1}(s)\}) \\ &= G_u(\Phi^{-1}(s)). \end{aligned}$$

This leads to $\Phi^{-1}(s) = G_u^{-1}(s \mu(\Omega))$ and finally to

$$\Phi(s) = G_u(s)/\mu(\Omega).$$

Hence, the gray value transformation is the distribution function itself.

If the range of gray values is discrete, i.e., $u : \Omega \rightarrow \{0, \dots, n\}$, for instance, we have to round in some meaningful way:

$$\Phi(s_0) = \text{round}\left(\frac{n}{\mu(\Omega)} G_u(s_0)\right) = \text{round}\left(\frac{n}{\mu(\Omega)} \sum_{s=0}^{s_0} H_u(s)\right). \quad (3.1)$$

Of course, in a discrete setting, an equalized histogram cannot be achieved, since equal gray values are mapped to equal gray values again. Nevertheless, this gray value transformation often yields good results, see Fig. 3.3.

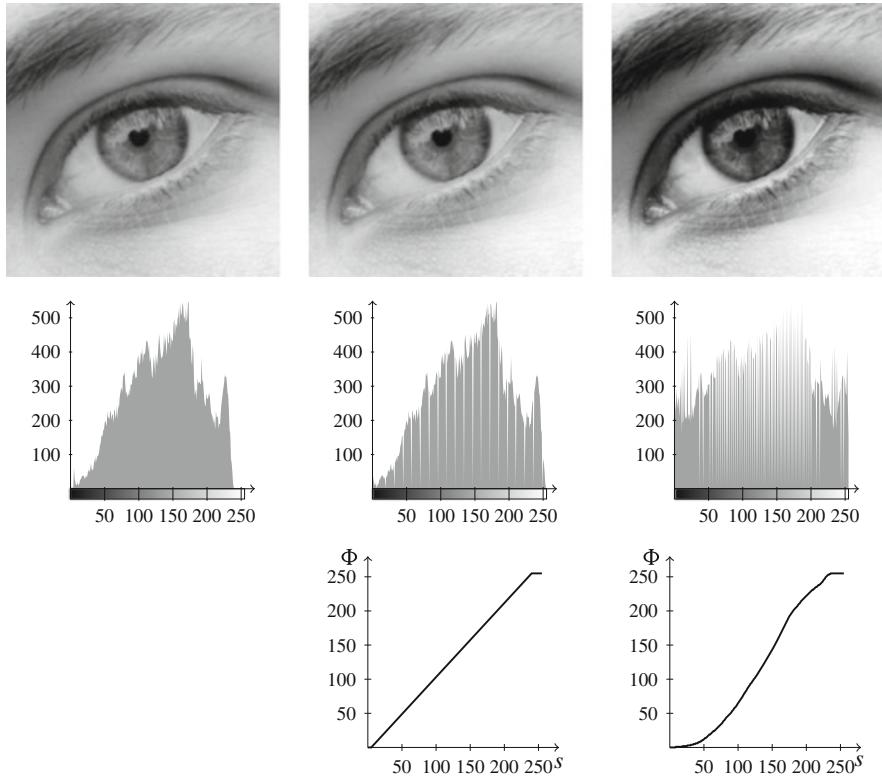


Fig. 3.3 Histogram equalization according to Application 3.10. Left column: Original image with histogram. Middle column: Spreading of the gray values to the full range. Right column: Image with equalized histogram. Lowest row: Respective transformation Φ

Application 3.11 (Segmentation Through Thresholding) A scan $u : \Omega \rightarrow [0, S]$ of a handwritten note or a black-and-white printout typically contains noise. In particular, in most cases, nearly all gray levels occur. If this scan is to be printed or archived, for instance, it is reasonable to first reduce the levels of gray to two again, namely black and white. For this purpose, it is advisable to use a *threshold*: All gray values below the threshold s_0 become black; all gray values above become white,

$$\tilde{u}(x) \mapsto \begin{cases} 0 & \text{if } u(x) \leq s_0, \\ 1 & \text{if } u(x) > s_0. \end{cases}$$

The question remains how to determine the threshold. There are numerous corresponding methods. In this simple case, the following idea often works well:

We select the threshold such that it corresponds to the arithmetic mean of the centers of mass of the histogram above and below the threshold.

Since the center of mass corresponds to the normed first moment, we can write this formula as follows: The threshold s_0 satisfies the equation

$$s_0 = \frac{1}{2} \left(\frac{\int_0^{s_0} s \, dH_u(s)}{\int_0^{s_0} 1 \, dH_u(s)} + \frac{\int_{s_0}^S s \, dH_u(s)}{\int_{s_0}^S 1 \, dH_u(s)} \right).$$

This equation can be solved by a fixed-point iteration, for instance

$$s_0^{n+1} = f(s_0^n) \text{ with } f(s_0) = \frac{1}{2} \left(\frac{\int_0^{s_0} s \, dH_u(s)}{\int_0^{s_0} 1 \, dH_u(s)} + \frac{\int_{s_0}^S s \, dH_u(s)}{\int_{s_0}^S 1 \, dH_u(s)} \right).$$

Why does this fixed-point iteration converge? The iteration map f , as a sum of two increasing functions, is monotonically increasing. Furthermore, we have

$$f(0) = \frac{1}{2} \frac{\int_0^S s H_u(s) \, ds}{\int_0^S H_u(s) \, ds} \geq 0, \quad f(S) = \frac{1}{2} \left(\frac{\int_0^S s H_u(s) \, ds}{\int_0^S H_u(s) \, ds} + S \right) \leq S.$$

Due to the monotonicity, there exists at least one fixed point with a slope of less than one. Thus, the fixed-point iteration converges.

This method is also known as the *isodata algorithm* and has been used since the 1970s (cf. [117]). An example is given in Fig. 3.4.

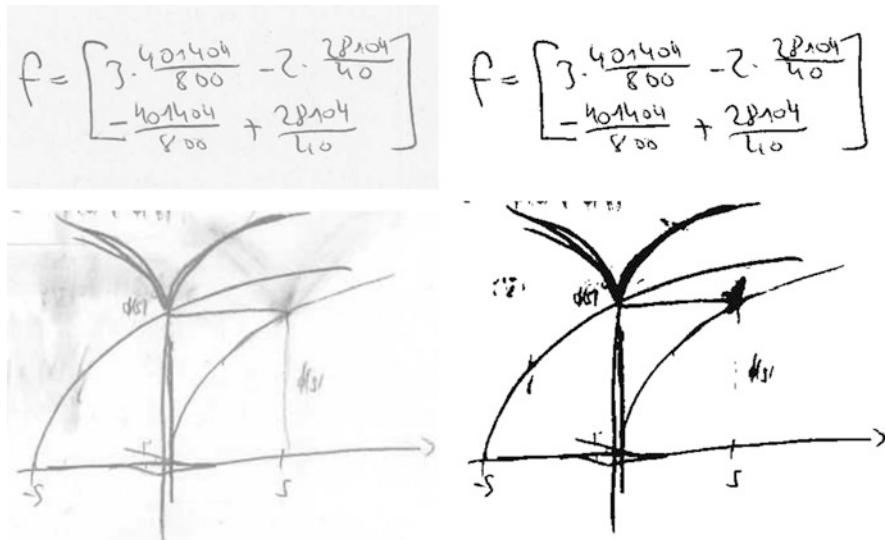


Fig. 3.4 Segmentation through thresholding according to Application 3.11. Left: Scanned handwritings. Right: Segmentation

3.3 Linear Filters

Linear filters belong to the oldest tools in digital image processing. We first consider an introductory example:

Example 3.12 (Denoising with the Moving Average) We consider a continuous image $u : \mathbf{R}^d \rightarrow \mathbf{R}$ and expect that this image exhibits some form of noise, i.e., we assume that the image u results from a real u^\dagger by adding noise n :

$$u = u^\dagger + n.$$

Furthermore, we assume that the noise is distributed in some way uniformly around zero (this is not a precise mathematical assumption, but we do without a more precise formulation here). In order to reduce the noise, we take averages over neighboring values and hope that this procedure will suppress the noise. In formulas, this reads: for a radius $r > 0$, we compute

$$M_r u(x) = \frac{1}{\mathcal{L}^d(B_r(0))} \int_{B_r(x)} u(y) dy.$$

We can equivalently write this by means of the indicator function as

$$M_r u(x) = \frac{1}{\mathcal{L}^d(B_r(0))} \int_{\mathbf{R}^d} u(y) \chi_{B_r(0)}(x + y) dy.$$

This operation is also called *moving average*, see Fig. 3.5.

Let us further remark that the operation M_r also models the “out-of-focus” blur, i.e., the blurring that occurs if the object does not lie in the plane of focus of the camera.

The example above describes what in digital image processing is called a *filter*. The function $\chi_{B_r(0)}$ is called a *filter function*. The mathematical structure that

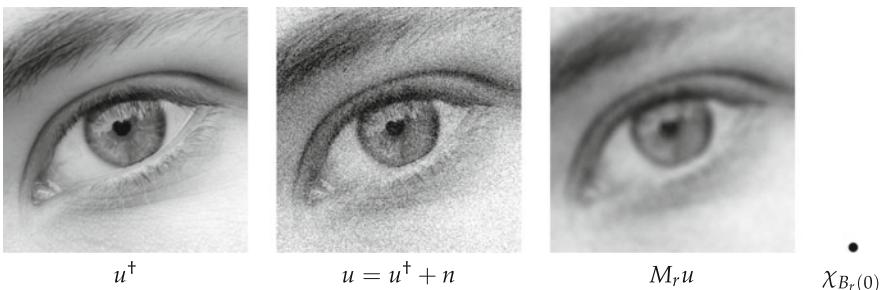


Fig. 3.5 Denoising with moving average. Left: Original image. Middle: Original image with noise. Right: Application of the moving average. Next to this: The indicator function used

underlies filtering is the convolution (apart from the sign in the argument of the filter function).

3.3.1 Definition and Properties

The restriction to indicator functions in Example 3.12 is not necessary, of course. We can consider more general weighted moving averages. For this purpose, let $u, h : \mathbf{R}^d \rightarrow \mathbf{R}$ be measurable. Then we define the *convolution of u with h* by

$$(u * h)(x) = \int_{\mathbf{R}^d} u(y)h(x - y) dy.$$

The function h is also called a convolution kernel. Obviously, the convolution is linear in both arguments. Furthermore, the convolution is translation invariant in the following sense:

$$T_y(u * h) = (u * T_y h) = (T_y u * h).$$

Further properties of the convolution are listed in the following theorem:

Theorem 3.13 (Properties of the Convolution)

1. For $1 \leq p, q, r \leq \infty$, $\frac{1}{r} + 1 = \frac{1}{p} + \frac{1}{q}$, $u \in L^p(\mathbf{R}^d)$, and $v \in L^q(\mathbf{R}^d)$, we have $u * v = v * u \in L^r(\mathbf{R}^d)$, and in particular, Young's inequality holds:

$$\|u * v\|_r \leq \|u\|_p \|v\|_q.$$

2. For $1 \leq p \leq \infty$, $u \in L^p(\mathbf{R}^d)$, and $\phi : \mathbf{R}^d \rightarrow \mathbf{R}$ k -times continuously differentiable with compact support, the convolution $u * \phi$ is k -times continuously differentiable, and for all multi-indices α with $|\alpha| \leq k$, one has

$$\frac{\partial^\alpha (u * \phi)}{\partial x^\alpha} = u * \frac{\partial^\alpha \phi}{\partial x^\alpha}.$$

3. Let be $\phi \in L^1(\mathbf{R}^d)$ with

$$\phi \geq 0, \quad \int_{\mathbf{R}^d} \phi(x) dx = 1,$$

and for $\varepsilon > 0$, set

$$\phi_\varepsilon(x) = \frac{1}{\varepsilon^d} \phi\left(\frac{x}{\varepsilon}\right).$$

Then for a uniformly continuous and bounded function $u : \mathbf{R}^d \rightarrow \mathbf{R}$, one has

$$(u * \phi_\varepsilon)(x) \rightarrow u(x) \text{ for } \varepsilon \rightarrow 0.$$

Furthermore, $u * \phi_\varepsilon$ converges uniformly on every compact subset of \mathbf{R}^d .

Proof

1. The equality $u * v = v * u$ follows by integral substitution. We consider the case $q = 1$ (i.e., $r = p$). One has

$$|u * v(x)| \leq \int_{\mathbf{R}^d} |u(x - y)| |v(y)| dy.$$

For $p = \infty$, we can pull the supremum of $|u|$ out of the integral and obtain the required estimate. For $p < \infty$, we integrate the p th power of $u * v$ and get

$$\int_{\mathbf{R}^d} |u * v(x)|^p dx \leq \int_{\mathbf{R}^d} \left(\int_{\mathbf{R}^d} |u(x - y)| |v(y)| dy \right)^p dx.$$

For $p = 1$, according to Fubini's theorem, we have

$$\int_{\mathbf{R}^d} |u * v(x)| dx \leq \int_{\mathbf{R}^d} \int_{\mathbf{R}^d} |u(x - y)| dx |v(y)| dy,$$

which yields the assertion. For $1 < p < \infty$ and $\frac{1}{p} + \frac{1}{p^*} = 1$, we apply Hölder's inequality and obtain (using Fubini's theorem)

$$\begin{aligned} \int_{\mathbf{R}^d} |u * v(x)|^p dx &\leq \int_{\mathbf{R}^d} \left(\int_{\mathbf{R}^d} |u(x - y)| |v(y)|^{1/p} |v(y)|^{1/p^*} dy \right)^p dx \\ &\leq \int_{\mathbf{R}^d} \left(\left(\int_{\mathbf{R}^d} |u(x - y)|^p |v(y)| dy \right) \left(\int_{\mathbf{R}^d} |v(y)| dy \right)^{p/p^*} \right) dx \\ &= \int_{\mathbf{R}^d} |v(y)| \left(\int_{\mathbf{R}^d} |u(x - y)|^p dx \right) dy \|v\|_1^{p/p^*} \\ &\leq \int_{\mathbf{R}^d} |u(x)|^p dx \|v\|_1^{p/p^* + 1}. \end{aligned}$$

The application of Fubini's theorem is justified here in retrospect since the latter integrals exist. After taking the p th root, the assertion follows. For $q > 1$, the proof is similar but more complicated, cf. [91], for instance.

2. We prove the assertion only for the first partial derivatives; the general case then follows through repeated application. We consider the difference quotient in the

direction of the i th unit vector e_i :

$$\frac{(u * \phi)(x + te_i) - (u * \phi)(x)}{t} = \int_{\mathbf{R}^d} u(y) \frac{\phi(x + te_i - y) - \phi(x - y)}{t} dy.$$

Since the quotient $(\phi(x + te_i - y) - \phi(x - y))/t$ converges uniformly to $\frac{\partial \phi}{\partial x_i}(x - y)$, the assertion follows.

3. To begin with, we remark that

$$\int_{\mathbf{R}^d} \phi_\varepsilon(x) dx = 1, \quad \int_{\mathbf{R}^d \setminus B_\rho(0)} \phi_\varepsilon(x) dx \rightarrow 0 \text{ for } \varepsilon \rightarrow 0,$$

which can be found using the variable transformation $\xi = x/\varepsilon$. We conclude that

$$|(u * \phi_\varepsilon)(x) - u(x)| \leq \int_{\mathbf{R}^d} |u(x - y) - u(x)| |\phi_\varepsilon(y)| dy.$$

We choose $\rho > 0$ and split the integral into large and small y :

$$\begin{aligned} & |(u * \phi_\varepsilon)(x) - u(x)| \\ & \leq \underbrace{\int_{B_\rho(0)} |u(x - y) - u(x)| |\phi_\varepsilon(y)| dy}_{\leq 1} + \underbrace{\int_{\mathbf{R}^d \setminus B_\rho(0)} |u(x - y) - u(x)| |\phi_\varepsilon(y)| dy}_{\rightarrow 0 \text{ for } \rho \rightarrow 0} \\ & \leq \underbrace{\int_{B_\rho(0)} |\phi_\varepsilon(y)| dy}_{\leq 1} \sup_{|y| \leq \rho} |u(x - y) - u(x)| \\ & \quad + \underbrace{\int_{\mathbf{R}^d \setminus B_\rho(0)} |\phi_\varepsilon(y)| dy}_{\rightarrow 0 \text{ for } \varepsilon \rightarrow 0 \text{ and every } \rho > 0} \underbrace{\sup_{|y| \geq \rho} |u(x - y) - u(x)|}_{\leq 2 \sup |u(x)|}. \end{aligned}$$

On the one hand, this shows the pointwise convergence; on the other hand, it also shows the uniform convergence on compact sets. \square

Expressed in words, the properties of the convolution read:

1. The convolution is a linear and continuous operation if it is considered between the function spaces specified.
2. The convolution of functions inherits the smoothness of the smoother function of the two, and the derivative of the convolution corresponds to the convolution with the derivative.
3. The convolution of a function u with a “narrow” function ϕ approximates u in a certain sense.

For image processing, the second and third property are of particular interest: Convolving with a smooth function smooths the image. When convolving with a function ϕ_ε , for small ε , only a small error occurs.

In fact, the convolution is often even slightly smoother than the smoother one of the functions. A basic example for this case is given in the following proposition:

Theorem 3.14 *Let $p > 1$ and p^* the dual exponent, $u \in L^p(\mathbf{R}^d)$, and $v \in L^{p^*}(\mathbf{R}^d)$. Then $u * v \in \mathcal{C}(\mathbf{R}^d)$.*

Proof For $h \in \mathbf{R}^d$, we estimate by means of Hölder's inequality:

$$\begin{aligned} |u * v(x + h) - u * v(x)| &\leq \int_{\mathbf{R}^d} |u(y)| |v(x + h - y) - v(x - y)| dy \\ &\leq \left(\int_{\mathbf{R}^d} |u(y)|^p dy \right)^{1/p} \left(\int_{\mathbf{R}^d} |v(x + h - y) - v(x - y)|^{p^*} dy \right)^{1/p^*} \\ &= \|u\|_p \|Th v - v\|_{p^*}. \end{aligned}$$

The fact that the last integral converges to 0 for $h \rightarrow 0$ corresponds to the assertion that L^{p^*} -functions are continuous in the p^* -mean for $1 \leq p < \infty$, see Exercise 3.4.

□

Note that Theorem 3.13 yields only $u * v \in L^\infty(\mathbf{R}^d)$ in this case.

The smoothing properties of the convolution can also be formulated in several other ways than in Theorem 3.13, for instance, in Sobolev spaces $H^{m,p}(\mathbf{R}^d)$:

Theorem 3.15 *Let $m \in \mathbf{N}$, $1 \leq p, q \leq \infty$, $\frac{1}{r} + 1 = \frac{1}{p} + \frac{1}{q}$, $u \in L^p(\mathbf{R}^d)$, and $h \in H^{m,q}(\mathbf{R}^d)$. Then $u * h \in H^{m,r}(\mathbf{R}^d)$, and for the weak derivatives up to order m , we have*

$$\partial^\alpha(u * h) = u * \partial^\alpha h \quad \text{almost everywhere.}$$

Furthermore, let Ω_1 and Ω_2 be domains. The assertion holds with $u * h \in H^{m,r}(\Omega_2)$ if $u \in L^p(\Omega_1)$ and $h \in H^{m,q}(\Omega_2 - \Omega_1)$.

Proof We use the definition of the weak derivative and calculate, similarly to Theorem 3.13, using Fubini's theorem:

$$\begin{aligned} \int_{\mathbf{R}^d} \partial^\alpha(u * h)\phi(x) dx &= (-1)^{|\alpha|} \int_{\mathbf{R}^d} (u * h)(x) \frac{\partial^\alpha}{\partial x^\alpha} \phi(x) dx \\ &= (-1)^{|\alpha|} \int_{\mathbf{R}^d} \int_{\mathbf{R}^d} u(y) h(x - y) dy \frac{\partial^\alpha}{\partial x^\alpha} \phi(x) dx \\ &= (-1)^{|\alpha|} \int_{\mathbf{R}^d} u(y) \int_{\mathbf{R}^d} h(x - y) \frac{\partial^\alpha}{\partial x^\alpha} \phi(x) dx dy \\ &= \int_{\mathbf{R}^d} u(y) \int_{\mathbf{R}^d} \partial^\alpha h(x - y) \phi(x) dx dy \\ &= \int_{\mathbf{R}^d} (u * \partial^\alpha h)(x) \phi(x) dx. \end{aligned}$$

Now the asserted rule for the derivative follows from the fundamental lemma of the calculus of variations, Lemma 2.75. Due to $\partial^\alpha h \in L^q(\mathbf{R}^d)$, Theorem 3.13 yields $\partial^\alpha(u * h) = u * \partial^\alpha h \in L^r(\mathbf{R}^d)$ for $|\alpha| \leq m$, and hence we have $u * h \in H^{m,p}(\mathbf{R}^d)$. In particular, the existence of the integral on the left-hand side above is justified retrospectively.

The additional claim follows analogously; we remark, however, that for $\phi \in \mathcal{D}(\Omega_2)$, extending by zero and $y \in \Omega_1$, we have $T_y\phi \in \mathcal{D}(\Omega_2 - y)$. Thus, the definition of the weak derivative can be applied. \square

If the function ϕ of Theorem 3.13 (3) additionally lies in $\mathcal{D}(\mathbf{R}^d)$, the functions ϕ_ε are also called *mollifiers*. This is motivated by the fact that in this case, $u * \phi_\varepsilon$ is infinitely differentiable and for small ε , it differs only slightly from u . This situation can be expressed more precisely in various norms, such as the L^p -norms, for instance:

Lemma 3.16 *Let $\Omega \subset \mathbf{R}^d$ be a domain and $1 \leq p < \infty$. Then for every $u \in L^p(\Omega)$, mollifier ϕ , and $\delta > 0$, there exists some $\varepsilon > 0$ such that*

$$\|\phi_\varepsilon * u - u\|_p < \delta.$$

In particular, the space $\mathcal{D}(\Omega)$ is dense in $L^p(\Omega)$.

Proof For $\delta > 0$ and $u \in L^p(\Omega)$, there exists, according to Theorem 2.55, some $f \in \mathcal{C}_c(\Omega)$ such that $\|u - f\|_p < \delta/3$. Together with the triangle inequality and Young's inequality, this yields

$$\|\phi_\varepsilon * u - u\|_p \leq \|\phi_\varepsilon * u - \phi_\varepsilon * f\|_p + \|\phi_\varepsilon * f - f\|_p + \|f - u\|_p \leq \frac{2\delta}{3} + \|\phi_\varepsilon * f - f\|_p.$$

Since f has compact support and is continuous, it is uniformly continuous as well, and according to Theorem 3.13 (3), it follows that for sufficiently small $\varepsilon > 0$, one has $\|\phi_\varepsilon * f - f\|_p < \delta/3$. By means of the above estimate, the assertion follows.

The remaining claim can be shown analogously, using the modification

$$\|u - \phi_\varepsilon * f\|_p \leq \|u - f\|_p + \|\phi_\varepsilon * f - f\|_p < \frac{2\delta}{3} < \delta,$$

where $\varepsilon > 0$ is chosen sufficiently small such that we have $\overline{\text{supp } f - \text{supp } \phi_\varepsilon} \subset\subset \Omega$. Together with $\phi_\varepsilon * f \in \mathcal{D}(\Omega)$, this implies the density. \square

A similar density result also holds for Sobolev spaces:

Lemma 3.17 *Let $\Omega \subset \mathbf{R}^d$ be a domain, $m \in \mathbf{N}$ with $m \geq 1$, $1 \leq p < \infty$ and $u \in H^{m,p}(\Omega)$. Then for every $\delta > 0$ and subdomain Ω' with $\overline{\Omega'} \subset\subset \Omega$, there exists $f \in \mathcal{C}^\infty(\Omega)$ such that $\|u - f\|_{m,p} < \delta$ on Ω' .*

Proof We choose a mollifier ϕ and $\varepsilon_0 > 0$ such that $\overline{\Omega' - \text{supp } \phi_\varepsilon} \subset\subset \Omega$ holds for all $\varepsilon \in]0, \varepsilon_0[$. The function $f_\varepsilon = \phi_\varepsilon * u$ now lies in $\mathcal{C}^\infty(\Omega)$ and according

to Theorem 3.15, for every $\varepsilon \in]0, \varepsilon_0[$, one has $f_\varepsilon \in H^{m,p}(\Omega')$, where we have $\partial^\alpha f_\varepsilon = u * \partial^\alpha \phi_\varepsilon$ for every multi-index α with $|\alpha| \leq m$. By means of Lemma 3.16, we can choose ε sufficiently small such that for every multi-index with $|\alpha| \leq m$, one has

$$\|\partial^\alpha(u - f_\varepsilon)\|_p = \|\partial^\alpha u - \phi_\varepsilon * \partial^\alpha u\|_p < \frac{\delta}{M} \quad \text{in } \Omega',$$

where M denotes the number of multi-indices with $|\alpha| \leq m$. Setting $f = f_\varepsilon$ and using the Minkowski inequality yields the desired assertion. \square

Theorem 3.18 *Let $1 \leq p < \infty$ and $m \in \mathbf{N}$. Then the space $\mathcal{D}(\mathbf{R}^d)$ is dense in $H^{m,p}(\mathbf{R}^d)$.*

Proof We first show that the set $\mathcal{C}^\infty(\mathbf{R}^d) \cap H^{m,p}(\mathbf{R}^d)$ is dense. Let $u \in H^{m,p}(\Omega)$ and ϕ a mollifier. Since in particular $u \in L^p(\mathbf{R}^d)$, we obtain $u * \phi_\varepsilon \rightarrow u$ in $L^p(\mathbf{R}^d)$ and for $|\alpha| \leq m$, $(\partial^\alpha u) * \phi_\varepsilon = \partial^\alpha(u * \phi_\varepsilon)$ converges to $\partial^\alpha u$ in $L^p(\mathbf{R}^d)$, i.e., we have $u * \phi_\varepsilon \rightarrow u$ in $H^{m,p}(\mathbf{R}^d)$.

Now we show that the space $\mathcal{D}(\mathbf{R}^d)$ is dense in $\mathcal{C}^\infty(\mathbf{R}^d) \cap H^{m,p}(\mathbf{R}^d)$ (which will complete the proof). For this purpose, let $u \in \mathcal{C}^\infty(\mathbf{R}^d) \cap H^{m,p}(\mathbf{R}^d)$, which implies in particular that the classical derivatives $\partial^\alpha u$ up to order m are in $L^p(\mathbf{R}^d)$. Now let $\eta \in \mathcal{D}(\mathbf{R}^d)$ with $\eta \equiv 1$ in a neighborhood of zero. For $R > 0$, we consider the functions $u_R(x) = u(x)\eta(x/R)$. Then $u_R \in \mathcal{D}(\mathbf{R}^d)$, and according to the dominated convergence theorem, we also have $u_R \rightarrow u$ in $L^p(\mathbf{R}^d)$ for $R \rightarrow \infty$. For the partial derivatives, due to the Leibniz formula, we have

$$(\partial^\alpha u_R)(x) = \sum_{\beta \leq \alpha} \binom{\alpha}{\beta} (\partial^{\alpha-\beta} u)(x) R^{-|\beta|} (\partial^\beta \eta)(x/R).$$

The summand for $\beta = 0$ converges to $\partial^\alpha u$ in $L^p(\mathbf{R}^d)$ (again due to dominated convergence). The remaining summands even converge to zero. Therefore, we have $\partial^\alpha u_R \rightarrow \partial^\alpha u$ in $L^p(\mathbf{R}^d)$, and we conclude that $u_R \rightarrow u$ in $H^{m,p}(\mathbf{R}^d)$. \square

Remark 3.19 If $u \in H^{m,p}(\mathbf{R}^d)$ has compact support, then for every $\varepsilon > 0$ and Ω with $\text{supp } u \subset\subset \Omega$ we can find by means of the above argument some $\phi \in \mathcal{D}(\Omega)$ such that $\|u - \phi\|_{m,p} < \varepsilon$ in Ω . Note that we did not prove the density result for domains $\Omega \subset \mathbf{R}^d$. In general, $\mathcal{D}(\Omega)$ is not dense in $H^{m,p}(\Omega)$; note that the closure of $\mathcal{D}(\Omega)$ has been denoted by $H_0^{m,p}(\Omega)$. However, the space $\mathcal{C}^\infty(\overline{\Omega})$ is often dense in $H^{m,p}(\Omega)$ as we will demonstrate in Proposition 6.74.

The convolution is defined not only for suitable integrable functions. For instance, measures can also be convolved with integrable functions: if μ is a measure on \mathbf{R}^d and $f : \mathbf{R}^d \rightarrow \mathbf{R}$ an integrable function, then

$$\mu * f(x) = \int_{\mathbf{R}^d} f(x-y) d\mu(y).$$

Remark 3.20 (Interpolation as Convolution) By means of the convolution of a measure and a function, we can take a new look at the interpolation of images. According to Remark 3.3, we write a discrete image $(U_{i,j})$ on a regular quadratic grid as a delta comb:

$$U = \sum U_{i,j} \delta_{x_{i,j}}, \quad x_{i,j} = (i, j).$$

The interpolation of U with an interpolation function ϕ is then given by

$$u(x) = U * \phi(x) = \int_{\mathbf{R}^2} \phi(x - y) dU(y) = \sum U_{i,j} \phi(x - x_{i,j}).$$

Remark 3.21 In image processing, it is more common to speak of filters rather than convolutions. This corresponds to a convolution with a reflected convolution kernel: the *linear filter* for h is defined by $u * D_{-\text{id}}h$. If not stated otherwise, we will use the term “linear filter” for the operation $u * h$ in this book.

3.3.2 Applications

By means of linear filters, one can create interesting effects and also tackle some of the fundamental problems of image processing. Exemplarily, we here show three applications:

Application 3.22 (Effect Filters) Some effects of analog photography can be realized by means of linear filters:

Duto filter: The Duto filter overlays the image with a smoothed version of itself.

The result gives the impression of blur on the one hand, whereas on the other hand, the sharpness is maintained. This results in a “dream-like” effect. In mathematical terms, the Duto filter can be realized by means of a convolution with a Gaussian function, for instance. For this purpose, let

$$G_\sigma(x) = \frac{1}{(2\pi\sigma)^{d/2}} \exp\left(\frac{-|x|^2}{2\sigma}\right) \quad (3.2)$$

be the d -dimensional Gaussian function with variance σ . In the case of the Duto filter, the convolved image is linearly overlayed with the original image. This can be written as a convex combination with parameter $\lambda \in [0, 1]$:

$$\lambda u + (1 - \lambda)u * G_\sigma.$$

Motion blur: If an object (or the camera) moves during the exposure time, a point of the object is mapped onto a line. For a linear motion of length l in direction

$v \in \mathbf{R}^d$, we can write the motion blur as

$$\frac{1}{l} \int_0^l u(x + tv) dt.$$

This also constitutes a linear operation. However, it is not realized by a convolution with a function, but with a measure. For this purpose, we consider the line $L = \{tv \mid 0 \leq t \leq l\}$ of length l and define the measure

$$\mu = \frac{1}{l} \mathfrak{H}^1 \llcorner L$$

by means of the Hausdorff measure of Example 2.38. The motion blur is then given by

$$\int u(x + y) d\mu(y).$$

In Fig. 3.6, an example for the application of the Duto filter as well as motion blurring is given.

Application 3.23 (Edge Detection According to Canny) For now, we simply define an edge by saying that at an edge, the gray value changes abruptly. For a one-dimensional gray value distribution u , it is reasonable to locate an edge at the point with maximal slope. This point can also be determined as a root of the second derivative u'' ; cf. also Fig. 3.7. If the image is noisy, differentiability cannot be expected, and numerical differentiation leads to many local maxima of the derivative and, therefore, to many roots of the second derivative as well. Hence, it is advisable to smoothen the image beforehand by means of a convolution with a smooth function f and compute the derivative after this. For the derivative of the



Fig. 3.6 Effect filters of Application 3.22. Left: Original image. Center: Duto blurrer with $\lambda = 0.5$. Right: Motion blurring

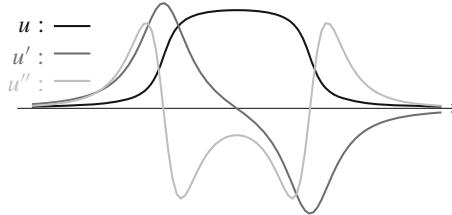


Fig. 3.7 A one-dimensional gray value distribution and its first two derivatives

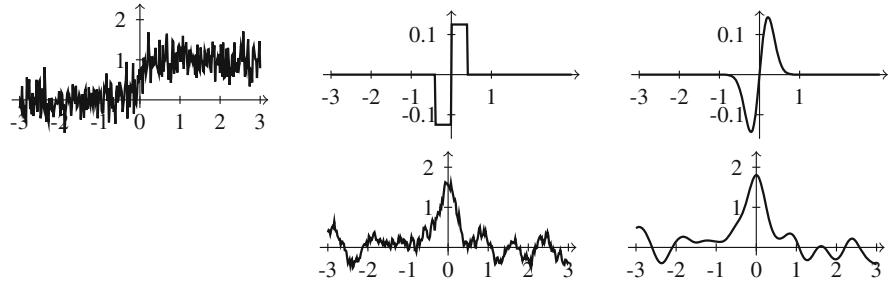


Fig. 3.8 Edge detection by determining the maxima of the smoothed derivative. Left: A noisy edge at zero. Center and right: On top a filter function, below the corresponding result of the filtering

convolution, we can write according to Proposition 3.13,

$$\frac{\partial(u * f)(x)}{\partial x} = (u * f')(x).$$

Different functions $h = f'$ lead to different qualities of results for the edge detection by determining the maxima of $(u * f')(x)$, cf. Fig. 3.8.

Canny [29] presents a lengthy derivation of a class of, in some sense, optimal functions f . Here, we present a heuristic variant, which leads to the same result. Edges exist on different scales, i.e., there are “coarse edges” and “fine edges.” Fine edges belong to small, delicately structured objects and are thus suppressed by a convolution with a function with a high variance. The convolution with a function with a small variance changes the image only slightly (cf. Proposition 3.13) and hence preserves all edges. Therefore, we consider rescaled versions $f_\sigma(x) = \sigma^{-1}f(\sigma^{-1}x)$ of a given convolution kernel f . If σ is large, f_σ is “wider” and if σ is small, f_σ is “narrower.” For one original image u_0 , we thus obtain a whole class of smoothed images:

$$u(x, \sigma) = u_0 * f_\sigma(x).$$

We now formulate requirements for finding a suitable f . We require that the location of the edges remain constant for different σ . Furthermore, no new edges shall appear

for larger σ . In view of Fig. 3.7, we hence require that at an edge point x_0 , we have

$$\begin{aligned}\frac{\partial^2}{\partial x^2}u(x_0, \sigma) > 0 &\implies \frac{\partial}{\partial \sigma}u(x_0, \sigma) > 0, \\ \frac{\partial^2}{\partial x^2}u(x_0, \sigma) = 0 &\implies \frac{\partial}{\partial \sigma}u(x_0, \sigma) = 0, \\ \frac{\partial^2}{\partial x^2}u(x_0, \sigma) < 0 &\implies \frac{\partial}{\partial \sigma}u(x_0, \sigma) < 0.\end{aligned}$$

In other words, if the second derivative in the x -direction of u is positive (or zero/negative), then u will increase (or remain constant/decrease) for increasing σ , i.e., for coarser scales. In order to ensure this, we require that the function u solve the following differential equation:

$$\frac{\partial}{\partial \sigma}u(x, \sigma) = \frac{\partial^2}{\partial x^2}u(x, \sigma). \quad (3.3)$$

Furthermore, $\sigma = 0$ should lead to the function u_0 , of course. Thus, we set the initial value for the differential equation as follows:

$$u(x, 0) = u_0(x). \quad (3.4)$$

The initial value problem (3.3), (3.4) is known from physics, where it models heat conduction in one dimension. The problem admits a unique solution, which is given by the convolution with the Gaussian function (3.2) with variance $\sqrt{2}\sigma$:

$$u(x, \sigma) = (u_0 * G_{\sqrt{2}\sigma})(x).$$

We have hence found a suitable function $f = G_1$.

Therefore, Canny edge detection starts with convolving a given image u with a Gaussian function G_σ . Then the gradient is computed, and its absolute value and direction are determined:

$$\begin{aligned}\rho(x) &= |\nabla(u * G_\sigma)(x)| = \sqrt{\partial_{x_1}(u * G_\sigma)(x)^2 + \partial_{x_2}(u * G_\sigma)(x)^2}, \\ \Theta(x) &= \measuredangle(\nabla(u * G_\sigma)(x)) = \arctan\left(\frac{\partial_{x_2}(u * G_\sigma)(x)}{\partial_{x_1}(u * G_\sigma)(x)}\right).\end{aligned}$$

The points x at which $\rho(x)$ exhibits a local maximum in the direction of the vector $(\sin(\Theta(x)), \cos(\Theta(x)))$ are then marked as edges. Afterward, a threshold is applied in order to suppress edges that are not important or induced by noise, i.e., if $\rho(x)$ is smaller than a given value τ , the corresponding x is removed. The result of the

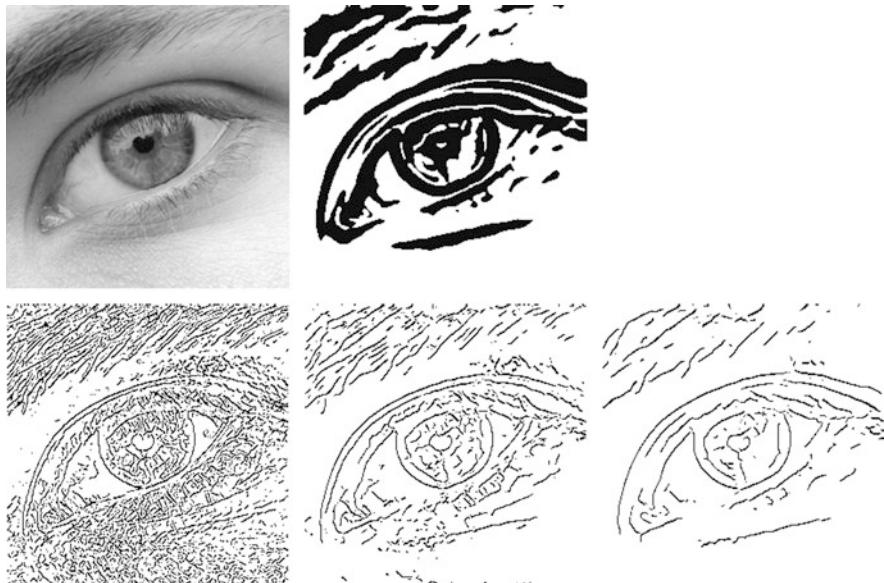
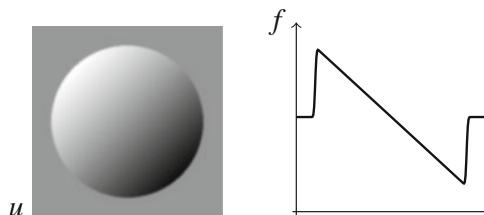


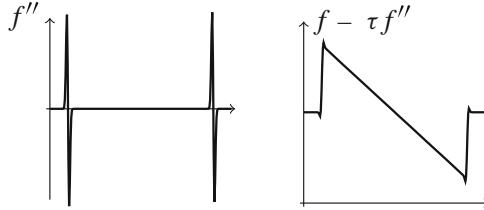
Fig. 3.9 Edge detection. Top left: Original image with 256×256 pixels. Top center: Edge detection through thresholding of the gradient after convolution with a Gaussian function with $\sigma = 3$. Bottom: Edge detection with Canny edge detector (from left to right: $\sigma = 1, 2, 3$)

Canny detector is shown in Fig. 3.9. For the sake of comparison, the result of a simple thresholding method is depicted as well. In this case, ρ is calculated as for the Canny edge detector, and then all points x for which $\rho(x)$ is larger than a given threshold are marked as edges.

Application 3.24 (Laplace Sharpening) Suppose we are given a blurred image u that we would like to sharpen. For this purpose, we consider a one-dimensional cross section f through the image u :



Furthermore, we consider the second derivative of the cross section f and remark (analogously to Fig. 3.7) that if we subtract from f a small multiple of the second derivative f'' , we obtain an image in which the edge is steeper than before:



Note, however, that edges occur in different directions. Thus, a rotationally invariant differential operator is necessary. The simplest one is given by the *Laplace operator*

$$\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}.$$

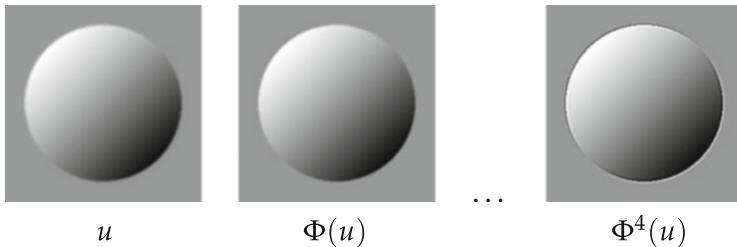
Therefore, the sharpening in 2D with a parameter $\tau > 0$ is performed by means of the operation

$$u - \tau \Delta u.$$

Note that this is a linear operation. In general, we cannot assume that the images u are sufficiently smooth, so that the Laplace operator may not be well defined. Again, a simple remedy is to smoothen the image beforehand—by a convolution with a Gaussian function, for instance. According to Proposition 3.13, we then obtain

$$\Phi(u) = (u - \tau \Delta u) * G_\sigma = u * (G_\sigma - \tau \Delta G_\sigma).$$

Successively applying this operation emphasizes the edges step by step:



In general, however, the edges are overemphasized after some time, i.e., the function values in a neighborhood of the edge are smaller or larger than in the original image. Furthermore, noise can be increased by this operation as well. These effects can be seen in Fig. 3.10.

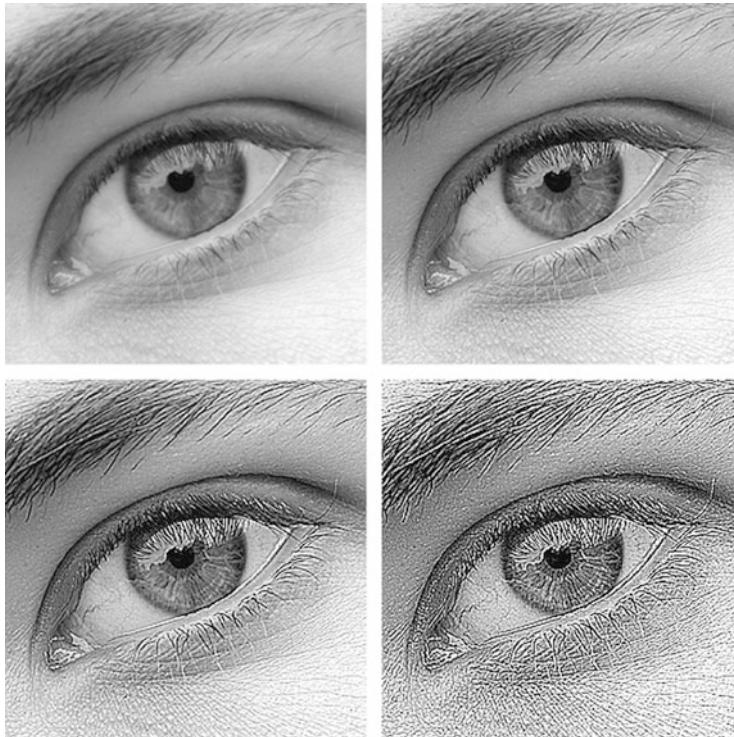


Fig. 3.10 Laplace sharpening. Top left: Original image with 256×256 pixels. Next to it: Successively applying Laplace sharpening of Application 3.24 (parameter: $\sigma = 0.25, \alpha = 1$)

3.3.3 Discretization of Convolutions

We now derive the standard discretization of convolutions. To start with, we consider a one-dimensional discrete image $U : \mathbf{Z} \rightarrow \mathbf{R}$ as well as a one-dimensional discrete convolution kernel $H : \mathbf{Z} \rightarrow \mathbf{R}$. We deduce the convolution of U with H from the continuous representation. By means of the piecewise constant interpolation in Sect. 3.1.1, we obtain, with $\phi = \chi_{[-1/2, 1/2]}$,

$$\begin{aligned}(u * h)(k) &= \int_{\mathbf{R}} u(y)h(k - y) \, dy = \int_{\mathbf{R}} \sum_{l \in \mathbf{Z}} U_l \phi(y - l) \sum_{m \in \mathbf{Z}} H_m \phi(k - y - m) \, dy \\ &= \sum_{l \in \mathbf{Z}} \sum_{m \in \mathbf{Z}} U_l H_m \int_{\mathbf{R}} \phi(x) \phi(k - l - m - x) \, dx.\end{aligned}$$

Since for the integral we have

$$\int_{\mathbf{R}} \phi(x)\phi(k-l-m-x) dx = \begin{cases} 1 & \text{if } m = k - l, \\ 0 & \text{otherwise,} \end{cases}$$

we arrive at the representation

$$(u * h)(k) = \sum_{l \in \mathbf{Z}} U_l H_{k-l}.$$

Therefore, we define the convolution of discrete images by

$$(U * H)_k = \sum_{l \in \mathbf{Z}} U_l H_{k-l}.$$

The generalization to higher dimensions is obvious. Dealing with finite images and finite convolution kernels, we obtain finite sums and are faced with the problem that we have to evaluate U or H at undefined points. This problem can be tackled by means of a boundary treatment or a boundary extension, in which it suffices to extend U . We extend a given discrete and finite image $U : \{0, \dots, N-1\} \rightarrow \mathbf{R}$ to an image $\tilde{U} : \mathbf{Z} \rightarrow \mathbf{R}$ in one of the following ways:

- Periodic extension: Tessellate \mathbf{Z} with copies of the original image

$$\tilde{U}_i = U_{i \bmod N}.$$

- Zero extension: Set $\tilde{U}_i = 0$ for $i < 0$ or $i \geq N$.
- Constant extension: Repeat the gray values at the boundary, i.e.,

$$\tilde{U}_i = U_{P_N(i)}, \quad P_N(i) = \max(\min(N-1, i), 0).$$

- Reflection and periodization or symmetric extension: Reflect the image successively at the edges, i.e.,

$$\tilde{U}_i = U_{Z_N(i \bmod 2N)}, \quad Z_N(i) = \min(i, 2N-1-i).$$

For the extension of images of multiple dimensions, the rules are applied to each dimension separately. Figure 3.11 shows an illustration of the different methods in two dimensions.

Note that the periodical and zero extensions induce unnatural jumps at the boundary of the image. The constant and the symmetric extensions do not produce such additional jumps.

Let us now consider two-dimensional images $U : \mathbf{Z}^2 \rightarrow \mathbf{R}$ and cover some classical methods that belong to the very first methods of digital image processing.

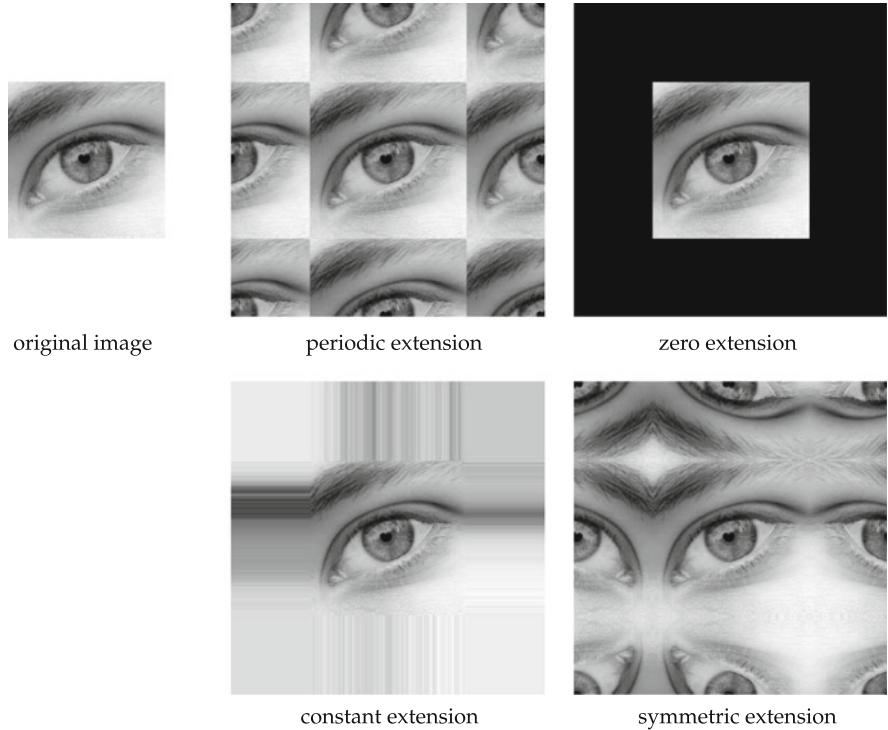


Fig. 3.11 Different methods of boundary extension

In this context, one usually speaks of filter masks rather than convolution kernels. A *filter mask* $H \in \mathbb{R}^{2r+1 \times 2s+1}$ defines a *filter* by

$$(U \boxtimes H)_{i,j} = \sum_{k=-r}^r \sum_{l=-s}^s H_{k,l} U_{i+k, j+l}.$$

We assume throughout that the filter mask is of odd size and is indexed in the following way:

$$H = \begin{bmatrix} H_{-r,-s} & \dots & H_{-r,s} \\ \vdots & H_{0,0} & \vdots \\ H_{r,-s} & \dots & H_{r,s} \end{bmatrix}.$$

Filtering corresponds to a convolution with the reflected filter mask. Due to the symmetry of the convolution, we observe that

$$(U \boxtimes H) \boxtimes G = U \boxtimes (H * G) = U \boxtimes (G * H) = (U \boxtimes G) \boxtimes H.$$

Therefore, the order of applying different filter masks does not matter. We now introduce some important filters:

Moving average: For odd n , the moving average is given by

$$M^n = \frac{1}{n} [1 \dots 1] \in \mathbf{R}^n.$$

A two-dimensional moving average is obtained by $(M^n)^T * M^n$.

Gaussian filter: The Gaussian filter G^σ with variance $\sigma > 0$ is a smoothing filter that is based on the Gaussian function. It is obtained by normalization:

$$\tilde{G}_{k,l}^\sigma = \exp\left(\frac{-(k^2 + l^2)}{2\sigma^2}\right), \quad G^\sigma = \frac{\tilde{G}^\sigma}{\sum_{k,l} \tilde{G}_{k,l}^\sigma}.$$

Binomial filter: The binomial filters B^n correspond to the normalized lines of Pascal's triangle:

$$\begin{aligned} B^1 &= \frac{1}{2} [1 1 0], \\ B^2 &= \frac{1}{4} [1 2 1], \\ B^3 &= \frac{1}{8} [1 3 3 1 0], \\ B^4 &= \frac{1}{16} [1 4 6 4 1], \\ &\vdots \end{aligned}$$

For large n , the binomial filters present good approximations to Gaussian filters. Two-dimensional binomial filters are obtained by $(B^n)^T * B^n$. An important property of binomial filters is the fact that B^{n+1} can be obtained by a convolution of B^n with B^1 (up to translation).

Derivative filter according to Prewitt and Sobel: In Application 3.23, we saw that edge detection can be realized by calculation of derivatives. Discretizing the derivative in the x and y direction by central difference quotients and normalizing the distance of two pixels to 1, we obtain the filters

$$D^x = \frac{1}{2} [-1 0 1], \quad D^y = (D^x)^T = \frac{1}{2} \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}.$$

Since derivatives amplify noise, it was suggested in the early days of image processing to complement these derivative filters with a smoothing into the

respectively opposite direction. In case of the *Prewitt filters* [116], a moving average $M^3 = [1 \ 1 \ 1]/3$ is used:

$$D_{\text{Prewitt}}^x = (M^3)^T * D^x = \frac{1}{6} \begin{bmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{bmatrix},$$

$$D_{\text{Prewitt}}^y = M^3 * D^y = \frac{1}{6} \begin{bmatrix} -1 & -1 & -1 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix}.$$

The *Sobel filters* [133] use the binomial filter B^2 as a smoothing filter:

$$D_{\text{Sobel}}^x = (B^2)^T * D^x = \frac{1}{8} \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix},$$

$$D_{\text{Sobel}}^y = B^2 * D^y = \frac{1}{8} \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}.$$

Laplace filter: We have already seen the Laplace operator in Application 3.24, where it was used for image sharpening. In this case, we need to discretize second-order derivatives. We realize this by successively applying forward and backward difference quotients:

$$\frac{\partial^2 u}{\partial x^2} \approx (U \boxtimes D_-^x) \boxtimes D_+^x = (U \boxtimes [-1 \ 1 \ 0]) \boxtimes [0 \ -1 \ 1] = U \boxtimes [1 \ -2 \ 1].$$

Therefore, the Laplace filter is obtained by

$$\Delta u \approx (U \boxtimes D_-^x) \boxtimes D_+^x + (U \boxtimes D_-^y) \boxtimes D_+^y = U \boxtimes \begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix}.$$

We now consider the numerical cost of filtering or convolutions: the complexity of calculating the convolution with a filter mask of size $2r + 1 \times 2s + 1$ in a pixel is of order $\mathcal{O}((2r + 1)(2s + 1))$. Since this has to be computed for every pixel, a better asymptotic behavior is desirable.

Separating filter masks: We call a filter mask $H \in \mathbf{R}^{2r+1 \times 2s+1}$ *separable* if it can be obtained from one-dimensional filter masks $F \in \mathbf{R}^{2r+1}$, $G \in \mathbf{R}^{2s+1}$ in the following way:

$$H = G^T * F.$$

With this decomposition, the filtering with H can be factorized into

$$U \boxtimes H = (U \boxtimes G^T) \boxtimes F$$

and therefore, the numerical cost reduces to $\mathcal{O}((2r+1)+(2s+1))$. The moving average, as well as the Laplace, Sobel, Prewitt, and binomial filters is separable.

Recursive implementation: The moving average can be implemented recursively: If $V_i = (U \boxtimes M^{2n+1})_i = \frac{1}{n} \sum_{k=-n}^n U_{i+k}$ is known, then

$$V_{i+1} = V_i + \frac{1}{n}(U_{i+1+n} - U_{i-n}).$$

Thus, the cost consists of two additions and one multiplication—independently of the size of the filter. Recursive filters play a major role in signal processing, in particular in real-time filtering of measured signals. In this case, only the already measured points are known, and the filter can use only these measurements.

Factorization/utilizing bit shifts: The binomial filters can be factorized into smaller filters; for instance,

$$\frac{1}{16} [1 \ 4 \ 6 \ 4 \ 1] = \frac{1}{16} [1 \ 1 \ 0] \boxtimes [0 \ 1 \ 1] \boxtimes [1 \ 1 \ 0] \boxtimes [0 \ 1 \ 1].$$

Note that each of the partial filters consists of only one addition. Furthermore, the multiplication by 1/16 presents a bit shift, which can be executed faster than a multiplication.

3.4 Morphological Filters

Morphological filters are the main tools of *mathematical morphology*, i.e., the theory of the analysis of spatial structures in images (the name is derived from the Greek word “morphe” = shape). The mathematical theory of morphological filters traces back to the engineers Georges Matheron and Jean Serra, cf. [134], for instance. Morphological methods aim mainly at the recognition and transformation of the shape of objects. We again consider an introductory example:

Example 3.25 (Denoising of Objects) Let us first assume that we have found an object in a discrete digital image—by means of a suitable segmentation method, for instance. For the mathematical description of the object, an obvious approach is to encode the object as a binary image, i.e., as a binary function $u : \mathbf{R}^d \rightarrow \{0, 1\}$ with

$$u(x) = \begin{cases} 1 & \text{if } x \text{ belongs to the object,} \\ 0 & \text{if } x \text{ does not belong to the object.} \end{cases}$$

Furthermore, we assume that the object is “perturbed,” i.e., that there are perturbations in the form of “small” objects. Since “1” typically encodes the color white and “0” the color black, the shape consists of the white part of the image u , and the perturbations are small additional white points.

Since we know that the perturbances are small, we define a “structure element” $B \subset \mathbf{R}^d$ of which we assume that it is just large enough to cover each perturbation. In order to eliminate the perturbances, we compute a new image v by

$$v(x) = \begin{cases} 1 & \text{if for all } y \in B, \text{ there holds } u(x + y) = 1 \\ 0 & \text{otherwise.} \end{cases}$$

This implies that only those points x for which the structure element B shifted by x lies within the old object completely are part of the new object v . This eliminates all parts of the objects that are smaller than the structuring element. However, the object is also changed significantly: the object is “thinner” than before. We try to undo the “thinning” by means of the following procedure: we compute another image w by

$$w(x) = \begin{cases} 1 & \text{if for a } y \in B, v(x + y) = 1, \\ 0 & \text{otherwise.} \end{cases}$$

Hence, a point x is part of the new object w if the structure element B shifted by x touches the old object v . This leads to an enlargement of the object. However, all remaining perturbations are increased again as well. All in all, we have reached our goal relatively well: the perturbations are eliminated to a large extent and the object is changed only slightly; cf. Fig. 3.12.

The methods used in this introductory example are the fundamental methods in mathematical morphology and will now be introduced systematically. The

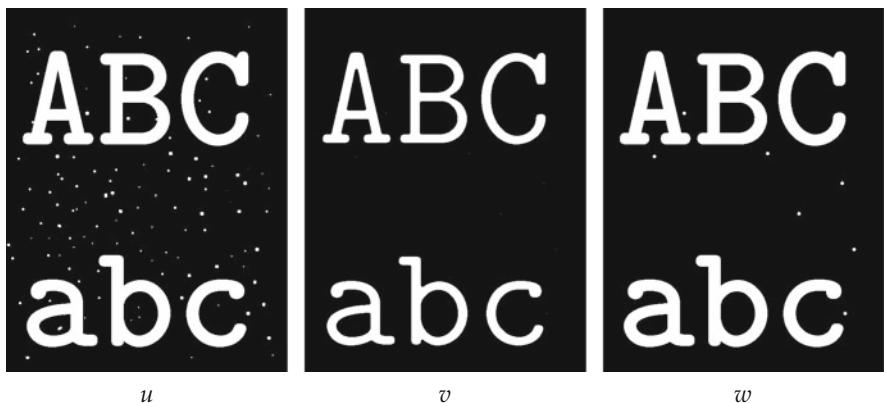


Fig. 3.12 Denoising of objects in Example 3.25

description of an object by means of a binary function $u : \mathbf{R}^d \rightarrow \{0, 1\}$ is equivalent to the description as a subset of \mathbf{R}^d . In the following, we will not distinguish between the two representations and use the slightly inexact notation $u \subset \mathbf{R}^d$ as long as it does not lead to any misunderstandings. Union and intersection of sets correspond to finding the maximum and minimum of functions, respectively:

$$(u \cup v)(x) = u(x) \vee v(x) = \max(u(x), v(x)),$$

$$(u \cap v)(x) = u(x) \wedge v(x) = \min(u(x), v(x)).$$

We will also use the notation $u \vee v$ and $u \wedge v$ for the supremum and infimum, respectively. Obtaining the complement corresponds to subtraction from one:

$$\complement u(x) = 1 - u(x).$$

With these operations, the binary functions are a *Boolean algebra*.

3.4.1 Fundamental Operations: Dilation and Erosion

We already became acquainted with the two fundamental operations of mathematical morphology in the introductory Example 3.25. They are called erosion and dilation.

Definition 3.26 Let $B \subset \mathbf{R}^d$ be a nonempty subset and $u \subset \mathbf{R}^d$. The *dilation* of u with a *structure element* B is defined by

$$(u \oplus B)(x) = \begin{cases} 1 & \text{if for some } y \in B \ u(x + y) = 1, \\ 0 & \text{otherwise.} \end{cases}$$

The *erosion* of u with the structure element B is defined by

$$(u \ominus B)(x) = \begin{cases} 1 & \text{if for all } y \in B \ u(x + y) = 1, \\ 0 & \text{otherwise.} \end{cases}$$

If we interpret B as a shape, the erosion of an object u provides an answer to the question, “Which translates of B fit into the object u ?”. Analogously, we can view the dilation of u as an answer to the question, “Which translates of B touch the object u ?”. Figure 3.13 shows an illustrative example of these operations.

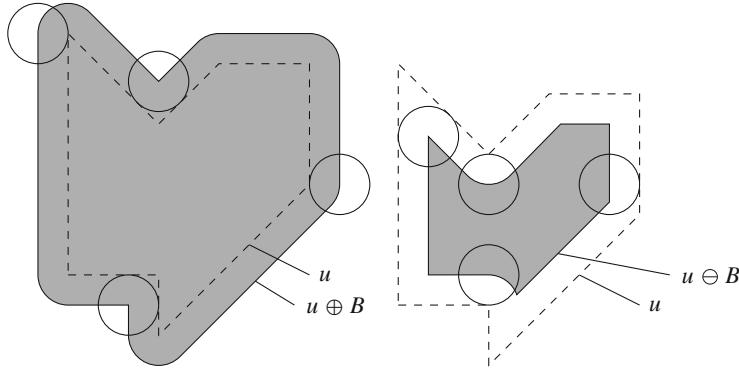


Fig. 3.13 Illustration of dilation (left) and erosion (right) of an object (dashed line) with a circular disk

Erosion and dilation can be extended to grayscale images in a natural way. The key to this is provided by the following simple lemma:

Lemma 3.27 *Let $u : \mathbf{R}^d \rightarrow \{0, 1\}$ and $B \subset \mathbf{R}^d$ nonempty. Then*

$$(u \oplus B)(x) = \sup_{y \in B} u(x + y), \quad (u \ominus B)(x) = \inf_{y \in B} u(x + y). \quad (3.5)$$

Proof The proof consists simply in carefully considering the definitions. For the dilation, we observe that we have $\sup_{y \in B} u(x + y) = 1$ if and only if $u(x + y) = 1$ for some $y \in B$. And for the erosion we have $\inf_{y \in B} u(x + y) = 1$ if and only if $u(x + y) = 1$ for all $y \in B$. \square

The formulation of erosion and dilation in this lemma does not use the fact that $u(x)$ attains only the values 0 and 1. Therefore, we can use the formulas for real-valued functions u analogously. In order to avoid the values $\pm\infty$ in the supremum and infimum, we assume in the following that u is bounded, i.e., we work in the vector space of bounded functions:

$$\mathcal{B}(\mathbf{R}^d) = \{u : \mathbf{R}^d \rightarrow \mathbf{R} \mid u \text{ bounded}\}.$$

Definition 3.28 (Erosion and Dilation of Grayscale Images) Let $B \subset \mathbf{R}^d$ be a nonempty subset and $u \in \mathcal{B}(\mathbf{R}^d)$. The *dilation* of u with a *structure element* B is defined by

$$(u \oplus B)(x) = \sup_{y \in B} u(x + y).$$

The *erosion* of u with the structure element B is defined by

$$(u \ominus B)(x) = \inf_{y \in B} u(x + y).$$

Erosion and dilation have a set of useful fundamental properties:

Theorem 3.29 *Let $u, v \in \mathcal{B}(\mathbf{R}^d)$, $B, C \subset \mathbf{R}^d$ nonempty structure elements, and $y \in \mathbf{R}^d$.*

Then the following properties hold

Duality

$$-(u \oplus B) = (-u) \ominus B.$$

Translation invariance

$$(T_y u) \ominus B = T_y(u \ominus B);$$

$$(T_y u) \oplus B = T_y(u \oplus B).$$

Monotonicity

$$u \leq v \quad \Rightarrow \quad \begin{cases} u \ominus B \leq v \ominus B, \\ u \oplus B \leq v \oplus B. \end{cases}$$

Distributivity

$$(u \wedge v) \ominus B = (u \ominus B) \wedge (v \ominus B),$$

$$(u \vee v) \oplus B = (u \oplus B) \vee (v \oplus B).$$

Composition *For $B + C = \{x + y \in \mathbf{R}^d \mid x \in B, y \in C\}$, one has*

$$(u \ominus B) \ominus C = u \ominus (B + C),$$

$$(u \oplus B) \oplus C = u \oplus (B + C).$$

Proof The proofs of these assertions rely on the respective properties of the supremum and infimum. For instance, we can show duality as follows:

$$-(u \oplus B)(x) = -\sup_{y \in B} u(x + y) = \inf_{y \in B} -u(x + y) = ((-u) \ominus B)(x).$$

The further proofs are a good exercise in understanding the respective notions. \square

Dilation and erosion obey a further fundamental property: among all operations on binary images, they are the only ones that are translation invariant and satisfy

the distributivity of Proposition 3.29, i.e., dilation and erosion are characterized by these properties, as the following theorem shows:

Theorem 3.30 *Let D be a translation invariant operator on binary images with $D(0) = 0$ such that for each set of binary images $u^i \subset \mathbf{R}^d$,*

$$D\left(\bigvee_i u^i\right) = \bigvee_i D(u^i).$$

Then there exists a set $B \subset \mathbf{R}^d$ such that

$$D(u) = u \oplus B.$$

If E is translation invariant with $E(\chi_{\mathbf{R}^d}) = \chi_{\mathbf{R}^d}$ and

$$E\left(\bigwedge_i u^i\right) = \bigwedge_i E(u^i),$$

then there exists $B \subset \mathbf{R}^d$ such that

$$E(u) = u \ominus B.$$

Proof Since D is translation invariant, translation invariant images are mapped onto translation invariant images again. Since 0 and $\chi_{\mathbf{R}^d}$ are the only translation invariant images, we must have either $D(0) = 0$ or $D(0) = \chi_{\mathbf{R}^d}$. The second case, in which D would not be a dilation, is excluded by definition.

Since we can write every binary image $u \subset \mathbf{R}^d$ as a union of its elements $\chi_{\{y\}}$, we have

$$Du = D\left(\bigvee_{y \in u} \chi_{\{y\}}\right) = \bigvee_{y \in u} D\chi_{\{y\}}.$$

Since D is translation invariant, we have $D\chi_{\{y\}} = D\chi_{\{0\}}(\cdot - y)$, which implies

$$Du(x) = \bigvee_{y \in u} (D\chi_{\{0\}})(x - y) = \begin{cases} 1 & \text{if } \exists y : u(y) = 1 \text{ and } D\chi_{\{0\}}(x - y) = 1, \\ 0 & \text{otherwise.} \end{cases}$$

On the other hand, we observe

$$(u \oplus B)(x) = \begin{cases} 1 & \text{if there holds } \exists y : B(y) = 1 \text{ and } u(x + y) = 1 \\ 0 & \text{else} \end{cases}$$

and we obtain

$$Du = u \oplus D\chi_{\{0\}}(-\cdot),$$

i.e., the operation D equals the dilation of u with the structure element $B = D\chi_{\{0\}}(-\cdot) = \{y \in \mathbf{R}^d \mid -y \in D\chi_{\{0\}}\}$.

The case of erosion can be deduced from the case of dilation. We define the operator \tilde{D} by $\tilde{D}u = \mathbb{C}E(\mathbb{C}u)$ and remark that it possesses the characteristic properties of a dilation. Thus, by means of the first part of the assertion, we can conclude that \tilde{D} is a dilation:

$$\tilde{D}(u) = u \oplus B,$$

where the structure element is given by $B = \tilde{D}(\chi_{\{0\}}(-\cdot)) = \mathbb{C}E(\mathbb{C}\chi_{\{0\}}(-\cdot)) = \mathbb{C}E(\chi_{\mathbf{R}^d \setminus \{0\}}(-\cdot))$. This implies

$$E(u) = \mathbb{C}\tilde{D}(\mathbb{C}u) = \mathbb{C}(\mathbb{C}u \oplus B) = u \ominus B.$$

□

A further property of erosion and dilation that makes them interesting for image processing is their *contrast invariance*:

Theorem 3.31 *Let $B \subset \mathbf{R}^d$ be nonempty. Erosion and dilation with structure element B are invariant with respect to change of contrast, i.e., for every continuous, monotonically increasing grayscale transformation $\Phi : \mathbf{R} \rightarrow \mathbf{R}$ and every $u \in \mathcal{B}(\mathbf{R}^d)$,*

$$\Phi(u) \ominus B = \Phi(u \ominus B) \text{ and } \Phi(u) \oplus B = \Phi(u \oplus B).$$

Proof The assertion is a direct consequence of the fact that continuous, monotonically increasing functions can be interchanged with supremum and infimum; shown exemplarily for the dilation,

$$\Phi(u) \oplus B(x) = \sup_{y \in B} \Phi(u(x + y)) = \Phi(\sup_{y \in B} u(x + y)) = \Phi(u \oplus B(x)). \quad \square$$

3.4.2 Concatenated Operations

Erosion and dilation can be combined in order to achieve specific effects. We already saw this procedure in the introductory Example 3.25: By means of erosion followed by a dilation we could remove small perturbations. By erosion of binary images, all objects that are “smaller” than the structure element B (in the sense of set inclusion) are eliminated. The eroded image contains only the larger structures, albeit “shrunk”

by B . A subsequent dilation with $-B$ then has the effect that the object is suitably enlarged again.

On the other hand, dilation fills up “holes” in the object that are smaller than B . However, the result is enlarged by B , which can analogously be corrected by a subsequent erosion with $-B$. In this way, we can “close” holes that are smaller than B . In fact, the presented procedures are well-known methods in morphology.

Definition 3.32 (Opening and Closing) Let $B \subset \mathbf{R}^d$ be a nonempty structure element and $u \in \mathcal{B}(\mathbf{R}^d)$ an image. Set $-B = \{-y \mid y \in B\}$. The operator

$$u \circ B = (u \ominus B) \oplus (-B)$$

is called *opening*, and the mapping

$$u \bullet B = (u \oplus B) \ominus (-B)$$

is called *closing*.

The operators inherit many properties from the basic operators, yet in contrast to erosion and dilation, it is less reasonable to iterate them.

Theorem 3.33 Let be B a nonempty structure element, $u, v \in \mathcal{B}(\mathbf{R}^d)$ images and $y \in \mathbf{R}^d$. Then the following properties hold

Translation invariance

$$(T_y u) \circ B = T_y(u \circ B),$$

$$(T_y u) \bullet B = T_y(u \bullet B).$$

Duality

$$-(u \bullet B) = (-u) \circ B.$$

Monotonicity

$$u \leq v \quad \Rightarrow \quad \begin{cases} u \circ B \leq v \circ B, \\ v \bullet B \leq v \bullet B. \end{cases}$$

Anti-extensionality and extensionality

$$u \circ B \leq u, \quad u \leq u \bullet B.$$

Idempotence

$$(u \circ B) \circ B = u \circ B,$$

$$(u \bullet B) \bullet B = u \bullet B.$$

Proof Translation invariance, duality, and monotonicity are a direct consequence of the properties of dilation and erosion in Theorem 3.29. For the anti-extensionality of the opening, we assume that the contrary holds, i.e., that for some x we had $(u \circ B)(x) = ((u \ominus B) \oplus (-B))(x) > u(x)$. Then there would be some $z \in B$ such that

$$\inf_{y \in B} u(x + y - z) > u(x).$$

This would imply that for all $y \in B$ we would have

$$u(x + y - z) > u(x),$$

which apparently does not apply for $y = z$. We obtain a contradiction and hence conclude that $u \circ B \leq u$. Analogously, we can deduce $u \bullet B \geq u$.

In order to show the idempotence of the opening, we remark that due to the anti-extensionality of the opening, we have

$$(u \circ B) \circ B \leq u \circ B.$$

On the other hand, the monotonicity of the erosion and the extensionality of the closing imply

$$\begin{aligned} (u \circ B) \ominus B &= ((u \ominus B) \oplus (-B)) \ominus B \\ &= (u \ominus B) \bullet (-B) \\ &\geq u \ominus B. \end{aligned}$$

By means of the monotonicity of the dilation, we obtain

$$(u \circ B) \circ B = ((u \circ B) \ominus B) \oplus (-B) \geq (u \ominus B) \oplus (-B) = u \circ B,$$

which implies $(u \circ B) \circ B = u \circ B$. For proving the idempotence of the closing, we can argue analogously. \square

Another commonly used combination of the fundamental morphological operators is the hit-or-miss operator. While in some sense, the erosion answers the question, “Does B fit into the object?” the hit-or-miss operator responds to “Does B fit into the object *exactly*?”. Mathematically, we can state this more precisely by introducing a miss mask $C \subset \complement B$ that describes the area in which the object does not fit (B is called a hitmask in this context).

Definition 3.34 Let $B, C \in \mathbf{R}^d$ be nonempty disjoint subsets. Then the *hit-or-miss operator* of a binary image $u \subset \mathbf{R}^d$ is defined by

$$u \odot (B, C) = (u \ominus B) \cap ((\complement u) \ominus C).$$

Remark 3.35 The definition can be generalized to grayscale images by introducing a reasonable complementation. If we assume that the entire range of gray values is given by $[0, 1]$, we can define, for instance,

$$u \odot (B, C) = (u \ominus B)((1 - u) \ominus C),$$

where pointwise multiplication implements the intersection of sets. For binary images $u \subset \mathbf{R}^d$, this expression is equivalent to Definition 3.34.

The last common class of morphological filters we cover is that of the *top-hat* operators, which aim at extracting information out of the image or object that is smaller than the structure element.

Definition 3.36 Let $B \subset \mathbf{R}^d$ be a nonempty structure element and $u \in \mathcal{B}(\mathbf{R}^d)$ an image. The *white top-hat operator* is defined by

$$u \sqcap B = u - u \circ B,$$

while

$$u \sqcup B = u \bullet B - u$$

defines the *black top-hat operator*.

Theorem 3.37 The white top-hat and black top-hat operators are translation invariant, idempotent, and positive in the sense of

$$u \sqcap B \geq 0, \quad u \sqcup B \geq 0,$$

for each image $u \in \mathcal{B}(\mathbf{R}^d)$.

Proof This is a direct consequence of the properties of opening and closing given in Theorem 3.33. \square

The top-hat operators can be used effectively to remove an irregular background or to correct nonuniform lighting, for instance; refer to Sect. 3.4.3.

3.4.3 Applications

Morphological operators can be used to solve a multitude of tasks. Often, the operators introduced here are combined in an intelligent and creative way. Therefore, it is difficult to tell in general for which type of problems morphology can provide a solution. The following application examples primarily illustrate the possibilities of applying morphological filters and serve as an inspiration for developing one's own methods.

Application 3.38 (Detection of Dominant Directions) Let a two-dimensional image be given in which the dominant directions are to be determined. This can be important in automatically verifying the orientation of a workpiece, for instance. By means of morphological methods, this task can be accomplished in the following way:

First, the edges are extracted from the image, by means of the algorithm according to Canny in Application 3.23, for instance. To determine the dominant directions in the edge image, the essential idea is to perform, for each $\alpha \in [0, \pi[$, an erosion with two points that are oriented by the angle α . We hence choose the structure element:

$$B_\alpha = \{(0, 0), (\sin \alpha, \cos \alpha)\}.$$

If there are many lines with direction α in the image, there will remain many points; otherwise, there will not. Considering the area of the remaining object, we can thus determine the dominant directions by means of the largest local maxima (also see Fig. 3.14).

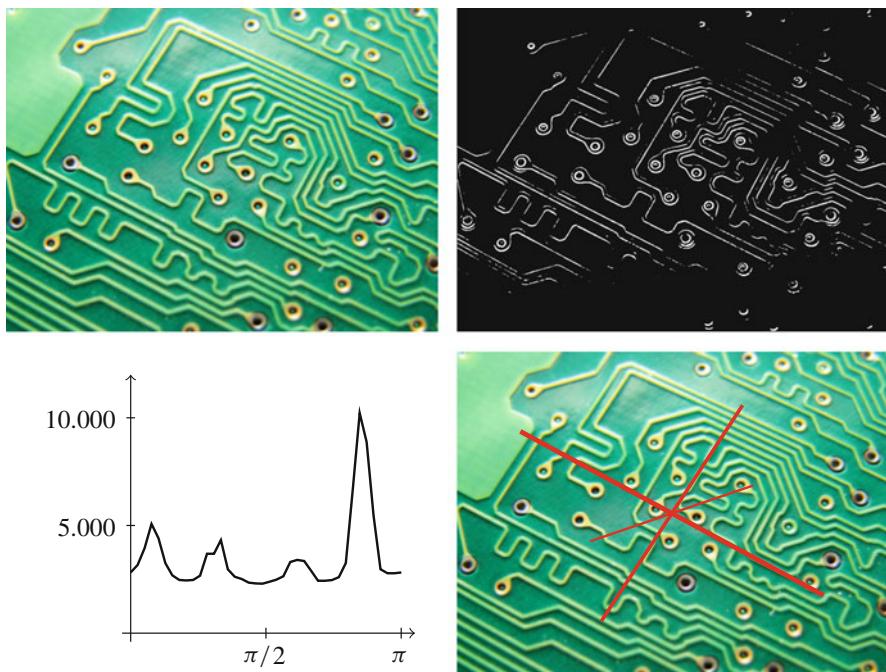
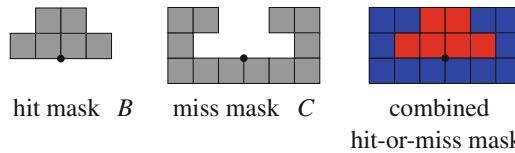


Fig. 3.14 Example for the detection of dominant directions. Upper left: Underlying image. Upper right: Result of edge detection by the algorithm according to Canny. Lower left: Amount of pixels in the eroded edge image depending on the angle α . Lower right: Image overlaid with the three dominant directions determined by maxima detection

- 1: Given: Finite binary image $u \subset \mathbf{R}^2$
- 2: **for** $\alpha \in [-\pi/2, \pi/2]$ **do**
- 3: Erode with two point mask B_α to angle α : $v_\alpha = u \ominus B_\alpha$
- 4: Calculate area of v_α : $N(\alpha) = \mathcal{L}^2(v_\alpha)$
- 5: **end for**
- 6: Determine the largest local maxima of N .

Application 3.39 (Extraction of Specific Forms) As an example for the application of the hit-or-miss operator we consider the following task: extract all letters from a scanned text that have a serif at the bottom. For this purpose, we define a hit mask in the form of the serif and a miss mask that describes the free space around the lower part of the serif as follows:



The application of the hit-or-miss operator now results in an image in which exactly those letters are marked that exhibit a serif at the bottom in the form of the hit mask; see Fig. 3.15.

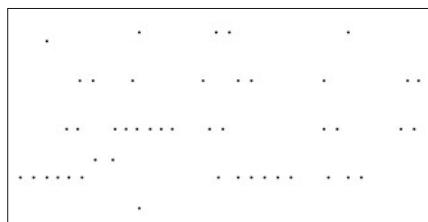
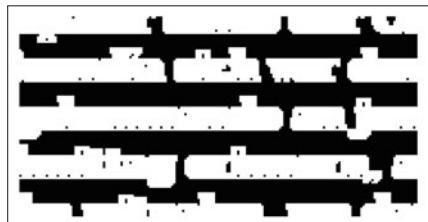
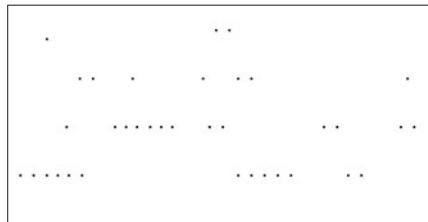
Application 3.40 (Correction of an Irregular Background) The segmentation of a grayscale image by a simple thresholding method is made considerably more difficult by an irregular background. In this case, the following correction of the background can remedy this:

By a closing with a large structure element, the dark parts that stand out from the light background are eliminated and the background remains. Subtracting the background from the image (which is the negative of the blacktop-hat operation), we obtain an image with a more even background, which is better suited for segmentation; see Fig. 3.16. The same method can be applied to images whose background is lit nonuniformly.

3.4.4 Discretization of Morphological Operators

We now describe the discretization of the fundamental morphological operators: erosion and dilation. For this purpose, we consider discrete images with continuous color space $u : \mathbf{Z}^2 \rightarrow \mathbf{R}$. For discrete images with finite support, we can perform a boundary extension analogously to Sect. 3.3.3.

später sollte der Or
Suendía sich vor den
gskommando an jen
hmittag erinnern, a:
hn mitnahm um das

 u  $u \ominus B$  $Cu \ominus C$  $u \odot (B, C)$

später sollte der Or
Suendía sich vor den
gskommando an jen
hmittag erinnern, a:
hn mitnahm um das

selections of the hit-mask (red)

später sollte der Or
Suendía sich vor den
gskommando an jen
hmittag erinnern, a:
hn mitnahm um das

selections of the miss-mask (blue)

später sollte der Or
Suendía sich vor den
gskommando an jen
hmittag erinnern, a:
hn mitnahm um das

selections of the hit-or-miss mask
(red/blue)

Fig. 3.15 Example for a simple application of the hit-or-miss operator to select a downward oriented serif B . The “hit operator” $u \ominus B$ also selects the cross bar of “t”, which is excluded by applying the “miss operator” $Cu \ominus C$. The combination finally leads to exactly the serif described by B and C

We encode the structure element B by a binary matrix $B \in \{0, 1\}^{2r+1 \times 2s+1}$ of odd size, which is indexed in the same way as in Sect. 3.3.3:

$$B = \begin{bmatrix} B_{-r,-s} & \cdots & B_{-r,s} \\ \vdots & B_{0,0} & \vdots \\ B_{r,-s} & \cdots & B_{r,s} \end{bmatrix}$$

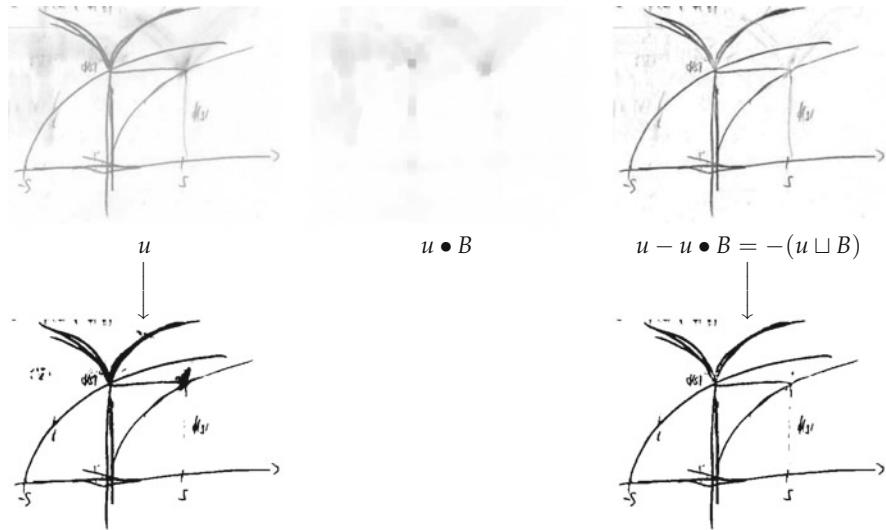


Fig. 3.16 Correction of an irregular background to improve segmentation. The structure element is chosen to be a square. The lower two images show the respective results of the automatic segmentation introduced in Application 3.11

where $B_{i,j} = 1$ denotes that (i, j) belongs to the structure element and $(i, j) = 0$ implies that (i, j) is not part of the structure element. Erosion and dilation are then defined as follows:

$$(u \ominus B)_{i,j} = \min\{u_{i+k,j+l} \mid (k, l) \text{ with } B_{k,l} = 1\},$$

$$(u \oplus B)_{i,j} = \max\{u_{i+k,j+l} \mid (k, l) \text{ with } B_{k,l} = 1\}.$$

The discrete erosion and dilation satisfy all properties of their continuous variants presented in Theorem 3.29. For the composition property

$$(u \ominus B) \ominus C = u \ominus (B + C), \quad (u \oplus B) \oplus C = u \oplus (B + C),$$

the sum $B + C \in \{0, 1\}^{2(r+u)+1 \times 2(s+v)+1}$ of structure elements $B \in \{0, 1\}^{2r+1 \times 2s+1}$ and $C \in \{0, 1\}^{2u+1 \times 2v+1}$ is defined by

$$(B + C)_{i,j} = \begin{cases} 1 & \text{if } (k, l) \text{ with } B_{k,l} = 1 \text{ and } (n, m) \text{ with } C_{n,m} = 1 \text{ and} \\ & (k+n, l+m) = (i, j) \text{ exist,} \\ 0 & \text{otherwise.} \end{cases}$$

Remark 3.41 (Increasing Efficiency of Erosion and Dilation) In order to evaluate the erosion (or dilation) at a pixel for a structure element with n elements, we have

to find the minimum (or maximum) of n numbers; which can be achieved with $n - 1$ pairwise comparisons. Let us further assume that the image consists of NM pixels and that the boundary extension is of negligible cost. Then we observe that the application of the erosion or dilation requires $(n - 1)NM$ pairwise comparisons. Due to the composition property in Theorem 3.29, we can increase the efficiency in certain cases:

If B and C consist of n and m elements, respectively, then $B + C$ can have at most nm elements. Hence, for the calculation of $u \ominus (B + C)$, we need $(nm - 1)NM$ pairwise comparisons in the worst case. However, to compute $(u \ominus B) \ominus C$, we need only $(n + m - 2)NM$ pairwise comparisons. Already for moderately large structure elements, this can make a significant difference, as the following example demonstrates. Therein, we omit the zeros at the boundary of the structure element and denote the center of a structure element by an underscore:

$$\underbrace{[1 \ 1]}_{B_1} + \underbrace{[1 \ 0 \ 1]}_{B_2} + \underbrace{[1 \ 0 \ 0 \ 0 \ 1]}_{B_3} = [1 \ 1 \ 1 \ 1 \ \underline{1} \ 1 \ 1 \ 1],$$

$$B_1 + B_2 + B_3 + (B_1^T + B_2^T + B_3^T) = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & \underline{1} & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}.$$

The erosion with an 8×8 square (64 elements, 63 pairwise comparisons) can hence be reduced to 6 erosions with structure elements containing two elements each (6 pairwise comparisons).

The other morphological operators (opening, closing, hit-or-miss, top-hat transformations) are obtained by combination. Their properties are analogous to the continuous versions.

In the discrete variant, there is an effective generalization of erosion and dilation: for erosion with a structure element with n elements, for each pixel of the image, the smallest gray value is taken that is hit by the structure element (for dilation, we take the largest one). The idea of this generalization is to sort the image values masked by the structure element and the subsequent replacement by the n th value within this rank order. These filters are called rank-order filters.

Definition 3.42 Let $B \in \{0, 1\}^{2r+1 \times 2s+1}$ be a structure element with $n \geq 1$ elements. The elements will be indexed by $I = \{(k_1, l_1), \dots, (k_n, l_n)\}$, i.e., we have $B_{k,l} = 1$ if and only if $(k, l) \in I$. By $\text{sort}(a_1, \dots, a_n)$ we denote the nondecreasing reordering of the vector $(a_1, \dots, a_n) \in \mathbf{R}^n$. The m th rank-order filter of a bounded

image $u : \mathbf{Z}^2 \rightarrow \mathbf{R}$ is then given by

$$(u \diamondsuit_m B)_{i,j} = \text{sort}(u_{i+k_1, j+l_1}, \dots, u_{i+k_n, j+l_n})_m.$$

For $k = 1$ and $k = n$, we respectively obtain for erosion and dilation

$$u \diamondsuit_1 B = u \ominus B, \quad u \diamondsuit_n B = u \oplus B.$$

For odd n , $\text{sort}(a_1, \dots, a_n)_{(n+1)/2}$ is called the *median* of the values a_1, \dots, a_n , i.e., the associated rank-order filter forms the median pointwise over the elements masked by B . Hence, it is similar to the moving average in Sect. 3.3.3, but uses the median instead of the average and is thus also called the \sim *median filter*. For median filters with structure elements B with an even number of elements, we use the common definition of the median:

$$\text{med}_B(u) = \begin{cases} u \diamondsuit_{(n+1)/2} B & \text{if } n \text{ odd,} \\ \frac{1}{2}((u \diamondsuit_{n/2} B) + (u \diamondsuit_{n/2+1} B)) & \text{if } n \text{ even.} \end{cases}$$

The operator performs a type of averaging that is more robust with respect to outliers than the computation of the arithmetic mean. It is thus well suited for denoising in case of so-called *impulsive noise*, i.e., the pixels are not all perturbed additively, but random pixels have a random value that is independent of the original gray value; refer to Fig. 3.17.

3.5 Further Developments

Even though this chapter treats basic methods, it is worthwhile to cover further developments in this field in the following. In particular, in the area of linear filters, there are noteworthy further developments. An initial motivation for linear filters was the denoising of images. However, we quickly saw that this does not work particularly well with linear filters, since in particular, edges cannot be preserved. In order to remedy this, we recall the idea in Example 3.12. Therein, the noise should be reduced by computing local averages. The blurring of the edges can then be explained by the fact that the average near an edge considers the gray values on both sides of the edge. The influence of the pixels on “the other side of the edge” can be resolved by considering not only the spatial proximity, but also the proximity of the gray values during the averaging process. The so-called \sim *bilateral filter* achieves this as follows: for a given image $u : \mathbf{R}^d \rightarrow \mathbf{R}$ and two functions $h : \mathbf{R}^d \rightarrow \mathbf{R}$ and $g : \mathbf{R} \rightarrow \mathbf{R}$, it computes a new image by

$$B_{h,g}u(x) = \frac{1}{\int_{\mathbf{R}^d} h(x-y)g(u(x)-u(y)) \, dy} \int_{\mathbf{R}^d} u(y)h(x-y)g(u(x)-u(y)) \, dy.$$



Fig. 3.17 Denoising in case of impulsive noise. Upper left: Original image. Upper right: Image perturbed by impulsive noise; 10% of the pixels were randomly replaced by black or white pixels (PSNR of 8.7 db). Lower left: Application of the moving average with a small circular disk with a radius of seven pixels (PSNR of 22.0 db). Lower right: Application of the median filter with a circular structure element with a radius of two pixels (PSNR of 31.4 db). The values of the radii were determined such that the respective PSNR is maximal

We observe that the function h denotes the weight of the gray value $u(y)$ depending on the *distance* $x - y$, while the function g presents the weight of the gray value $u(y)$ depending on the *similarity of the gray values* $u(x)$ and $u(y)$. The factor $(\int_{\mathbf{R}^d} h(x - y)g(u(x) - u(y)) dy)^{-1}$ is a normalization factor that ensures that the weights integrate to one in every point x . For linear filters (in this case, when there is no function g), it does not depend on x , and usually $\int_{\mathbf{R}^d} h(y) dy = 1$ is required. The name “bilateral filter” traces back to Tomasi and Manduchi [136]. If we chose h and g to be Gaussian functions, the filters are also called “nonlinear Gaussian filters” after Aurich and Weule [11]. In the case of characteristic functions h and g , the filter is also known as SUSAN [132]. The earliest reference for this kind of filters is probably Yaroslavski [145]. The bilateral filter exhibits excellent properties for edge-preserving denoising; see Fig. 3.18. A naive discretization of the integrals, however, reveals a disadvantage: the numerical cost is significantly higher than for linear filters, since the normalization factor has to be recalculated for every point

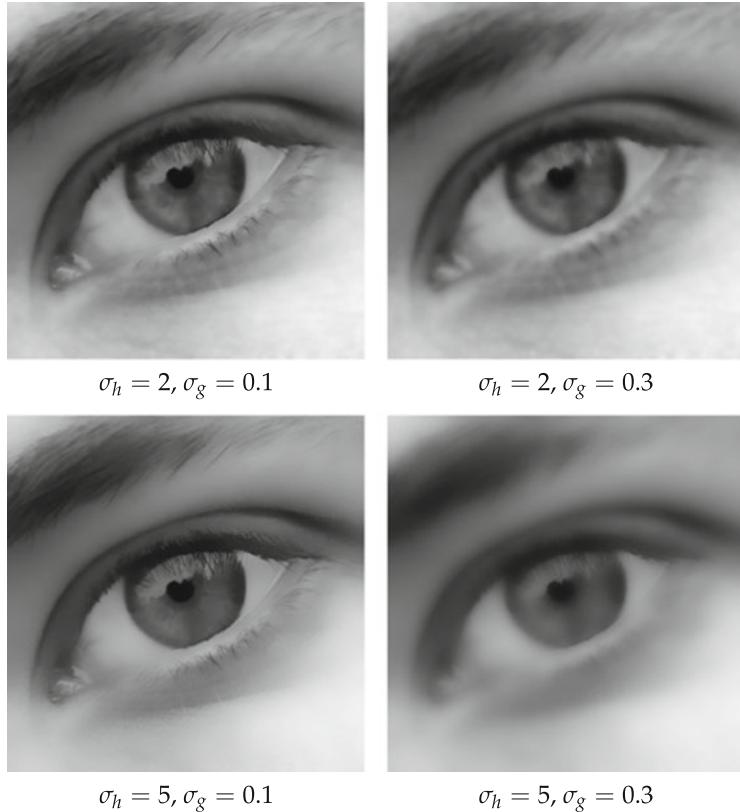


Fig. 3.18 The bilateral filter with Gaussian functions h and g , i.e., $h(x) = \exp(-|x|^2/(2\sigma_h^2))$ and $g(x) = \exp(-|x|^2/(2\sigma_g^2))$ applied to the original image in Fig. 3.17. The range of gray values is $[0, 1]$, i.e., the distance of black to white amounts to one

x . Methods to increase the efficiency are covered in the overview article [108], for instance.

Progressing a step further than bilateral filters, the so-called *nonlocal averages* [25] take averages not over values that are close and have a similar gray value, but over values that have a *similar neighborhood*. Mathematically, this is realized as follows: for an image $u : \mathbf{R}^d \rightarrow \mathbf{R}$ and a function $g : \mathbf{R}^d \rightarrow \mathbf{R}$, we define the function

$$h(x, y) = \int_{\mathbf{R}^d} g(t)|u(x+t) - u(y+t)|^2 dt.$$

For the choice of the function g , again a Gaussian function or the characteristic function of a ball around the origin is suitable, for instance. The function h exhibits a small value in (x, y) if the functions $t \mapsto u(x+t)$ and $t \mapsto u(y+t)$ are similar

in a neighborhood of the origin. If they are dissimilar, the value is large. Hence, the values x and y have similar neighborhoods in this sense if $h(x, y)$ is small. This motivates the following definition of the nonlocal averaging filter:

$$NL u(x) = \frac{1}{\int_{\mathbf{R}^d} e^{-h(x,y)} dy} \int_{\mathbf{R}^d} u(y) e^{-h(x,y)} dy.$$

Nonlocal averaging filters are particularly well suited for denoising of regions with textures. Their naive discretization is even more costly than for the bilateral filter, since for every value x , we first have to determine $h(x, y)$ by means of an integration. For ideas regarding an efficient implementation, we refer to [23].

The median filter that we introduced in Sect. 3.4.4 was defined only for images with discrete image domain. It was based on the idea of ordering the neighboring pixels. A generalization to images $u : \mathbf{R}^d \rightarrow \mathbf{R}$ is given by the following: for a measurable set $B \subset \mathbf{R}^d$ let

$$\text{med}_B u(x) = \inf_{B' \subset B} \sup_{|B'| \geq \frac{|B|}{2}}_{y \in B'} u(x + y).$$

For this filter, there is a connection to the mean curvature flow, which we will cover in Sect. 5.2.2. Roughly speaking, the iterated application of the median filter with $B = B_h(0)$ asymptotically corresponds (for $h \rightarrow 0$) to a movement of the level lines of the image into the direction of the normal with a velocity proportional to the average curvature of the contour lines. For details, we refer to [69].

In this form, the median filter is based on the ordering of the real numbers. For an image $u : \Omega \rightarrow F$ with discrete support Ω but non-ordered color space F , the concept of the median based on ordering cannot be transferred. Non-ordered color spaces, for instance, arise in the case of color images (e.g. $F = \mathbf{R}^3$) or in so-called diffusion tensor imaging; in which F is the set of symmetric matrices. If there is a distance defined on F , we can use the following idea: the median of real numbers a_1, \dots, a_n is a minimizer of the functional

$$F(a) = \sum_{i=1}^n |a - a_i|$$

(cf. Exercise 3.13). If there is a distance $\|\cdot\|$ defined on F , we can define the median of n “color values” A_1, \dots, A_n as a minimizer of

$$F(A) = \sum_{i=1}^n \|A - A_i\|.$$

In [143], it was shown that this procedure defines reasonable “medians” in case of matrices A_i depending on the matrix norm, for instance.

3.6 Exercises

Exercise 3.1 (Translation and Linear Coordinate Transformation on $L^p(\mathbf{R}^d)$)

Show that the operators T_y and D_A , for invertible matrices A , are well defined as linear operators on $L^p(\mathbf{R}^d)$ with $1 \leq p \leq \infty$. Calculate the adjoint operators. Are the operators continuous?

Exercise 3.2 (Commutativity of T_y and D_A) Show the following commutativity relation of linear coordinate transformations and translations:

$$T_y D_A = D_A T_{Ay}.$$

Exercise 3.3 (Average of an Image in the Histogram) Let $(\Omega, \mathfrak{F}, \mu)$ be a σ -finite measure space and $u : \Omega \rightarrow [0, 1]$ a measurable image. Show that

$$\int_{\Omega} u(x) dx = \int_0^{\mu(\Omega)} s dH_u.$$

Exercise 3.4 (L^p -Functions Are Continuous in the p th Mean) Let $1 \leq p < \infty$ and $u \in L^p(\mathbf{R}^d)$. Show that

$$\|T_h u - u\|_p \xrightarrow{h \rightarrow 0} 0.$$

In other words, the translation operator T_h is continuous in the argument h on L^p .

Hint: Use the fact that according to Theorem 2.55, the continuous functions with compact support are dense in $L^p(\mathbf{R}^d)$.

Exercise 3.5 (Solution of the Heat Equation) Let G_σ be the d -dimensional Gaussian function defined in (3.2) and

$$F(t, x) = G_{\sqrt{2}t}(x).$$

1. Show that for $t > 0$, the function F solves the equation

$$\partial_t F = \Delta F.$$

2. Let $u_0 : \mathbf{R}^d \rightarrow \mathbf{R}$ be bounded and continuous. Show that the function

$$u(t, x) = (u_0 * F(t, \cdot))(x)$$

solves the initial-boundary value problem

$$\begin{aligned}\partial_t u(t, x) &= \Delta u(t, x) \quad \text{for } t > 0, \\ u(0, x) &= u_0(x) \quad \text{for } x \in \mathbf{R}^d,\end{aligned}$$

the latter in the sense of $u(0, x) = \lim_{t \rightarrow 0} u(t, x)$.

Exercise 3.6 (Rotational Invariance of the Laplace Operator) Show that the Laplace operator

$$\Delta = \sum_{i=1}^d \frac{\partial^2}{\partial x_i^2}$$

is rotationally invariant in \mathbf{R}^d , i.e., for every twice continuously differentiable function $u : \mathbf{R}^d \rightarrow \mathbf{R}$ and every rotation $R \in O_d(\mathbf{R})$, one has $(\Delta u) \circ R = \Delta(u \circ R)$.

Furthermore, show that every rotationally invariant linear differential operator D of order $K \geq 1$ of the form

$$D = \sum_{|\alpha| \leq K} c_\alpha \frac{\partial^\alpha}{\partial x^\alpha}$$

does not exhibit any terms of odd order of differentiation, i.e., one has $c_\alpha = 0$ for all multi-indices α with $|\alpha|$ odd.

Exercise 3.7 (Separability Test) We call a discrete two-dimensional filter mask $H \in \mathbf{R}^{(2r+1) \times (2r+1)}$ *separable* if for some one-dimensional filter masks $F, G \in \mathbf{R}^{2r+1}$, one has $H = F \otimes G$ (i.e., $H_{i,j} = F_i G_j$).

Derive a method that for every $H \in \mathbf{R}^{(2r+1) \times (2r+1)}$, provides an $n \geq 0$ as well as separable filter masks $H_k \in \mathbf{R}^{(2r+1) \times (2r+1)}$, $1 \leq k \leq n$, such that there holds

$$H = H_1 + H_2 + \cdots + H_n$$

and n is minimal.

Exercise 3.8 (Proofs in the Morphology Section) Prove the remaining parts of Theorems 3.29, 3.33, and 3.37.

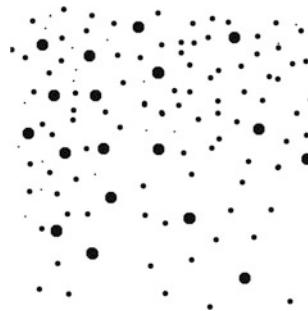
Exercise 3.9 (Lipschitz Constants for Erosion and Dilation) Let $B \subset \mathbf{R}^d$ be a nonempty structure element and $u \in \mathcal{B}(\mathbf{R}^d)$ Lipschitz continuous with constant $L > 0$. Show that $u \ominus B$ and $u \oplus B$ are also Lipschitz continuous and that their Lipschitz constants are less than or equal to L .

Exercise 3.10 (Non-expansiveness of Erosion and Dilation) Let $B \subset \mathbf{R}^d$ be a nonempty structure element. Show that the operations erosion and dilation are non-expansive with respect to the ∞ -norm, i.e., for $u, v \in \mathcal{B}(\mathbf{R}^d)$, one has

$$\|u \oplus B - v \oplus B\|_\infty \leq \|u - v\|_\infty$$

$$\|u \ominus B - v \ominus B\|_\infty \leq \|u - v\|_\infty.$$

Exercise 3.11 (Counting Circles by Means of Morphological Methods) Suppose an image contains circular objects of varying sizes:



Describe an algorithm (based on morphological operations) that returns the number and sizes of the circles. Implement the algorithm.

Exercise 3.12 (Decomposition of Structure Elements)

1. Let a diamond-shaped structure element of size n be given by the set

$$D_n = \{(i, j) \in \mathbf{Z}^2 \mid |i| + |j| \leq n\}.$$

How many elements does D_n contain? How can D_n be expressed as a sum of $\mathcal{O}(\log_2 |D_n|)$ two-point structure elements?

2. Show that, if a structure element B is invariant with respect to opening, i.e., $B = B \circ B_1$ for some B_1 , then the element B can be decomposed as $B = (-B_1) + B_2$ for some structure element B_2 .

Based on this observation, develop and implement a “greedy” algorithm for the decomposition of a given structure element B_0 :

- (a) Find, if possible, a two-point element Z_1 such that $B_0 = Z_1 + B_1$ with a minimal number of elements in B_1 .
- (b) As long as possible, continue the decomposition of the remainder according to the scheme above such that we finally obtain $B_0 = Z_1 + Z_2 + \dots + Z_n + B_n$.

Apply the algorithm to the set

$$K_8 = \{(i, j) \in \mathbf{Z}^2 \mid i^2 + j^2 \leq 8^2\}.$$

Exercise 3.13 (Description of the Median) Let be $a_1, \dots, a_n \in \mathbf{R}$. Show that the median of these values is a solution to the minimization problem

$$\min_{a \in \mathbf{R}} \sum_{i=1}^n |a - a_i|.$$

Furthermore, show that the arithmetic mean $\bar{a} = (a_1 + \dots + a_n)/n$ is the unique solution to the minimization problem

$$\min_{a \in \mathbf{R}} \sum_{i=1}^n |a - a_i|^2.$$

Chapter 4

Frequency and Multiscale Methods



Like the methods covered in Chap. 3, the methods based on frequency or scale-space decompositions belong to the older methods in image processing. In this case, the basic idea is to transform an image into a different representation in order to determine its properties or carry out manipulations. In this context, the Fourier transformation plays an important role.

4.1 The Fourier Transform

Since Fourier transforms are naturally complex-valued, as we will see shortly, it is reasonable to first introduce complex-valued images. Thus, in the following, we assume

$$u : \mathbf{R}^d \rightarrow \mathbf{C}.$$

In this section, we will define the Fourier transform for certain Lebesgue spaces and measures of Sect. 2.2.2 as well as distributions of Sect. 2.3. We begin by defining the Fourier transform on the space $L^1(\mathbf{R}^d)$.

4.1.1 The Fourier Transform on $L^1(\mathbf{R}^d)$

Definition 4.1 Let be $u \in L^1(\mathbf{R}^d)$ and $\xi \in \mathbf{R}^d$. Then the *Fourier transform* of u at ξ is defined by

$$(\mathcal{F}u)(\xi) = \widehat{u}(\xi) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbf{R}^d} u(x) e^{-ix \cdot \xi} dx.$$

The mapping $\mathcal{F} : u \mapsto \widehat{u}$ is called the *Fourier transform* as well.

Lemma 4.2 As a mapping $\mathcal{F} : L^1(\mathbf{R}^d) \rightarrow \mathcal{C}(\mathbf{R}^d)$, the Fourier transform is well defined, linear, and continuous.

Proof The integrand in the Fourier transform is continuous in ξ for almost all x and bounded by $|u(x)|$ for almost all ξ . By means of the dominated convergence theorem, we obtain for $\xi_n \rightarrow \xi$,

$$\lim_{n \rightarrow \infty} \widehat{u}(\xi_n) = \widehat{u}(\xi),$$

and hence the continuity of \widehat{u} . The linearity of \mathcal{F} is obvious, and the continuity results from the estimate

$$|\widehat{u}(\xi)| = \frac{1}{(2\pi)^{d/2}} \left| \int_{\mathbf{R}^d} u(x) e^{-ix \cdot \xi} dx \right| \leq \frac{1}{(2\pi)^{d/2}} \int_{\mathbf{R}^d} |u(x)| dx = \frac{1}{(2\pi)^{d/2}} \|u\|_1,$$

which implies $\|\widehat{u}\|_\infty \leq \frac{1}{(2\pi)^{d/2}} \|u\|_1$. \square

The above lemma implies in particular that Fourier transforms of L^1 -functions are bounded.

Remark 4.3 (Alternative Definitions of the Fourier Transform) In other books, other definitions of the Fourier transform are used. The following variants are common

$$\begin{aligned} (\mathcal{F}u)(\xi) &= \frac{1}{(2\pi)^d} \int_{\mathbf{R}^d} u(x) e^{-ix \cdot \xi} dx, \\ (\mathcal{F}u)(\xi) &= \int_{\mathbf{R}^d} u(x) e^{-ix \cdot \xi} dx, \\ (\mathcal{F}u)(\xi) &= \int_{\mathbf{R}^d} u(x) e^{-2\pi ix \cdot \xi} dx. \end{aligned}$$

Furthermore, the minus sign in the exponent may be omitted. Therefore, caution is advised when using tables of Fourier transforms or looking up calculation rules.

The Fourier transform goes well with translations T_y and linear coordinate transformations D_A . Furthermore, it also goes well with modulations, which we will define now.

Definition 4.4 For $y \in \mathbf{R}^d$, we set

$$m_y : \mathbf{R}^d \rightarrow \mathbf{C}, \quad m_y(x) = e^{ix \cdot y}$$

and thereby define the *modulation* of u by pointwise multiplication by m_y :

$$M_y : L^1(\mathbf{R}^d) \rightarrow L^1(\mathbf{R}^d), \quad M_y u = m_y u.$$

In the following lemma, we collect some elementary properties of the Fourier transform that will be helpful in the following.

Lemma 4.5 *Let $u \in L^1(\mathbf{R}^d)$, $y \in \mathbf{R}^d$, and $A \in \mathbf{R}^{d \times d}$ a regular matrix. Then we have the following equalities*

$$\begin{aligned}\mathcal{F}(T_y u) &= M_y(\mathcal{F}u), \\ \mathcal{F}(M_y u) &= T_{-y}(\mathcal{F}u), \\ \mathcal{F}(D_A u) &= |\det A|^{-1} D_{A^{-\top}}(\mathcal{F}u), \\ \mathcal{F}(\overline{u}) &= \overline{\mathcal{D}_{-\text{id}}(\mathcal{F}u)}.\end{aligned}$$

Proof One should first assure oneself that the operators T_y , M_y , and D_A map both $L^1(\mathbf{R}^d)$ and $\mathcal{C}(\mathbf{R}^d)$ onto themselves, i.e., all occurring terms are well defined. According to the transformation formula for integrals,

$$\begin{aligned}(\mathcal{F}M_\omega T_y u)(\xi) &= \frac{1}{(2\pi)^{d/2}} \int_{\mathbf{R}^d} u(x+y) e^{-ix \cdot (\xi - \omega)} dx \\ &= \frac{1}{(2\pi)^{d/2}} e^{i(\xi - \omega) \cdot y} \int_{\mathbf{R}^d} u(z) e^{-iz \cdot (\xi - \omega)} dz \\ &= (T_{-\omega} M_y \mathcal{F}u)(\xi).\end{aligned}$$

For $\omega = 0$, we obtain the translation formula, for $y = 0$, the modulation formula. The formula for the linear coordinate transformation follows directly from the transformation formula integrals as well, and the formula for the conjugation can be obtained elementarily. \square

The following symmetry properties are a direct consequence:

Corollary 4.6 *For $u \in L^1(\mathbf{R}^d)$*

$$\begin{aligned}u \text{ real-valued} &\implies \overline{\widehat{u}(\xi)} = \widehat{u}(-\xi), \\ u \text{ imaginary valued} &\implies \overline{\widehat{u}(\xi)} = -\widehat{u}(-\xi), \\ \widehat{u} \text{ real-valued} &\implies \overline{u(x)} = u(-x), \\ \widehat{u} \text{ imaginary valued} &\implies \overline{u(x)} = -u(-x).\end{aligned}$$

An important and surprisingly elementary property of the Fourier transform is its effect on convolutions introduced in Sect. 3.3.1. The Fourier transform maps convolutions into pointwise multiplications:

Theorem 4.7 (Convolution Theorem) *For $u, v \in L^1(\mathbf{R}^d)$*

$$\mathcal{F}(u * v) = (2\pi)^{d/2} \mathcal{F}(u) \mathcal{F}(v).$$

Proof Applying Fubini's theorem, we obtain

$$\begin{aligned}
\mathcal{F}(u * v)(\xi) &= \frac{1}{(2\pi)^{d/2}} \int_{\mathbf{R}^d} \int_{\mathbf{R}^d} u(y)v(x-y) dy e^{-ix \cdot \xi} dx \\
&= \frac{1}{(2\pi)^{d/2}} \int_{\mathbf{R}^d} \int_{\mathbf{R}^d} u(y)e^{-iy \cdot \xi} v(x-y)e^{-i(x-y) \cdot \xi} dx dy \\
&= \int_{\mathbf{R}^d} u(y)e^{-iy \cdot \xi} dy \mathcal{F}(v)(\xi) \\
&= (2\pi)^{d/2} \mathcal{F}(u)(\xi) \mathcal{F}(v)(\xi). \quad \square
\end{aligned}$$

Analogously to the convolution theorem, we can prove the following lemma:

Lemma 4.8 *For $u, v \in L^1(\mathbf{R}^d)$*

$$\int_{\mathbf{R}^d} \widehat{u}(\xi)v(\xi) d\xi = \int_{\mathbf{R}^d} u(\xi)\widehat{v}(\xi) d\xi.$$

At this point, it is tempting to state the assertion of the lemma as an equation of inner products. According to Lemma 4.5, we would have

$$(\widehat{u}, v)_2 = \int_{\mathbf{R}^d} \widehat{u}(\xi)\overline{v}(\xi) d\xi = \int_{\mathbf{R}^d} u(\xi)\widehat{\overline{v}}(\xi) d\xi = \int_{\mathbf{R}^d} u(\xi)\overline{\widehat{v}(-\xi)} d\xi = (u, D_{-\text{id}}\widehat{v})_2.$$

However, this is not allowed at this instance, since in Definition 4.1, we defined the Fourier transform for L^1 -functions only. This was for a good reason, since for L^2 -functions, it cannot be readily ensured that the defining integral exists. Anyhow, it appears desirable and will prove to be truly helpful to have access to the Fourier transform not only on the (not even reflexive) Banach space $L^1(\mathbf{R}^d)$, but also on the Hilbert space $L^2(\mathbf{R}^d)$.

4.1.2 The Fourier Transform on $L^2(\mathbf{R}^d)$

The extension of the Fourier transform to the space $L^2(\mathbf{R}^d)$ requires some further work. As a first step, we define a “small” function space on which the Fourier transform exhibits some further interesting properties—the Schwartz space:

Definition 4.9 The *Schwartz space of rapidly decreasing functions* is defined by

$$\mathcal{S}(\mathbf{R}^d) = \left\{ u \in \mathcal{C}^\infty(\mathbf{R}^d) \mid \forall \alpha, \beta \in \mathbf{N}^d : C_{\alpha, \beta}(u) = \sup_{x \in \mathbf{R}^d} |x^\alpha \frac{\partial^\beta}{\partial x^\beta} u(x)| < \infty \right\}.$$

A function $u \in \mathcal{S}(\mathbf{R}^d)$ is also called a *Schwartz function*.

Roughly speaking, the Schwartz space contains smooth functions that tend to zero faster than polynomials tend to infinity. It can be verified elementarily that the Schwartz space is a vector space. In order to make it accessible for analytical methods, we endow it with a topology. We describe this topology by defining a notion of convergence for sequences of functions.

Definition 4.10 A sequence (u_n) in the Schwartz space converges to u if and only if for all multi-indices α, β , one has

$$C_{\alpha,\beta}(u_n - u) \rightarrow 0 \quad \text{for } n \rightarrow \infty.$$

Convergence in the Schwartz space is very restrictive: a sequence of functions converges if it and all its derivatives multiplied by arbitrary monomials converge uniformly.

Remark 4.11 For our purposes, the description of the topology $\mathcal{S}(\mathbf{R}^d)$ by convergence of sequences suffices. Let us remark that the functionals $C_{\alpha,\beta}$ are so-called seminorms on the Schwartz space and thereby turn it into a metrizable, locally convex space; refer to [122], for instance.

Lemma 4.12 *The Schwartz space is nonempty and closed with respect to derivatives of arbitrary order as well as pointwise multiplication.*

Proof An example of a function in $\mathcal{S}(\mathbf{R}^d)$ is given by $u(x) = \exp(-|x|^2)$, as one can show elementarily.

For $u \in \mathcal{S}(\mathbf{R}^d)$, one has for every multi-index γ that

$$C_{\alpha,\beta}\left(\frac{\partial^\gamma}{\partial x^\gamma} u\right) = C_{\alpha,\beta+\gamma}(u) < \infty,$$

and hence we have $\frac{\partial^\gamma}{\partial x^\gamma} u \in \mathcal{S}(\mathbf{R}^d)$.

The fact that for $u, v \in \mathcal{S}(\mathbf{R}^d)$, the product uv lies in the Schwartz space as well can be proved by means of the Leibniz rule for multi-indices (see Sect. 2.1.1). \square

The Schwartz space is closely connected to the Fourier transform. The following lemma presents further calculation rules for Fourier transforms of Schwartz functions.

Lemma 4.13 *Let $u \in \mathcal{S}(\mathbf{R}^d)$, $\alpha \in \mathbf{N}^d$ a multi-index, and define $p^\alpha(x) = x^\alpha$. Then, one has the equalities*

$$\mathcal{F}\left(\frac{\partial^\alpha u}{\partial x^\alpha}\right) = i^{|\alpha|} p^\alpha \mathcal{F}(u),$$

$$\mathcal{F}(p^\alpha u) = i^{|\alpha|} \frac{\partial^\alpha}{\partial x^\alpha} \mathcal{F}(u).$$

Proof We begin with the following auxiliary calculations:

$$\frac{\partial^\alpha}{\partial x^\alpha} (e^{-ix \cdot \xi}) = (-i)^{|\alpha|} \xi^\alpha e^{-ix \cdot \xi} \quad \text{and} \quad x^\alpha e^{ix \cdot \xi} = i^{|\alpha|} \frac{\partial^\alpha}{\partial \xi^\alpha} (e^{-ix \cdot \xi}).$$

By means of integration by parts, we obtain

$$\begin{aligned}\mathcal{F}(\frac{\partial^\alpha}{\partial x^\alpha} u)(\xi) &= \frac{1}{(2\pi)^{d/2}} \int_{\mathbf{R}^d} \frac{\partial^\alpha}{\partial x^\alpha} u(x) e^{-ix \cdot \xi} dx \\ &= \frac{1}{(2\pi)^{d/2}} i^{|\alpha|} \xi^\alpha \int_{\mathbf{R}^d} u(x) e^{-ix \cdot \xi} dx \\ &= i^{|\alpha|} p^\alpha(\xi) \mathcal{F}u(\xi).\end{aligned}$$

By interchanging the order of integration and differentiation, we arrive at

$$\begin{aligned}\mathcal{F}(p^\alpha u)(\xi) &= \frac{1}{(2\pi)^{d/2}} \int_{\mathbf{R}^d} u(x) x^\alpha e^{-ix \cdot \xi} dx \\ &= \frac{1}{(2\pi)^{d/2}} i^{|\alpha|} \int_{\mathbf{R}^d} u(x) \frac{\partial^\alpha}{\partial \xi^\alpha} e^{-ix \cdot \xi} dx \\ &= i^{|\alpha|} (\frac{\partial^\alpha}{\partial \xi^\alpha} \mathcal{F}u)(\xi).\end{aligned}$$

Both of the previous arguments are valid, since the integrands are infinitely differentiable with respect to ξ and integrable with respect to x . \square

We thus observe that the Fourier transform transforms a differentiation into a multiplication and vice versa. This lets us assume already that the Schwartz space $\mathcal{S}(\mathbf{R}^d)$ is mapped onto itself by the Fourier transform. In order to show this, we state the following lemma:

Lemma 4.14 *For the Gaussian function $G(x) = e^{-\frac{|x|^2}{2}}$, one has*

$$\widehat{G}(\xi) = G(\xi),$$

i.e., the Gaussian function is an eigenfunction of the Fourier transform corresponding to the eigenvalue one.

Proof The Gaussian function can be written as a tensor product of one-dimensional Gaussian functions $g : \mathbf{R} \rightarrow \mathbf{R}$, $g(t) = \exp(-t^2/2)$ such that $G(x) = \prod_{k=1}^d g(x_k)$. By Fubini's theorem, we obtain

$$\widehat{G}(\xi) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbf{R}^d} \prod_{k=1}^d g(x_k) e^{-ix_k \xi_k} dx = \prod_{k=1}^d \widehat{g}(\xi_k).$$

In order to determine the Fourier transform of g , we remark that g satisfies the differential equation $g'(t) = -tg(t)$. Applying the Fourier transform to this equation, we by means of Lemma 4.13 obtain the differential equation $-\omega \widehat{g}(\omega) = \widehat{g}'(\omega)$. Furthermore, $\widehat{g}(0) = \int_{\mathbf{R}} g(t) dt = 1 = g(0)$. Therefore, the functions g and

\widehat{g} satisfy the same differential equation with the same initial value. By the Picard-Lindelöf theorem on uniqueness of solutions of initial value problems, they thus have to coincide, which proves the assertion. \square

Theorem 4.15 *The Fourier transform is a continuous and bijective mapping of the Schwartz space into itself. For $u \in \mathcal{S}(\mathbf{R}^d)$, we have the inversion formula*

$$(\mathcal{F}^{-1}\mathcal{F}u)(x) = \check{u}(x) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbf{R}^d} \widehat{u}(\xi) e^{ix \cdot \xi} d\xi = u(x).$$

Proof According to Lemma 4.13, we have for every $\xi \in \mathbf{R}^d$ that

$$|\xi^\alpha \frac{\partial^\beta}{\partial \xi^\beta} \widehat{u}(\xi)| = |\mathcal{F}(\frac{\partial^\alpha}{\partial x^\alpha} p^\beta u)(\xi)| \leq \frac{1}{(2\pi)^{d/2}} \|\frac{\partial^\alpha}{\partial x^\alpha} p^\beta u\|_1. \quad (4.1)$$

Therefore, for $u \in \mathcal{S}(\mathbf{R}^d)$, we also have $\widehat{u} \in \mathcal{S}(\mathbf{R}^d)$. Since the Fourier transform is linear, it is sufficient to show the continuity at zero. We thus consider a null sequence (u_n) in the Schwartz space, i.e., as $n \rightarrow \infty$, $C_{\alpha,\beta}(u_n) \rightarrow 0$. That is, (u_n) , as well as $(\partial^\alpha p^\beta u_n)$ for all α, β , converges to zero uniformly. This implies that the right-hand side in (4.1) tends to zero. In particular, we obtain that $C_{\alpha,\beta}(\widehat{u}_n) \rightarrow 0$, which implies that (\widehat{u}_n) is a null sequence, proving continuity.

In order to prove the inversion formula, we for now consider two arbitrary functions $u, \phi \in \mathcal{S}(\mathbf{R}^d)$. By means of Lemma 4.8 and the calculation rules for translation and modulation given in Lemma 4.5, we infer for the convolution of \widehat{u} and ϕ that

$$\begin{aligned} (\widehat{u} * \phi)(x) &= \int_{\mathbf{R}^d} \widehat{u}(y) \phi(x - y) dy = \int_{\mathbf{R}^d} \widehat{u}(y) e^{ix \cdot y} \widehat{\phi}(-y) dy \\ &= \int_{\mathbf{R}^d} u(y) \widehat{\phi}(-x - y) dy = (u * \widehat{\phi})(-x). \end{aligned}$$

Now we choose ϕ to be a rescaled Gaussian function:

$$\phi_\varepsilon(x) = \varepsilon^{-d} (D_{\varepsilon^{-1} \text{id}} G)(x) = \varepsilon^{-d} e^{-\frac{|x|^2}{2\varepsilon^2}}.$$

According to the calculation rule for linear coordinate transformations of Lemma 4.5, we infer that $\widehat{\phi}_\varepsilon = D_{\varepsilon \text{id}} \widehat{G}$, and hence we have $\widehat{\phi}_\varepsilon = \varepsilon^{-d} D_{\varepsilon^{-1} \text{id}} \widehat{G}$ as well. According to Lemma 4.14, $\widehat{G} = G$ and thus $\widehat{\phi}_\varepsilon = \phi_\varepsilon$. Since u is in particular bounded and continuous, and furthermore, G is positive and its integral is normalized to one, we can apply Theorem 3.13 and obtain that for $\varepsilon \rightarrow 0$,

$$\widehat{u} * \phi_\varepsilon(x) \rightarrow \widehat{u}(x) \quad \text{and} \quad u * \phi_\varepsilon(-x) \rightarrow u(-x).$$

We hence conclude that

$$\widehat{\tilde{u}}(x) = u(-x).$$

Note that we can state the inversion formula for the Fourier transform in the following way as well:

$$\check{u} = \overline{\mathcal{F}u}.$$

According to the calculation rule for conjugation in Lemma 4.5, we infer that $\check{u} = D_{-\text{id}}\widehat{u}$, and substituting \widehat{u} for u , this altogether results in

$$\check{\check{u}} = D_{-\text{id}}\widehat{\widehat{u}} = u.$$

□

Theorem 4.16 *There is a unique continuous operator $\mathcal{F} : L^2(\mathbf{R}^d) \rightarrow L^2(\mathbf{R}^d)$ that extends the Fourier transform \mathcal{F} to $\mathcal{S}(\mathbf{R}^d)$ and satisfies the equation $\|u\|_2 = \|\mathcal{F}u\|_2$ for all $u \in L^2(\mathbf{R}^d)$.*

Furthermore, this operator \mathcal{F} is bijective, and its inverse \mathcal{F}^{-1} is a continuous extension of \mathcal{F}^{-1} onto $\mathcal{S}(\mathbf{R}^d)$.

Proof For two functions $u, v \in \mathcal{S}(\mathbf{R}^d)$, we infer according to Lemma 4.8 that

$$(\widehat{u}, \widehat{v})_2 = (u, v)_2$$

and in particular $\|u\|_2 = \|\mathcal{F}u\|_2$. Thus, the Fourier transform is an isometry defined on a dense subset of $L^2(\mathbf{R}^d)$. Hence, there exists a unique continuous extension onto the whole space. Due to the symmetry between \mathcal{F} and \mathcal{F}^{-1} , an analogous argument yields the remainder of the assertion. □

Remark 4.17 As remarked earlier, the formula

$$\mathcal{F}(u)(\xi) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbf{R}^d} u(x) e^{-i\xi \cdot x} dx$$

cannot be applied to a function $u \in L^2(\mathbf{R}^d)$, since the integral does not necessarily exist. However, for $u \in L^2(\mathbf{R}^d)$, there holds that the function

$$\psi_R(\xi) = \frac{1}{(2\pi)^{d/2}} \int_{|x| \leq R} u(x) e^{-i\xi \cdot x} dx$$

converges to \widehat{u} for $R \rightarrow \infty$ in the sense of L^2 convergence. An analogous statement holds for the inversion formula. In the following, we will neglect this distinction and use the integral representation for L^2 -functions as well.

The isometry property $\|u\|_2 = \|\mathcal{F}u\|_2$ also implies that

$$(u, v)_2 = (\mathcal{F}u, \mathcal{F}v)_2, \quad (4.2)$$

which is known as *Plancherel's formula*.

The calculation rules in Lemma 4.5, the symmetry relations in Corollary 4.6, and the convolution Theorem 4.7 also hold for the Fourier transform on $L^2(\mathbf{R}^d)$, of course; refer also to Sect. 4.1.3. The inversion formula enables the following interpretation of the Fourier transform:

Example 4.18 (Frequency Representation of a Function) For $u \in L^2(\mathbf{R}^d)$, according for the inversion formula we have

$$u(x) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbf{R}^d} \widehat{u}(\xi) e^{ix \cdot \xi} d\xi.$$

Thus, we can say that in some sense, u can be expressed as a superposition of complex exponential functions and that furthermore, $\widehat{u}(\xi)$ indicates the extend to which the corresponding exponential function $x \mapsto e^{ix \cdot \xi}$ contributes to u . For this reason, \widehat{u} is also called the *frequency representation* of u (in this context, we also call u itself the *spatial representation*, or for $d = 1$, the *time representation*).

The convolution theorem now facilitates a new interpretation of the linear filters in Sect. 3.3:

Example 4.19 (Interpretation of Linear Filters in Frequency Representation) For a function $u \in L^2(\mathbf{R}^d)$ and a convolution kernel $h \in L^2(\mathbf{R}^d)$, one has

$$\mathcal{F}(u * h) = (2\pi)^{d/2} \mathcal{F}(u) \mathcal{F}(h).$$

This is to say that the Fourier transform of h indicates in what way the frequency components of u are damped, amplified, or modulated. We also call \widehat{h} the *transfer function* in this context. A convolution kernel h whose transfer function \widehat{h} is zero (or attains small values) for large ξ is called a *low-pass filter*, since it lets low frequencies pass. Analogously, we call h a *high-pass filter* if $\widehat{h}(\xi)$ is zero (or small) for small ξ . Since noise contains many high-frequency components, one can try to reduce noise by a low-pass filter. For image processing, it is a disadvantage in this context that edges also exhibit many high-frequency components. Hence, a low-pass filter necessarily blurs the edges as well. It turns out that edge-preserving denoising cannot be accomplished with linear filters; cf. Fig. 4.1 as well.

Example 4.20 (Image Decomposition into High- and Low-Frequency Components) By means of high- and low-pass filters, we can decompose a given image into its high- and low-frequency components. for this purpose, let h be the so-called *perfect*

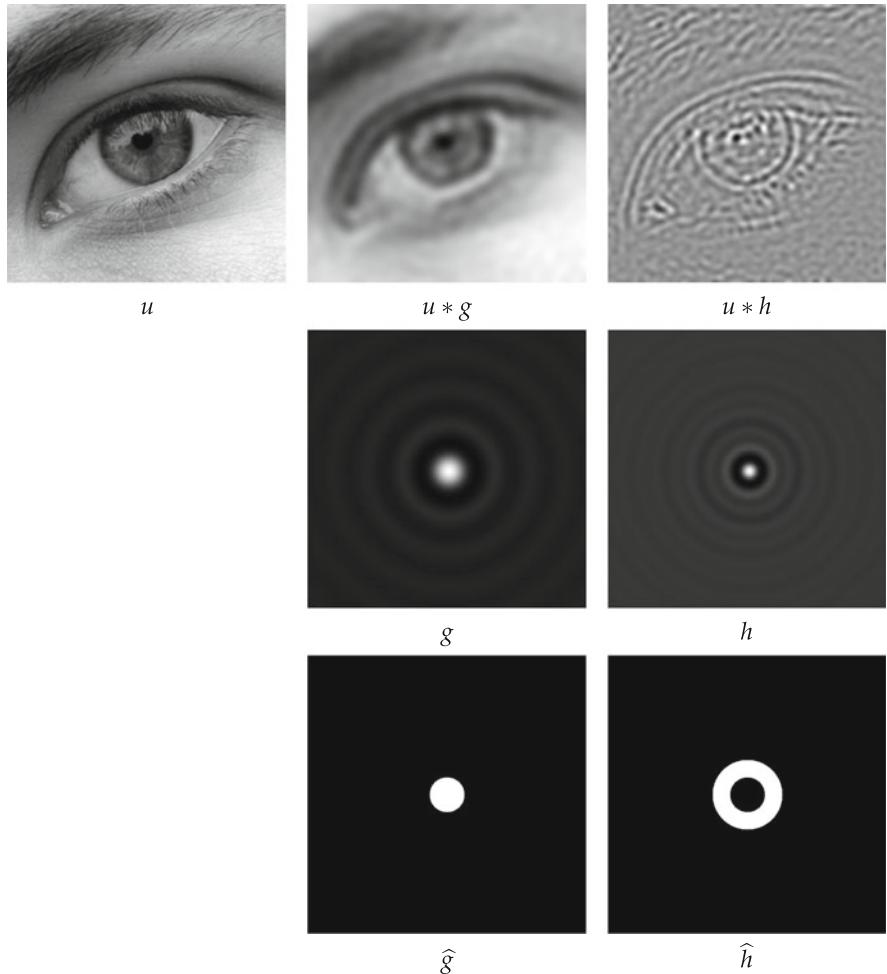


Fig. 4.1 High- and low-pass filters applied to an image. The Fourier transform of the low-pass filter g is a characteristic function of a ball around the origin, and the Fourier transform of the high-pass filter h is a characteristic function of an annulus around the origin. Note that the filters oscillate slightly, which is noticeable in the images as well

low-pass filter, i.e., for a radius $r > 0$, the Fourier transform of h is given by

$$\widehat{h} = \frac{1}{(2\pi)^{d/2}} \chi_{B_r(0)}.$$

The low- and high-frequency components of the image u are then respectively given by

$$u^{\text{low}} = u * h \quad \text{and} \quad u^{\text{high}} = u - u^{\text{low}}.$$

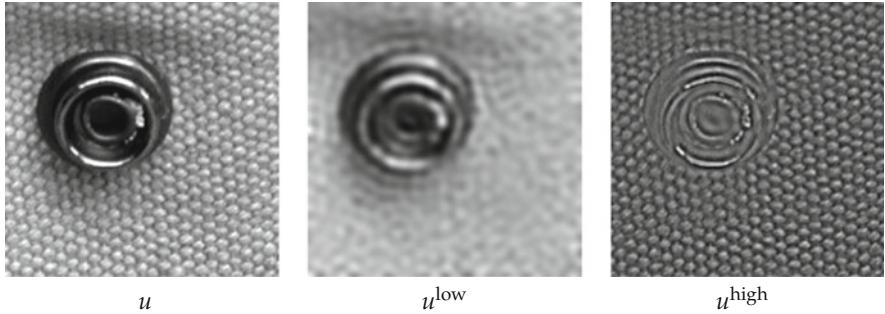


Fig. 4.2 Image decomposition into low- and high-frequency components

In particular,

$$\widehat{u^{\text{high}}} = \widehat{u}(1 - (2\pi)^{d/2}\widehat{h}) = \widehat{u} \cdot \chi_{\{|\xi| > r\}}.$$

In Fig. 4.2, we observe that the separation of textured components by this method has its limitations: the low-frequency component now contains almost no texture of the fabric, whereas this is contained in the high-frequency component. However, we also find essential portions of the edges, i.e., the separation of texture from nontextured components is not very good.

Remark 4.21 (Deconvolution with the Fourier Transform) The “out-of-focus” and motion blurring as well as other models of blurring assume that the blurring is modeled by a linear filter, i.e., by a convolution. A deblurring can in this case be achieved by a deconvolution: For a blurred image u given by

$$u = u_0 * h$$

with an unknown image u_0 and a known convolution kernel h , one has $\widehat{u} = 2\pi\widehat{u}_0\widehat{h}$ (in the two-dimensional case), and we obtain the unknown image by

$$u_0 = \mathcal{F}^{-1}\left(\frac{\widehat{u}}{2\pi\widehat{h}}\right).$$

If u is measured exactly, the model for h is accurate, and $\widehat{h}(\xi) \neq 0$, then it is actually possible in this way to eliminate the blurring exactly. However, if u is not available exactly, also for accurate h , a problem arises: typically, \widehat{h} exhibits zeros and (somehow more severe) arbitrarily small values. If now instead of u only \tilde{u} is given, which is endowed with an error, then $\widehat{\tilde{u}}$ exhibits an error as well. By division by \widehat{h} , this error can be magnified arbitrarily. To observe this, it is sufficient in this case to use a fine quantization of the gray values as an error in \tilde{u} ; cf. Fig. 4.3.

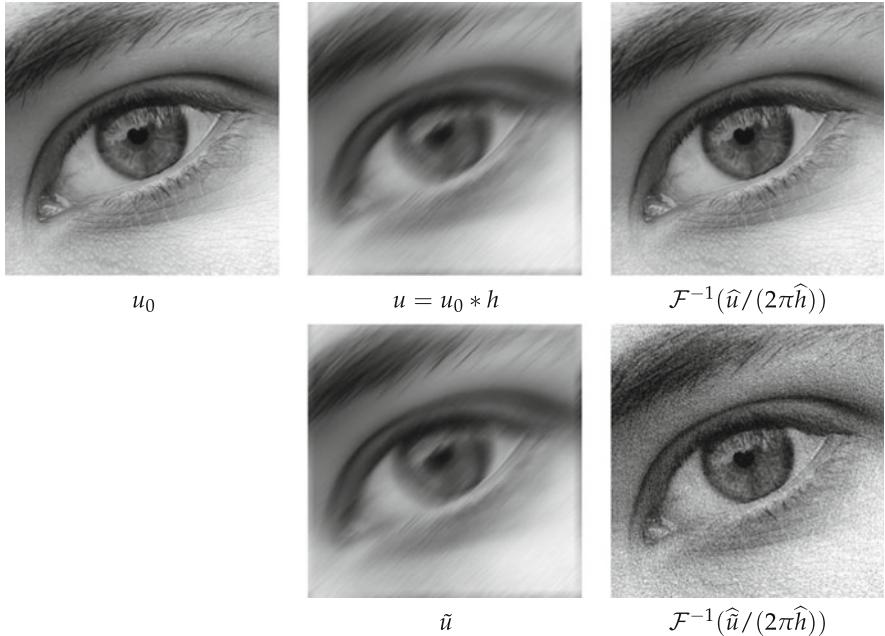


Fig. 4.3 Deconvolution with the Fourier transform. The convolution kernel h models a motion blurring. The degraded image \tilde{u} results from a quantization of u into 256 gray values (a difference the eye cannot perceive). After the deconvolution, the error becomes unpleasantly apparent

4.1.3 The Fourier Transform for Measures and Tempered Distributions

We can not only extend the Fourier transform from the Schwartz space $\mathcal{S}(\mathbf{R}^d)$ to $L^2(\mathbf{R}^d)$, but even define it for certain distributions as well. For this purpose, we define a specific space of distributions:

Definition 4.22 By $\mathcal{S}(\mathbf{R}^d)^*$ we denote the dual space of $\mathcal{S}(\mathbf{R}^d)$, i.e., the space of all linear and continuous functionals $T : \mathcal{S}(\mathbf{R}^d) \rightarrow \mathbf{C}$. We call this space the space of *tempered distributions*.

Tempered distributions are distributions in the sense of Sect. 2.3. Furthermore, there are both regular and non-regular tempered distributions. The delta-distribution is a non-regular tempered distribution, for example. In particular, every function $u \in \mathcal{S}(\mathbf{R}^d)$ induces a regular tempered distribution T_u :

$$T_u(\phi) = \int_{\mathbf{R}^d} u(x)\phi(x) \, dx,$$

but every polynomial function u also does so.

Remark 4.23 We use the notation T_u for the distribution induced by u and the similar notation T_y for the translation by y as long as there cannot be any confusion.

Our goal is to define a Fourier transform for tempered distributions. Since one often does not distinguish between a function and the induced distribution, it is reasonable to denote the Fourier transform of T_u by $\widehat{T}_u = T_{\widehat{u}}$. According to Lemma 4.8,

$$\widehat{T}_u(\phi) = \int_{\mathbf{R}^d} \widehat{u}(\xi) \phi(\xi) d\xi = \int_{\mathbf{R}^d} u(\xi) \widehat{\phi}(\xi) d\xi = T_u(\widehat{\phi}).$$

This motivates the following definition:

Definition 4.24 The Fourier transform of $T \in \mathcal{S}(\mathbf{R}^d)^*$ is defined by

$$\widehat{T}(\phi) = T(\widehat{\phi}).$$

Analogously, the inverse Fourier transform of T is given by

$$\check{T}(\phi) = T(\check{\phi}).$$

Since the Fourier transform is bijective from the Schwartz space to itself, the same holds if we view the Fourier transform as a map from the space of tempered distributions to itself.

Theorem 4.25 As a mapping of the space of tempered distributions into itself, the Fourier transform $T \mapsto \widehat{T}$ is bijective and is inverted by $T \mapsto \check{T}$.

Since according to the Riesz-Markov representation theorem (Theorem 2.62), Radon measures are elements of the dual space of continuous functions, they are in particular tempered distributions as well. Hence, by Definition 4.24, we have defined a Fourier transform for Radon measures as well.

Example 4.26 The distribution belonging to the Dirac measure δ_x of Example 2.38 is the *delta distribution*, denoted by δ_x as well:

$$\delta_x(\phi) = \int_{\mathbf{R}^d} \phi d\delta_x = \phi(x).$$

Its Fourier transform is given by

$$\widehat{\delta}_x(\phi) = \delta_x(\widehat{\phi}) = \widehat{\phi}(x) = \int_{\mathbf{R}^d} \frac{1}{(2\pi)^{d/2}} e^{-ix \cdot y} \phi(y) dy.$$

Therefore, the Fourier transform of δ_x is regular and represented by the function $y \mapsto \frac{1}{(2\pi)^{d/2}} e^{-ix \cdot y}$. In particular, the Fourier transform of δ_0 is given by the constant function $1/(2\pi)^{d/2}$.

Calculation with tempered distributions in the context of the Fourier transform does not usually constitute a significant difficulty. We illustrate this using the example of the convolution theorem on $L^2(\mathbf{R}^d)$:

Theorem 4.27 *For $u, v \in L^2(\mathbf{R}^d)$, one has for almost all $\xi \in \mathbf{R}^d$ that*

$$\widehat{u * v}(\xi) = (2\pi)^{d/2} \widehat{u}(\xi) \widehat{v}(\xi).$$

Proof We calculate “distributionally” and show the equality $\widehat{T_{u*v}} = T_{(2\pi)^{d/2}\widehat{u}\widehat{v}}$:

$$\begin{aligned} \int_{\mathbf{R}^d} (u * v)(\xi) \widehat{\phi}(\xi) d\xi &= \int_{\mathbf{R}^d} \int_{\mathbf{R}^d} u(y) v(\xi - y) dy \widehat{\phi}(\xi) d\xi \\ &= \int_{\mathbf{R}^d} u(y) \int_{\mathbf{R}^d} v(\xi - y) \widehat{\phi}(\xi) d\xi dy \\ &= \int_{\mathbf{R}^d} u(y) \int_{\mathbf{R}^d} \frac{1}{(2\pi)^{d/2}} \int_{\mathbf{R}^d} v(\xi - y) \phi(\xi) e^{-i\xi \cdot x} dx d\xi dy \\ &= \int_{\mathbf{R}^d} u(y) \int_{\mathbf{R}^d} \widehat{v}(x) e^{-iy \cdot x} \phi(x) dx dy \\ &= \int_{\mathbf{R}^d} (2\pi)^{d/2} \widehat{u}(x) \widehat{v}(x) \phi(x) dx. \end{aligned} \quad \square$$

The computation rules for Fourier transforms and derivatives in Lemma 4.13 hold analogously for weak derivatives:

Lemma 4.28 *Let $u \in L^2(\mathbf{R}^d)$ and $\alpha \in \mathbf{N}^d$ be such that the weak derivative $\partial^\alpha u$ lies in $L^2(\mathbf{R}^d)$ as well. Then*

$$\mathcal{F}(\partial^\alpha u) = i^{|\alpha|} p^\alpha \mathcal{F}(u).$$

For $p^\alpha u \in L^2(\mathbf{R}^d)$, there holds

$$\mathcal{F}(p^\alpha u) = i^{|\alpha|} \partial^\alpha \mathcal{F}(u).$$

Proof As above, we show the equation in the distributional sense. We use integration by parts, Lemma 4.13, and the Plancherel formula (4.2), to obtain for a Schwartz function ϕ ,

$$\begin{aligned} \widehat{T_{\partial^\alpha u}}(\phi) &= T_{\partial^\alpha u}(\widehat{\phi}) = \int_{\mathbf{R}^d} \partial^\alpha u(x) \widehat{\phi}(x) dx \\ &= (-1)^{|\alpha|} \int_{\mathbf{R}^d} u(x) \partial^\alpha \widehat{\phi}(x) dx \end{aligned}$$

$$\begin{aligned}
&= (-1)^{|\alpha|} \int_{\mathbf{R}^d} u(x) (-\widehat{i^{|\alpha|} p^\alpha \phi})(x) dx \\
&= (-1)^{|\alpha|} \int_{\mathbf{R}^d} \widehat{u}(x) (-i^{|\alpha|} p^\alpha(x) \phi(x)) dx = T_{i^{|\alpha|} p^\alpha} \widehat{u}(\phi).
\end{aligned}$$

The second assertion is left as Exercise 4.8. \square

This lemma yields the following characterization of the Sobolev spaces $H^k(\mathbf{R}^d)$:

Theorem 4.29 *For $k \in \mathbf{N}$,*

$$u \in H^k(\mathbf{R}^d) \iff \int_{\mathbf{R}^d} (1 + |\xi|^2)^k |\widehat{u}(\xi)|^2 d\xi < \infty.$$

Proof The Sobolev space $H^k(\mathbf{R}^d)$ consists of those $L^2(\mathbf{R}^d)$ functions u for which the corresponding norm $\|u\|_{k,2}^2 = \sum_{|\alpha| \leq k} \|\partial^\alpha u\|_2^2$ is finite. By means of the Plancherel formula, this translates into

$$\begin{aligned}
\|u\|_{k,2}^2 &= \sum_{|\alpha| \leq k} \|\widehat{\partial^\alpha u}\|_2^2 \\
&= \sum_{|\alpha| \leq k} \int_{\mathbf{R}^d} |\xi^\alpha \widehat{u}(\xi)|^2 d\xi \\
&= \int_{\mathbf{R}^d} \sum_{|\alpha| \leq k} |\xi^\alpha|^2 |\widehat{u}(\xi)|^2 d\xi.
\end{aligned}$$

The asserted equivalence now follows from the fact that the functions $h(\xi) = \sum_{|\alpha| \leq k} |\xi^\alpha|^2$ and $g(\xi) = (1 + |\xi|^2)^k$ are comparable, i.e., they can be estimated against each other by constants that depend on k and d only, cf. Exercise 4.9. This shows in particular that $(\int_{\mathbf{R}^d} (1 + |\xi|^2)^k |\widehat{u}(\xi)|^2 d\xi)^{1/2}$ is an equivalent norm on $H^k(\mathbf{R}^d)$. \square

Another way to put the previous theorem is that Sobolev space $H^k(\mathbf{R}^d)$ is the Fourier transform of the weighted Lebesgue space $L^2_{(1+|\cdot|^2)^k}(\mathbf{R}^d)$.

Example 4.30 Roughly speaking, (weak) differentiability of a function is reflected in a rapid decay of the Fourier transform at infinity. Concerning this, we exemplarily consider the Fourier transforms of the $L^2(\mathbf{R})$ -functions

$$\begin{aligned}
u(x) &= \chi_{[-1,1]}(x), \\
v(x) &= \exp(-x^2), \\
w(x) &= (1 + x^2)^{-1}
\end{aligned}$$

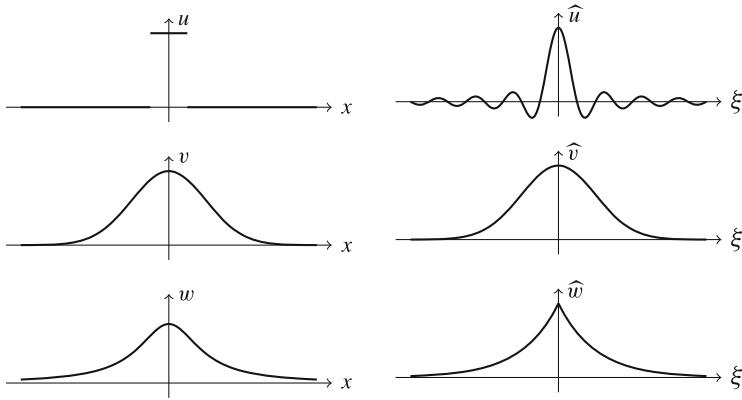


Fig. 4.4 Illustration to Example 4.30: the smoothness of a function is reflected in the rapid decay of the Fourier transform (and vice versa)

depicted in Fig. 4.4. The Fourier transform of u exhibits a decay rate at infinity like $|\xi|^{-1}$; in particular, the function $\xi \mapsto |\xi|^2 \hat{u}(\xi)$ is not in $L^2(\mathbf{R})$. For v and w , however, the Fourier transforms decay exponentially (cf. Exercise 4.4); in particular, $\xi \mapsto |\xi|^k \hat{v}(\xi)$ is an $L^2(\mathbf{R})$ -function for every $k \in \mathbf{N}$ (just as it is for w). Conversely, the slow decay of w is reflected in non-differentiability of \hat{w} .

The relationship between smoothness and decay is of fundamental importance for image processing: images with discontinuities never have a rapidly decaying Fourier transform. This demonstrates again that in filtering with low-pass filters, edges are necessarily smoothed and hence become blurred (cf. Example 4.19).

The equivalence in Theorem 4.29 motivates us to define Sobolev spaces for arbitrary smoothness $s \in \mathbf{R}$ as well:

Definition 4.31 The *fractional Sobolev space* to $s \in \mathbf{R}$ is defined by

$$u \in H^s(\mathbf{R}^d) \iff \int_{\mathbf{R}^d} (1 + |\xi|^2)^s |\hat{u}(\xi)|^2 d\xi < \infty.$$

The space is endowed with the following inner product:

$$(u, v)_{H^s(\mathbf{R}^d)} = \int_{\mathbf{R}^d} (1 + |\xi|^2)^s \hat{u}(\xi) \overline{\hat{v}(\xi)} d\xi.$$

Remark 4.32 In fractional Sobolev spaces, “nonsmooth” functions can still exhibit a certain smoothness. For instance, the characteristic function $u(x) = \chi_{[-1,1]}(x)$ lies in the space $H^s(\mathbf{R})$ for every $s \in [0, 1/2[$ as one should convince oneself in Exercise 4.10.

4.2 Fourier Series and the Sampling Theorem

Apart from the Fourier transform on $L^1(\mathbf{R}^d)$, $L^2(\mathbf{R}^d)$, $\mathcal{S}(\mathbf{R}^d)$, and $\mathcal{S}(\mathbf{R}^d)^*$, analogous transformations for functions f on rectangles $\prod_{k=1}^d [a_k, b_k] \subset \mathbf{R}^d$ are of interest for image processing as well. This leads to the so-called Fourier series. By means of these, we will prove the sampling theorem, which explains the connection of a continuous image to its discrete sampled version. Furthermore, we will be able to explain the aliasing in Fig. 3.1 that results from incorrect sampling.

4.2.1 Fourier Series

For now, we consider one-dimensional signals $u : [-\pi, \pi] \rightarrow \mathbf{C}$. We will obtain signals on general bounded intervals through scaling, and we will cover higher dimensional mappings by forming tensor products. In contrast to the Fourier transform situation, we can begin directly in the Hilbert space $L^2([-\pi, \pi])$ immediately in the case of Fourier series. We endow it with the normalized inner product

$$(u, v)_{[-\pi, \pi]} = \frac{1}{2\pi} \int_{-\pi}^{\pi} u(x)\bar{v}(x) dx.$$

The result of relevance for us is given in the following theorem:

Theorem 4.33 *For $k \in \mathbf{Z}$, we set $e_k(x) = e^{ikx}$. These functions $(e_k)_{k \in \mathbf{Z}}$ form an orthonormal basis of $L^2([-\pi, \pi])$. In particular, every function $u \in L^2([-\pi, \pi])$ can be expressed as a Fourier series*

$$u = \sum_{k \in \mathbf{Z}} (u, e_k)_{[-\pi, \pi]} e_k.$$

Proof The orthonormality of the functions $(e_k)_k$ can be verified elementarily. In order to show that $(e_k)_k$ forms a basis, we will show that the linear span of $(e_k)_k$ is dense in $L^2([-\pi, \pi])$. According to the Weierstrass approximation theorem for trigonometric polynomials (cf. [123], for instance), for every continuous function $u : [-\pi, \pi] \rightarrow \mathbf{C}$ and every $\varepsilon > 0$, there exists a trigonometric polynomial $P_k(x) = \sum_{n=-k}^k a_n e_n(x)$ such that $|u(x) - P_k(x)| \leq \varepsilon$. This implies

$$\|u - P_k\|_2^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} |u(x) - P_k(x)|^2 dx \leq \varepsilon^2.$$

Since the continuous functions are dense in $L^2([-\pi, \pi])$, every L^2 -function can also be approximated by trigonometric polynomials arbitrarily well, and we conclude that $(e_k)_k$ forms a basis. \square

The values

$$(u, e_k)_{[-\pi, \pi]} = \frac{1}{2\pi} \int_{-\pi}^{\pi} u(x) e^{-ikx} dx$$

are called Fourier coefficients of u .

Remark 4.34 For functions in $L^2([-B, B])$, we define the inner product

$$(u, v)_{[-B, B]} = \frac{1}{2B} \int_{-B}^B u(x) \bar{v}(x) dx.$$

In this case, the functions $e_k(x) = e^{ik\frac{\pi}{B}x}$ form an orthonormal basis and together with the Fourier coefficients

$$(u, e_k)_{[-B, B]} = \frac{1}{2B} \int_{-B}^B u(x) e^{-ik\frac{\pi}{B}x} dx$$

of $u \in L^2([-B, B])$, one has

$$u = \sum_{k \in \mathbf{Z}} (u, e_k)_{[-B, B]} e_k.$$

On a d -dimensional rectangle $\Omega = \prod_{l=1}^d [-B_l, B_l]$, we define the functions $(e_{\vec{k}})_{\vec{k} \in \mathbf{Z}^d}$ by

$$e_{\vec{k}}(x) = \prod_{l=1}^d e^{ik_l \frac{\pi}{B_l} x_l}$$

and obtain an orthonormal basis in $L^2(\Omega)$ with respect to the inner product

$$(u, v)_{\Omega} = \frac{1}{2^d \prod_{l=1}^d B_l} \int_{\Omega} u(x) \bar{v}(x) dx.$$

4.2.2 The Sampling Theorem

Continuous one-dimensional signals $u : \mathbf{R} \rightarrow \mathbf{C}$ are usually sampled with a constant *sampling rate* $T > 0$, i.e., the values $(u(nT))_{n \in \mathbf{Z}}$ are measured. As illustrated in Figs. 3.1 and 4.5, the discrete sampling of a signal can have unexpected effects. In particular, the sampled function does not necessarily reflect the actual function well. The following theorem shows that under certain conditions, the discrete sampling points nevertheless carry the entire information of the signal.

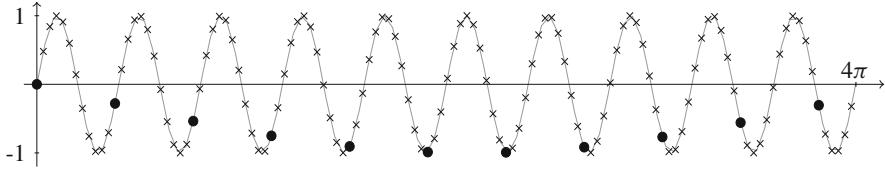


Fig. 4.5 Sampling of the function $u(x) = \sin(5x)$. While sampling with rate $T = 0.1$ (crosses) reflects the function well, the sampling rate $T = 1.2$ (dots) yields a totally wrong impression, since it suggests a much too low frequency

Theorem 4.35 (Shannon-Whittaker Sampling Theorem) *Let $B > 0$ and $u \in L^2(\mathbf{R})$ be such that $\widehat{u}(\xi) = 0$ for $|\xi| > B$. Then u is determined by the values $(u(k\pi/B))_{k \in \mathbf{Z}}$ and for all $x \in \mathbf{R}$, one has the reconstruction formula*

$$u(x) = \sum_{k \in \mathbf{Z}} u\left(\frac{k\pi}{B}\right) \operatorname{sinc}\left(\frac{B}{\pi}(x - \frac{k\pi}{B})\right).$$

Proof In this proof, we make use of the trick that \widehat{u} can be regarded as an element in $L^2(\mathbf{R})$ as well as in $L^2([-B, B])$. Thus, we can consider both the Fourier transform and the Fourier series of \widehat{u} .

Since \widehat{u} lies in $L^2([-B, B])$, it lies in $L^1([-B, B])$ as well. Thus, u is continuous and the evaluation of u at a point is well defined. We use the inversion formula of the Fourier transform, and with the basis functions $e_k(x) = e^{ik\frac{\pi}{B}x}$, we obtain

$$u\left(\frac{k\pi}{B}\right) = \frac{1}{\sqrt{2\pi}} \int_{-B}^B \widehat{u}(\xi) e^{i\xi \frac{k\pi}{B}} d\xi = \sqrt{\frac{2}{\pi}} B (\widehat{u}, e_{-k})_{[-B, B]}.$$

Hence, the values $u\left(\frac{k\pi}{B}\right)$ determine the coefficients $(\widehat{u}, e_{-k})_{[-B, B]}$, and due to $\widehat{u} \in L^2([-B, B])$, they actually determine the whole function \widehat{u} . This proves that u is determined by the values $(u(k\pi/B))_{k \in \mathbf{Z}}$.

In order to prove the reconstruction formula, we develop \widehat{u} into its Fourier series and note that for $\xi \in \mathbf{R}$, we need to restrict the result by means of the characteristic function $\chi_{[-B, B]}$:

$$\widehat{u}(\xi) = \sum_{k \in \mathbf{Z}} (\widehat{u}, e_k)_{[-B, B]} e_k(\xi) \chi_{[-B, B]}(\xi) = \sqrt{\frac{\pi}{2}} \frac{1}{B} \sum_{k \in \mathbf{Z}} u\left(-\frac{k\pi}{B}\right) e_k(\xi) \chi_{[-B, B]}(\xi).$$

Since the inverse Fourier transform is continuous, we can pull it inside the series and obtain

$$u = \sqrt{\frac{\pi}{2}} \frac{1}{B} \sum_{k \in \mathbf{Z}} u\left(-\frac{k\pi}{B}\right) \mathcal{F}^{-1}(e_k \chi_{[-B, B]}).$$

By means of the calculation rules in Lemma 4.5 and Exercise 4.3, we infer

$$\begin{aligned}\mathcal{F}^{-1}(e_k \chi_{[-B, B]})(x) &= \overline{\mathcal{F}(\overline{M}_k \frac{\pi}{B} \chi_{[-B, B]})(x)} \\ &= D_{-1} T_{-k \frac{\pi}{B}} \mathcal{F}(\chi_{[-B, B]})(x) = \sqrt{\frac{2}{\pi}} B \operatorname{sinc}\left(\frac{B}{\pi}(-x - \frac{k\pi}{B})\right) \\ &= \sqrt{\frac{2}{\pi}} B \operatorname{sinc}\left(\frac{B}{\pi}(x + \frac{k\pi}{B})\right).\end{aligned}$$

Inserting this expression into the previous equation yields the assertion. \square

Remark 4.36 In the above case, we call B the *bandwidth* of the signal. This bandwidth indicates the highest frequency contained in the signal. Expressed in words, the sampling theorem reads:

A signal with bandwidth B has to be sampled with sampling rate $\frac{\pi}{B}$ in order to store all information of the signal.

We here use the word “frequency” not in the sense in which it is often used in engineering. In this context, typically the angular frequency $f = 2\pi B$ is used. Also, the variant of the Fourier transform in Remark 4.3 including the term $e^{-2\pi i x \cdot \xi}$ is common there. For this situation, the assertion of the sampling theorem reads:

If a signal exhibits frequencies up to a maximal angular frequency f , it has to be sampled with the sampling rate $\frac{1}{2f}$ in order to store all information of the signal.

That is, one has to sample twice as fast as the highest angular frequency. The sampling frequency $\frac{1}{2f}$ is also called the *Nyquist rate* or *Nyquist frequency*.

4.2.3 Aliasing

Aliasing is what we observe in Figs. 3.1 and 4.5: the discrete image or signal does not match the original signal, since in the discrete version, frequencies arise that are not contained in the original. As “aliases,” they stand for the actual frequencies.

In the previous subsection, we saw that this effect cannot occur if the signal is sampled at a sufficiently high rate. In this subsection, we desire to understand how exactly aliasing arises and how we can eliminate it.

For this purpose, we need an additional tool:

Lemma 4.37 (Poisson Formula) *Let $u \in L^2(\mathbf{R})$ and $B > 0$ be such that either the function $\sum_{k \in \mathbf{Z}} \widehat{u}(\cdot + 2Bk) \in L^2([-B, B])$ or the series $\sum_{k \in \mathbf{Z}} |u(\frac{k\pi}{B})|^2$ converges. Then, for almost all $\xi \in \mathbf{R}$,*

$$\sum_{k \in \mathbf{Z}} \widehat{u}(\xi + 2Bk) = \frac{\sqrt{2\pi}}{2B} \sum_{k \in \mathbf{Z}} u\left(\frac{k\pi}{B}\right) e^{-i \frac{k\pi}{B} \xi}.$$

Proof We define the periodization of \widehat{u} by

$$\phi(\xi) = \sum_{k \in \mathbf{Z}} \widehat{u}(\xi + 2Bk).$$

For $\phi \in L^2([-B, B])$, we can represent the function by its Fourier series. The Fourier coefficients are given by

$$\begin{aligned} (\phi, e_k)_{[-B, B]} &= \frac{1}{2B} \int_{-B}^B \phi(\xi) e^{-i\frac{k\pi}{B}\xi} d\xi \\ &= \frac{1}{2B} \int_{-B}^B \sum_{l \in \mathbf{Z}} \widehat{u}(\xi + 2Bl) e^{-i\frac{k\pi}{B}\xi} d\xi \\ &= \frac{1}{2B} \int_{-B}^B \sum_{l \in \mathbf{Z}} \widehat{u}(\xi + 2Bl) e^{-i\frac{k\pi}{B}(\xi+2Bl)} d\xi \\ &= \frac{1}{2B} \int_{\mathbf{R}} \widehat{u}(\xi) e^{-i\frac{k\pi}{B}\xi} d\xi \\ &= \frac{\sqrt{2\pi}}{2B} u\left(-\frac{k\pi}{B}\right). \end{aligned}$$

Therefore, the Fourier series

$$\begin{aligned} \phi(\xi) &= \sum_{k \in \mathbf{Z}} (\phi, e_k)_{[-B, B]} e_k(\xi) \\ &= \frac{\sqrt{2\pi}}{2B} \sum_{k \in \mathbf{Z}} u\left(-\frac{k\pi}{B}\right) e^{i\frac{k\pi}{B}\xi} \end{aligned}$$

is convergent in the L^2 -sense, which implies the assertion.

Conversely, the above Fourier series converges if the coefficients $(u(\frac{k\pi}{B}))_k$ are square-summable, and the assertion follows as well. \square

Remark 4.38 For the spacial case $\xi = 0$, we obtain the remarkable formula

$$\sum_{k \in \mathbf{Z}} \widehat{u}(2Bk) = \frac{\sqrt{2\pi}}{2B} \sum_{k \in \mathbf{Z}} u\left(\frac{\pi}{B}k\right),$$

which relates the values of u and \widehat{u} .

We will now consider sampling in greater detail. Expressed by means of distributions, we can represent a signal, sampled discretely with rate $\frac{\pi}{B}$, as a delta

comb, as discussed in Remark 3.3:

$$u_d = \sum_{k \in \mathbf{Z}} u\left(\frac{k\pi}{B}\right) \delta_{k\frac{\pi}{B}}.$$

The connection between u and u_d can be clarified via the Fourier transform:

Lemma 4.39 *For almost all $\xi \in \mathbf{R}$, one has*

$$\widehat{u}_d(\xi) = \frac{B}{\pi} \sum_{k \in \mathbf{Z}} \widehat{u}(\xi + 2Bk).$$

Proof According to Example 4.26, the Fourier transform of $\delta_{k\frac{\pi}{B}}$ is given by

$$\mathcal{F}(\delta_{k\frac{\pi}{B}})(\xi) = \frac{1}{\sqrt{2\pi}} e^{-i\frac{k\pi}{B}\xi}.$$

Due to the Poisson formula in Lemma 4.37, we therefore have

$$\begin{aligned} \widehat{u}_d(\xi) &= \frac{1}{\sqrt{2\pi}} \sum_{k \in \mathbf{Z}} u\left(\frac{k\pi}{B}\right) e^{-i\frac{k\pi}{B}\xi} \\ &= \frac{B}{\pi} \sum_{n \in \mathbf{Z}} \widehat{u}(\xi + 2Bn). \end{aligned} \quad \square$$

Expressed in words, the lemma states that the Fourier transform of the sampled signal corresponds to a periodization of the Fourier transform of the original signal with period $2B$.

In this way of speaking, we can interpret the reconstruction formula in the sampling theorem (Theorem 4.35) as a convolution as well:

$$u(x) = \sum_{k \in \mathbf{Z}} u\left(\frac{k\pi}{B}\right) \operatorname{sinc}\left(\frac{B}{\pi}(x - \frac{k\pi}{B})\right) = u_d * \operatorname{sinc}\left(\frac{B}{\pi} \cdot\right)(x).$$

In the Fourier realm, this formally means

$$\widehat{u}(\xi) = \widehat{u}_d(\xi) \frac{B}{\pi} \chi_{[-B, B]}(\xi).$$

If the support of \widehat{u} is contained in the interval $[-B, B]$, then no overlap occurs during periodization, and $\widehat{u}_d \frac{B}{\pi} \chi_{[-B, B]}$ corresponds to \widehat{u} exactly. This procedure is depicted in Fig. 4.6.

However, if \widehat{u} has a larger support, then the support of $\widehat{u}(\cdot + 2Bk)$ exhibits a nonempty intersection with $[-B, B]$ for several k . This “folding” in the frequency domain is responsible for aliasing; cf. Fig. 4.7.

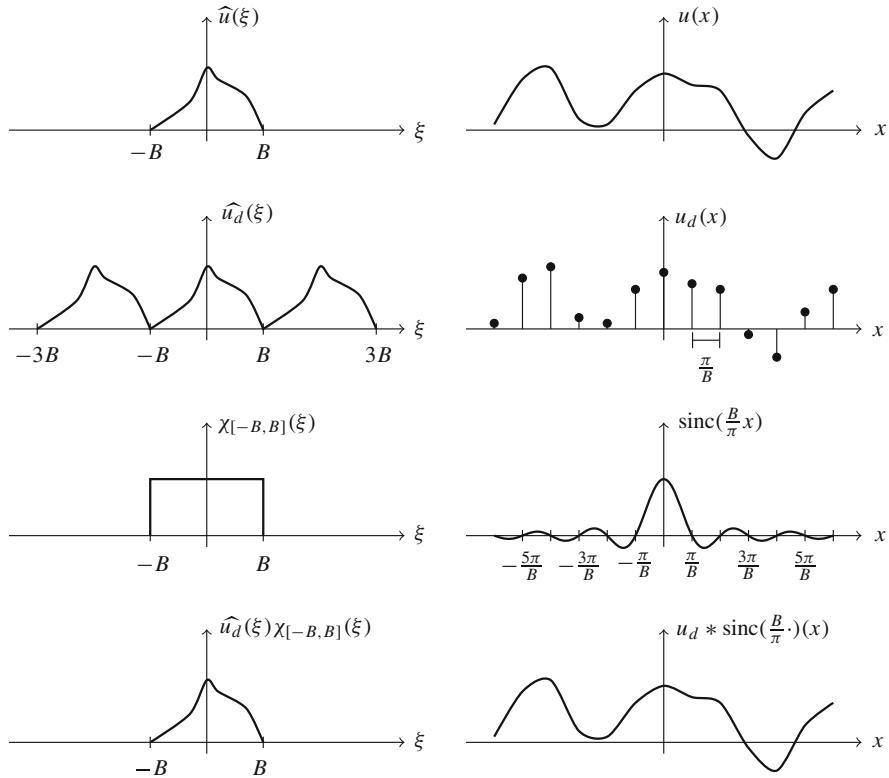


Fig. 4.6 Reconstruction of a discretized signal by means of the reconstruction formula in the sampling theorem. First row: A signal u and its Fourier transform. Second row: Sampling the function renders the Fourier transform periodic. Third row: The sinc convolution kernel and its Fourier transform. Fourth row: Convoluting with the sinc function reconstructs the signal perfectly

Example 4.40 (Sampling of Harmonic Oscillations) We consider a harmonic oscillation

$$u(x) = \cos(\xi_0 x) = \frac{e^{i\xi_0 x} + e^{-i\xi_0 x}}{2}.$$

Its Fourier transform is given by

$$\hat{u} = \sqrt{\frac{\pi}{2}}(\delta_{\xi_0} + \delta_{-\xi_0}).$$

Hence, the signal formally has bandwidth ξ_0 . If we assume another bandwidth B and accordingly sample the signal with rate π/B , we obtain in the Fourier realm

$$\hat{u}_d = \frac{\pi}{B} \sqrt{\frac{\pi}{2}} \sum_{k \in \mathbf{Z}} (\delta_{\xi_0 - 2kB} + \delta_{-\xi_0 - 2kB}).$$

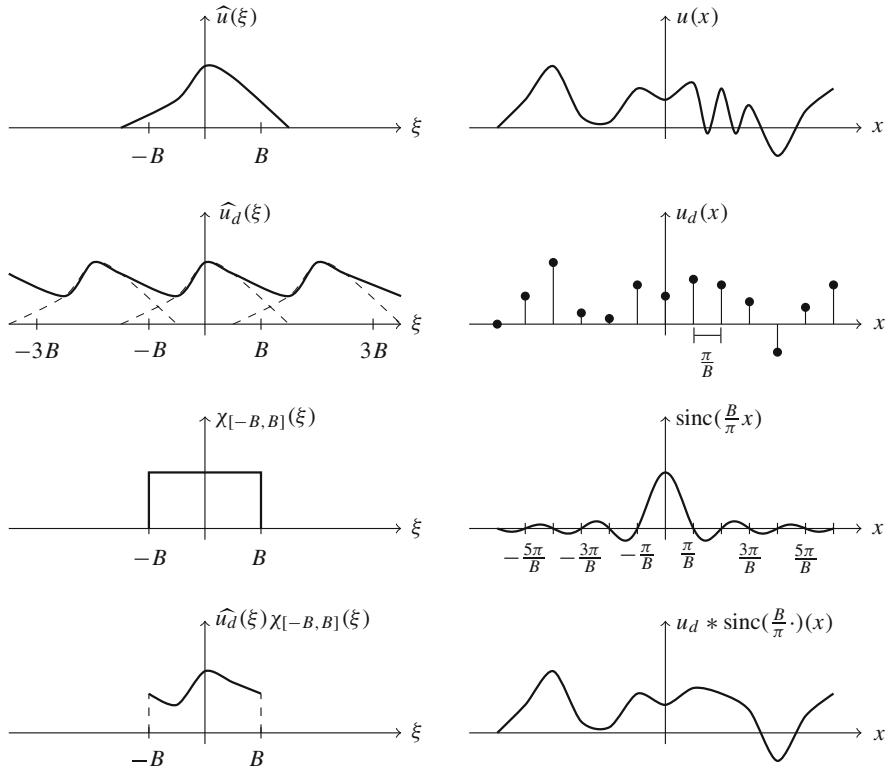


Fig. 4.7 Illustration of aliasing. First row: A signal u and its Fourier transform. Second row: Sampling of the function renders the Fourier transform periodic and produces an overlap. Third row: The sinc convolution kernel and its Fourier transform. Fourth row: Convoluting with the sinc function reconstructs a signal in which the high-frequency components are represented by low frequencies

Reconstructing according to the sampling theorem (Theorem 4.35) means to restrict the support of \hat{u}_d to the interval $[-B, B]$. In order to understand what this implies for the signal, we need to study how this restriction affects the series.

Oversampling: If we assume a bandwidth $B > \xi_0$ that is too large, we sample the signal too fast. Then, the terms in the series for \hat{u}_d that lie in the interval $[-B, B]$ are precisely those that belong to $k = 0$. One has

$$\begin{aligned}\hat{u}_d &= \frac{\pi}{B} \sqrt{\frac{\pi}{2}} \sum_{k \in \mathbf{Z}} (\delta_{\xi_0 - 2kB} + \delta_{-\xi_0 - 2kB}) \frac{B}{\pi} \chi_{[-B, B]} \\ &= \sqrt{\frac{\pi}{2}} (\delta_{\xi_0} + \delta_{-\xi_0}) = \hat{u}.\end{aligned}$$

This implies $u_d = u$, i.e., we reconstruct the signal perfectly.

Undersampling: If we take a bandwidth $B < \xi_0 < 3B$ that is too small, we sample the signal too slowly. Then again, exactly two terms in the series for \widehat{u}_d lie in the interval $[-B, B]$, namely δ_{ξ_0-2B} and $\delta_{-\xi_0+2B}$, i.e., we have

$$\widehat{u}_d \frac{B}{\pi} \chi_{[-B, B]} = \sqrt{\frac{\pi}{2}} (\delta_{\xi_0-2B} + \delta_{-\xi_0+2B}).$$

We hence reconstruct the signal

$$u_{\text{rek}}(x) = \cos((\xi_0 - 2B)x).$$

The reconstruction is again a harmonic oscillation, but it exhibits a different frequency.

Through undersampling, high frequencies ξ_0 are represented by low frequencies in $[-B, B]$. As an exercise, one can calculate the frequency observed from undersampling in Fig. 4.5.

Remark 4.41 (Sampling in 2D) We obtain a simple generalization of the sampling theorem and the explanation of aliasing for two dimensions by forming the tensor product. Let $u : \mathbf{R}^2 \rightarrow \mathbf{C}$ be such that its Fourier transform \widehat{u} has its support in the rectangle $[-B_1, B_1] \times [-B_2, B_2]$. In this case, u is determined by the values $u(k_1\pi/B_1, k_2\pi/B_2)$, and

$$u(x_1, x_2) = \sum_{k \in \mathbf{Z}^2} u\left(\frac{k_1\pi}{B_1}, \frac{k_2\pi}{B_2}\right) \operatorname{sinc}\left(\frac{B_1}{\pi}(x_1 - \frac{k_1\pi}{B_1})\right) \operatorname{sinc}\left(\frac{B_2}{\pi}(x_2 - \frac{k_2\pi}{B_2})\right).$$

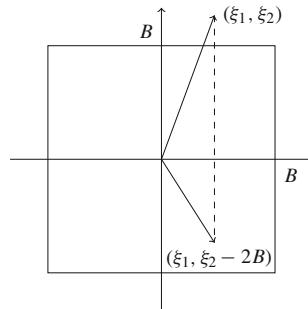
We denote by

$$u_d = \sum_{k \in \mathbf{Z}^2} u(k_1 T_1, k_2 T_2) \delta_{(k_1 T_1, k_2 T_2)}$$

an image that is discretely sampled on a rectangular grid with sampling rates T_1 and T_2 . By means of the Fourier transform, the connection to the continuous image u can be expressed as

$$\widehat{u}_d(\xi) = \frac{B_1 B_2}{\pi^2} \sum_{k \in \mathbf{Z}^2} \widehat{u}(\xi_1 + 2B_1 k_1, \xi_2 + 2B_2 k_2).$$

Also in this case, aliasing occurs if the image does not have finite bandwidth or is sampled too slowly. In addition to the change of frequency, a change in the direction also may occur here:



Example 4.42 (Undersampling, Preventing Aliasing) If we want reduce the size of a given discrete image $u_d = \sum_{k \in \mathbb{Z}^2} u_k \delta_k$ by the factor $l \in \mathbf{N}$, we obtain $u_d^l = \sum_{k \in \mathbb{Z}^2} u_{lk} \delta_{lk}$. Also during this undersampling, aliasing arises; cf. Fig. 3.1. In order to prevent this, a low-pass filter h should be applied before the undersampling in order to eliminate those frequencies that due to the aliasing would be reconstructed as incorrect frequencies. It suggests itself to choose this filter as the *perfect low-pass filter* with width π/l , i.e., we have $\hat{h} = \chi_{[-\pi/l, \pi/l]^2}$. This prevents aliasing; cf. Fig. 4.8.

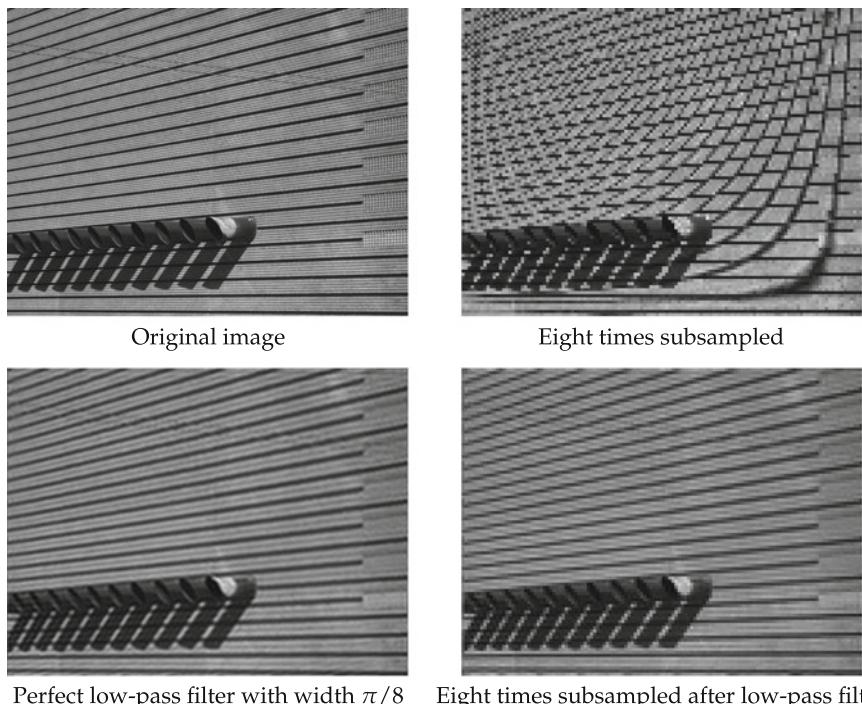


Fig. 4.8 Preventing aliasing by low-pass filtering. For better comparability, the subsampled images are rescaled to the original size

4.3 The Discrete Fourier Transform

For the numerical realization of frequency methods, we need the discrete Fourier transform. Also in this case, we shall study one-dimensional discrete images for now and will obtain the higher-dimensional version later as a tensor product. Hence, we consider

$$u : \{0, \dots, N - 1\} \rightarrow \mathbf{C}.$$

These images form an N -dimensional vector space \mathbf{C}^N , which by means of the inner product

$$(u, v) = \sum_{n=0}^{N-1} u_n \bar{v}_n$$

turns into a Hilbert space.

Definition 4.43 The one-dimensional *discrete Fourier transform* $\widehat{u} \in \mathbf{C}^N$ of $u \in \mathbf{C}^N$ is defined by

$$\widehat{u}_k = \frac{1}{N} \sum_{n=0}^{N-1} u_n \exp\left(\frac{-2\pi i n k}{N}\right).$$

By means of the vectors

$$b^n = \left(\exp\left(\frac{-2\pi i n k}{N}\right) \right)_{k=0, \dots, N-1}$$

and the resulting matrix

$$B = [b^0 \dots b^{N-1}],$$

we can express the discrete Fourier transform as a matrix vector product:

$$\widehat{u} = \frac{1}{N} Bu.$$

Theorem 4.44 *The vectors b^n are orthogonal. In particular,*

$$(b^n, b^{n'}) = N \delta_{n,n'}.$$

Furthermore, the Fourier transform is inverted by

$$u_n = \sum_{k=0}^{N-1} \hat{u}_k \exp\left(\frac{2\pi i n k}{N}\right).$$

Proof The orthogonality relation of the vectors b^n can be verified elementarily. In particular, this implies that the matrix B is orthogonal and $B^*B = N \text{id}$, which yields $B^{-1} = \frac{1}{N} B^*$; i.e., we have

$$u = NB^{-1}\hat{u} = B^*\hat{u}.$$

For the adjoint matrix B^* , we have

$$(B^*)_{k,l} = \exp\left(\frac{2\pi i k l}{N}\right)$$

and in particular, we obtain

$$B^* = \begin{bmatrix} \overline{b^0} & \dots & \overline{b^{N-1}} \end{bmatrix}. \quad \square$$

In other words, the discrete Fourier transform expresses a vector u in terms of the orthogonal basis

$$(\overline{b^k})_{k=0,\dots,N-1}, \quad \text{i.e.,} \quad u = \sum_k \hat{u}_k \overline{b^k}.$$

Also in the discrete case, we denote the inverse of the Fourier transform by \check{u} .

The generalization to two-dimensional images is simple:

Remark 4.45 (Two-Dimensional Discrete Fourier Transform) The two-dimensional discrete Fourier transform $\hat{u} \in \mathbf{C}^{N \times M}$ of $u \in \mathbf{C}^{N \times M}$ is defined by

$$\hat{u}_{k,l} = \frac{1}{MN} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} u_{n,m} \exp\left(\frac{-2\pi i n k}{N}\right) \exp\left(\frac{-2\pi i m l}{M}\right)$$

and is inverted by

$$u_{n,m} = \sum_{k=0}^{N-1} \sum_{l=0}^{M-1} \hat{u}_{k,l} \exp\left(\frac{2\pi i n k}{N}\right) \exp\left(\frac{2\pi i m l}{M}\right).$$

Remark 4.46 (Periodicity of the Discrete Fourier Transform) The vectors b^n can also be regarded as N -period, i.e.,

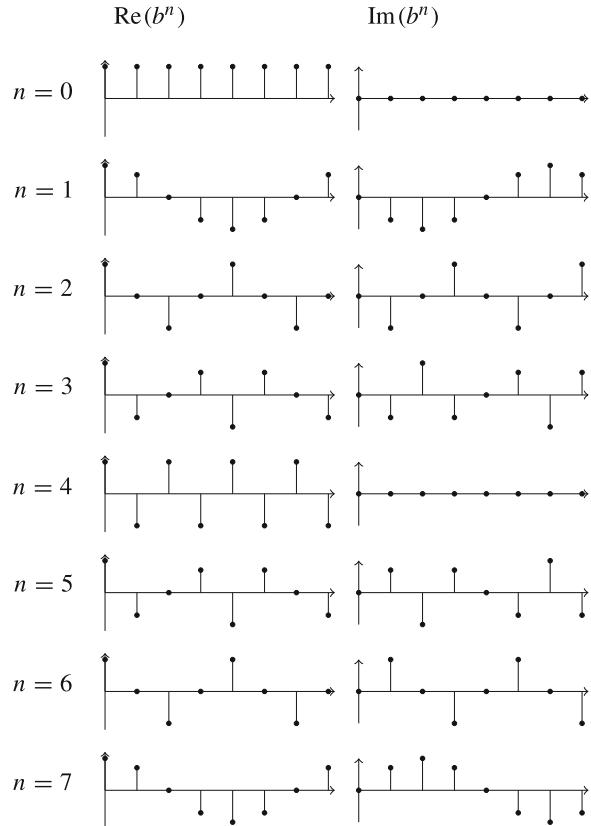
$$b_{k+N}^n = \exp\left(\frac{2\pi i n(k+N)}{N}\right) = \exp\left(\frac{2\pi i nk}{N}\right) = b_k^n.$$

Furthermore, one has also $b^{n+N} = b^n$. In other words, for the discrete Fourier transform, all signals are N -periodic, since we have

$$\hat{u}_{k+N} = \hat{u}_k.$$

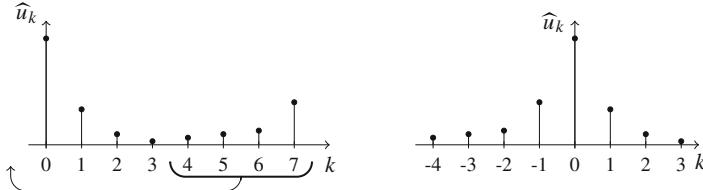
This observation is important for the interpretation of the Fourier coefficients. The basis vectors b^{N-n} correspond to the vectors b^{-n} , i.e., the entries \hat{u}_{N-k} with small k correspond to the low-frequency vectors b^{-k} . This can also be observed in Fig. 4.9: the “high” basis vectors have low frequencies, while the “middle” basis vectors have the highest frequencies. Another explanation for this situation is given by the sampling theorem as well as the alias effect: when sampling with rate 1, the highest

Fig. 4.9 Basis vectors of the discrete Fourier transform for $N = 8$ (left: real part, right: imaginary part)



frequency that can be represented is given by π , which corresponds to the vector $b^{n/2}$. The higher frequencies are represented as low frequencies due to aliasing.

If we want to consider the frequency representation of a signal or an image, it is hence reasonable to reorder the values of the discrete Fourier transform such that the coefficient belonging to the frequency zero is placed in the center:



There is also a convolution theorem for the discrete Fourier transform. In view of the previous remark, it is not surprising that it holds for periodic convolution:

Definition 4.47 Let $u, v \in \mathbf{C}^N$. The *periodic convolution* of u with v is defined by

$$(u \circledast v)_n = \sum_{k=0}^{N-1} v_k u_{(n-k) \bmod N}.$$

Theorem 4.48 For $u, v \in \mathbf{C}^N$,

$$(\widehat{u \circledast v})_n = N \widehat{u}_n \widehat{v}_n.$$

Proof Using the periodicity of the complex exponential function, the equation can be verified directly:

$$\begin{aligned} (\widehat{u \circledast v})_n &= \frac{1}{N} \sum_{k=0}^{N-1} \sum_{l=0}^{N-1} v_l u_{(k-l) \bmod N} \exp\left(\frac{-2\pi i nk}{N}\right) \\ &= \frac{1}{N} \sum_{l=0}^{N-1} v_l \exp\left(\frac{-2\pi i nl}{N}\right) \sum_{k=0}^{N-1} u_{(k-l) \bmod N} \exp\left(\frac{-2\pi i n(k-l)}{N}\right) \\ &= N \widehat{v}_n \widehat{u}_n. \end{aligned}$$

□

Also in the discrete case, convolution can be expressed via the multiplication of the Fourier transform. And here we again call the Fourier transform of a convolution kernel the *transfer function*:

Definition 4.49 The *transfer function* of a convolution kernel $h \in \mathbf{R}^{2r+1}$ is defined by

$$\widehat{h}_k = \frac{1}{N} \sum_{n=-r}^r h_n \exp\left(-\frac{2\pi i n k}{N}\right).$$

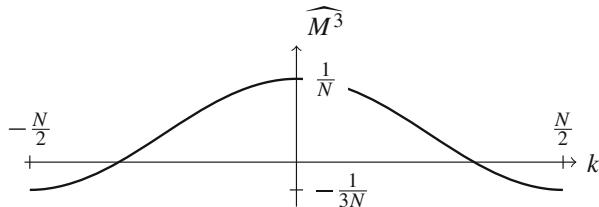
Note that the Fourier transform of a convolution kernel depends on the period N of the signal! Furthermore, it is important that we use here the periodic boundary extension of Sect. 3.3.3 throughout.

Example 4.50 We consider some of the filters introduced in Sect. 3.3.3.

For the moving average filter $M^3 = [1 \ 1 \ 1]/3$, the transfer function is given by

$$\begin{aligned} \widehat{M^3}_k &= \frac{1}{3N} \sum_{n=-1}^1 \exp\left(-\frac{2\pi i n k}{N}\right) \\ &= \frac{1}{3N} \left(1 + \exp\left(\frac{2\pi i k}{N}\right) + \exp\left(-\frac{2\pi i k}{N}\right) \right) \\ &= \frac{1}{3N} \left(1 + 2 \cos\left(\frac{2\pi k}{N}\right) \right). \end{aligned}$$

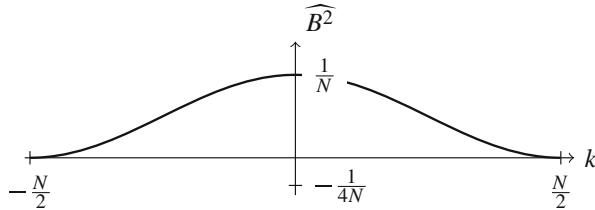
- For $k \approx \pm N/3$, the transfer function is almost 0, i.e., the corresponding frequencies vanish from the signal.
- For $k \approx N/2$, the transfer function exhibits negative values, i.e., the corresponding frequencies change their sign.



We now consider the binomial filter $B^2 = [1 \ 2 \ 1]/4$. The transfer function is given by

$$\begin{aligned} \widehat{B^2}_k &= \frac{1}{4N} \left(2 + \exp\left(\frac{2\pi i k}{N}\right) + \exp\left(-\frac{2\pi i k}{N}\right) \right) \\ &= \frac{1}{4N} \left(2 + 2 \cos\left(\frac{2\pi k}{N}\right) \right). \end{aligned}$$

The transfer function of this filter is nonnegative throughout, i.e., no frequencies are “flipped.”



In this light, Sobel filters appear more reasonable than Prewitt filters.

Remark 4.51 (Fast Fourier Transform and Fast Convolution) Evaluating the sums directly, the discrete Fourier transform needs $\mathcal{O}(N^2)$ operations. By making use of symmetries, the cost can be significantly reduced to $\mathcal{O}(N \log_2 N)$, cf. [97], for instance. By means of the convolution theorem, this can be utilized for a fast convolution; cf. Exercises 4.11 and 4.12.

Remark 4.52 The discrete Fourier transform satisfies the following symmetry relation:

$$\hat{u}_k = \frac{1}{N} \sum_{n=0}^{N-1} u_n \exp\left(-\frac{2\pi i n k}{N}\right) = \frac{1}{N} \overline{\sum_{n=0}^{N-1} \bar{u}_n \exp\left(-\frac{2\pi i n (N-k)}{N}\right)} = \bar{\bar{u}}_{N-k}.$$

If u is real, then one also has $\hat{u}_k = \widehat{u_{N-k}}$. That is, the Fourier transform contains redundant data – for N even, knowing \hat{u}_k for $k = 0, \dots, N/2-1$ suffices, for N odd, knowing \hat{u}_k for $k = 0, \dots, (N-1)/2$ suffices. Hence, in the case of real signals, more quantities are computed than necessary. These unnecessary calculations can be remedied by means of the discrete cosine transform (DCT). With

$$\lambda_k = \begin{cases} 1/\sqrt{2} & \text{if } k = 0, \\ 1 & \text{otherwise,} \end{cases}$$

we define for a real signal $u \in \mathbf{R}^N$,

$$\text{DCT}(u)_k = \frac{2\lambda_k}{N} \sum_{n=0}^{N-1} u_n \cos\left(\frac{k\pi}{N}(n + \frac{1}{2})\right).$$

The DCT is an orthogonal transformation, and we have the inversion formula

$$u_n = \text{IDCT}(\text{DCT}(u))_n = \lambda_n \sum_{k=0}^{N-1} \text{DCT}(u)_k \cos\left(\frac{k\pi}{N}(n + \frac{1}{2})\right).$$

Like the discrete Fourier transform, the DCT can be computed with complexity $\mathcal{O}(N \log_2 N)$. There are three further variants of the discrete cosine transform; cf. [97] for details, for instance.

Application 4.53 (Compression with the DCT: JPEG) The discrete cosine transform is a crucial part of the JPEG compression standard, which is based on the idea of transform coding, whereby the image is transformed into a different representation that is better suited for compression through quantization. One can observe that the discrete cosine transform of an image typically exhibits many small coefficients, which mostly belong to the high frequencies. Additionally, there is a physiological observation that the eye perceives gray value variations less well for higher frequencies. Together with further techniques, this constitutes the foundation of the JPEG standard. For grayscale images, this standard consists of the following steps:

- Partition the image into 8×8 -blocks and apply the discrete cosine transform to these blocks.
- Quantize the transformed values (this is the lossy part).
- Reorder the quantized values.
- Apply entropy coding to the resulting numerical sequence.

The compression potential of the blockwise two-dimensional DCT is illustrated in Fig. 4.10. For a detailed description of the functioning of JPEG, we refer to [109].

4.4 The Wavelet Transform

In the previous sections, we discussed in detail that the Fourier transform yields the frequency representation of a signal or an image. However, any information about location is not encoded in an obvious way. In particular, a local alteration of the signal or image at only one location results in a global change of the whole Fourier transform. In other words, the Fourier transform is a global transformation in the sense that the value $\widehat{u}(\xi)$ depends on all values of u . In several circumstances, transformations that are “local” in a certain sense appear desirable. Before we introduce the wavelet transform, we present another alternative to obtain “localized” frequency information: the windowed Fourier transform.

4.4.1 The Windowed Fourier Transform

An intuitive idea for localizing the Fourier transform is the following modification. We use a *window function* $g : \mathbf{R}^d \rightarrow \mathbf{C}$, which is nothing more than a function that is “localized around the origin,” i.e., a function that for large $|x|$, assumes small values. For $\sigma > 0$, two such examples are given by the following functions, which

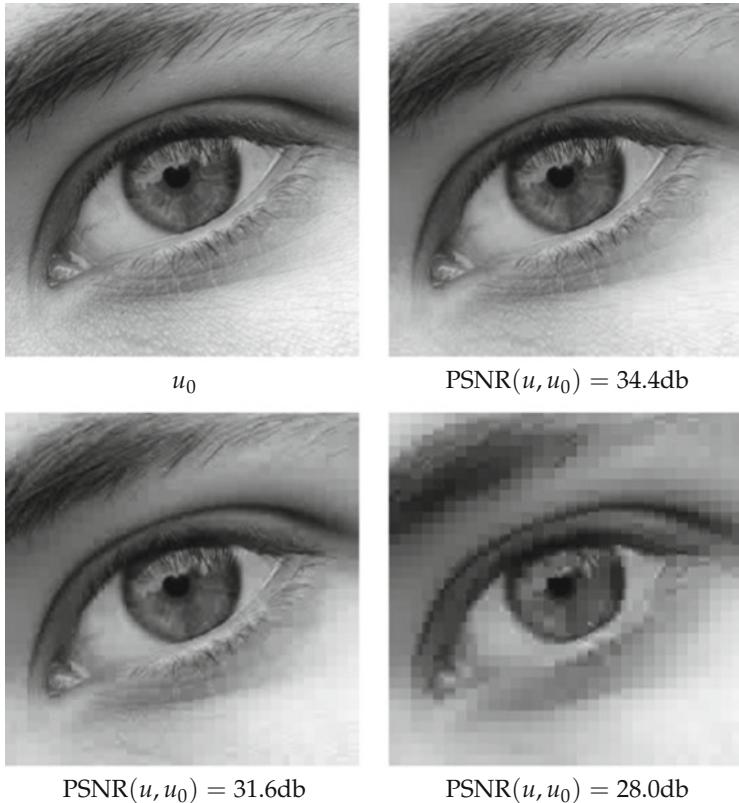


Fig. 4.10 Illustration of the compression potential of the two-dimensional DCT on 8×8 blocks. From upper left to lower right: Original image, reconstruction based on 10%, 5% and 2% of the DCT coefficients, respectively. The resulting artifacts are disturbing only in the last case

are normalized in $L^2(\mathbf{R}^d)$:

$$g(x) = \frac{\Gamma(1 + d/2)}{\sigma^d \pi^{d/2}} \chi_{B_\sigma(0)}(x) \quad \text{and} \quad g(x) = \frac{1}{(2\pi\sigma)^{d/2}} \exp\left(-\frac{|x|^2}{2\sigma}\right).$$

We localize the function of interest u around a point t by multiplying u by $g(\cdot - t)$. After that, we compute the Fourier transform of the product:

Definition 4.54 Let $u, g \in L^2(\mathbf{R}^d)$. The *windowed Fourier transform* of u with *window function* g is defined by

$$(\mathcal{G}_g u)(\xi, t) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbf{R}^d} u(x) \overline{g(x-t)} e^{-ix \cdot \xi} dx.$$

The windowed Fourier transform depends on a *frequency parameter* ξ as well as a *spatial parameter* t , and there are several other possibilities to represent it, for instance

$$\begin{aligned} (\mathcal{G}_g)u(\xi, t) &= \mathcal{F}(uT_{-t}\bar{g})(\xi) \\ &= \frac{1}{(2\pi)^{d/2}}(u, M_\xi T_{-t}g)_2 \\ &= \frac{1}{(2\pi)^{d/2}}(M_{-\xi}u * D_{-\text{id}}\bar{g})(t). \end{aligned}$$

The first alternative again explains the name “windowed” Fourier transform: through multiplication by the shifted window g , the function u is localized prior to the Fourier transform.

Note that the windowed Fourier transform is a function of $2d$ variables: $\mathcal{G}_g u : \mathbf{R}^{2d} \rightarrow \mathbf{C}$. If the window function g is a Gaussian function, i.e., $g(x) = (2\pi\sigma)^{-d/2} \exp(-|x|^2/(2\sigma))$ for some $\sigma > 0$, the transform is also called the Gabor transform.

Thanks to the previous work in Sects. 4.1.1–4.1.3, the analysis of the elementary properties of the windowed Fourier transform is straightforward.

Lemma 4.55 *Let $u, v, g \in L^2(\mathbf{R}^d)$. Then we have $\mathcal{G}_g u \in L^2(\mathbf{R}^{2d})$ and*

$$(\mathcal{G}_g u, \mathcal{G}_g v)_{L^2(\mathbf{R}^{2d})} = \|g\|_2^2(u, v)_2.$$

Proof In order to prove the equality of the inner products, we use the isometry property of the Fourier transform and in particular the Plancherel formula (4.2). With \mathcal{F}_t denoting the Fourier transform with respect to the variable t , we use one of the alternative representations of the windowed Fourier transform as well as the convolution theorem to obtain

$$\begin{aligned} \mathcal{F}_t(\mathcal{G}_g u(\xi, \cdot))(\omega) &= \mathcal{F}_t((2\pi)^{-d/2}(M_{-\xi}u * D_{-\text{id}}\bar{g}))(\omega) \\ &= \mathcal{F}(M_{-\xi}u)(\omega)(\mathcal{F}D_{-\text{id}}\bar{g})(\omega) \\ &= \widehat{u}(\omega + \xi)\overline{\widehat{g}}(\omega). \end{aligned}$$

We then obtain the assertion by computing

$$\begin{aligned} (\mathcal{G}_g u, \mathcal{G}_g v)_{L^2(\mathbf{R}^{2d})} &= (\mathcal{F}_t(\mathcal{G}_g u), \mathcal{F}_t(\mathcal{G}_g v))_{L^2(\mathbf{R}^{2d})} \\ &= \int_{\mathbf{R}^d} \int_{\mathbf{R}^d} \widehat{u}(\omega + \xi)\overline{\widehat{g}}(\omega)\overline{\widehat{v}}(\omega + \xi)\widehat{g}(\omega) d\xi d\omega \\ &= \int_{\mathbf{R}^d} |\widehat{g}(\omega)|^2 \int_{\mathbf{R}^d} \widehat{u}(\omega + \xi)\overline{\widehat{v}}(\omega + \xi) d\xi d\omega \end{aligned}$$

$$\begin{aligned}
&= \|\widehat{g}\|_2^2 (\widehat{u}, \widehat{v})_2 \\
&= \|g\|_2^2 (u, v)_2.
\end{aligned}$$

□

We thereby see that the windowed Fourier transform is an isometry, and as such, it is inverted on its range by its adjoint (up to a constant). In particular, we have the following.

Corollary 4.56 *For $u, g \in L^2(\mathbf{R}^d)$ with $\|g\|_2 = 1$, we have the inversion formula*

$$u(x) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbf{R}^d} \int_{\mathbf{R}^d} \mathcal{G}_g u(\xi, t) g(x - t) e^{ix \cdot \xi} d\xi dt \quad \text{for almost all } x.$$

Proof Since g is normalized, \mathcal{G}_g is an isometry and after the previous remarks, it remains only to compute the adjoint operator. For $u \in L^2(\mathbf{R}^d)$ and $F \in L^2(\mathbf{R}^{2d})$, we have

$$\begin{aligned}
(u, \mathcal{G}_g^* F)_{L^2(\mathbf{R}^d)} &= (\mathcal{G}_g u, F)_{L^2(\mathbf{R}^{2d})} \\
&= \int_{\mathbf{R}^{2d}} \mathcal{G}_g u(\xi, t) \overline{F(\xi, t)} d\xi dt \\
&= \int_{\mathbf{R}^{2d}} \frac{1}{(2\pi)^{d/2}} \int_{\mathbf{R}^d} u(x) \overline{g(x-t)} e^{-ix \cdot \xi} dx \overline{F(\xi, t)} d\xi dt \\
&= \int_{\mathbf{R}^d} u(x) \overline{\frac{1}{(2\pi)^{d/2}} \int_{\mathbf{R}^{2d}} F(\xi, t) e^{ix \cdot \xi} g(x-t) d\xi dt} dx,
\end{aligned}$$

which implies

$$\mathcal{G}_g^* F(x) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbf{R}^{2d}} F(\xi, t) e^{ix \cdot \xi} g(x-t) d\xi dt. \quad \square$$

The windowed Fourier transform is not applied in image processing on a large scale. This is due to several reasons: On the one hand, the transformation of an image yields a function in four variables. This leads to a large memory consumption and is also no longer easily visually accessible. On the other hand, the discretization of the windowed Fourier transform is not obvious, and there is no direct analogue to the Fourier series or the discrete Fourier transform. For further discussion of the windowed Fourier transform, we refer to [68].

4.4.2 The Continuous Wavelet Transform

Another transformation that analyzes local behavior is given by the wavelet transform. This transform has found broad applications in signal and image processing, in particular due to its particularly elegant discretization and its numerical efficiency.

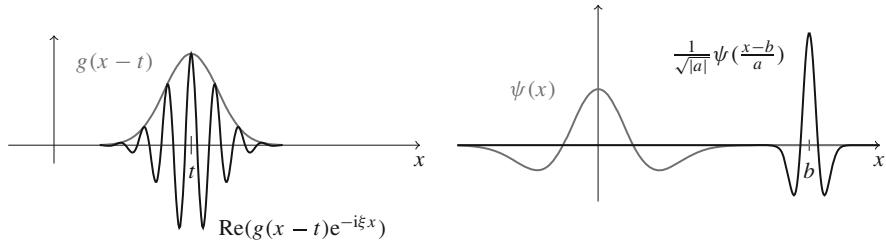


Fig. 4.11 Localization during windowed Fourier transform (left) and during wavelet transform (right)

We will follow a similar path as for the Fourier transform: we first introduce the continuous wavelet transform as well as the wavelet series, and finally, we cover the discrete wavelet transform.

While the windowed Fourier transform uses a fixed window in order to localize the function of interest, the wavelet transform uses functions of varying widths, cf. Fig. 4.11. In case of dimensions higher than one, the wavelet transform can be defined in various ways. We here cover the one-dimensional case of real-valued functions:

Definition 4.57 Let $u, \psi \in L^2(\mathbf{R}, \mathbf{R})$. For $b \in \mathbf{R}$ and $a > 0$, the wavelet *transform wavelet transform* of u with ψ is defined by

$$L_\psi u(a, b) = \int_{\mathbf{R}} u(x) \frac{1}{\sqrt{a}} \psi\left(\frac{x-b}{a}\right) dx.$$

The wavelet transform depends on a *spatial parameter* b and a *scale parameter* a . As in the case of the windowed Fourier transform, we have several representation possibilities by means of the inner product as well as the convolution:

$$\begin{aligned} L_\psi u(a, b) &= \frac{1}{\sqrt{a}} (u, T_{-b} D_{1/a} \psi)_{L^2(\mathbf{R})} \\ &= \frac{1}{\sqrt{a}} (u * D_{-1/a} \psi)(b). \end{aligned}$$

In the second representation, we see a relationship to Application 3.23 (edge detection according to Canny), since there the image of interest was convolved with different scaled convolution kernels.

Like the windowed Fourier transform, the wavelet transform exhibits a certain isometry property, however, with respect to a weighted measure. For this purpose, we introduce

$$L^2([0, \infty[\times \mathbf{R}, \frac{da db}{a^2}) = \left\{ F : [0, \infty[\times \mathbf{R} \rightarrow \mathbf{R} \mid \int_{\mathbf{R}} \int_0^\infty |F(a, b)|^2 \frac{da db}{a^2} < \infty \right\},$$

where we have omitted the usual transition to equivalence classes (cf. Sect. 2.2.2). The inner product in this space is given by

$$(F, G)_{L^2([0, \infty[\times \mathbf{R}, \frac{da db}{a^2})} = \int_{\mathbf{R}} \int_0^\infty F(a, b) G(a, b) \frac{da db}{a^2}.$$

Theorem 4.58 *Let $u, \psi \in L^2(\mathbf{R})$ with*

$$0 < c_\psi = 2\pi \int_0^\infty \frac{|\widehat{\psi}(\xi)|^2}{\xi} d\xi < \infty.$$

Then

$$L_\psi : L^2(\mathbf{R}) \rightarrow L^2([0, \infty[\times \mathbf{R}, \frac{da db}{a^2})$$

is a linear mapping, and one has

$$(L_\psi u, L_\psi v)_{L^2([0, \infty[\times \mathbf{R}, \frac{da db}{a^2})} = c_\psi(u, v)_{L^2(\mathbf{R})}.$$

Proof We use the inner product representation of the wavelet transform, the calculation rules for the Fourier transform, and the Plancherel formula (4.2) to obtain

$$\begin{aligned} L_\psi u(a, b) &= \frac{1}{\sqrt{a}}(u, T_{-b}D_{1/a}\psi)_{L^2(\mathbf{R})} \\ &= \frac{1}{\sqrt{a}}(\widehat{u}, \mathcal{F}(T_{-b}D_{1/a}\psi))_{L^2(\mathbf{R})} \\ &= \frac{1}{\sqrt{a}}(\widehat{u}, aM_{-b}D_a\widehat{\psi})_{L^2(\mathbf{R})} \\ &= \sqrt{a} \int_{\mathbf{R}} \widehat{u}(\xi) e^{ib\xi} \overline{\widehat{\psi}(a\xi)} d\xi \\ &= \sqrt{a} 2\pi \mathcal{F}^{-1}(\widehat{u} \overline{D_a \widehat{\psi}})(b). \end{aligned}$$

Now we compute the inner product of $L_\psi u$ and $L_\psi v$, again using the calculation rules for the Fourier transform and the Plancherel formula with respect to the variable b :

$$\begin{aligned} (L_\psi u, L_\psi v)_{L^2([0, \infty[\times \mathbf{R}, \frac{da db}{a^2})} &= \int_{\mathbf{R}} \int_0^\infty L_\psi u(a, b) L_\psi v(a, b) \frac{da db}{a^2} \\ &= 2\pi \int_0^\infty \int_{\mathbf{R}} a \mathcal{F}^{-1}(\widehat{u} \overline{D_a \widehat{\psi}})(b) \overline{\mathcal{F}^{-1}(\widehat{v} \overline{D_a \widehat{\psi}})(b)} db \frac{da}{a^2} \end{aligned}$$

$$\begin{aligned}
&= 2\pi \int_0^\infty \int_{\mathbf{R}} a\hat{u}(\xi) \overline{\widehat{\psi}(a\xi)} \overline{\widehat{v}(\xi)} \overline{\widehat{\psi}(a\xi)} d\xi \frac{da}{a^2} \\
&= 2\pi \int_{\mathbf{R}} \widehat{u}(\xi) \overline{\widehat{v}(\xi)} \int_0^\infty \frac{|\widehat{\psi}(a\xi)|^2}{a} da d\xi.
\end{aligned}$$

Then a change of variables and $|\widehat{\psi}(-\xi)| = |\widehat{\psi}(\xi)|$ leads to

$$\int_0^\infty \frac{|\widehat{\psi}(a\xi)|^2}{a} da = \int_0^\infty \frac{|\widehat{\psi}(a|\xi|)|^2}{a} da = \int_0^\infty \frac{|\widehat{\psi}(\omega)|^2}{\omega} d\omega = \frac{c_\psi}{2\pi}.$$

Applying the Plancherel formula again yields the assertion. \square

The condition $c_\psi < \infty$ ensures that L_ψ is a continuous mapping, while $c_\psi > 0$ guarantees the stable invertibility on the range of L_ψ .

Definition 4.59 The condition

$$0 < c_\psi = 2\pi \int_0^\infty \frac{|\widehat{\psi}(\xi)|^2}{\xi} d\xi < \infty \quad (4.3)$$

is called *admissibility condition*, and the functions ψ that satisfy it are called *wavelets*.

The admissibility condition says in particular that around zero, the Fourier transform of a wavelet must tend to zero sufficiently fast, roughly speaking, $\widehat{\psi}(0) = 0$. This implies that the average of a wavelet vanishes.

Analogously to Corollary 4.56 for the windowed Fourier transform, we derive that the wavelet transform is inverted on its range by its adjoint (up to a constant).

Corollary 4.60 Let $u, \psi \in L^2(\mathbf{R})$ and $c_\psi = 1$. Then,

$$u(x) = \int_{\mathbf{R}} \int_0^\infty L_\psi u(a, b) \frac{1}{\sqrt{a}} \psi\left(\frac{x-b}{a}\right) \frac{da db}{a^2}.$$

Proof Due to the normalization, we have only to compute the adjoint of the wavelet transform. For $u \in L^2(\mathbf{R})$ and $F \in L^2([0, \infty[\times \mathbf{R}, \frac{da db}{a^2})$,

$$\begin{aligned}
(L_\psi u, F)_{L^2([0, \infty[\times \mathbf{R}, \frac{da db}{a^2})} &= \int_{\mathbf{R}} \int_0^\infty \int_{\mathbf{R}} u(x) \frac{1}{\sqrt{a}} \psi\left(\frac{x-b}{a}\right) dx F(a, b) \frac{da db}{a^2} \\
&= \int_{\mathbf{R}} u(x) \int_{\mathbf{R}} \int_0^\infty F(a, b) \frac{1}{\sqrt{a}} \psi\left(\frac{x-b}{a}\right) \frac{da db}{a^2} dx.
\end{aligned}$$

This implies

$$L_\psi^* F(x) = \int_{\mathbf{R}} \int_0^\infty F(a, b) \frac{1}{\sqrt{a}} \psi\left(\frac{x-b}{a}\right) \frac{da db}{a^2},$$

which yields the assertion. \square

Example 4.61 (Continuous Wavelets) For a continuous wavelet transform, the following wavelets are employed, for instance:

Derivatives of the Gaussian function: For $G(x) = e^{-x^2/2}$, we set $\psi(x) = -\frac{d}{dx}G(x) = xe^{-x^2/2}$. This function is a wavelet, since $\widehat{\psi}(\xi) = i\xi e^{-\xi^2/2}$, which implies

$$c_\psi = 2\pi \int_0^\infty \frac{|\widehat{\psi}(\xi)|^2}{\xi} d\xi = \pi.$$

This wavelet and its Fourier transform look as follows:

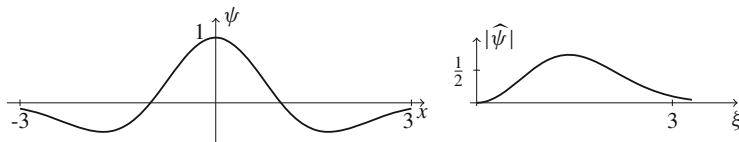


We can express the wavelet transform by means of the convolution and obtain in this case

$$\begin{aligned} L_\psi u(a, b) &= \frac{1}{\sqrt{a}} (u * D_{-1/a}\psi)(b) \\ &= \frac{1}{\sqrt{a}} (u * D_{-1/a}G')(b) \\ &= \sqrt{a} (u * (D_{-1/a}G'))(b) \\ &= \sqrt{a} \frac{d}{db} (u * (D_{-1/a}G))(b). \end{aligned}$$

Here we can recognize an analogy to edge detection according to Canny in Application 3.23, whereby the image was convolved with scaled Gaussian functions as well.

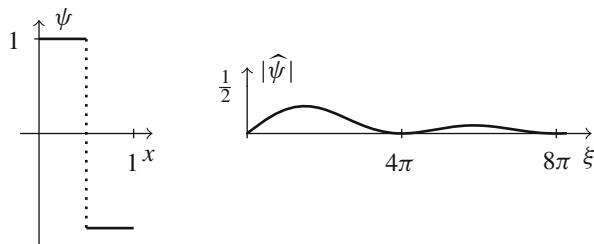
A similar construction leads to the so-called Mexican hat function: $\psi(x) = -\frac{d^2}{dx^2}G(x) = (1-x^2)e^{-x^2/2}$. Here we have $\widehat{\psi}(\xi) = \xi^2 e^{-\xi^2/2}$ and $c_\psi = \sqrt{\pi/2}$. The Mexican hat function is named after its shape:



Haar wavelet: A different kind of wavelet is given by the Haar wavelet, defined by

$$\psi(x) = \begin{cases} 1 & \text{if } 0 \leq x < \frac{1}{2}, \\ -1 & \text{if } \frac{1}{2} \leq x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

It is discontinuous, but it exhibits compact support:



The wavelets that result from derivatives of the Gaussian function are very smooth (infinitely differentiable) and decay rapidly. In particular, they (as well as their Fourier transforms) are well localized. For a discrete implementation, compact support would furthermore be desirable, since then, the integrals that need to be computed would be finite. As stated above, the Haar wavelet exhibits compact support, but it is discontinuous. In the following subsection, we will see that it is particularly well suited for the discrete wavelet transform. In this context, we will also encounter further wavelets.

4.4.3 The Discrete Wavelet Transform

The continuous wavelet transform is a redundant representation. The question arises whether it would be sufficient to know the wavelet transform of a function on a subset of $[0, \infty[\times \mathbf{R}$. This is actually the case for certain discrete subsets— independent of the function and the wavelet. This will enable us to derive an analogue to the Fourier series. It will turn out that under certain conditions, the functions

$$\{\psi_{j,k}(x) = 2^{-j/2}\psi(2^{-j}x - k) \mid j, k \in \mathbf{Z}\}$$

form an orthonormal basis of $L^2(\mathbf{R})$. To comprehend this will require quite a bit of work. We first introduce the central notion for wavelet series and the discrete wavelet transform, namely the notion of a “multiscale analysis.”

Definition 4.62 (Multiscale Analysis) A sequence $(V_j)_{j \in \mathbf{Z}}$ of closed subspaces of $L^2(\mathbf{R})$ is called a *multiscale analysis* if it satisfies the following conditions:

Translation invariance: For all $j, k \in \mathbf{Z}$,

$$u \in V_j \iff T_{2^j k} u \in V_j.$$

Inclusion: For all $j \in \mathbf{Z}$, we have

$$V_{j+1} \subset V_j.$$

Scaling: For all $j \in \mathbf{Z}$,

$$u \in V_j \iff D_{1/2} u \in V_{j+1}.$$

Trivial intersection:

$$\bigcap_{j \in \mathbf{Z}} V_j = \{0\}.$$

Completeness:

$$\overline{\bigcup_{j \in \mathbf{Z}} V_j} = L^2(\mathbf{R}).$$

Orthonormal basis: There is a function $\phi \in V_0$ such that the functions $\{T_k \phi \mid k \in \mathbf{Z}\}$ form an orthonormal basis of V_0 .

The function ϕ is called a *generator* or *scaling function* of the multiscale analysis.

Let us make some remarks regarding this definition: The spaces V_j are translation invariant with respect to the *dyadic translations* by 2^j . Furthermore, they are nested into each other and become smaller with increasing j . If we denote the orthogonal projection onto V_j by P_{V_j} , then we have for every u that

$$\lim_{j \rightarrow \infty} P_{V_j} u = 0 \quad \text{and} \quad \lim_{j \rightarrow -\infty} P_{V_j} u = u.$$

Remark 4.63 In most cases, it is required in the definition of a multiscale analysis only that the functions $(T_k \phi)$ form a Riesz basis, i.e., that its linear span is dense in the space V_0 and that there exist $0 < A \leq B$ such that for every $u \in V_0$, one has

$$A \|u\|^2 \leq \sum_{k \in \mathbf{Z}} |(u, T_k \phi)|^2 \leq B \|u\|^2.$$

We do not cover this construction here and refer to [94, 97].

Next, we present the standard example for a multiscale analysis:

Example 4.64 (Piecewise Constant Multiscale Analysis) Let V_j be the set of functions $u \in L^2(\mathbf{R})$ that are constant on the dyadic intervals $[k2^j, (k+1)2^j[$, $k \in \mathbf{Z}$. Translation invariance, inclusion, and scaling are obvious. The trivial intersection of the V_j is implied by the circumstance that the zero function is the only constant L^2 -function. Completeness follows from the fact that every L^2 -function can be approximated by piecewise constant functions. As a generator of this multiscale analysis, we can choose $\phi = \chi_{[0,1]}$, for instance.

The scaling property implies that the functions $\phi_{j,k}(x) = 2^{-j/2}\phi(2^{-j}x - k)$, $k \in \mathbf{Z}$, form an orthonormal basis of V_j (Exercise 4.14). In the sense of the previous example, we can say, roughly speaking, that $P_{V_j}u$ is the representation of u “on the scale V_j ” and contains details of u “up to size 2^j ”; cf. Fig. 4.12.

Due to the scaling property, ϕ does not lie only in V_0 , but also in V_{-1} . Since the functions $\phi_{-1,k}$ form an orthonormal basis of V_{-1} , one has

$$\phi(x) = \sum_{k \in \mathbf{Z}} h_k \sqrt{2} \phi(2x - k) \quad (4.4)$$

with $h_k = (\phi, \phi_{-1,k})$. Equation (4.4) is called the *scaling equation*, and it explains the name *scaling function* for ϕ . The functions $\phi_{j,k}$ already remind us of the continuous wavelet transform with discrete values $a = 2^j$ and $b = 2^j k$. We find the wavelets in the following construction again:

Definition 4.65 (Approximation and Detail Spaces) Let $(V_j)_{j \in \mathbf{Z}}$ be a multiscale analysis. Let the spaces W_j be defined as orthogonal complements of V_j in V_{j-1} , i.e.,

$$V_{j-1} = V_j \oplus W_j, \quad V_j \perp W_j.$$

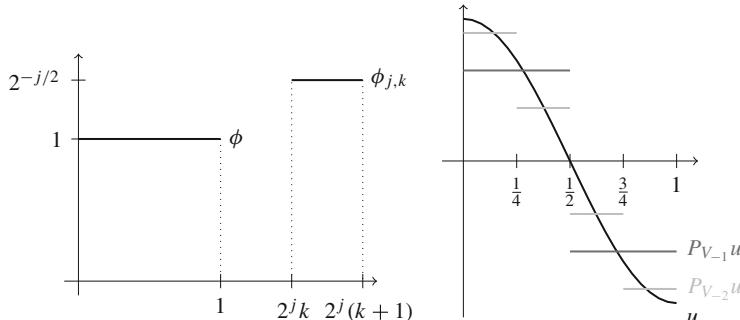


Fig. 4.12 Representation of a function in the piecewise constant multiscale analysis. Left: Generator function ϕ and $\phi_{j,k}$ for $j = -1$, $k = 3$. Right: The function $u(x) = \cos(\pi x)$ and its representation in the spaces V_{-1} and V_{-2}

The space V_j is called the *approximation space to the scale j*; the space W_j is called the *detail space* or *wavelet space to the scale j*.

The definition of the spaces W_j immediately implies

$$V_j = \bigoplus_{m \geq j+1} W_m$$

and due to the completeness of V_j , there also follows

$$L^2(\mathbf{R}) = \bigoplus_{m \in \mathbf{Z}} W_m.$$

Furthermore, we have $P_{V_{j-1}} = P_{V_j} + P_{W_j}$ and hence

$$P_{W_j} = P_{V_{j-1}} - P_{V_j}.$$

We can now represent every $u \in L^2(\mathbf{R})$ by means of the spaces V_j and W_j in different ways:

$$u = \sum_{j \in \mathbf{Z}} P_{W_j} u = P_{V_m} u + \sum_{j \leq m} P_{W_j} u.$$

These equations justify the name “multiscale analysis”: the spaces V_j allow a systematic approximation of functions on different scales.

Example 4.66 (Detail Spaces W_j for Piecewise Constant Multiscale Analysis) Let us investigate what the spaces W_j look like in Example 4.64. We construct the spaces by means of the projection P_{V_j} . For $x \in [k2^j, (k+1)2^j[$, we have

$$P_{V_j} u(x) = 2^{-m} \int_{k2^j}^{(k+1)2^j} u(x) dx.$$

In particular, this implies

$$P_{V_j} u = \sum_{k \in \mathbf{Z}} (u, \phi_{j,k}) \phi_{j,k}.$$

In order to obtain $P_{W_{j+1}} = P_{V_j} - P_{V_{j+1}}$, we use the scaling equation for ϕ . In this case, we have

$$\phi(x) = \chi_{[0,1]}(x) = \phi(2x) + \phi(2x-1)$$

or expressed differently, $\phi_{0,0} = (\phi_{-1,0} + \phi_{-1,1})/\sqrt{2}$. Slightly more generally, we also have $\phi_{j+1,k} = (\phi_{j,2k} + \phi_{j,2k+1})/\sqrt{2}$. By splitting up even and odd indices and

using the properties of the inner product, we obtain

$$\begin{aligned}
P_{W_{j+1}} u &= P_{V_j} u - P_{V_{j+1}} u \\
&= \sum_{k \in \mathbf{Z}} (u, \phi_{j,k}) \phi_{j,k} - \sum_{k \in \mathbf{Z}} (u, \phi_{j+1,k}) \phi_{j+1,k} \\
&= \sum_{k \in \mathbf{Z}} (u, \phi_{j,2k}) \phi_{j,2k} + \sum_{k \in \mathbf{Z}} (u, \phi_{j,2k+1}) \phi_{j,2k+1} \\
&\quad - \frac{1}{2} \sum_{k \in \mathbf{Z}} (u, \phi_{j,2k} + \phi_{j,2k+1})(\phi_{j,2k} + \phi_{j,2k+1}) \\
&= \frac{1}{2} \sum_{k \in \mathbf{Z}} (u, \phi_{j,2k} - \phi_{j,2k+1})(\phi_{j,2k} - \phi_{j,2k+1}).
\end{aligned}$$

The projection onto W_{j+1} hence corresponds to the expansion with respect to the functions $(\phi_{j,2k} - \phi_{j,2k+1})/\sqrt{2}$. This can be simplified by means of the following notation:

$$\psi(x) = \phi(2x) - \phi(2x - 1), \quad \psi_{j,k}(x) = 2^{-j/2} \psi(2^{-j}x - k).$$

This implies $\psi_{j+1,k} = (\phi_{j,2k} - \phi_{j,2k+1})/\sqrt{2}$, and we have

$$P_{W_j} u = \sum_{k \in \mathbf{Z}} (u, \psi_{j,k}) \psi_{j,k};$$

cf. Fig. 4.13. Therefore, also the spaces W_j have orthonormal bases, namely $(\psi_{j,k})_{k \in \mathbf{Z}}$. The function ψ is just the Haar wavelet in Example 4.61 again.

The above example shows that for piecewise constant multiscale analysis, there is actually a wavelet (the Haar wavelet) that yields an orthonormal basis of the wavelet spaces W_j . A similar construction also works in general, as the following theorem demonstrates:

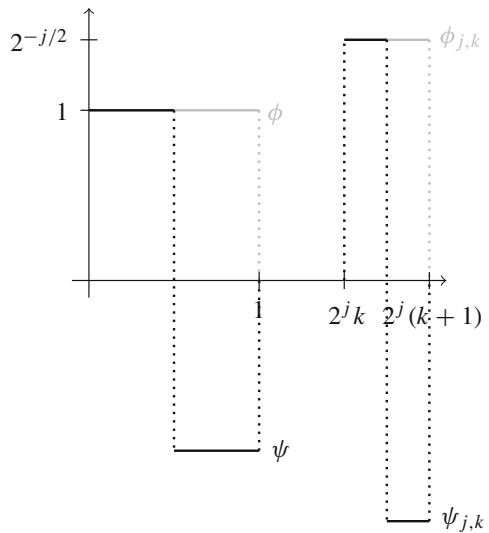
Theorem 4.67 *Let (V_j) be a multiscale analysis with generator ϕ such that ϕ satisfies the scaling equation (4.4) with a sequence (h_k) . Furthermore, let $\psi \in V_{-1}$ be defined by*

$$\psi(x) = \sqrt{2} \sum_{k \in \mathbf{Z}} (-1)^k h_{1-k} \phi(2x - k).$$

Then:

1. *The set $\{\psi_{j,k} \mid k \in \mathbf{Z}\}$ is an orthonormal basis of W_j .*
2. *The set $\{\psi_{j,k} \mid j, k \in \mathbf{Z}\}$ is an orthonormal basis of $L^2(\mathbf{R})$.*
3. *The function ψ is a wavelet with $c_\psi = 2 \ln 2$.*

Fig. 4.13 Scaling function and wavelets for piecewise constant multiscale analysis



Proof We first verify that for all $k \in \mathbf{Z}$, we have

$$(\psi, \phi_{k,0}) = 0,$$

$$(\psi, \psi_{k,0}) = \delta_{0,k}$$

(Exercise 4.16). The first equation implies $\psi \perp V_0$, and hence we have $\psi \in W_0$. The second equation yields the orthonormality of the translates of ψ .

We now demonstrate that the system $\{\psi_{k,0} \mid k \in \mathbf{Z}\}$ is complete in W_0 . Due to $V_{-1} = V_0 \oplus W_0$, it is equivalent to show that the system $\{\phi_{k,0}, \psi_{k,0} \mid k \in \mathbf{Z}\}$ is complete in V_{-1} . We derive the latter by showing that $\phi_{-1,0}$ can be represented by $\{\phi_{k,0}, \psi_{k,0} \mid k \in \mathbf{Z}\}$. For this purpose, we calculate by means of the scaling equation (4.4) and the definition of ψ :

$$\begin{aligned} & \sum_{k \in \mathbf{Z}} |(\phi_{-1,0}, \psi_{k,0})|^2 + |(\phi_{-1,0}, \phi_{k,0})|^2 \\ &= \sum_{k \in \mathbf{Z}} \left| \sum_{l \in \mathbf{Z}} h_l \underbrace{(\phi_{-1,0}, \phi_{-1,l+2k})}_{=\delta_{0,l+2k}} \right|^2 + \left| \sum_{l \in \mathbf{Z}} (-1)^l h_{1-l} \underbrace{(\phi_{-1,0}, \phi_{-1,l+2k})}_{=\delta_{0,l+2k}} \right|^2 \\ &= \sum_{k \in \mathbf{Z}} h_{-2k}^2 + h_{1+2k}^2 = \sum_{k \in \mathbf{Z}} h_k^2. \end{aligned}$$

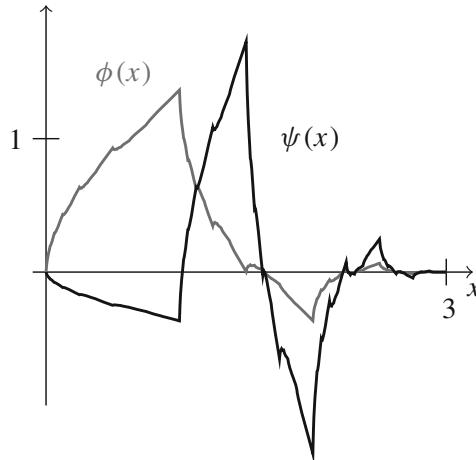
In Exercise 4.16, you will prove that in particular, $\sum_{k \in \mathbf{Z}} h_k^2 = 1$, and due to $\|\phi_{-1,0}\| = 1$, it follows that the system $\{\phi_{k,0}, \psi_{k,0} \mid k \in \mathbf{Z}\}$ is complete in V_{-1} . The assertions 1 and 2 can now be shown by simple arguments (cf. Exercise 4.14). For assertion 3, we refer to [94]. \square

In the context of Theorem 4.67, we also note that the set $\{\psi_{j,k} \mid j, k \in \mathbf{Z}\}$ is an orthonormal wavelet basis of $L^2(\mathbf{R})$.

Given a sequence of subspaces (V_j) that satisfies all the further requirements for a multiscale analysis, it is in general not easy to come up with an orthonormal basis for V_0 (another example is given in Example 4.15). However, it can be shown that under the assumption in Remark 4.63 (the translates of ϕ form a Riesz basis of V_0), a generator can be constructed whose translates form an orthonormal basis of V_0 , cf. [94], for instance. Theorem 4.67 expresses what the corresponding wavelet looks like. By now, a large variety of multiscale analyses and wavelets with different properties have been constructed. Here, we only give some examples:

Example 4.68 (Daubechies Wavelets and Symlets) We consider two important examples of multiscale analysis:

Daubechies Wavelets: The Daubechies wavelets (named after Ingrid Daubechies, cf. [48]) are wavelets with compact support, a certain degree of smoothness and a certain number of vanishing moments, i.e., for $l = 0, \dots, k$ up to a certain k , the integrals $\int_{\mathbf{R}} x^l \psi(x) dx$ are all zero. There is a whole scale of these wavelets, and for a given support, these wavelets are those that exhibit the most vanishing moments, i.e., k is maximal. The so-called db2-wavelet ψ (featuring two vanishing moments) and the corresponding scaling function ϕ look as follows:

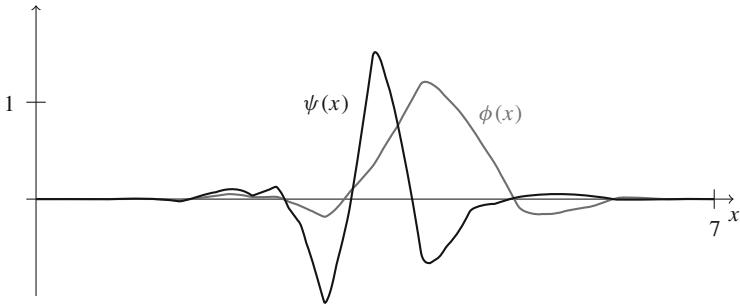


For the db2-wavelet, the analytical coefficients of the scaling equation are given by

$$h_0 = \frac{1 - \sqrt{3}}{8\sqrt{2}}, \quad h_1 = \frac{3 - \sqrt{3}}{8\sqrt{2}}, \quad h_2 = \frac{3 + \sqrt{3}}{8\sqrt{2}}, \quad h_3 = \frac{1 + \sqrt{3}}{8\sqrt{2}}.$$

For the other Daubechies wavelets, the values are tabulated in [48], for instance.

Symlets: Symlets also trace back to Ingrid Daubechies. They are similar to Daubechies wavelets, but more “symmetric.” Also in this case, there is a scale of symlets. The coefficients of the scaling equation are tabulated but not available in analytical form. The so-called `sym4`-wavelet (exhibiting four vanishing moments) and the corresponding scaling function look as follows:



4.4.4 Fast Wavelet Transforms

The scaling equation (4.4) and the definition of the wavelet in Theorem 4.67 are the key to a fast wavelet transform. We recall that since $\{\psi_{j,k} \mid j, k \in \mathbf{Z}\}$ forms an orthonormal basis of $L^2(\mathbf{R})$, one has

$$u = \sum_{j,k \in \mathbf{Z}} (u, \psi_{j,k}) \psi_{j,k}.$$

In other words: the wavelet decomposition of a signal $u \in L^2(\mathbf{R})$ consists in computing the inner products $(u, \psi_{j,k})$. The following lemma shows that these products can be calculated recursively:

Lemma 4.69 *Let ϕ be a generator of a multiscale analysis and ψ the corresponding wavelet of Theorem 4.67. Then the following equalities hold:*

$$\begin{aligned} \phi_{j,k} &= \sum_{l \in \mathbf{Z}} h_l \phi_{j-1, l+2k}, \\ \psi_{j,k} &= \sum_{l \in \mathbf{Z}} (-1)^l h_{1-l} \phi_{j-1, l+2k}. \end{aligned}$$

Furthermore, there is an analogous relationship for the inner products:

$$(u, \phi_{j,k}) = \sum_{l \in \mathbf{Z}} h_l (u, \phi_{j-1,l+2k}), \quad (4.5)$$

$$(u, \psi_{j,k}) = \sum_{l \in \mathbf{Z}} (-1)^l h_{1-l} (u, \phi_{j-1,l+2k}). \quad (4.6)$$

Proof Using the scaling equation (4.4) for ϕ , we obtain

$$\begin{aligned} \phi_{j,k}(x) &= 2^{-j/2} \phi(2^{-j}x - k) \\ &= \sum_l h_l \sqrt{2} 2^{-j/2} \phi(2(2^{-j}x - k) - l) \\ &= \sum_l h_l \phi_{j-1,l+2k}(x). \end{aligned}$$

The equation for $\psi_{j,k}$ can be shown analogously, and the equations for the inner products are an immediate consequence. \square

Starting from the values $(u, \phi_{0,k})$, we now have recurrence formulas for the coefficients on coarser scales $j > 0$. By means of the abbreviations

$$c^j = ((u, \phi_{j,k}))_{k \in \mathbf{Z}} \quad \text{and} \quad d^j = ((u, \psi_{j,k}))_{k \in \mathbf{Z}},$$

the recurrence formulas (4.5) and (4.6) take the form

$$c_k^j = \sum_{l \in \mathbf{Z}} h_l c_{2k+l}^{j-1} \quad \text{and} \quad d_k^j = \sum_{l \in \mathbf{Z}} (-1)^l h_{1-l} c_{2k+l}^{j-1}.$$

Based on the projection $P_{V_j} u$, the fast wavelet transform computes the coarser projection $P_{V_{j+1}} u$ in the approximation space V_{j+1} and the wavelet component $P_{W_{j+1}} u$ in the detail space W_{j+1} . Note that in the case of a finite coefficient sequence h , the summation processes are finite. In the case of short coefficient sequences, only few calculations are necessary in each recursive step.

For the reconstruction, the projection $P_{V_j} u$ is computed based on the coarser approximation $P_{V_{j+1}} u$ and the details $P_{W_{j+1}} u$, as the following lemma describes:

Lemma 4.70 *For the coefficient sequences $d^j = ((u, \psi_{j,k}))_{k \in \mathbf{Z}}$ and $c^j = ((u, \phi_{j,k}))_{k \in \mathbf{Z}}$, we have the following recurrence formula:*

$$c_k^j = \sum_{l \in \mathbf{Z}} c_l^{j+1} h_{k-2l} + \sum_{l \in \mathbf{Z}} d_l^{j+1} (-1)^{k-2l} h_{1-(k-2l)}.$$

Proof Since the space V_j is orthogonally decomposed into the spaces V_{j+1} and W_{j+1} , one has $P_{V_j} u = P_{V_{j+1}} u + P_{W_{j+1}} u$. Expressing the projections by means of

the orthonormal bases, we obtain

$$\begin{aligned}
\sum_{k \in \mathbf{Z}} c_k^j \phi_{j,k} &= P_{V_j} u \\
&= P_{V_{j+1}} u + P_{W_{j+1}} u \\
&= \sum_{l \in \mathbf{Z}} c_l^{j+1} \phi_{j+1,l} + \sum_{l \in \mathbf{Z}} d_l^{j+1} \psi_{j+1,l} \\
&= \sum_{l \in \mathbf{Z}} c_l^{j+1} \sum_{n \in \mathbf{Z}} h_n \phi_{j,n+2l} + \sum_{l \in \mathbf{Z}} d_l^{j+1} \sum_{n \in \mathbf{Z}} (-1)^n h_{1-n} \phi_{j,n+2l} \\
&= \sum_{l \in \mathbf{Z}} c_l^{j+1} \sum_{k \in \mathbf{Z}} h_{k-2l} \phi_{j,k} + \sum_{l \in \mathbf{Z}} d_l^{j+1} \sum_{k \in \mathbf{Z}} (-1)^{k-2l} h_{1-(k-2l)} \phi_{j,k}.
\end{aligned}$$

Swapping the sums and comparing the coefficients yields the assertion. \square

In order to denote the decomposition and the reconstruction in a concise way, we introduce the following operators:

$$\begin{aligned}
H : \ell^2(\mathbf{Z}) &\rightarrow \ell^2(\mathbf{Z}), \quad (Hc)_k = \sum_{l \in \mathbf{Z}} h_l c_{2k+l}, \\
G : \ell^2(\mathbf{Z}) &\rightarrow \ell^2(\mathbf{Z}), \quad (Gc)_k = \sum_{l \in \mathbf{Z}} (-1)^l h_{1-l} c_{2k+l}.
\end{aligned}$$

By means of this notation, the fast wavelet transform reads:

Input: $c^0 \in \ell^2(\mathbf{Z})$ and decomposition depth M .

for $m = 1$ to M **do**

Compute $d^m = Gc^{m-1}$ and $c^m = Hc^{m-1}$.

end for

Output: d^1, \dots, d^M and c^M .

Remark 4.71 (Decomposition as Convolution with Downsampling) A decomposition step of the wavelet transform exhibits a structure that is similar to a convolution. A change of indices yields

$$(Hc)_k = \sum_{n \in \mathbf{Z}} h_{n-2k} c_n = (c * D_{-1}h)_{2k}.$$

Using the abbreviation $g_l = (-1)^l h_{1-l}$, we analogously obtain

$$(Gc)_k = \sum_{n \in \mathbf{Z}} g_{n-2k} c_n = (c * D_{-1}g)_{2k}.$$

Hence, a decomposition step consists of a convolution with the mirrored filters as well as a downsampling.

The fast wavelet reconstruction can be expressed by means of the adjoint operators of H and G :

$$H^* : \ell^2(\mathbf{Z}) \rightarrow \ell^2(\mathbf{Z}), \quad (H^*c)_k = \sum_{l \in \mathbf{Z}} h_{k-2l} c_l,$$

$$G^* : \ell^2(\mathbf{Z}) \rightarrow \ell^2(\mathbf{Z}), \quad (G^*c)_k = \sum_{l \in \mathbf{Z}} (-1)^{k-2l} h_{1-(k-2l)} c_l$$

and thereby, the algorithm reads

Input: Decomposition depth M , d^1, \dots, d^M and c^M .
for $m = M$ to 1 **do**
 Compute $c^{m-1} = H^*c^m + G^*d^m$.
end for
Output: $c^0 \in \ell^2(\mathbf{Z})$.

Remark 4.72 (Reconstruction as Upsampling with Convolution) Also the wavelet reconstruction exhibits the structure of a convolution. However, we need here an upsampling. For a sequence c , we define

$$(\tilde{c})_n = \begin{cases} c_l & \text{if } n = 2l, \\ 0 & \text{if } n = 2l + 1. \end{cases}$$

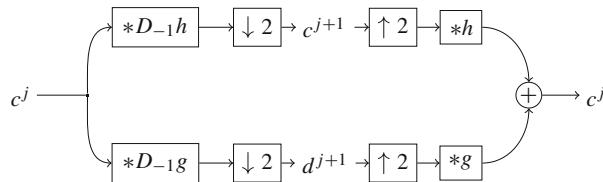
Thereby, we obtain

$$(H^*c)_k = \sum_{l \in \mathbf{Z}} h_{k-2l} c_l = \sum_{n \in \mathbf{Z}} h_{k-n} \tilde{c}_n = (\tilde{c} * h)_k.$$

By means of the abbreviation $g_l = (-1)^l h_{1-l}$, one has analogously

$$(G^*d)_k = (\tilde{d} * g)_k.$$

Schematically, decomposition and reconstruction are often depicted as a so-called *filter bank* as follows:



Here, $\downarrow 2$ refers to downsampling with the factor 2, i.e., omitting every second value. Analogously, $\uparrow 2$ denotes upsampling with the factor 2, i.e., the extension of the vector by filling in zeros at every other place.

It remains to discuss how to obtain an initial sequence c^J . Let us assume that the signal of interest u lies in a certain approximation space V_J , i.e.,

$$u = \sum_{k \in \mathbf{Z}} c_k^J \phi_{J,k},$$

where the coefficients are given by

$$c_k^J = 2^{J/2} \int_{\mathbf{R}} u(x) 2^{-J} \phi(2^{-J}x - k) dx.$$

Since $\int_{\mathbf{R}} 2^{-J} \phi(2^{-J}x - k) dx = 1$, we find that $2^{-J/2} c_k^J$ is a weighted average of u in a neighborhood of $2^J k$ whose size is proportional to 2^{-J} . For u regular (continuous, for instance), we hence have

$$c_k^J \approx 2^{J/2} u(2^J k).$$

If u is available at sampling points kT , it is thus reasonable to interpret the sampled values $u(kT)$ with $2^J = T$ as

$$c_k^J = \sqrt{T} u(Tk).$$

An example of a one-dimensional wavelet decomposition is presented in Fig. 4.14.

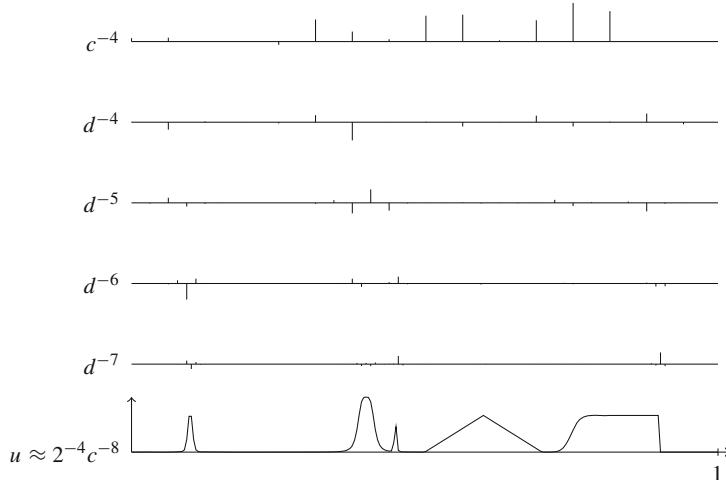


Fig. 4.14 One-dimensional wavelet transform of a signal with the wavelet `sym4` (cf. Example 4.68). Bottom row: Signal of interest $u : [0, 1] \rightarrow \mathbf{R}$, sampled with $T = 2^{-8}$ corresponding to approximately $2^{-4} c^{-8}$. Upper graphs: Wavelet and approximation coefficients, respectively. Note that the jumps and singularities of the signal evoke large coefficients on the fine scales. In contrast, the smooth parts of the signal can be represented nearly exclusively by the approximation coefficients

For a finite signal u , the sampling results in a finite sequence. Since the discrete wavelet transform consists of convolutions, again the problem arises that the sequences have to be evaluated at undefined points. Also in this case, boundary extension strategies are of help. The simplest method is given by the periodic boundary extension: the convolutions are replaced by periodic convolutions. This corresponds to the periodization of the wavelet functions $\psi_{j,k}$. A drawback of this method is the fact that the periodization typically induces a jump discontinuity that leads to unnaturally large wavelet coefficients at the boundary. Other methods such as symmetric extension or zero extension are somewhat more complex to implement; cf. [97], for instance.

The numerical cost of a decomposition or reconstruction step is proportional to the length of the filter h and the length of the signal. For a finite sequence c^0 , the length of the sequence is cut by half due to the downsampling in every decomposition step (up to boundary extension effects). Since the boundary extension effects are of the magnitude of the filter length, the total complexity of the decomposition of a signal of length $N = 2^M$ into M levels with a filter h of length n is given by $\mathcal{O}(nN)$. The same complexity true for the reconstruction. For short filters h , this cost is even lower than for the fast Fourier transform.

4.4.5 The Two-Dimensional Discrete Wavelet Transform

Based on an orthonormal wavelet basis $\{\psi_{j,k} \mid j, k \in \mathbf{Z}\}$ of $L^2(\mathbf{R})$, we can construct an orthonormal basis of $L^2(\mathbf{R}^2)$ by collecting all tensor products: The functions

$$(x_1, x_2) \mapsto \psi_{j_1, k_1}(x_1)\psi_{j_2, k_2}(x_2), \quad j_1, j_2, k_1, k_2 \in \mathbf{Z}$$

form an orthonormal basis of $L^2(\mathbf{R}^2)$.

In the same way, we can constitute a multiscale analysis of $L^2(\mathbf{R}^2)$: for a multiscale analysis (V_j) of $L^2(\mathbf{R})$, we set up the spaces

$$V_j^2 = V_j \otimes V_j \subset L^2(\mathbf{R}^2)$$

that are defined by the fact that the functions

$$\Phi_{j,k} : (x_1, x_2) \mapsto \phi_{j,k_1}(x_1)\phi_{j,k_2}(x_2), \quad k = (k_1, k_2) \in \mathbf{Z}^2$$

form an orthonormal basis of V_j^2 . This construction is also called a *tensor product* of separable Hilbert spaces; cf. [141], for instance.

In the two-dimensional case, the wavelet spaces, i.e., the orthogonal complements of V_j^2 in V_{j-1}^2 , exhibit a little more structure. We define the wavelet space W_j^2 by

$$V_{j-1}^2 = V_j^2 \oplus W_j^2$$

(where the superscripted number two in one case denotes a tensor product and in the other just represents a name). On the other hand, $V_{j-1} = V_j \oplus W_j$, and we obtain

$$\begin{aligned} V_{j-1}^2 &= (V_j \oplus W_j) \otimes (V_j \oplus W_j) \\ &= (V_j \otimes V_j) \oplus (V_j \otimes W_j) \oplus (W_j \otimes V_j) \oplus (W_j \otimes W_j), \end{aligned}$$

which by means of the notation

$$H_j^2 = V_j \otimes W_j, \quad S_j^2 = W_j \otimes V_j, \quad D_j^2 = W_j \otimes W_j$$

can be expressed as

$$V_{j-1}^2 = V_j^2 \oplus H_j^2 \oplus S_j^2 \oplus D_j^2.$$

Denoting the scaling function of (V_j) by ϕ and the corresponding wavelet by ψ , we define three functions:

$$\psi^1(x_1, x_2) = \phi(x_1)\psi(x_2), \quad \psi^2(x_1, x_2) = \psi(x_1)\phi(x_2), \quad \psi^3(x_1, x_2) = \psi(x_1)\psi(x_2).$$

For $m \in \{1, 2, 3\}$, $j \in \mathbf{Z}$, and $k \in \mathbf{Z}^2$, we set

$$\psi_{j,k}^m(x_1, x_2) = 2^{-j}\psi^m(2^{-j}x_1 - k_1, 2^{-j}x_2 - k_2).$$

Now one can prove that the functions $\{\psi_{j,k}^1 \mid k \in \mathbf{Z}^2\}$ form an orthonormal basis of H_j^2 , the functions $\{\psi_{j,k}^2 \mid k \in \mathbf{Z}^2\}$ constitute an orthonormal basis of S_j^2 , and the functions $\{\psi_{j,k}^3 \mid k \in \mathbf{Z}^2\}$ are an orthonormal basis of D_j^2 . Then

$$\{\psi_{j,k}^m \mid m = 1, 2, 3, k \in \mathbf{Z}^2, j \in \mathbf{Z}\}$$

naturally forms an orthonormal basis of $L^2(\mathbf{R}^2)$.

We observe that the wavelet spaces W_j^2 are generated by the three wavelets ψ^1 , ψ^2 , and ψ^3 together with their scalings and translations. The spaces H_j^2 contain the *horizontal* details on the scale j (i.e., the details in the x_1 -direction), the spaces S_j^2 the *vertical* details (in the x_2 -direction) and the spaces D_j^2 the *diagonal* details; cf. Fig. 4.15.

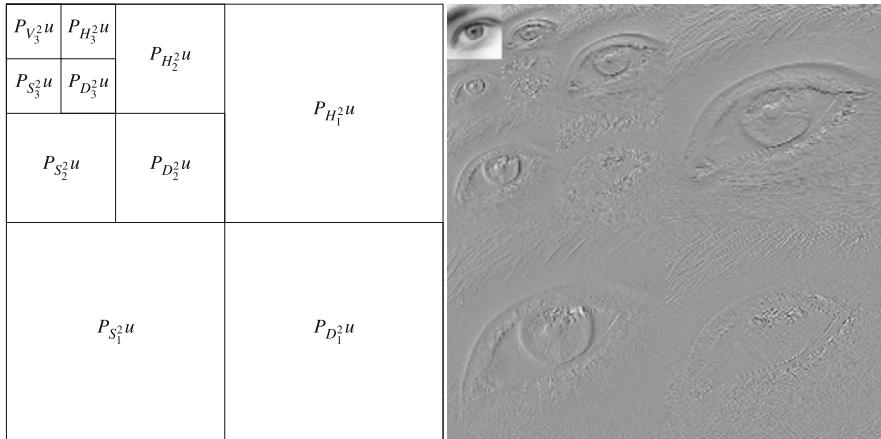
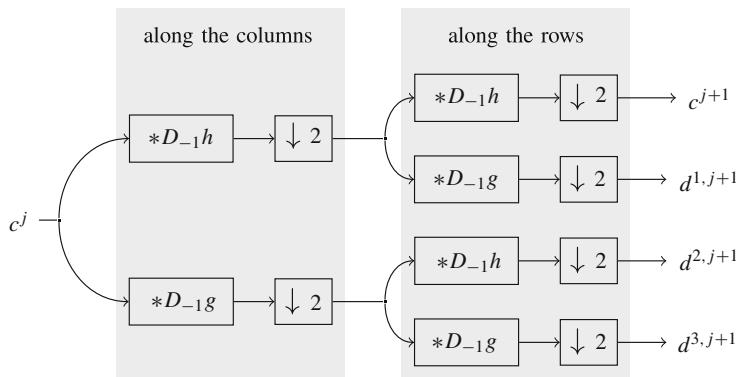


Fig. 4.15 Two-dimensional wavelet transform of an image by means of the Haar wavelet. The image itself is interpreted as the finest wavelet representation in the space V_0 . Based on this, the components in the coarser approximation and detail spaces are computed

This kind of two-dimensional multiscale analysis is rather obvious and is in particular simple to implement algorithmically: based on the approximation coefficients c^j , we compute the approximation coefficients c^{j+1} on a coarser scale as well as the three detail coefficients $d^{1,j+1}$, $d^{2,j+1}$, and $d^{3,j+1}$ (belonging to the spaces H^{j+1} , S^{j+1} , and D^{j+1} , respectively). In practice, this can be accomplished by the concatenation of the one-dimensional wavelet decomposition along the columns as well as the rows. Illustrated as a filter bank, this looks as follows:



A reconstruction step is performed analogously according to the scheme of the one-dimensional case. A drawback of the tensor product approach is the poor resolution of directions. Essentially, only vertical, horizontal, and diagonal structures can be recognized. Other procedures are possible and are described in [94], for instance.

Application 4.73 (JPEG2000) The DCT being a foundation of the JPEG standard, the discrete wavelet transform constitutes a foundation of the JPEG2000 standard. Apart from numerous further differences between JPEG and JPEG2000, using the wavelet transform instead of the blockwise DCT is the most profound distinction. This procedure has several advantages:

- Higher compression rates, but preserving the same subjective visual quality.
- More “pleasant” artifacts.
- Stepwise image buildup through stepwise decoding of the scales (of advantage in transferring with a low data rate).

Figure 4.16 shows the compression potential of the discrete wavelet transform (compare to the DCT case in Fig. 4.10). For a detailed description of JPEG2000, we refer to [109] again.

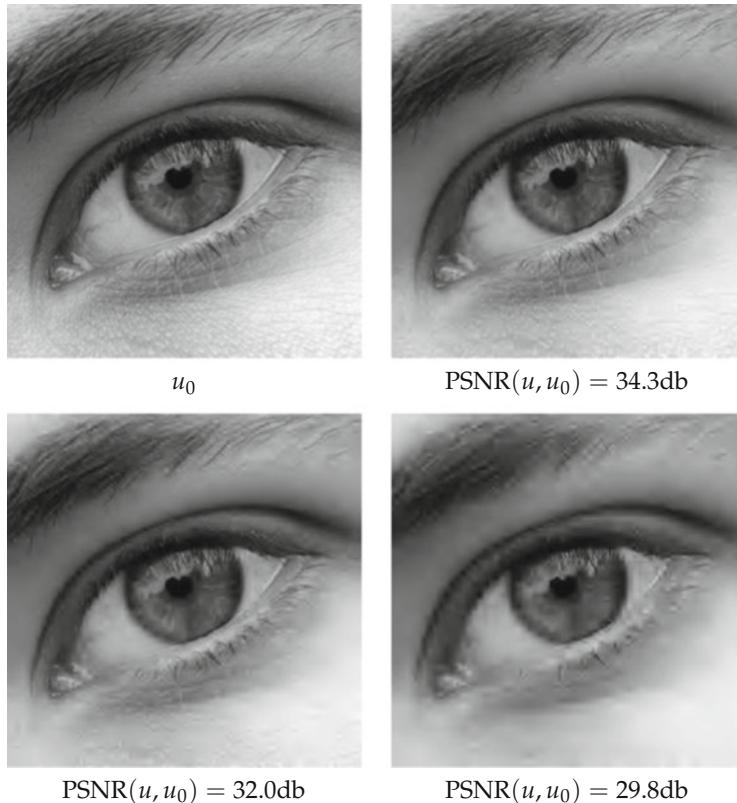


Fig. 4.16 Illustration of the compression potential of the two-dimensional wavelet transform. From upper left to lower right: Original image, reconstruction based on 10%, 5% and 2% of the wavelet coefficients, respectively

4.5 Further Developments

The representation of multidimensional functions, in particular of images, is currently a highly active field of research. Some powerful tools are already available in the form of Fourier analysis (especially Fourier series) and orthogonal wavelet bases. As seen above, a disadvantage of two-dimensional wavelets is their comparably poor adaptation to directional structures. For this case, newer systems of functions try to provide a remedy. A prominent example is given by the so-called *curvelets*; cf. [27], for instance. Curvelets result from a “mother function” $\gamma : \mathbf{R}^2 \rightarrow \mathbf{R}$ by scaling, rotation, and translation. More precisely, let $a > 0$, $\theta \in [0, 2\pi[$, and $b \in \mathbf{R}^2$, and set

$$S_a = \begin{bmatrix} a^2 & 0 \\ 0 & a \end{bmatrix}, \quad R_\theta = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}$$

in order to define

$$\gamma_{a,\theta,b}(x) = a^{\frac{3}{2}} \gamma(S_a R_\theta x - b).$$

The functions $\gamma_{a,\theta,b}$ hence result from γ through translation, rotation, and *parabolic scaling*. The continuous curvelet transform is then given by

$$\Gamma_\gamma u(a, \theta, b) = \int_{\mathbf{R}^2} u(x) \gamma_{a,\theta,b}(x) dx;$$

cf. [96] as well. It is remarkable about the construction of the curvelets that they allow a discretization that nearly results in an orthonormal basis, cf. [27]. Apart from that, curvelets are in a certain sense nearly optimally suited to represent functions that are piecewise twice continuously differentiable and whose discontinuities occur on sets that can be parameterized twice continuously differentiably. A curvelet decomposition and reconstruction can be implemented efficiently, cf. [28]. In comparison to Fourier or wavelet decompositions, however, it is still quite involved.

Another ansatz is given by the so-called *shearlets*; cf. [90], for instance. These functions are also based on translations and parabolic scalings. In contrast to curvelets, however, *shearings* are used instead of rotations. Specifically, for $a, s > 0$, we set

$$M_{a,s} = \begin{bmatrix} 1 & s \\ 0 & 1 \end{bmatrix} \begin{bmatrix} a & 0 \\ 0 & \sqrt{a} \end{bmatrix} = \begin{bmatrix} a & \sqrt{as} \\ 0 & \sqrt{a} \end{bmatrix},$$

and for $b \in \mathbf{R}^2$ as well as a mother function ψ , we define

$$\psi_{a,s,b}(x) = a^{-\frac{3}{4}} \psi(M_{a,s}^{-1}(x - b)),$$

which yields the continuous shearlet transform by

$$\mathcal{S}_\psi u(a, s, b) = \int_{\mathbf{R}^2} u(x) \psi_{a,s,b}(x) dx.$$

Also, the shearlet transform allows a systematic discretization and can be implemented by efficient algorithms.

Apart from curvelets and shearlets, there are numerous other procedures to decompose images into elementary components that reflect the structure of the image as well as possible: Ridgelets, edgelets, bandlets, brushlets, beamlets, or platelets are just a few of these approaches. In view of the abundance of “-lets,” some people also speak of $*$ -lets (read “starlets”).

4.6 Exercises

Exercise 4.1 (Commutativity of Modulation) For $\xi, y \in \mathbf{R}^d$ and $A \in \mathbf{R}^{d \times d}$, prove the following commutativity relation for translation, modulation and linear coordinate transformation:

$$M_\xi T_y = e^{i\xi \cdot y} T_y M_\xi, \quad M_\xi D_A = D_A M_{A^\top \xi}.$$

Exercise 4.2 Elaborate the proof of Lemma 4.8.

Exercise 4.3 (The sinc Function as Fourier Transform) Show that the Fourier transform of the characteristic function $\chi_{[-B,B]}$ is given by

$$\mathcal{F}\chi_{[-B,B]}(\xi) = \sqrt{\frac{2}{\pi}} B \operatorname{sinc}(\frac{B\xi}{\pi}).$$

Exercise 4.4 (Fourier Transform and Semigroup Property)

1. Compute the Fourier transform of $f : \mathbf{R} \rightarrow \mathbf{R}$, defined by

$$f(x) = \frac{1}{1+x^2}.$$

2. Show that for the family of functions $(f_a)_{a>0}$ defined by

$$f_a(x) = \frac{1}{a\pi} \frac{1}{1+\frac{x^2}{a^2}},$$

we have the following semigroup property with respect to convolution:

$$f_a * f_b = f_{a+b}.$$

3. Show that the family of scaled Gaussian functions $g_a : \mathbf{R}^d \rightarrow \mathbf{R}$, defined by

$$g_a(x) = \frac{1}{(4\pi a)^{d/2}} e^{-\frac{|x|^2}{4a}}$$

also satisfies the following semigroup property:

$$g_a * g_b = g_{a+b}.$$

Exercise 4.5 (Convergence in Schwartz Space) Let $\gamma \in \mathbf{N}^d$ be a multi-index. Show that the derivative operator $\frac{\partial^\gamma}{\partial x^\gamma} : \mathcal{S}(\mathbf{R}^d) \rightarrow \mathcal{S}(\mathbf{R}^d)$ is continuous.

Exercise 4.6 (Fourier Transform of a Distribution) Show that the mapping $\Phi_k(\phi) = \phi^{(k)}(0)$ that assigns to a Schwartz function ϕ the value of its k th derivative at 0 is tempered distribution and compute its Fourier transform.

Exercise 4.7 (Fourier Transform of Polynomials) Show that every polynomial $p : \mathbf{R} \rightarrow \mathbf{R}$, $p(x) = \sum_{n=0}^n a_n x^n$ induces a regular distribution and compute its Fourier transform.

Exercise 4.8 (Fourier Transform and Weak Derivative) Let $u \in L^2(\mathbf{R}^d)$, $\alpha \in \mathbf{N}^d$ a multi-index, and $p^\alpha u \in L^2(\mathbf{R}^d)$. Show the following relationship between the Fourier transform of $p^\alpha u$ and the weak derivative of the Fourier transform:

$$\mathcal{F}(p^\alpha u) = i^{|\alpha|} \partial^\alpha \mathcal{F}(u).$$

Exercise 4.9 (Regarding Equivalent Sobolev Norms, cf. Theorem 4.29) Let $k, d \in \mathbf{N}$ and define $f, g : \mathbf{R}^d \rightarrow \mathbf{R}$ by

$$f(\xi) = \sum_{|\alpha| \leq k} |\xi^\alpha|^2, \quad g(\xi) = (1 + |\xi|^2)^k.$$

Show that there exist constants $c, C > 0$ (which may depend on k and d) such that

$$cf \leq g \leq Cf.$$

[Hint:] One has the following multinomial theorem:

$$(x_1 + \cdots + x_d)^k = \sum_{|\alpha|=k} \frac{k!}{\alpha!} x^\alpha.$$

Exercise 4.10 (Smoothness of Characteristic Functions) Show that the function $u(x) = \chi_{[-1,1]}(x)$ lies in the space $H^s(\mathbf{R})$ for every $s \in [0, 1/2[$. What is the situation in the limiting case $s = 1/2$?

Addition: For which $s \in \mathbf{R}$ and $d \in \mathbf{N}$ does the delta distribution lie in the Sobolev space $H^s(\mathbf{R}^d)$?

Exercise 4.11 (Recursive Representation of the Discrete Fourier Transform)

- Let N be an even number. Show that the elements of the vector \hat{y} have the following representations of even entries,

$$\hat{y}_{2k} = \frac{1}{N} \sum_{n=0}^{N/2-1} (y_n + y_{n+N/2}) e^{-\frac{2\pi i kn}{N/2}},$$

and odd entries,

$$\hat{y}_{2k+1} = \frac{1}{N} \sum_{n=0}^{N/2-1} e^{-\frac{2\pi in}{N}} (y_n - y_{n+N/2}) e^{-\frac{2\pi i kn}{N/2}}.$$

(That is, the values \hat{y}_{2k} result from the Fourier transform of the $N/2$ -periodic signal $y_n + y_{n+N/2}$ and the values \hat{y}_{2k+1} are obtained by the Fourier transform of the $N/2$ -periodic signal $e^{-\frac{2\pi in}{N}} (y_n - y_{n+N/2})$.)

- Show that there exists an algorithm that computes the discrete Fourier transform of a signal of length $N = 2^n$ by means of $\mathcal{O}(n2^n)$ operations (in comparison to $\mathcal{O}(2^n 2^n)$ operations for a direct implementation of the sums).

Exercise 4.12 (Fast Convolution with the Discrete Fourier Transform) The convolution of $u, v \in \mathbf{C}^{\mathbf{Z}}$ is defined by

$$(u * v)_k = \sum_{n \in \mathbf{Z}} u_k v_{n-k}.$$

The *support* of $u \in \mathbf{C}^{\mathbf{Z}}$ is given by

$$\text{supp } u = \{k \in \mathbf{Z} \mid u_k \neq 0\}.$$

- Let $u, h \in \mathbf{C}^{\mathbf{Z}}$ with $\text{supp } u = \{0, \dots, N-1\}$ and $\text{supp } h = \{-r, \dots, r\}$. Then, $\text{supp } u * v \subset \{-r, \dots, N+r-1\}$ (why?). Develop and implement an algorithm `fftconv` that computes the convolution of u and v on the whole support by means of the Fourier transform. Input: Two vectors $u \in \mathbf{C}^N$ and $h \in \mathbf{C}^{2r+1}$. Output: The result $w \in \mathbf{C}^{N+2r}$ of the convolution of u and v .
- Analogously to the previous task, develop and implement a two-dimensional fast convolution `fft2conv`. Input: A grayscale image $u \in \mathbf{R}^{N \times M}$ and a convolution kernel $h \in \mathbf{R}^{2r+1 \times 2s+1}$. Output: The convolution $u * h \in \mathbf{R}^{N+2r, M+2s}$.
- What is the complexity for the fast convolution in contrast to the direct evaluation of the sums? For which sizes of u and h does the fast convolution pay off in this regard?

4. Test the algorithm `fft2conv` with convolution kernels of your choice. Compare the results and execution times with an implementation of the direct calculation of the sums according to Sect. 3.3.3 (also in the light of the complexity estimates).

Exercise 4.13 (Overlap-Add Convolution) For a signal $u \in \mathbf{C}^N$ and a convolution kernel $h \in \mathbf{C}^M$ that is significantly shorter than the signal, the convolution of u with h can be computed considerably more efficiently than in Exercise 4.12. For M a factor of N , we partition u into N/M blocks of length M :

$$u_n = \sum_{r=0}^{N/M-1} u_{n-rM}^r \quad \text{with} \quad u_n^r = \begin{cases} u_{(r-1)M+n} & \text{if } 0 \leq n \leq M-1, \\ 0 & \text{otherwise.} \end{cases}$$

$$\underbrace{u_0 \ u_1 \ \dots \ u_{M-1}}_{u^1} \underbrace{u_M \ u_{M+1} \ \dots \ u_{2M-1} \ \dots}_{u^2}$$

(without explicitly mentioning it, we have considered all vectors to be vectors on the whole of \mathbf{Z} by zero-extension.) The convolution of u with h can now be realized by means of a fast convolution of the parts u^r with h : we compute $v^r = u^r * h$ (which according to Exercise 4.12, requires $\mathcal{O}(M \log M)$ operations) and due to the linearity of the convolution, we obtain

$$u * h = \sum_{r=0}^{N/M-1} v^r.$$

Note that v^r and v^{r+1} overlap. This procedure is also called the overlap-add method.

1. Show that the complexity of the overlap-add convolution is given by $\mathcal{O}(N \log M)$.
2. Develop, implement, and document an algorithm `fftconv_oa` that computes the convolution of u and v on the whole support by means of the overlap-add method. Input: Two vectors $u \in \mathbf{C}^N$ and $h \in \mathbf{C}^M$ where M is a factor of N . Output: The result $w \in \mathbf{C}^{N+M-1}$ of the convolution of u and v .
3. For suitable test examples, compare the results and execution times of your algorithm with those of the algorithm `fftconv` in Exercise 4.12.

Exercise 4.14 (Scaled Bases of a Multiscale Analysis) Let (V_j) be a multiscale analysis with generator ϕ . Show that the set $\{\phi_{j,k} \mid k \in \mathbf{Z}\}$ forms an orthonormal basis of V_j .

Exercise 4.15 (Multiscale Analysis of Bandlimited Functions) Let

$$V_j = \{u \in L^2(\mathbf{R}) \mid \text{supp } \widehat{u} \subset [-2^{-j}\pi, 2^{-j}\pi]\}.$$

1. Show that (V_j) together with the generator $\phi(x) = \text{sinc}(x)$ forms a multiscale analysis of $L^2(\mathbf{R})$.
2. Determine the coefficient sequence (h_k) with which ϕ satisfies the scaling equation (4.4) and calculate the corresponding wavelet.

Exercise 4.16 (Properties of ψ in Theorem 4.67) Let $\phi : \mathbf{R} \rightarrow \mathbf{R}$ be the generator of a multiscale analysis and let ψ be defined as in Theorem 4.67.

1. Show that the coefficients (h_k) of the scaling equation are real and satisfy the following condition: For all $l \in \mathbf{Z}$,

$$\sum_{k \in \mathbf{Z}} h_k h_{k+2l} = \begin{cases} 1 & \text{if } l = 0, \\ 0 & \text{if } l \neq 0. \end{cases}$$

(Use the fact that the function ϕ is orthogonal to the functions $\phi(\cdot + m)$ for $m \in \mathbf{Z}, m \neq 0$.)

2. Show:

- (a) For all $l \in \mathbf{Z}$, $(\psi, \psi(\cdot - l)) = \begin{cases} 1 & \text{if } l = 0, \\ 0 & \text{if } l \neq 0. \end{cases}$
- (b) For all $l \in \mathbf{Z}$, $(\phi, \psi(\cdot - l)) = 0$.

Chapter 5

Partial Differential Equations in Image Processing



Our first encounter with a partial differential equation is this book was Application 3.23 on edge detection according to Canny: we obtained a smoothed image by solving the heat equation. The underlying idea was that images contain information on different spatial scales and one should not fix one scale a priori. The perception of an image depends crucially on the resolution of the image. If you consider a satellite image, you may note the shape of coastlines or mountains. For an aerial photograph taken from a plane, these features are replaced by structures on smaller scales such as woods, settlements, or roads. We see that there is no notion of an absolute scale and that the scale depends on the aims of the analysis. Hence, we ask ourselves whether there is a mathematical model of this concept of scale. Our aim is to develop a scale-independent representation of an image. This aim is the motivation behind the notion of *scale space* and the multiscale description of images [88, 98, 144]. The notion “scale space” does not refer to a vector space or a similar structure, but to a *scale space representation* or *multiscale representation*: for a given image $u_0 : \Omega \rightarrow \mathbf{R}$ one defines a function $u : [0, \infty[\times \Omega \rightarrow \mathbf{R}$, and the new positive parameter describes the *scale*. The scale space representation for scale parameter equal to 0 will be the original image:

$$u(0, x) = u_0(x).$$

For larger scale parameters we should obtain images on “coarser scales.” We could also view the introduction of the new scale variable as follows: We consider the original image u_0 as an element in a suitable space X (e.g., a space of functions $\Omega \rightarrow \mathbf{R}$). The scale space representation is then a map $u : [0, \infty[\rightarrow X$, i.e. a path through the space X . This view is equivalent to the previous (setting $u(0) = u_0$ and “ $u(\sigma)(x) = u(\sigma, x)$ ”).

We can imagine many different representations of an image on different scales. Starting from the methods we already know, we can define:

Convolution with scaled kernels: As in Application 3.23 we can define, for an image $u_0 \in L^2(\mathbf{R}^d)$ and a kernel $h \in L^2(\mathbf{R}^d)$,

$$u(\sigma, x) = (u_0 * \sigma^{-d} D_{\sigma^{-1} \text{id}} h)(x).$$

Hence, the function $u : [0, \infty[\times \mathbf{R}^d \rightarrow \mathbf{R}$ consists of convolutions of u with kernels $y \mapsto \sigma^{-d} h(\sigma^{-1} y)$ that are obtained by scaling a given kernel to different widths but keeping their integrals constant. This is similar to what happens in the continuous wavelet transformation.

Alternatively, we could define $u(\sigma) = (u_0 * \sigma^{-d} D_{\sigma^{-1} \text{id}} h)$ and obtain $u : [0, \infty[\rightarrow L^2(\mathbf{R}^d)$.

Morphological operators with scaled structure elements: For a bounded image $u_0 : \mathbf{R}^d \rightarrow \mathbf{R}$ and a structure element $B \subset \mathbf{R}^d$ we can define

$$u(\sigma, x) = (u_0 \ominus \sigma B)(x) \quad \text{and} \quad u(\sigma, x) = (u_0 \oplus \sigma B)(x) \quad \text{respectively.}$$

An alternative way would be to set $u(\sigma) = (u_0 \ominus \sigma B)$ and $u(\sigma) = (u_0 \oplus \sigma B)$, respectively, to get $u : [0, \infty[\rightarrow \mathcal{B}(\mathbf{R}^d)$.

One could easily produce more examples of scale spaces but one should ask, which of these are meaningful. There is an axiomatic approach to this question from [3], which, starting from a certain set of axioms, arrives at a restricted set of scale spaces or *multiscale analyses*. In this chapter we will take this approach, which will lead us to partial differential equations. This allows for a characterization of scale spaces and to build further methods on top of these.

5.1 Axiomatic Derivation of Partial Differential Equations

The idea behind an axiomatic approach is to characterize and build methods for image processing by specifying certain obvious properties that can be postulated. In the following we will develop a theory, starting from a fairly small set off axioms, and this theory will show that the corresponding methods indeed correspond to the solution of certain partial differential equations. This provides the foundation of the widely developed theory of partial differential equations in image processing. The starting point for scale space theory is the notion of multiscale analysis according to [3].

5.1.1 Scale Space Axioms

We start with the notion of scale space. Roughly speaking, a scale space is a family of maps. We define the following spaces of functions:

$$\begin{aligned}\mathcal{C}_b^\infty(\mathbf{R}^d) &= \{u : \mathbf{R}^d \rightarrow \mathbf{R} \mid u \in \mathcal{C}^\infty(\mathbf{R}^d), \partial^\alpha u \text{ bounded for all } \alpha \in \mathbf{N}^d\}, \\ \mathcal{BC}(\mathbf{R}^d) &= \{u : \mathbf{R}^d \rightarrow \mathbf{R} \mid u \in \mathcal{C}^0(\mathbf{R}^d), u \text{ bounded}\}.\end{aligned}$$

Definition 5.1 A *scale space* is a family of transformations $(\mathcal{T}_t)_{t \geq 0}$ with the property that

$$\mathcal{T}_t : \mathcal{C}_b^\infty(\mathbf{R}^d) \rightarrow \mathcal{BC}(\mathbf{R}^d) \quad \text{for all } t \geq 0.$$

We call $t \geq 0$ the *scale*.

This definition is very general, and without further assumptions we cannot interpret the notion of scale any further. The intuition should be that $\mathcal{T}_t u$ represents an image u on all possible “scales,” and with larger t contains only coarser elements of the image. This intuition is reflected in the following axioms, which we group into three different classes:

Architectural axioms: These axioms provide the foundation and mainly model the basic intuition behind a “scale space.”

Stability: The axioms that provides stability will be the so-called comparison principle, also called the maximum principle.

Morphological axioms: These axioms describe properties that are particular for the description of images. They say how shapes should behave in a scale space.

We use the abbreviation $X = \mathcal{C}_b^\infty(\mathbf{R}^d)$ and postulate:

Architectural Axioms

Recursivity: (Semigroup property)

For all $u \in X, s, t \geq 0 :$

$$\mathcal{T}_0(u) = u, \tag{REC}$$

$$\mathcal{T}_s \circ \mathcal{T}_t(u) = \mathcal{T}_{s+t}(u).$$

The concatenation of two transforms in a scale space should give another transformation of said scale space (associated with the sum of the scale parameters). This implies that one can obtain $\mathcal{T}_t(u)$ from any representation $\mathcal{T}_s(u)$ with $s < t$. Hence, $\mathcal{T}_s(u)$ contains all information to generate the coarser representations

$\mathcal{T}_t(u)$ in some sense. Put differently, the amount of information in the images decreases. Moreover, one can calculate all representations on an equidistant discrete scale by iterating a single operator $\mathcal{T}_{t/N}$.

There is a small technical problem: the range of the operators \mathcal{T}_t may not be contained in the respective domains of definition. Hence, the concatenation of \mathcal{T}_t and \mathcal{T}_s is not defined in general. For now we resort to saying that [REC] will be satisfied whenever $\mathcal{T}_s \circ \mathcal{T}_t(u)$ is defined. In Lemma 5.8 we will see a more elegant solution of this problem.

Regularity:

$$\begin{aligned} \text{For all } u, v \in X \text{ and } h, t \in [0, 1], \text{ there exists } C(u, v) > 0 : \\ \|\mathcal{T}_t(u + hv) - \mathcal{T}_t u - hv\|_\infty \leq C(u, v)ht. \end{aligned} \quad [\text{REG}]$$

This is an assumption on the boundedness of difference quotients in the direction v . In the case of linear operators \mathcal{T}_t , this becomes

$$\text{For all } v \in X \text{ and } t \in [0, 1] \text{ there exists } C(v) > 0 : \|\mathcal{T}_t v - v\|_\infty \leq C(v)t.$$

Locality:

$$\begin{aligned} \text{For all } u, v \in X, x \in \mathbf{R}^d \text{ with } \partial^\alpha u(x) = \partial^\alpha v(x) \text{ for all } \alpha \in \mathbf{N}^d, \\ (\mathcal{T}_t u - \mathcal{T}_t v)(x) = o(t) \quad \text{for } t \rightarrow 0. \end{aligned} \quad [\text{LOC}]$$

Roughly speaking this axiom says that the value $\mathcal{T}_t u(x)$ depends only on the behavior of u in a neighborhood of x if t is small.

Stability

Comparison principle: (Monotonicity)

$$\begin{aligned} \text{For all } u, v \in X \text{ with } u \leq v \text{ one has} \\ \mathcal{T}_t u \leq \mathcal{T}_t v. \end{aligned} \quad [\text{COMP}]$$

If one image is brighter than another, this will be preserved under the scale space. If \mathcal{T}_t is linear, this is equivalent to $\mathcal{T}_t u \geq 0$ for $u \geq 0$.

Morphological Axioms

The architectural axioms and stability do not say much about what actually happens to images under the scale space. The morphological axioms describe properties that are natural from the point of view of image processing.

Gray-level-shift invariance:

For all $t \geq 0$, $c \in \mathbf{R}$, $u \in X$, one has

$$\mathcal{T}_t(0) = 0 \quad [\text{GLSI}]$$

$$\mathcal{T}_t(u + c) = \mathcal{T}_t(u) + c.$$

This axiom says the one does not have any a priori assumption on the range of gray values of the image.

Gray-scale invariance: (also contrast invariance; contains [GLSI], but is stronger)

For all $t \geq 0$, $u \in X$, $h : \mathbf{R} \rightarrow \mathbf{R}$ nondecreasing and $h \in \mathcal{C}^\infty(\mathbf{R})$, one has

$$\mathcal{T}_t(h(u)) = h(\mathcal{T}_t(u)). \quad [\text{GSI}]$$

The map h rescales the gray values but preserves their order. The axiom says that the scale space will depend only on the shape of the levelsets and not on the contrast.

Translation invariance:

For all $t \geq 0$, $u \in X$, $h \in \mathbf{R}^d$, one has
[TRANS]

$$\mathcal{T}_t(T_h u) = T_h(\mathcal{T}_t u).$$

All points in \mathbf{R}^d are treated equally, i.e., the action of the operators \mathcal{T}_t does not depend on the location of objects.

Isometry invariance:

For all $t \geq 0$, $u \in X$, and all orthonormal transformations $R \in O(\mathbf{R}^d)$, one has

$$\mathcal{T}_t(D_R u) = D_R(\mathcal{T}_t u) \quad [\text{ISO}]$$

Scale invariance:

For all $\lambda \in \mathbf{R}$ and $t \geq 0$ there exists $t'(t, \lambda) \geq 0$ such that
[SCALE]

$$\mathcal{T}_t(D_{\lambda \text{id}} u) = D_{\lambda \text{id}}(\mathcal{T}_{t'} u) \text{ for all } u \in X.$$

In some sense, the scale space should be invariant with respect to zooming of the images. Otherwise, it would depend on unknown distance of the object to the camera.

5.1.2 Examples of Scale Spaces

The notion of scale space and its axiomatic description contains a broad class of the methods that we already know. We present some of these as examples.

Example 5.2 (Coordinate Transformations) Let $A \in \mathbf{R}^{d \times d}$ be a matrix. Then the linear coordinate transformation

$$(\mathcal{T}_t u)(x) = u(\exp(At)x) \quad (5.1)$$

is a scale space.

The operators \mathcal{T}_t are linear and the scale space satisfies the axioms [REC], [REG], [LOC], [COMP], [GSI], and [SCALE], but not [TRANS] or [ISO] (in general). This can be verified by direct computation. We show two examples. For [REC]: Let $s, t \geq 0$. Then for $u_t(x) = (\mathcal{T}_t u)(x) = u(\exp(At)x)$, one has

$$\begin{aligned} (\mathcal{T}_s \mathcal{T}_t u)(x) &= (\mathcal{T}_s u_t)(x) = u_t(\exp(As)x) = u(\exp(As)\exp(At)x) \\ &= u(\exp(A(s+t))x) = (\mathcal{T}_{s+t} u)(x). \end{aligned}$$

Moreover, $\exp(A0) = \text{id}$, and hence $(\mathcal{T}_0 u) = u$. For [LOC] we consider the Taylor expansion for $u \in X$:

$$u(y) = u(x) + (y - x) \cdot \nabla u(x) + \mathcal{O}(|x - y|^2).$$

Hence, we obtain for $u, v \in X$ with $\partial^\alpha u(x) = \partial^\alpha v(x)$ and $y = \exp(At)$:

$$\begin{aligned} (\mathcal{T}_t u - \mathcal{T}_t v)(x) &= u(\exp(At)x) - v(\exp(At)x) \\ &= \mathcal{O}(|(\exp(At) - \text{id})x|^2) \leq \mathcal{O}(\|\exp(At) - \text{id}\|^2). \end{aligned}$$

The properties of the matrix exponential imply that $\|\exp At - \text{id}\| = \mathcal{O}(t)$ and thus $(\mathcal{T}_t u - \mathcal{T}_t v)(x) = o(t)$.

In general, one gets semigroups of transformations also by solving ordinary differential equations. We consider a vector field $v \in C^\infty(\mathbf{R}^d, \mathbf{R}^d)$. The corresponding integral curves j are defined by

$$\partial_t j(t, x) = v(j(t, x)), \quad j(0, x) = x.$$

A scale space is then given by

$$(\mathcal{T}_t u)(x) = u(j(t, x)). \quad (5.2)$$

In the special case $v(x) = Ax$ this reduces to the above coordinate transformation (5.1). Analogously, this class of scale space inherits the listed properties of (5.1). The action of (5.1) and (5.2) on an image is shown in Fig. 5.1.

Example 5.3 (Convolution with Dilated Kernels) Let $\varphi \in L^1(\mathbf{R}^d)$ be a convolution kernel with $\int_{\mathbf{R}^d} \varphi(x) dx = 1$, and $\tau : [0, \infty[\rightarrow [0, \infty[$ a continuous and increasing time scaling with $\tau(0) = 0$ and $\lim_{t \rightarrow \infty} \tau(t) = \infty$. We define the dilated kernels $\varphi_t(x) = \tau(t)^{-d} \varphi\left(\frac{x}{\tau(t)}\right)$ and define the operators \mathcal{T}_t by

$$(\mathcal{T}_t u) = \begin{cases} u * \varphi_t & \text{if } t > 0, \\ u & \text{if } t = 0, \end{cases} \quad (5.3)$$

and obtain a scale space. Its action is shown in Fig. 5.2.

The operators are, like those in Example 5.2, linear, and we check which axioms are satisfied. We go a little more deeply into detail and discuss every axiom.

[REC]: In general it is not true that $\varphi_t * \varphi_s = \varphi_{t+s}$ for all $s, t > 0$, which would be equivalent to [REC]. However, Exercise 4.4 shows that this can indeed be satisfied, as shown by the Gaussian kernel

$$\varphi(x) = \frac{1}{(4\pi)^{d/2}} e^{-\frac{|x|^2}{4}}, \quad \tau(t) = \sqrt{t}.$$

Hence, the validity of this axiom depends crucially on the kernel and the time scaling.

[REG]: To check this axiom, we assume further properties of the kernel and the time scaling:

$$\int_{\mathbf{R}^d} x \varphi(x) dx = 0, \quad \int_{\mathbf{R}^d} |x|^2 |\varphi(x)| dx < \infty, \quad \tau(t) \leq C\sqrt{t}. \quad (5.4)$$

Linearity of the convolution shows that

$$\mathcal{T}_t(u + hv) - \mathcal{T}_t u - hv = h\mathcal{T}_t v - hv = h(\mathcal{T}_t - \text{id})v.$$

By the first condition in (5.4), we get, for all $z \in \mathbf{R}^d$,

$$\int_{\mathbf{R}^d} z \cdot y \varphi_t(y) dy = 0.$$

The Taylor approximation up to second order of v shows that

$$|v(y) - v(x) - \nabla v(x) \cdot (x - y)| \leq C \|\nabla^2 v\|_\infty |x - y|^2.$$

Original image u

$$A = \begin{bmatrix} -1 & 2 \\ 2 & -3 \end{bmatrix}$$

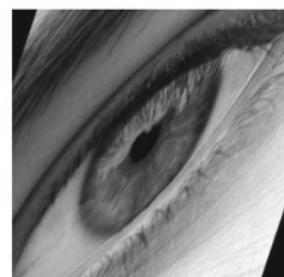
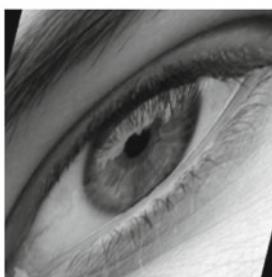
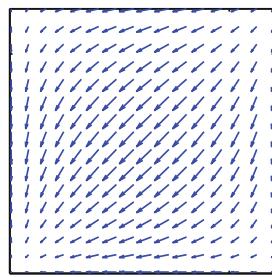
Application of (5.1) for some scales t Original image u vector field v Application of (5.2) for some scales t

Fig. 5.1 Example of a scale space given by a coordinate transformation. The first row shows the image and the matrix that generates the scale space according to (5.1), as shown in the second row. Similarly, the third row shows the image and the vector field, and the fourth row shows the applications of \mathcal{T}_t according to (5.2)

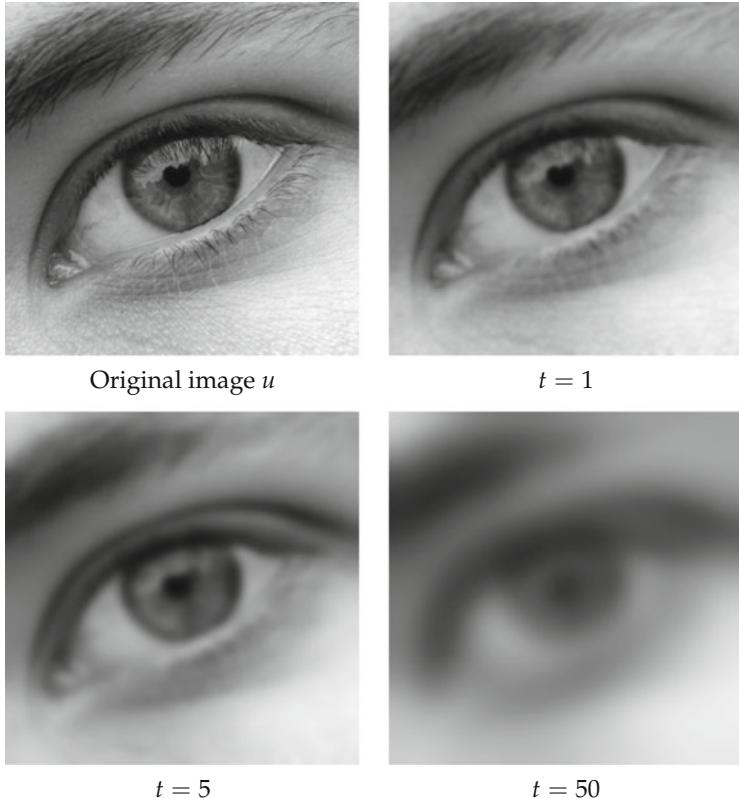


Fig. 5.2 Illustration of multiscale convolution with a Gaussian kernel on different scales

We combine the two observations and get (using a generic constant C)

$$\begin{aligned}
 |(\mathcal{T}_t v - v)(x)| &= \left| \int_{\mathbf{R}^d} (v(y) - v(x) - \nabla v(x) \cdot (y - x)) \varphi_t(x - y) dy \right| \\
 &\leq C \|\nabla^2 v\|_\infty \int_{\mathbf{R}^d} |x - y|^2 \varphi_t(x - y) dx \\
 &\leq C \int_{\mathbf{R}^d} \frac{|x|^2}{\tau(t)^d} \left| \varphi\left(\frac{x}{\tau(t)}\right) \right| dx = C \int_{\mathbf{R}^d} \tau(t)^2 |x|^2 |\varphi(x)| dx \\
 &= C \tau(t)^2 \leq Ct .
 \end{aligned}$$

Since this estimate is independent of x , we see that [REG] holds.

With assumptions for higher moments of the convolution kernel φ one could use a similar argument to allow time scalings of the form $\tau(t) \leq Ct^{1/n}$.

[LOC]: To check locality, we can use linearity and restrict our attention to the case of u with $\partial^\alpha u(x) = 0$ for all α . Without loss of generality we can also assume $x = 0$. Moreover, we assume that the kernel φ and the time scaling τ satisfy the assumptions in (5.4) and, on top of that, also

$$\int_{|x|>R} |\varphi(x)| dx \leq CR^{-\alpha}, \quad \alpha > 2.$$

Now we estimate:

$$\begin{aligned} |(\mathcal{T}_t u)(0)| &= \left| \int_{\mathbf{R}^d} u(y) \varphi_t(y) dy \right| \\ &\leq \sup_{|y|\leq\varepsilon} |\nabla^2 u(y)| \int_{|y|\leq\varepsilon} |y|^2 |\varphi_t(y)| dy + \|u\|_\infty \int_{|y|>\varepsilon} |\varphi_t(y)| dy \\ &\leq \sup_{|y|\leq\varepsilon} |\nabla^2 u(y)| \tau(t)^2 \int_{\mathbf{R}^d} |x|^2 |\varphi(x)| dx \\ &\quad + C \|u\|_\infty \int_{|x|\geq\varepsilon/\tau(t)} |\varphi(x)| dx \\ &\leq C \left(\sup_{|y|\leq\varepsilon} |\nabla^2 u(y)| t + \varepsilon^{-\alpha} t^{\alpha/2} \right). \end{aligned}$$

Now let $\delta > 0$ be given. Since $\nabla^2 u(0) = 0$ and $\nabla^2 u$ is continuous, we see that $\sup_{|y|\leq\varepsilon} |\nabla^2 u(y)| \rightarrow 0$ for $\varepsilon \rightarrow 0$. Hence, we can choose $\varepsilon > 0$ small enough to ensure that

$$C \sup_{|y|\leq\varepsilon} |\nabla^2 u(y)| t \leq \frac{\delta}{2} t.$$

Now for t small enough, we have

$$C \varepsilon^{-\alpha} t^{\alpha/2} \leq \frac{\delta}{2} t.$$

Putting things together, we see that for every $\delta > 0$, we have (if t is small enough)

$$|(\mathcal{T}_t u)(0)| \leq \delta t$$

and this means that $|(\mathcal{T}_t u)(0)| = o(t)$, which was our aim.

[COMP]: Again we can use the linearity of \mathcal{T}_t to show that this axiom is satisfied: It is enough to show $\mathcal{T}_t u \geq 0$ for all $u \geq 0$. This is fulfilled if $\varphi \geq 0$ almost everywhere, since then,

$$(\mathcal{T}_t u)(x) = \int_{\mathbf{R}^d} \underbrace{u(y)}_{\geq 0} \underbrace{\varphi_t(x-y)}_{\substack{\geq 0 \text{ a.e.} \\ \geq 0 \text{ a.e.}}} dx \geq 0.$$

[GLSI]: A simple calculation shows that gray-value shift invariance is satisfied:

$$\begin{aligned} (\mathcal{T}_t(u + c))(x) &= ((u + c) * \varphi_t)(x) = (u * \varphi_t)(x) + \int_{\mathbf{R}^d} c \varphi_t(y) dy \\ &= (u * \varphi_t)(x) + c = (\mathcal{T}_t u)(x) + c. \end{aligned}$$

[GSI]: Gray-value scaling invariance is not satisfied for convolution operators. It is simple to construct counterexamples (Exercise 5.2).

[TRANS]: Earlier consideration already showed that convolution with φ_t is a translation invariant operation, and hence every \mathcal{T}_t , is translation invariant as well.

[ISO]: If φ is rotationally invariant, then so is φ_t , and we see that for every rotation R , we have

$$\begin{aligned} (\mathcal{T}_t(D_R u))(x) &= \int_{\mathbf{R}^d} u(Ry) \varphi_t(x - y) dy = \int_{\mathbf{R}^d} u(z) \varphi_t(R^T(Rx - z)) dz \\ &= \int_{\mathbf{R}^d} u(z) \varphi_t(Rx - z) dz = (D_R(\mathcal{T}_t u))(x). \end{aligned}$$

Hence, isometry invariance holds for kernels that are rotationally invariant.

[SCALE]: For $\lambda \geq 0$ we can write

$$\begin{aligned} (\mathcal{T}_t(D_{\lambda \text{id}} u))(x) &= \int_{\mathbf{R}^d} u(\lambda y) \frac{1}{\tau(t)^d} \varphi\left(\frac{x - y}{\tau(t)}\right) dy \\ &= \int_{\mathbf{R}^d} u(z) \frac{1}{(\lambda \tau(t))^d} \varphi\left(\frac{\lambda x - z}{\lambda \tau(t)}\right) dz = (D_{\lambda \text{id}}(\mathcal{T}_{t'} u))(x), \end{aligned}$$

where t' is chosen such that $\tau(t') = \lambda \tau(t)$ (this is possible, since $\tau : [0, \infty[\rightarrow [0, \infty[$ is bijective by assumption). For $\lambda \leq 0$ one argues similarly, and hence scale invariance is satisfied.

Example 5.4 (Erosion and Dilation) For a nonempty structure element $B \subset \mathbf{R}^d$ we define $tB = \{ty \mid y \in B\}$. Based on this scaling we construct a scale space related to B as

$$(\mathcal{T}_t u) = u \oplus tB. \quad (5.5)$$

Similarly one could define a multiscale erosion related to B ; see Fig. 5.3. We restrict our attention to the case of dilation.

In contrast to Examples 5.2 and 5.3 above, the scale space in (5.5) is in general *nonlinear*. We discuss some axioms in greater detail:

[REC]: The axiom [REC] is satisfied if B is convex. You should check this in Exercise 5.4.

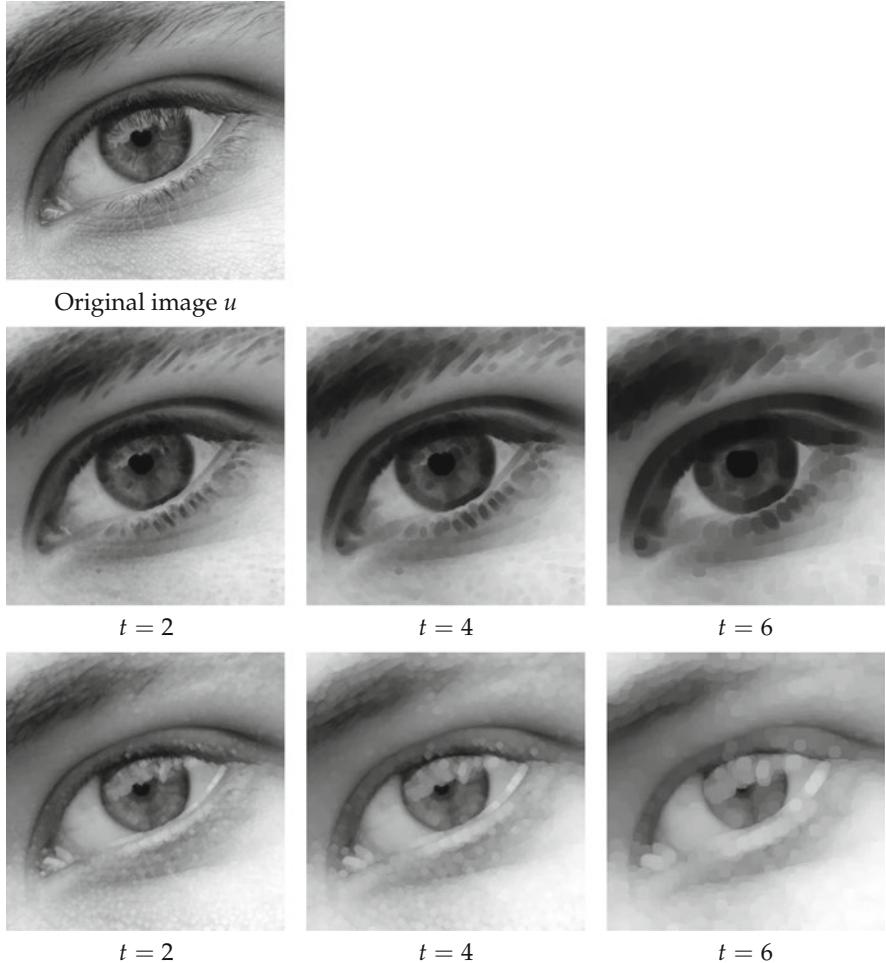


Fig. 5.3 Example of multiscale erosion (second row) and dilation (third row). The structure element is an octagon centered at the origin

[REG]: We estimate

$$\begin{aligned}
 & ((u + hv) \oplus tB)(x) - (u \oplus tB)(x) - hv(x) \\
 &= \sup_{y \in tB} [u(x+y) + hv(x+y)] - \sup_{y \in tB} [u(x+y)] - hv(x) \\
 &\leq \sup_{y \in tB} [u(x+y) + hv(x+y) - u(x+y)] - hv(x) \\
 &\leq h \sup_{y \in tB} [v(x+y) - v(x)].
 \end{aligned}$$

A similar computation with $-u$ instead of u and $-v$ instead of v shows that

$$((u + hv) \oplus tB)(x) - (u \oplus tB)(x) - hv(x) \geq -h \sup_{y \in tB} [v(x + y) - v(x)].$$

If we assume that B is bounded, the assertion follows by Lipschitz continuity of v . We conclude that The multiscale dilation satisfied the axiom [REG] if the structure element is bounded.

[LOC]: For locality we estimate as follows:

$$\begin{aligned} (u \oplus tB)(x) - (v \oplus tB)(x) &= \sup_{y \in tB} [u(x + y)] - \sup_{y \in tB} [v(x + y)] \\ &\leq \sup_{y \in tB} [u(x + y) - v(x + y)]. \end{aligned}$$

Replacing u by $-u$ and v by $-v$, we obtain

$$(u \oplus tB)(x) - (v \oplus tB)(x) \geq \sup_{y \in tB} [v(x + y) - u(x + y)].$$

If we assume again that B is bounded and $\partial^\alpha u(x) = \partial^\alpha v(x)$ for $|\alpha| \leq 1$, we get

$$\sup_{y \in tB} [u(x + y) - v(x + y)] = o(t) \quad \text{and} \quad \sup_{y \in tB} [v(x + y) - u(x + y)] = o(t).$$

Again we deduce that multiscale dilation satisfies the axiom [LOC] if the structure element is bounded.

[COMP]: We have already seen that the comparison principle is satisfied in Theorem 3.29 (under the name “monotonicity”).

[TRANS]: Translation invariance has also been shown in Theorem 3.29.

[GSI]: Gray-scale invariance has been shown in Theorem 3.31 under the name “contrast invariance.”

[ISO]: Isometry invariance is satisfied if the structure element is invariant under rotations.

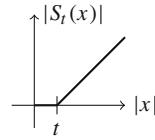
[SCALE]: Scale invariance can be seen as follows:

$$(D_{\lambda \text{id}} u \oplus tB)(x) = \sup_{y \in tB} u(\lambda(x + y)) = \sup_{z \in \lambda tB} u(\lambda x + z) = D_{\lambda \text{id}}(u \oplus \lambda tB)(x).$$

Hence, scale invariance holds for a symmetric structure element (i.e., $-B = B$) with $t' = |\lambda|t$.

Example 5.5 (Fourier Soft Thresholding) A somewhat unusual scale space is given by the following construction. Let \mathbf{S}_t be the operator that applies the complex *soft thresholding function*

$$S_t(x) = \begin{cases} \frac{x}{|x|} (|x| - t) & \text{if } |x| > t, \\ 0 & \text{if } |x| \leq t, \end{cases}$$



pointwise, i.e., $(\mathbf{S}_t(u))(x) = S_t(u(x))$. We apply soft thresholding to the Fourier transform, i.e.,

$$\mathcal{T}_t(u) = \mathcal{F}^{-1}(\mathbf{S}_t(\mathcal{F}u)). \quad (5.6)$$

Since all symmetries in Corollary 4.6 are preserved by pointwise thresholding, $\mathcal{T}_t u$ is again a real-valued function if it is defined. Here is a technical problem: unfortunately, strictly speaking, $(\mathcal{T}_t)_{t \geq 0}$ is not a scale space according to Definition 5.1, since the Fourier transform of a function in $C_b^\infty(\mathbf{R}^d)$ is merely a distribution, and hence pointwise operations are not defined. One can show that (5.6) can be defined on $L^2(\mathbf{R}^d)$. In that sense *Fourier soft thresholding* is a scale space on $L^2(\mathbf{R}^d)$.

We will not check which axioms are satisfied and only mention that Fourier soft thresholding is a nonlinear semi-group. Locality, translation, scaling and gray-level-shift invariance are not satisfied; only isometry invariance can be deduced. An illustration of the effect of Fourier soft thresholding is shown in Fig. 5.4. One may notice that the images reduce to the dominant oscillating structures (regardless of their frequency) with increasing scale parameter. (This is in contrast to “coarse scale structure” in the case of dilated convolution kernels.)

Similarly to Fourier soft thresholding one can define wavelet soft thresholding. We use the two-dimensional discrete wavelet transform from Sect. 4.4.3 to generate the approximation coefficients c^J and the detail coefficients $d^{1,J}, d^{2,J}, d^{3,J}, \dots, d^{1,1}, d^{2,1}, d^{3,1}$ from an image u . Then we apply the soft thresholding function with parameter t to the detail coefficient and reconstruct. This scale space has similar properties to those of Fourier soft thresholding, but also isometry invariance is not satisfied. This scale space is widely used in practice to denoise images [32, 35, 54, 139]. Fourier soft thresholding is also suited for denoising. Figure 5.5 compares both methods. Wavelet soft thresholding leads to slightly higher PSNR values than Fourier soft thresholding, and also the subjective impression seems to be a little bit better.

Example 5.6 (Solutions of Partial Differential Equations of Second Order) Let $F : \mathbf{R}^d \times \mathbf{R} \times \mathbf{R}^d \times \mathbf{R}^{d \times d} \rightarrow \mathbf{R}$ be a smooth function and consider the Cauchy problem

$$\partial_t u = F(x, u(x), \nabla u(x), \nabla^2 u(x)), \quad u(0) = u_0. \quad (5.7)$$

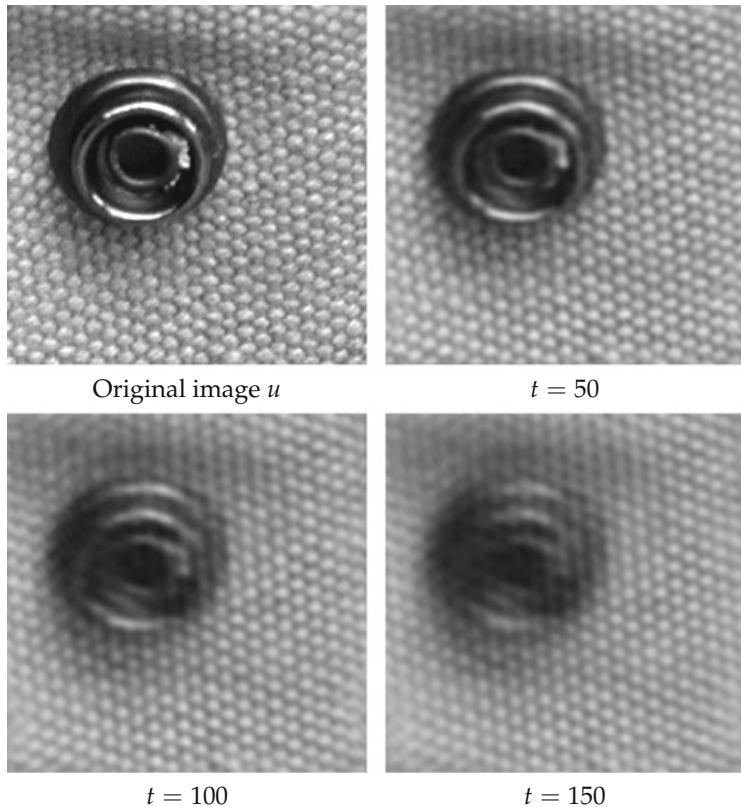


Fig. 5.4 Illustration of Fourier soft thresholding from Example 5.5 and different scales t

If we assume that there exists a unique solution of (5.7) for every initial value $u_0 \in \mathcal{C}_b^\infty(\mathbf{R}^d)$, then we can define

$$\mathcal{T}_t u_0 = u(t, \cdot),$$

which is (obviously) a scale space.

The next section will show that many of our previous examples can be written, in some sense, as solutions of (5.7). This is a central result in scale space theory and illustrates the central role of partial differential equations in image processing. Also, this shows that partial differential equations of this type are among the most general type of multiscale analyses.

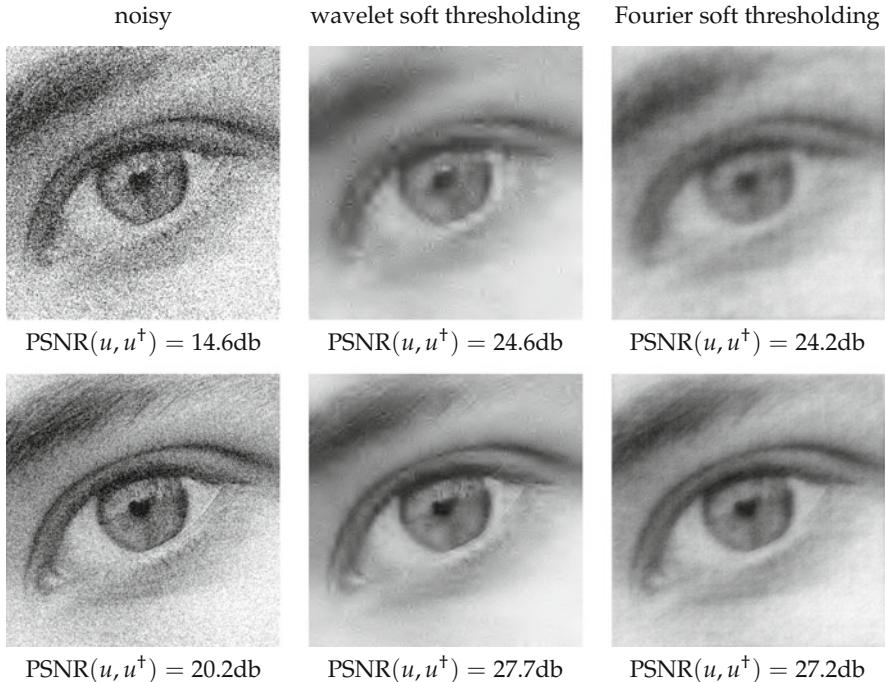


Fig. 5.5 Denoising by Fourier and wavelet soft thresholding from Example 5.5. The parameter t has been chosen to maximize the PSNR

5.1.3 Existence of an Infinitesimal Generator

The axioms of a scale space immediately imply a number of further properties:

Lemma 5.7 *If a scale space $(\mathcal{T}_t)_{t \geq 0}$ satisfies the axioms [COMP] and [GLSI], then it also satisfies*

$$\text{for all } u, v \in X \text{ and } t \geq 0 \text{ one has} \quad \|\mathcal{T}_t u - \mathcal{T}_t v\|_\infty \leq \|u - v\|_\infty. \quad [\text{CONT}]$$

In other words, it is Lipschitz continuous with constant no larger than 1 (a property also called non-expansivity).

Proof From $u \leq v + \|u - v\|_\infty$ we deduce, using [COMP], that $\mathcal{T}_t u \leq \mathcal{T}_t(v + \|v - u\|_\infty)$. From [GLSI] it follows that $\mathcal{T}_t u \leq \mathcal{T}_t v + \|v - u\|_\infty$ and

$$\mathcal{T}_t u - \mathcal{T}_t v \leq \|v - u\|_\infty.$$

holds. Swapping the roles of u and v we obtain the claimed property [CONT]. \square

As we have seen in Example 5.4, the scale space for dilation satisfies the axioms [COMP] and [GLSI]. So we have shown again that [CONT] holds in this case, a result we already derived in Exercise 3.10.

The next lemma allows us to extend a scale space from $\mathcal{C}_b^\infty(\mathbf{R}^d)$ to a larger space, namely to the space

$$\mathcal{BUC}(\mathbf{R}^d) = \{u : \mathbf{R}^d \rightarrow \mathbf{R} \mid u \text{ bounded and uniformly continuous}\}.$$

Lemma 5.8 *If [CONT] and [TRANS] hold for (\mathcal{T}_t) , one can extend every \mathcal{T}_t uniquely to a mapping*

$$\mathcal{T}_t : \mathcal{BUC}(\mathbf{R}^d) \rightarrow \mathcal{BUC}(\mathbf{R}^d).$$

Proof By Lipschitz continuity, [CONT], and the density of $\mathcal{C}_b^\infty(\mathbf{R}^d)$ in the space $\mathcal{BUC}(\mathbf{R}^d)$ we can extend \mathcal{T}_t to a map $\mathcal{T}_t : \mathcal{BUC}(\mathbf{R}^d) \rightarrow \mathcal{BC}(\mathbf{R}^d)$ uniquely. It remains to show the uniform continuity of $\mathcal{T}_t u$, $u \in \mathcal{BC}(\mathbf{R}^d)$.

We choose for arbitrary $\varepsilon > 0$ a $\delta > 0$ such that for all $x \in \mathbf{R}^d$ and $|h| < \delta$, one has $|u(x) - u(x + h)| < \varepsilon$. With $v = T_h u$ and because of [TRANS] and [CONT] we get

$$\begin{aligned} |(\mathcal{T}_t u)(x) - (\mathcal{T}_t u)(x + h)| &= |(\mathcal{T}_t u)(x) - (\mathcal{T}_t v)(x)| \\ &\leq \|u - v\|_\infty = \sup_{x \in \mathbf{R}^d} |u(x) - u(x + h)| \leq \varepsilon, \end{aligned}$$

which shows that $\mathcal{T}_t u$ is indeed uniformly continuous. \square

This lemma retroactively justifies the formulation of the axiom [REC]: if [CONT] and [TRANS] hold, we can consider the operators \mathcal{T}_t as mappings from $\mathcal{BUC}(\mathbf{R}^d)$ to itself, and the concatenation of two of these operators makes sense.

Now we turn toward a central result in scale space theory, namely the relation between a scale space and solutions of partial differential equations of second order. To that end, we are going to show the existence of an *infinitesimal generator* that can be written as a differential operator and that acts on the spatial dimensions. One obtains this generator, if it exists, by a simple limit.

Theorem 5.9 *Let $(\mathcal{T}_t)_{t \geq 0}$ be a scale space that satisfies [TRANS], [COMP], [GLSI], [REC], and [REG]. Moreover, let the constant $C(u, v)$ in [REG] be independent of $u, v \in Q$ for every set Q of the form*

$$Q = \{u \in \mathcal{C}_b^\infty(\mathbf{R}^d) \mid \|\partial^\alpha u\|_\infty \leq C_\alpha \text{ for all } \alpha \in \mathbf{N}^d\}. \quad (5.8)$$

Then we have the following existence result:

Existence of an infinitesimal generator:

There exists an $A : \mathcal{C}_b^\infty(\mathbf{R}^d) \rightarrow \mathcal{BUC}(\mathbf{R}^d)$ such that

$$A[u] = \lim_{t \rightarrow 0} \frac{\mathcal{T}_t u - u}{t} \quad \text{uniformly on } \mathbf{R}^d. \quad [\text{GEN}]$$

The operator A is continuous in the following sense: if $\partial^\alpha u_n \rightarrow \partial^\alpha u$ uniformly on \mathbf{R}^d for all $\alpha \in \mathbf{N}^d$, then $A[u_n] \rightarrow A[u]$ uniformly on \mathbf{R}^d .

Proof We give only a proof sketch. Details can be found in the original work [3].

Let $\delta_t(v) = (\mathcal{T}_t v - v)/t$ denote the difference quotient and deduce from [REG] that

$$\begin{aligned} \|\delta_t(u + hv) - \delta_t(u)\|_\infty &= \frac{1}{t} \|\mathcal{T}_t(u + hv) - u - hv - \mathcal{T}_t(u) + u\|_\infty \\ &= \frac{1}{t} \|\mathcal{T}_t(u + hv) - \mathcal{T}_t(u) - hv\|_\infty \\ &\leq C(u, v)h. \end{aligned} \quad (5.9)$$

Using $\mathcal{T}_t(0) = 0$, [GLSI], and $v = u$, $u = 0$, $h = 1$ in [REG], one gets

$$\|\delta_t(u)\|_\infty = \frac{1}{t} \|\mathcal{T}_t(0 + 1u) - \mathcal{T}_t(0) - 1u\|_\infty \leq C(u),$$

i.e., $\delta_t(u)$ is, for $t \rightarrow 0$, a bounded sequence in $\mathcal{BC}(\mathbf{R}^d)$.

The next step is to show that $\delta_t(u)$ has a limit for $t \rightarrow 0$. To that end, one shows Lipschitz continuity for every $\delta_t(u)$ with a Lipschitz constant independent of t , again for $t \rightarrow 0$. Choose $|y| = 1$, $h \in [0, 1]$ and note that by [TRANS], one has $(msa_t u)(x + hy) = (\mathcal{T}_t u(\cdot + hy))(x)$. Moreover,

$$u(x + hy) = u(x) + h \int_0^1 \nabla u(x + shy) \cdot y \, ds = u(x) + hv_h(x),$$

where, for $h \rightarrow 0$, one has $v_h \in \mathcal{C}_b^\infty(\mathbf{R}^d)$. Now one easily sees that all v_h are in a suitable set Q from (5.8). The estimate (5.9) gives the desired Lipschitz inequality

$$\|(\delta_t(u))(\cdot + hy) - \delta_t(u)\|_\infty = \|\delta_t(u + hv_h) - \delta_t(u)\|_\infty \leq Ch,$$

where by assumption, C is independent of v_h .

The theorem of Arzelà and Ascoli on compact subsets of spaces of continuous functions now implies the existence of a convergent subsequence of $\{\delta_{t_n}(u)\}$ for $t_n \rightarrow 0$. The operator A is now defined as the limit $A[u] = \lim_{t \rightarrow 0} \delta_t(u)$, if it is unique. To show this, one shows that the whole sequence converges, which is quite some more work. Hence, we refer to the original paper for the rest of the proof.

(which also contains a detailed argument for the uniform convergence $A[u_n] \rightarrow A[u]$). \square

Our next step is to note that the operator A can be written as a (degenerate) elliptic differential operator of second order.

Definition 5.10 Denote by $S^{d \times d}$ the space of symmetric $d \times d$ matrices. We write $X - Y \succcurlyeq 0$ or $X \succcurlyeq Y$ if $X - Y$ is positive semi-definite. A function $f : S^{d \times d} \rightarrow \mathbf{R}$ is called *elliptic* if $f(X) \geq f(Y)$ for $X \succcurlyeq Y$. If $f(X) > f(Y)$ for $X \succcurlyeq Y$ with $X \neq Y$, f is called *strictly elliptic*, and *degenerate elliptic* otherwise.

Theorem 5.11 Let $(\mathcal{T}_t)_{t \geq 0}$ be a scale space that satisfies the axioms [GEN], [COMP], and [LOC]. Then there exists a continuous function $F : \mathbf{R}^d \times \mathbf{R} \times \mathbf{R}^d \times S^{d \times d} \rightarrow \mathbf{R}$ such that $F(x, c, p, \cdot)$ is elliptic for all $(x, c, p) \in \mathbf{R}^d \times \mathbf{R} \times \mathbf{R}^d$ and

$$A[u](x) = F(x, u(x), \nabla u(x), \nabla^2 u(x))$$

for every $u \in \mathcal{C}_b^\infty(\mathbf{R}^d)$ and $x \in \mathbf{R}^d$.

Proof From [LOC] and by the definition of A we immediately get that $A[u](x) = A[v](x)$ if $\partial^\alpha u(x) = \partial^\alpha v(x)$ for all $\alpha \in \mathbf{N}^d$. We aim to show that this holds even if only $u(x) = v(x)$, $\nabla u(x) = \nabla v(x)$ as well as $\nabla^2 u(x) = \nabla^2 v(x)$ is satisfied: We consider $x_0 \in \mathbf{R}^d$ and u, v with $u(x_0) = v(x_0) = c$, and

$$\nabla u(x_0) = \nabla v(x_0) = p, \quad \nabla^2 u(x_0) = \nabla^2 v(x_0) = X.$$

Now we construct $\eta \in \mathcal{C}_b^\infty(\mathbf{R}^d)$ such that $\eta(x) = |x - x_0|^2$ in a neighborhood of x_0 and that there exists a constant $m > 0$ with $u_\varepsilon = u + \varepsilon\eta \geq v$ on $B_{m\varepsilon}(x_0)$ (see also Fig. 5.6). Such an η exists, since $v - u = o(|x - x_0|^2)$ for $|x - x_0| \rightarrow 0$ (recall that the derivatives coincide up to second order).

Moreover, choose a $w \in \mathcal{D}(B_m(x_0))$ with

$$w \geq 0, \quad w = 1 \text{ on } B_\sigma(x_0), \tag{5.10}$$

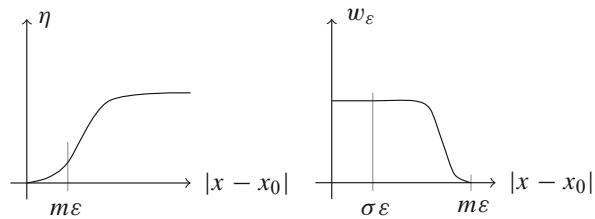
and use $w_\varepsilon(x) = w((x - x_0)/\varepsilon + x_0)$ (see Fig. 5.6) to construct the functions

$$\bar{u}_\varepsilon(x) = w_\varepsilon(x)u_\varepsilon(x) + (1 - w_\varepsilon(x))v(x).$$

These functions have the property that $\partial^\alpha \bar{u}_\varepsilon(x) = \partial^\alpha u_\varepsilon(x)$ for all $\alpha \in \mathbf{N}^d$ as well as $\bar{u}_\varepsilon(x) \geq v(x)$ on the whole \mathbf{R}^d . By [COMP] this implies $\mathcal{T}_t \bar{u}_\varepsilon(x_0) \geq \mathcal{T}_t v(x_0)$, and by the monotonicity of the limit also $A[\bar{u}_\varepsilon](x_0) \geq A[v](x_0)$. Moreover, $A[\bar{u}_\varepsilon](x_0) = A[u_\varepsilon](x_0)$ by construction of w and, again, by [LOC]. The continuity of A gives

$$\lim_{\varepsilon \rightarrow 0} A[\bar{u}_\varepsilon](x_0) = A[u](x_0)$$

Fig. 5.6 Schematic visualization of the auxiliary functions η and w_ε used in the proof of Theorem 5.11



and $A[u](x_0) \geq A[v](x_0)$. Switching the sign in the previous argument, we also get $A[u](x_0) \leq A[v](x_0)$ and hence $A[u](x_0) = A[v](x_0)$. We conclude that

$$A[u](x) = F(x, u(x), \nabla u(x), \nabla^2 u(x)),$$

as desired.

It remains to show the continuity of F and that F is elliptic in its last component. The latter follows from the following consideration: Construct, using w from (5.10), the functions

$$u(x) = (c + p \cdot (x - x_0) + \frac{1}{2}(x - x_0)^T X(x - x_0))w(x),$$

$$v(x) = (c + p \cdot (x - x_0) + \frac{1}{2}(x - x_0)^T Y(x - x_0))w(x),$$

which satisfy $u \geq v$ by construction. Moreover, $u(x_0) = v(x_0)$, $\nabla u(x_0) = \nabla v(x_0)$, and $\nabla^2 u(x_0) = X$, $\nabla^2 v(x_0) = Y$. Hence by [COMP], we have $\mathcal{T}_t u \geq \mathcal{T}_t v$, and for $t \rightarrow 0$ we obtain

$$F(x_0, c, p, X) = \lim_{t \rightarrow 0} \frac{(\mathcal{T}_t u - u)(x_0)}{t} \geq \lim_{t \rightarrow 0} \frac{(\mathcal{T}_t v - v)(x_0)}{t} = F(x_0, c, p, Y).$$

The continuity of F can be seen similarly: For sequences $x_n \rightarrow x_0$, $c_n \rightarrow c$, $p_n \rightarrow p$, and $X_n \rightarrow X$, the functions

$$u_n(x) = (c_n + p_n \cdot (x - x_n) + \frac{1}{2}(x - x_n)^T X_n(x - x_n))w(x)$$

converge uniformly to

$$u(x) = (c + p \cdot (x - x_0) + \frac{1}{2}(x - x_0)^T X(x - x_0))w(x),$$

and all their derivatives converge to the respective derivatives, too. By the conclusion of Theorem 5.9 we get that $A[u_n] \rightarrow A[u]$ uniformly. This implies

$$F(x_0, c_n, p_n, X_n) = A[u_n](x_0) \rightarrow A[u](x_0) = F(x_0, c, p, X). \quad \square$$

Remark 5.12 The proof also reveals the reason behind the fact that the order of the differential operator has to be two. The auxiliary function η from the proof is, in a neighborhood of zero, a polynomial of degree two. Every other positive polynomial of higher degree would also work, but would imply a dependence on higher derivatives. However, there is no polynomial of degree one that is strictly positive in a pointed neighborhood of zero. Hence, degree two is the lowest degree for which the argumentation in the proof works, and hence the order of the differential operator is two.

If we add further morphological axioms to the setting of Theorem 5.11, we obtain an even simpler form of F .

Lemma 5.13 *Assume that the assumptions in Theorem 5.11 hold.*

1. *If, additionally, [TRANS] holds, then*

$$A[u](x) = F(u(x), \nabla u(x), \nabla^2 u(x)).$$

2. *If, additionally, [GLSI] holds, then*

$$A[u](x) = F(x, \nabla u(x), \nabla^2 u(x)).$$

Proof The proof is based on the fact that the properties [TRANS] and [GLSI] are transferred from \mathcal{T}_t to A , and you should work out the rest in Exercise 5.5. \square

5.1.4 Viscosity Solutions

By Theorem 5.11 one may say that for $u_0 \in \mathcal{C}_b^\infty(\mathbf{R}^d)$, one has that $u(t, x) = (\mathcal{T}_t u_0)(x)$ solves the Cauchy problem

$$\frac{\partial u}{\partial t}(t, x) = F(x, u(t, x), \nabla u(t, x), \nabla^2 u(t, x)), \quad u(0, x) = u_0(x)$$

in some sense, but only *at time $t = 0$* (and the time derivative is only a one-sided limit).

Remark 5.14 We used in the above formula for the Cauchy problem the widely used convention that for function with distinguished “time coordinate” (t in this case), the operators ∇ and ∇^2 act only on the “spatial variable” x . We will keep using this convention in the following.

To show that the equation is also satisfied for $t > 0$, we can argue as follows: By [REC], we should have

$$\begin{aligned}\frac{\partial u}{\partial t}(t, x) &= \lim_{s \rightarrow 0^+} \frac{\mathcal{T}_{t+s}(u_0)(x) - \mathcal{T}_t(u_0)(x)}{s} = \lim_{s \rightarrow 0^+} \frac{\mathcal{T}_s(\mathcal{T}_t(u_0))(x) - \mathcal{T}_t(u_0)(x)}{s} \\ &= A[\mathcal{T}_t(u_0)](x) = A[u(t, \cdot)](x) = F(x, u(t, x), \nabla u(t, x), \nabla^2 u(t, x)).\end{aligned}$$

This would imply that u satisfies the differential equation for all times. However, there is a problem with this argument: $\mathcal{T}_t(u_0)$ is not necessarily an element in $C_b^\infty(\mathbf{R}^d)$, and the conclusion $\lim_{s \rightarrow 0^+} \frac{1}{s} (\mathcal{T}_s(\mathcal{T}_t(u_0)) - \mathcal{T}_t(u_0)) = A[\mathcal{T}_t(u_0)]$ is not valid.

The lack of regularity is a central problem in the theory of partial differential equations. An approach that often helps is to introduce a suitable notion of *weak solutions*. This means a generalized notion of solution that requires less regularity than the original equation requires. In the context of scale space theory, the notion of *viscosity solutions* is appropriate to define weak solutions with the desired properties. In the following we give a short introduction to the wide field of viscosity solutions but do not go into great detail.

The notion of viscosity solution is based on the following important observation:

Theorem 5.15 *Let $F : [0, \infty[\times \mathbf{R}^d \times \mathbf{R} \times \mathbf{R}^d \times S^{d \times d} \rightarrow \mathbf{R}$ be a continuous function that is elliptic, i.e., $F(t, x, u, p, X) \geq F(t, x, u, p, Y)$ for $X \succcurlyeq Y$. Then $u \in C^2([0, \infty[\times \mathbf{R}^d)$ is a solution of the partial differential equation*

$$\frac{\partial u}{\partial t}(t, x) = F(t, x, u(t, x), \nabla u(t, x), \nabla^2 u(t, x)) \quad (5.11)$$

in $]0, \infty[\times \mathbf{R}^d$ if and only if the following conditions are satisfied:

1. For all $\varphi \in C^2([0, \infty[\times \mathbf{R}^d)$ and for all local maxima (t_0, x_0) of the function $u - \varphi$,

$$\frac{\partial \varphi}{\partial t}(t_0, x_0) \leq F(t_0, x_0, u(t_0, x_0), \nabla \varphi(t_0, x_0), \nabla^2 \varphi(t_0, x_0)).$$

2. For all $\varphi \in C^2([0, \infty[\times \mathbf{R}^d)$ and for all local minima (t_0, x_0) of the function $u - \varphi$,

$$\frac{\partial \varphi}{\partial t}(t_0, x_0) \geq F(t_0, x_0, u(t_0, x_0), \nabla \varphi(t_0, x_0), \nabla^2 \varphi(t_0, x_0)).$$

Proof Let $\varphi \in \mathcal{C}^2([0, \infty[\times \mathbf{R}^d)$ and let (t_0, x_0) be a local maximum of the function $u - \varphi$. By the classical necessary conditions for a maximum, we obtain

$$\begin{aligned}\nabla(u - \varphi)(t_0, x_0) &= 0, & \text{i.e.,} & \quad \nabla u(t_0, x_0) = \nabla \varphi(t_0, x_0), \\ \frac{\partial(u - \varphi)}{\partial t}(t_0, x_0) &= 0, & \text{i.e.,} & \quad \frac{\partial u}{\partial t}(t_0, x_0) = \frac{\partial \varphi}{\partial t}(t_0, x_0), \\ \nabla^2(u - \varphi)(t_0, x_0) &\preccurlyeq 0, & \text{i.e.,} & \quad \nabla^2 u(t_0, x_0) \preccurlyeq \nabla^2 \varphi(t_0, x_0).\end{aligned}$$

Hence, by ellipticity,

$$\begin{aligned}\frac{\partial \varphi}{\partial t}(t_0, x_0) &= \frac{\partial u}{\partial t}(t_0, x_0) = F(t_0, x_0, u(t_0, x_0), \nabla u(t_0, x_0), \nabla^2 u(t_0, x_0)) \\ &\leq F(t_0, x_0, u(t_0, x_0), \nabla u(t_0, x_0), \nabla^2 \varphi(t_0, x_0)) \\ &= F(t_0, x_0, u(t_0, x_0), \nabla \varphi(t_0, x_0), \nabla^2 \varphi(t_0, x_0)).\end{aligned}$$

Claim 2 is proven similarly.

Conversely, we assume 1 and 2. Then we set $u = \varphi$ and get local maxima and minima in the whole domain $[0, \infty[\times \mathbf{R}^d$. Consequently,

$$\begin{aligned}\frac{\partial u}{\partial t}(t, x) &\leq F(t, x, u(t, x), \nabla u(t, x), \nabla^2 u(t, x)), \\ \frac{\partial u}{\partial t}(t, x) &\geq F(t, x, u(t, x), \nabla u(t, x), \nabla^2 u(t, x)),\end{aligned}$$

and this implies

$$\frac{\partial u}{\partial t}(t, x) = F(t, x, u(t, x), \nabla u(t, x), \nabla^2 u(t, x))$$

on the whole of $[0, \infty[\times \mathbf{R}^d$. □

The characterization of solutions in Theorem 5.15 has the notable feature that claims 1 and 2 do not need any differentiability of u . If u is merely continuous, we can still decide whether a point is a local maximum or minimum of $u - \varphi$. Moreover, the respective inequalities use only derivatives of the test functions φ . Put differently, one implication of the theorem says if 1 and 2 hold for continuous u , then u is also a solution of equation (5.11) if u has the additional regularity $u \in \mathcal{C}^2([0, \infty[\times \mathbf{R}^d)$. If we weaken the regularity assumption, we obtain the definition of viscosity solutions.

Definition 5.16 Let F be as in Theorem 5.15 and let $u \in \mathcal{C}([0, \infty[\times \mathbf{R}^d)$. Then we have the following:

1. u is a *viscosity sub-solution* if for every $\varphi \in \mathcal{C}^2([0, \infty[\times \mathbf{R}^d)$ and every local maximum (t_0, x_0) of $u - \varphi$, one has

$$\frac{\partial \varphi}{\partial t}(t_0, x_0) \leq F(t_0, x_0, u(t_0, x_0), \nabla \varphi(t_0, x_0), \nabla^2 \varphi(t_0, x_0)),$$

2. u is a *viscosity super-solution* if for every $\varphi \in \mathcal{C}^2([0, \infty[\times \mathbf{R}^d)$ and every local minimum (t_0, x_0) of $u - \varphi$, one has

$$\frac{\partial \varphi}{\partial t}(t_0, x_0) \geq F(t_0, x_0, u(t_0, x_0), \nabla \varphi(t_0, x_0), \nabla^2 \varphi(t_0, x_0)),$$

3. u is a *viscosity solution* if u is both a viscosity sub-solution and viscosity super-solution.

For the special case that the function F depends only on ∇u and $\nabla^2 u$, we have the following helpful lemma:

Lemma 5.17 *Let $F : \mathbf{R}^d \times S^{d \times d} \rightarrow \mathbf{R}$ be continuous and elliptic, i.e., $F(p, X) \geq F(p, Y)$ for $X \succcurlyeq Y$. A function $u \in \mathcal{C}([0, \infty[\times \mathbf{R}^d)$ is a viscosity sub- or super-solution, respectively, if for all $f \in \mathcal{C}_b^\infty(\mathbf{R}^d)$ and $g \in \mathcal{C}_b^\infty(\mathbf{R})$, the respective part in Definition 5.16 holds for $\varphi(t, x) = f(x) + g(t)$.*

Proof We show the case of a viscosity sub-solution and assume that Definition 5.16 holds for all φ of the form $\varphi(t, x) = f(x) + g(t)$ with f, g as given. Without loss of generality we assume that $(t_0, x_0) = (0, 0)$ and consider a function $\varphi \in \mathcal{C}^2([0, \infty[\times \mathbf{R}^d)$ such that $u - \varphi$ has a maximum in $(0, 0)$. Hence, we have to show that

$$\frac{\partial \varphi}{\partial t}(0, 0) \leq F(\nabla \varphi(0, 0), \nabla^2 \varphi(0, 0)).$$

We consider the Taylor expansion of φ at $(0, 0)$ and get with $a = \varphi(0, 0)$, $b = \frac{\partial \varphi}{\partial t}(0, 0)$, $p = \nabla \varphi(0, 0)$, $c = \frac{1}{2} \frac{\partial^2 \varphi}{\partial t^2}(0, 0)$, $Q = \frac{1}{2} \nabla^2 \varphi(0, 0)$, and $q = \frac{1}{2} \frac{\partial}{\partial t} \nabla \varphi(0, 0)$ that

$$\varphi(t, x) = a + bt + p \cdot x + ct^2 + x^T Qx + tq \cdot x + o(|x|^2 + t^2).$$

We define, for all $\varepsilon > 0$, the functions $f \in \mathcal{C}_b^\infty(\mathbf{R}^d)$ and $g \in \mathcal{C}_b^\infty(\mathbf{R})$ for small values of x and t by

$$\begin{aligned} f(x) &= a + p \cdot x + x^T Qx + \varepsilon(1 + \frac{|q|}{2})|x|^2, \\ g(t) &= bt + (\frac{|q|}{2\varepsilon} + \varepsilon + c)t^2. \end{aligned}$$

The boundedness of f and g can be ensured with the help of a suitable cutoff function w as in (5.10) in the proof of Theorem 5.11. Hence, for small values of x and t , we have

$$\varphi(t, x) = f(x) + g(t) - \left(\frac{\varepsilon|q|}{2}|x|^2 + \frac{|q|}{2\varepsilon}t^2 - tq \cdot x + \varepsilon(|x|^2 + t^2) \right) + o(|x|^2 + t^2).$$

Because of

$$\begin{aligned} \frac{\varepsilon|q|}{2}|x|^2 + \frac{|q|}{2\varepsilon}t^2 - tq \cdot x &\geq \varepsilon \frac{|q|}{2}|x|^2 + \frac{|q|}{2\varepsilon}t^2 - t|q||x| \\ &= \left(\sqrt{\frac{|q|\varepsilon}{2}}|x| - \sqrt{\frac{|q|}{2\varepsilon}}t \right)^2 \geq 0 \end{aligned}$$

we obtain, for small x and t , that $\varphi(t, x) \leq f(x) + g(t)$. Hence, in a neighborhood of $(0, 0)$, we also get $u(t, x) - \varphi(t, x) \geq u(t, x) - f(x) - g(t)$, and in particular we see that $u - f - g$ has a local maximum at $(0, 0)$. By assumption we get

$$\frac{\partial(f+g)}{\partial t}(0, 0) \leq F(\nabla(f+g)(0, 0), \nabla^2(f+g)(0, 0)).$$

Now we note that

$$\begin{aligned} \frac{\partial(f+g)}{\partial t}(0, 0) &= \frac{\partial\varphi}{\partial t}(0, 0), & \nabla(f+g)(0, 0) &= \nabla\varphi(0, 0), \\ \nabla^2(f+g)(0, 0) &= \nabla^2\varphi(0, 0) + 2\varepsilon(1 + |q|) \text{id}, \end{aligned}$$

and conclude that

$$\frac{\partial\varphi}{\partial t}(0, 0) \leq F(\nabla\varphi(0, 0), \nabla^2\varphi(0, 0) + 2\varepsilon(1 + |q|) \text{id}).$$

Since this inequality holds for every $\varepsilon > 0$, we can, thanks to the continuity of F , pass to the limit $\varepsilon = 0$ and obtain the claim. The case of viscosity super-solutions is proved in the same way. \square

Now we are able to prove that the scale spaces from Theorem 5.11 are viscosity solutions of partial differential equations.

Theorem 5.18 *Let (\mathcal{T}_t) be a scale space that satisfies the axioms [TRANS], [COMP], [GLSI], [REC], [REG], and [LOC]. Let the infinitesimal generator of (\mathcal{T}_t) be denoted by $A[u] = F(\nabla u, \nabla^2 u)$ and let $u_0 \in C_b^\infty(\mathbf{R}^d)$. Then $u(t, x) = (\mathcal{T}_t u_0)(x)$ is a viscosity solution of*

$$\frac{\partial u}{\partial t}(t, x) = F(\nabla u(t, x), \nabla^2 u(t, x))$$

with initial condition $u(0, x) = u_0(x)$.

Proof Theorem 5.11 and Lemma 5.13 ensure that the generator has the stated form. Now we show that u is a viscosity sub-solution. The proof that u is also a viscosity super-solution is similar. Let $\varphi \in C^2([0, \infty[\times \mathbf{R}^d)$ be such that (t_0, x_0) with $t_0 > 0$ is a local maximum of $u - \varphi$. Without loss of generality we can assume that $u(t_0, x_0) =$

$\varphi(t_0, x_0), u \leq \varphi$ and by Lemma 5.17 it is enough to consider $\varphi(t, x) = f(x) + g(t)$ with $f \in \mathcal{C}_b^\infty(\mathbf{R}^d)$ and $g \in \mathcal{C}_b^\infty(\mathbf{R})$.

By [REC] we can write for $0 < h \leq t_0$,

$$\varphi(t_0, x_0) = u(t_0, x_0) = \mathcal{T}_h(u(t_0 - h, \cdot))(x_0).$$

By [COMP] and [GLSI] we get

$$\begin{aligned} f(x_0) + g(t_0) &= \varphi(t_0, x_0) \\ &= \mathcal{T}_h(u(t_0 - h, \cdot))(x_0) \\ &\leq \mathcal{T}_h(\varphi(t_0 - h, \cdot))(x_0) \\ &= \mathcal{T}_h(f)(x_0) + g(t_0 - h). \end{aligned}$$

Rearranging gives

$$\frac{g(t_0) - g(t_0 - h)}{h} \leq \frac{\mathcal{T}_h(f) - f}{h}(x_0).$$

Since $f \in \mathcal{C}_b^\infty(\mathbf{R}^d)$ and g are differentiable, we can pass to the limit $h \rightarrow 0$ and get, by Theorem 5.11, $g'(t_0) \leq F(\nabla f(x_0), \nabla^2 f(x_0))$. Since $\varphi(t, x) = f(x) + g(t)$, we have also

$$\frac{\partial \varphi}{\partial t}(t_0, x_0) \leq F(\nabla \varphi(t_0, x_0), \nabla^2 \varphi(t_0, x_0)).$$

And hence, by definition, u is a viscosity sub-solution. □

5.2 Standard Models Based on Partial Differential Equations

We have now completed our structural analysis of scale spaces. We have seen that scale spaces naturally (i.e., given the specific axioms) lead to functions $F : \mathbf{R}^d \times \mathbf{R} \times \mathbf{R}^d \times S^{d \times d} \rightarrow \mathbf{R}$ and that these functions characterize the respective scale space. In this section we will see how the morphological axioms influence F and discover some differential equations that are important in imaging.

5.2.1 Linear Scale Spaces: The Heat Equation

The main result of this chapter will be that among the linear scale spaces there is essentially only the heat equation.

Theorem 5.19 (Uniqueness of the Heat Equation) *Let \mathcal{T}_t be a scale space that satisfies [TRANS], [COMP], [GLSI], [REC], [LOC], [ISO], and [REG] (with uniform constant $C(u, v)$ in [REG]). If the maps \mathcal{T}_t are in addition linear, then there exists $c > 0$ such that $F(\nabla u, \nabla^2 u) = c\Delta u$. In other words, $u(t, x) = (\mathcal{T}_t u_0)(x)$ satisfies the heat equation*

$$\begin{aligned}\partial_t u - c\Delta u &= 0 \quad \text{in } \mathbf{R}^+ \times \mathbf{R}^d, \\ u(0, \cdot) &= u_0 \quad \text{in } \mathbf{R}^d.\end{aligned}$$

Proof By Theorem 5.9 the infinitesimal generator A exists, and by Theorem 5.11 and Lemma 5.13 it has the form

$$A[u] = F(\nabla u, \nabla^2 u).$$

We aim to prove that $F : \mathbf{R}^d \times S^{d \times d} \rightarrow \mathbf{R}$ is actually

$$F(p, X) = c \operatorname{trace} X \quad \text{for some } c > 0.$$

1. Linearity of \mathcal{T}_t implies that the infinitesimal generator is also linear, i.e.,

$$F(\lambda p + \mu q, \lambda X + \mu Y) = \lambda F(p, X) + \mu F(q, Y).$$

Especially, we have

$$\begin{aligned}F(p, X) &= F(p + 0, 0 + X) = F(p, 0) + F(0, X) \\ &= F_1(p) + F_2(X).\end{aligned}$$

2. By [ISO] we have for every isometry $R \in \mathbf{R}^{d \times d}$ that

$$A[D_R u] = D_R A[u],$$

since a similar equation holds for every \mathcal{T}_t . For F this means

$$F((\nabla(D_R u))(x), (\nabla^2(D_R u))(x)) = F((D_R \nabla u)(x), (D_R \nabla^2 u)(x)).$$

Since $(\nabla(D_R u))(x) = R^T \nabla u(Rx)$ and $(\nabla^2(D_R u))(x) = R^T \nabla^2 u(Rx)R$, we obtain $F(R^T p, R^T X R) = F(p, X)$. Referring to the first step, we have arrived at

$$F_1(R^T p) = F_1(p), \tag{A}$$

$$F_2(R X R^T) = F_2(X). \tag{B}$$

3. Now (A) implies that $F_1(p) = f(|p|)$. Linearity of F implies F_1 is also linear and thus, $F_1(p) \equiv 0$. From (B) we deduce that F_2 depends only on quantities that are invariant under similarity transforms. This is the set of eigenvalues of X with their multiplicities. Since all eigenvalues have the same role, linearity leads to

$$F_2(X) = h(\text{trace } X) = c \text{ trace } X$$

for some $c \in \mathbf{R}$.

4. By [COMP] we get that for $X \succcurlyeq Y$, we have $F(p, X) \geq F(p, Y)$. This implies that for $X \succcurlyeq Y$, we have

$$\begin{aligned} c \text{ trace } X &\geq c \text{ trace } Y, \\ \text{i.e. } c \text{ trace}(X - Y) &\geq 0. \end{aligned}$$

Hence $c \geq 0$, and the proof is complete. \square

The heat equation plays a central role in the area of image processing based on partial differential equations. One can say that the heat equation is, essentially, the only linear partial differential equation that is used.

Remark 5.20 The scale space based on the heat equation does not satisfy [GSI], i.e., it is not contrast invariant. This can be seen as follows: Let u be a solution of the differential equation $\partial_t u - \Delta u = 0$ and let $u = h(v)$ with a differentiable and increasing gray value transformation $h : \mathbf{R} \rightarrow \mathbf{R}$. Then

$$0 = \partial_t(h(v)) - \Delta(h(v)) = h' \partial_t v - h' \Delta v - h'' |\nabla v|^2.$$

If [GSI] were satisfied, then we would have $\partial_t v - \Delta v = 0$, which holds only if $h'' = 0$ (i.e., linear gray value transformations are allowed).

Thus, linearity and [GSI] contradict each other: A scale space that satisfies [GSI] has to be nonlinear. We will see examples for contrast invariant scale spaces in the next sections.

Remark 5.21 The heat equation has very good smoothing properties in theory (for $t > 0$ we have that $u(t, \cdot)$ is infinitely differentiable), but it is not well suited to denoise images. On the one hand, this is because all information in the image is smoothed in the same way, including edges. Another important point is the dislocation of information, i.e., that location of coarse scale structure (i.e., large t) does not correspond to the location of the related fine-scale structure; see Fig. 5.7.

One way around this drawback is to modify the differential equation such that the process preserves edges better. This will be accomplished by the Perona-Malik equation, which we will introduce and study in Sect. 5.3.1.

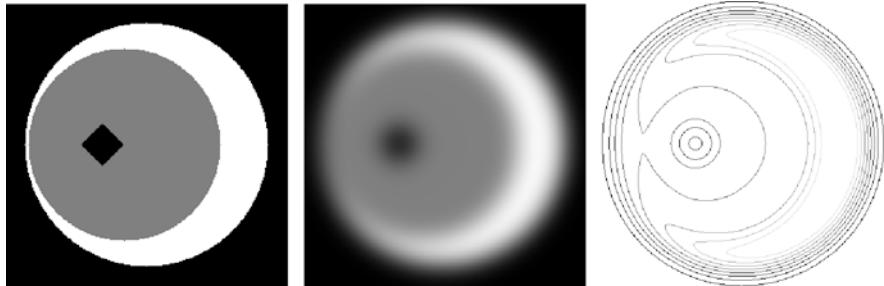


Fig. 5.7 Dislocation of information by the heat equation illustration by the movement of edges. Left: original image, middle: image after application of the heat equation, right: level lines of the middle image

5.2.2 Morphological Scale Space

Among the linear scale spaces, there is essentially only the heat equation. In Remark 5.20 we saw that linearity and contrast invariance are mutually exclusive. In this section we study the consequences of contrast invariance, i.e., of the axiom [GSI]. This means that we look for scale spaces that are invariant under contrast changes. In other words, the scale space depends only on the level sets, i.e., only on the shape of objects and not on the particular gray values, hence the name “morphological equations.”

Lemma 5.22 *Let $F : \mathbf{R}^d \setminus \{0\} \times S^{d \times d} \rightarrow \mathbf{R}$ be continuous. A scale space*

$$\partial_t u = F(\nabla u, \nabla^2 u)$$

satisfies [GSI] if and only if F satisfies the following invariance: for all $p \neq 0$, $X \in S^{d \times d}$, and all $\lambda \in \mathbf{R}$, $\mu \geq 0$, one has (with $p \otimes p = pp^T$) that

$$F(\mu p, \mu X + \lambda p \otimes p) = \mu F(p, X). \quad (*)$$

Proof ([GSI] \implies ()):* Let $h : \mathbf{R} \rightarrow \mathbf{R}$ be a twice continuously differentiable and non-decreasing gray value transformation. Moreover, let $\mathcal{T}_t(h(u_0)) = h(\mathcal{T}_t(u_0))$ be satisfied. Then $u = \mathcal{T}_t(u_0)$ is a solution of

$$\begin{aligned} \partial_t u &= F(\nabla u, \nabla^2 u), \\ u(0) &= u_0, \end{aligned} \quad (\text{I})$$

and $h(u)$ a solution of

$$\begin{aligned} \partial_t(h \circ u) &= F(\nabla(h \circ u), \nabla^2(h \circ u)), \\ (h \circ u)(0) &= h(u_0). \end{aligned} \quad (\text{II})$$

Using the chain rule, we obtain

$$\nabla(h \circ u) = h' \nabla u$$

$$\nabla^2(h \circ u) = h' \nabla^2 u + h'' \nabla u \otimes \nabla u$$

(which you should derive in Exercise 5.7). Equation (II) becomes

$$h' \partial_t u = F(h' \nabla u, h' \nabla^2 u + h'' \nabla u \otimes \nabla u).$$

With (I) we obtain

$$h' F(\nabla u, \nabla^2 u) = F(h' \nabla u, h' \nabla^2 u + h'' \nabla u \otimes \nabla u).$$

Since u is an arbitrary solution and h is also arbitrary and non-decreasing, we can choose $\nabla u = p$, $\nabla^2 u = X$, $h' = \mu \geq 0$ and $h'' = \lambda$ (cf. the techniques in the proof of Theorem 5.11).

((*) \implies [GSI]): Conversely, assume that (*) is satisfied. Moreover, let u be a solution of (I) and h a non-decreasing gray value transformation. We have to show that $v = h \circ u$ is also a solution of (I) with initial value $h \circ u_0$. By the chain rule one gets

$$\partial_t v = h' \partial_t u = h' F(\nabla u, \nabla^2 u).$$

By (*) we get, again using the chain rule (similarly to the previous part), that

$$h' F(\nabla u, \nabla^2 u) = F(\nabla(h \circ u), \nabla^2(h \circ u)),$$

which proves the assertion. \square

Theorem 5.23 Let \mathcal{T}_t be a scale space that satisfies [GSI] and [COMP] and that has an infinitesimal generator. For $p \in \mathbf{R}^d \setminus \{0\}$ let $Q_p = (\text{id} - \frac{p \otimes p}{|p|^2})$ be the projection onto the subspace perpendicular to p . Then for $X \in S^{d \times d}$ and $p \in \mathbf{R}^d \setminus \{0\}$, one has

$$F(p, X) = F(p, Q_p X Q_p).$$

Proof Since X can be chosen independently of $p \neq 0$, we first prove the claim in the special case $p = c e_d$, where $e_d = (0, \dots, 0, 1)$ denotes the d th unit vector and $c \neq 0$. In this case, we have

$$p \otimes p = c^2 \begin{pmatrix} 0 & \dots & 0 & 0 \\ \vdots & & \vdots & \vdots \\ 0 & \dots & 0 & 0 \\ 0 & \dots & 0 & 1 \end{pmatrix}, \quad Q_p = \begin{pmatrix} 1 & & & \\ & \ddots & & \\ & & 1 & \\ & & & 0 \end{pmatrix}.$$

In particular, we get from Lemma 5.22, that $F(p, \cdot)$ does not depend on the entry $X_{d,d}$. It remains to show that the entries $X_{d,i}$ with $1 \leq i \leq d-1$ also do not play a role. To that end, we define $M = x_{d,1}^2 + \cdots + x_{d,d-1}^2$ and

$$I_\varepsilon = \begin{pmatrix} \varepsilon & & & \\ & \ddots & & \\ & & \varepsilon & \\ & & & \frac{M}{\varepsilon} \end{pmatrix}.$$

In Exercise 5.8 you will show that

$$\begin{aligned} Q_p X Q_p &\preccurlyeq X + I_\varepsilon \\ X &\preccurlyeq Q_p X Q_p + I_\varepsilon. \end{aligned}$$

Axiom [COMP] implies ellipticity of F . Since F does not depend on the entry $X_{d,d}$, we get, on the one hand, that

$$F(p, Q_p X Q_p) \leq F(p, X + I_\varepsilon) = F(p, X + \varepsilon \text{id}),$$

and on the other hand, that

$$F(p, X) \leq F(p, Q_p X Q_p + I_\varepsilon) = F(p, Q_p X Q_p + \varepsilon \text{id}).$$

Sending ε to zero in both equations, we get, as claimed,

$$F(p, X) = F(p, Q_p X Q_p).$$

The general case $p \neq 0$ follows similarly with an additional orthogonal change of coordinates that sends p to a multiple of e_d . \square

The matrix $Q_p X Q_p$ maps p to zero and leaves the space perpendicular to p invariant. Since $Q_p X Q_p$ is also symmetric, one eigenvalue is zero and all $d-1$ other eigenvalues are also real. If we now add the axiom [ISO] we obtain that the infinitesimal generator depends only on the magnitude of p and the eigenvalues:

Theorem 5.24 *In addition to the assumptions in Theorem 5.23 let [ISO] be satisfied. For $p \in \mathbf{R}^d \setminus \{0\}$ and $X \in S^{d \times d}$ denote by $\lambda_1, \dots, \lambda_{d-1}$ the nonzero eigenvalues of $Q_p X Q_p$. Then*

$$F(p, X) = F_1(|p|, \lambda_1, \dots, \lambda_{d-1}).$$

Moreover, F_1 is symmetric in the eigenvalues and non-decreasing.

Proof First we note that we have seen in step 2 of the proof of Theorem 5.19 that [ISO] implies that for every isometry $R \in \mathbf{R}^{d \times d}$,

$$F(p, X) = F(R^T p, R^T X R). \quad (*)$$

1. We fix p and choose an isometry that leaves p invariant, i.e., $Rp = p$ and $R^T p = p$. With Lemma 5.22 and $(*)$ we get

$$F(p, X) = F(p, Q_p X Q_p) = F(R^T p, R^T Q_p X Q_p R) = F(p, R^T Q_p X Q_p R).$$

Hence, the function F can depend only on the eigenvalues of $Q_p X Q_p$ on the space orthogonal to p . Since p is an eigenvector of Q_p for the eigenvalue zero, there is a function G such that

$$F(p, X) = G(p, \lambda_1, \dots, \lambda_{d-1}).$$

2. Now let R denote any isometry and set $q = R^T p$. Then also $Rq = p$ and $|p| = |q|$. We calculate

$$\begin{aligned} RQ_q &= R(\text{id} - \frac{qq^T}{|q|^2}) \\ &= R - \frac{pq^T}{|q|^2} \\ &= (\text{id} - \frac{pp^T}{|p|^2})R = Q_p R, \end{aligned}$$

i.e., $RQ_q = Q_p R$. This implies $Q_q R^T X R Q_q = R^T Q_p X Q_p R$, and we see that $Q_q R^T X R Q_q$ and $Q_p X Q_p$ have the same eigenvalues. By $(*)$ and the first point we obtain

$$G(p, \lambda_1, \dots, \lambda_{d-1}) = G(R^T p, \lambda_1, \dots, \lambda_{d-1}).$$

Thus, G depends only on the magnitude of p , as claimed. \square

To get a better understanding of the consequences of axiom [GSI], we shall have a look at some examples:

The Equations for Erosion and Dilation

We consider the unit ball $B_1(0) \subset \mathbf{R}^d$ as a structure element and the corresponding scale spaces for erosion and dilation:

$$\mathcal{E}_t u_0 = u_0 \ominus tB,$$

$$\mathcal{D}_t u_0 = u_0 \oplus tB.$$

The infinitesimal generators are

$$F(p, X) = -|p| \text{ for erosion and } F(p, X) = |p| \text{ for dilation.}$$

Hence, the corresponding differential equations are

$$\begin{aligned} \partial_t u &= -|\nabla u| \text{ for erosion,} \\ \partial_t u &= |\nabla u| \text{ for dilation.} \end{aligned} \tag{5.12}$$

By Theorem 5.15 we immediately get that these differential equations are solved by erosion and dilation, respectively, in the viscosity sense. Since erosion and dilation produce non-differentiable functions in general, one cannot define classical solutions here, and we see how helpful the notion of viscosity solutions is.

Remark 5.25 (Interpretation as a Transport Equation) To understand the equations in a qualitative way, we interpret them as *transport equations*. We write the infinitesimal generator for the dilation as

$$|\nabla u| = \frac{\nabla u}{|\nabla u|} \cdot \nabla u.$$

Recall that the equation $\partial_t u = v \cdot \nabla u$ with initial value $u(0) = u_0$ is solved by $u(t, x) = u_0(x + tv)$. This means that the initial value u_0 is shifted in the direction $-v$. Similarly, one can say that erosion and dilation shift the initial value in the direction of the negative gradient and that the velocity is either -1 or 1 , respectively. Hence, we can describe both examples by the movement of the level sets. For dilation, the level lines are shifted in the direction of the negative gradient. The corresponding equation is also called the *grassfire equation*, since the level lines move like the contours of a grass fire; the burning/bright regions expand. Erosion behaves similarly: the dark areas expand uniformly in all directions.

The Mean Curvature Motion

Another important contrast-invariant scale space is given by the generator

$$F(p, X) = \text{trace}(Q_p X). \tag{5.13}$$

Since $Q_{\mu p} = Q_p$ and $Q_p p = 0$, we conclude that

$$F(\mu p, \mu X + \lambda p \otimes p) = \text{trace}(Q_{\mu p}(\mu X + \lambda pp^T)) = \mu \text{trace}(Q_p X) = \mu F(p, X).$$

We see that the invariance from Lemma 5.22 is satisfied. The differential operator related to F is

$$F(\nabla u, \nabla^2 u) = \text{trace}\left((\text{id} - \frac{\nabla u \otimes \nabla u}{|\nabla u|^2}) \nabla^2 u\right).$$

To better understand this quantity we use $\nabla u \otimes \nabla u = \nabla u \nabla u^T$, linearity of the trace, and the formula $\text{trace}(A B C) = \text{trace}(C A B)$ (invariance under cyclic shifts, if the dimensions fit):

$$\begin{aligned} F(\nabla u, \nabla^2 u) &= \Delta u - \frac{1}{|\nabla u|^2} \text{trace}(\nabla u \nabla u^T \nabla^2 u) \\ &= \Delta u - \frac{1}{|\nabla u|^2} \text{trace}(\nabla u^T \nabla^2 u \nabla u) = \Delta u - \frac{1}{|\nabla u|^2} \nabla u^T \nabla^2 u \nabla u. \end{aligned}$$

We can describe this expression in so-called *local coordinates* in a simpler way. We define the direction

$$\eta = \frac{\nabla u}{|\nabla u|}$$

and denote by $\partial_\eta u$ the first derivative of u in the direction η and by $\partial_{\eta\eta} u$ the second derivative, i.e.,

$$\partial_\eta u = |\nabla u|, \quad \partial_{\eta\eta} u = \eta^T \nabla^2 u \eta.$$

Hence,

$$F(\nabla u, \nabla^2 u) = \Delta u - \partial_{\eta\eta} u.$$

We recall that the Laplace operator is rotationally invariant and note that $\Delta u - \partial_{\eta\eta} u$ is the sum of the second derivatives of u in the $d - 1$ directions perpendicular to the gradient. Since the gradient is normal to the level sets of u , we see that $\Delta u - \partial_{\eta\eta} u$ is the projection of the Laplace operator onto the space tangential to the level sets. Hence, the differential equation

$$\partial_t u = \text{trace}(Q_{\nabla u} \nabla^2 u) = \Delta u - \partial_{\eta\eta} u$$

is also called the *heat equation on the tangent space* or the *morphological equivalent of the heat equation*.

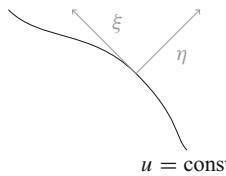
The case $d = 2$ is even easier, since the space orthogonal to η is one-dimensional. We define

$$\xi = \eta^\perp = \begin{pmatrix} -\eta_2 \\ \eta_1 \end{pmatrix}$$

and note that

$$\partial_\xi u = 0, \quad \partial_{\xi\xi} u = \xi^T \nabla^2 u \xi.$$

Since η is the normalized gradient, it is orthogonal to the level sets $\{u = \text{const}\}$, and the situation in local coordinates (η, ξ) looks as follows:



In particular, we have $\Delta u = \partial_{\eta\eta} u + \partial_{\xi\xi} u$ and thus

$$\text{trace}(Q_{\nabla u} \nabla^2 u) = \partial_{\xi\xi} u.$$

In Exercise 5.9 you shall show that $\partial_{\xi\xi} u$ is related to the curvature κ of the levelset; more precisely,

$$\partial_{\xi\xi} u = |\nabla u| \kappa.$$

We summarize: the differential equation of the generator (5.13) has the form

$$\partial_t u = |\nabla u| \kappa.$$

In analogy to Remark 5.25, we may interpret this equation as a transport equation

$$\partial_t u = \kappa \frac{\nabla u}{|\nabla u|} \nabla u,$$

and we see that the initial value (as in the case of dilation and erosion) is shifted in the direction of the negative gradient. In this case, the velocity is proportional to the curvature of the levelsets, and the resulting scale space is also called *curvature motion*. As a result, the level sets are “straightened”; this explains the alternative name *curve shortening flow*. One can even show that the boundaries of convex and compact sets shrink to a point in finite time and that they look like a circle, asymptotically; see, for example, [79]. The action of the curvature motion on an image is shown in Fig. 5.8.

In higher dimensions, similar claims are true: here $\text{trace}(Q_{\nabla u} \nabla^2 u) = |\nabla u| \kappa$ with the *mean curvature* κ ; this motivates the name *mean curvature motion*. For the definition and further properties of the mean curvature we refer to [12].

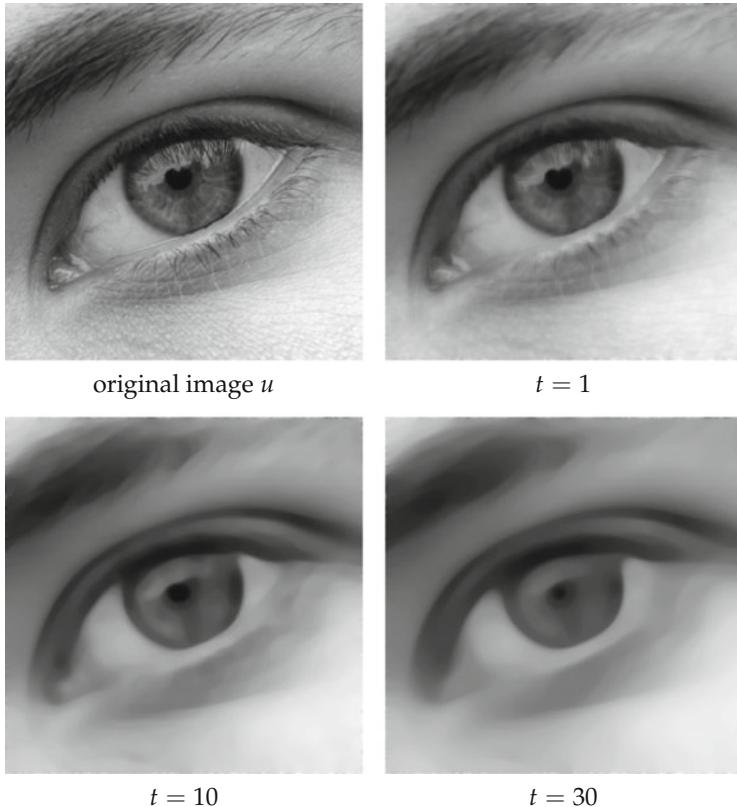


Fig. 5.8 Illustration of the curvature motion at different scales t

5.3 Nonlinear Diffusion

The heat equation satisfies the most axioms among the linear methods, but it is not well suited to denoise images. Its main drawback is the heavy smoothing of edges, see Remark 5.21. In this section we will treat denoising methods that are variations of the heat equation. We build the modifications of the heat equation on its interpretation as a diffusion process: Let u describe some quantity (e.g., heat of metal or the concentration of salt in a liquid) in d dimensions. A concentration gradient induces a flow j from high to low concentration:

$$j = -A \nabla u. \quad (\text{Fick's law})$$

The matrix A is called a *diffusion tensor*. If A is a multiple of the identity, one speaks of *isotropic* diffusion, if not, of *anisotropic* diffusion. The diffusion tensor controls how fast and in which direction the flow goes.

Moreover, we assume that the overall concentration remains constant, i.e., that no quantity appears or vanishes. Mathematically we describe this as follows. For some volume V we consider the change of the overall concentration in this volume:

$$\partial_t \int_V u(x) dx.$$

If the total concentration stays the same, this change has to be equal to the flow across the boundary of V by the flow j , i.e.,

$$\partial_t \int_V u(x) dx = \int_{\partial V} j \cdot (-v) d\tilde{\mathfrak{H}}^{d-1}.$$

Interchanging integration and differentiation on the left-hand side and using the divergence theorem (Theorem 2.81) on the right-hand side, we obtain

$$\int_V \partial_t u(x) dx = - \int_V (\operatorname{div} j)(x) dx.$$

Since this holds for all volumes, we get that the integrands are equal at almost every point:

$$\partial_t u = - \operatorname{div} j. \quad (\text{continuity equation})$$

Plugging the continuity equation into Fick's law ,we obtain the following equation for u :

$$\partial_t u = \operatorname{div}(A \nabla u).$$

This equation is called a *diffusion equation*. If A is independent of u , the left-hand side of the diffusion equation is linear, and we speak of linear diffusion.

5.3.1 The Perona-Malik Equation

The idea of Perona and Malik in [110] was to slow down diffusion at edges. As we have seen in Application 3.23, edges can be described as points where the gradient has a large magnitude. Hence, we steer the diffusion tensor A in such a way that it slows down diffusion where gradients are large. Since we don't have any reason for anisotropic diffusion yet, we set

$$A = g(|\nabla u|) \operatorname{id}$$

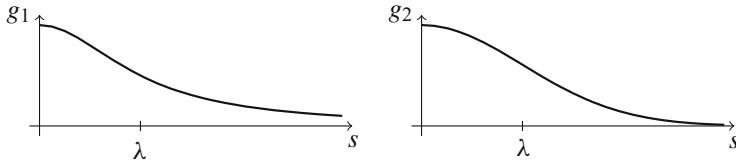


Fig. 5.9 Functions g from (5.14) in the diffusion coefficient of the Perona-Malik equation

with some function $g : [0, \infty[\rightarrow [0, \infty[$ that is close to one for small arguments and monotonically decreasing to zero. Consequently, diffusion acts as it does in the heat equation at places with small gradient, and diffusion is slower at places with large gradient. Two widely used examples of such functions, depending on a parameter $\lambda > 0$, are

$$g_1(s) = \frac{1}{1 + \frac{s^2}{\lambda^2}}, \quad g_2(s) = e^{-\frac{s^2}{2\lambda^2}}, \quad (5.14)$$

see Fig. 5.9.

The parameter λ says how fast the function tends to zero. In conclusion, the Peron-Malik equation reads as

$$\begin{aligned} \partial_t u &= \operatorname{div}(g(|\nabla u|)\nabla u), \\ u(0, x) &= u_0(x). \end{aligned} \quad (5.15)$$

Figure 5.10 illustrates that the Perona-Malik equation indeed has the desired effect.

We begin our analysis of the Perona-Malik equation with the following observation:

Lemma 5.26 *We have*

$$\operatorname{div}(g(|\nabla u|)\nabla u) = \frac{g'(|\nabla u|)}{|\nabla u|} \nabla u^T \nabla^2 u \nabla u + g(|\nabla u|) \Delta u,$$

and thus the Perona-Malik equation has the following infinitesimal generator:

$$F(p, X) = \frac{g'(|p|)}{|p|} p^T X p + g(|p|) \operatorname{trace} X.$$

Proof You shall check this in Exercise 5.10. □

Remark 5.27 The Perona-Malik equation is an example of non-linear isotropic diffusion, since the diffusion tensor is a multiple of the identity. The decomposition in Lemma 5.26, however, shows that diffusion acts differently in the directions η

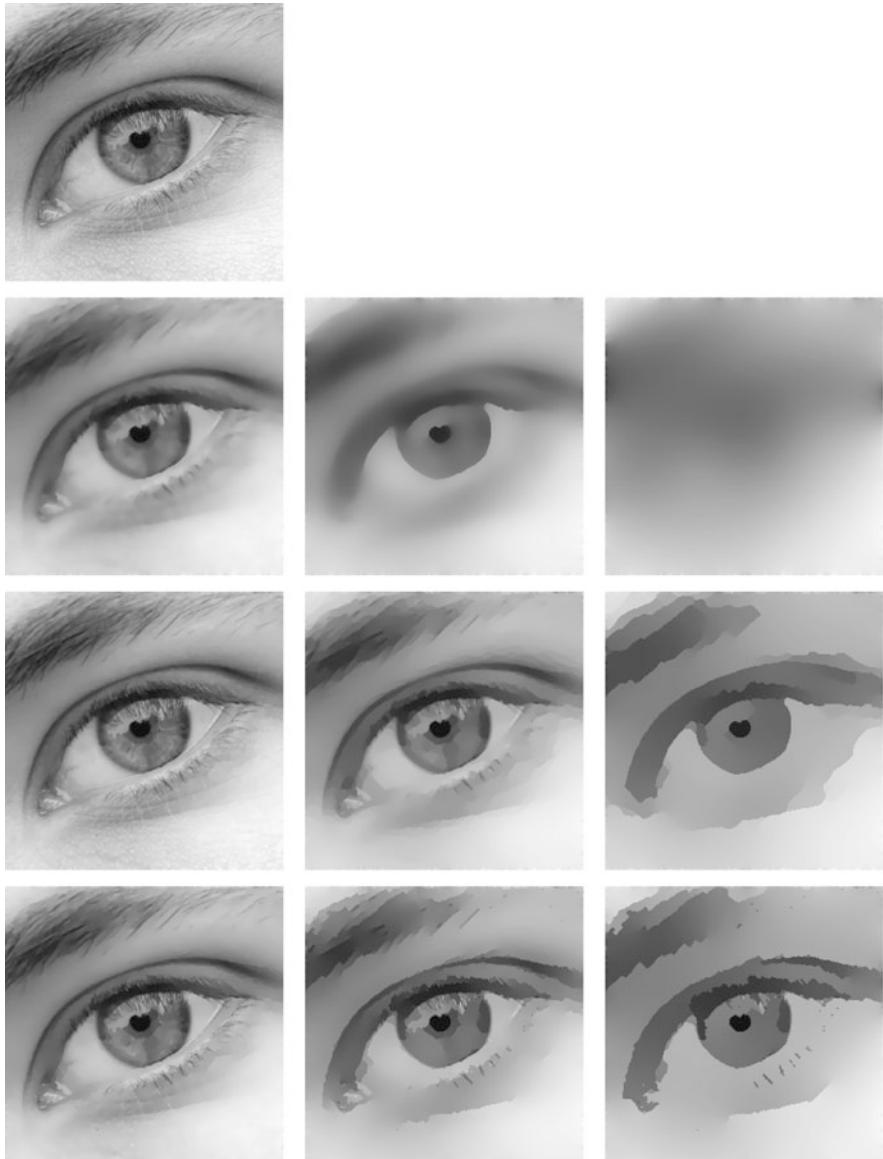


Fig. 5.10 Results for the Perona-Malik equation. Top left: original image. Second row: Function g_1 with $\lambda = 0.02$. Third row: function g_1 with $\lambda = 0.005$. Fourth row: Function g_2 with $\lambda = 0.02$. The images show the results at times $t = 5, 50, 500$, respectively

and orthogonal to η . For that reason some authors call the Perona-Malik equation “anisotropic diffusion.”

The function F with g_1 and g_2 from (5.14) is not elliptic. This leads to a problem, since our current results do not imply that the Perona-Malik equation has any solution, even in the viscosity sense. In fact, one can even show that the equation does not possess any solution for certain initial values [83, 86]. Roughly speaking, the reason for this is that one cannot expect or show that the gradients remain bounded, and hence the diffusion coefficient cannot be bounded away from zero. Some experience with parabolic equations hints that this should lead to problems with existence of solutions.

We leave the problem aside for now and assume that the Perona-Malik equation has solutions, at least for small time.

Theorem 5.28 *The Perona-Malik method satisfies the axioms [REC], [GLSI], [TRANS], [ISO]. The axiom [SCALE] holds if g is a monomial.*

Proof Recursivity [REC] holds, since the operators \mathcal{T}_t are solution operators of a differential equation, and hence possess the needed semi-group property.

For gray value shift invariance [GLSI] we note, that if u solves the differential equation (5.15), then $u + c$ solves the same differential equation with initial value $u_0(x) + c$. This implies $\mathcal{T}_t(u_0 + c) = \mathcal{T}_t(u_0) + c$ as desired. (In other words, the differential operator is invariant with respect to linear gray level shifts.)

Translation invariance [TRANS] and isometry invariance [ISO] can be seen similarly (the differential operator is invariant with respect to translations and rotations).

If g is a monomial, i.e., $g(s) = s^p$, then $v(t, x) = u(t, \lambda x)$ satisfies

$$\operatorname{div}(g(|\nabla v|)\nabla v)(x) = |\lambda|^p \lambda \operatorname{div}(g(|\nabla u|)\nabla u)(\lambda x).$$

Thus, with $t' = t/(|\lambda|^p \lambda)$, we have

$$\mathcal{T}_t(u_0(\lambda \cdot)) = (\mathcal{T}_{t'}u_0)(\lambda \cdot),$$

as claimed. □

Since the map F is not elliptic in general, we cannot use the theory of viscosity solutions in this case. However, we can consider the equation on some domain $\Omega \subset \mathbf{R}^d$ (typically on a rectangle in \mathbf{R}^2) and add boundary conditions, which leads us to a so-called *boundary initial value problem*. (One can also consider viscosity solutions for differential equations on domains; however, the formulation of boundary values is intricate.) For the boundary initial value problem for the Perona-Malik equation one can show that a maximum principle is satisfied:

Theorem 5.29 Let $\Omega \subset \mathbf{R}^d$, $u_0 \in L^2(\Omega)$, $g : [0, \infty[\rightarrow [0, \infty[$ continuous, and $u : [0, \infty[\times \Omega \rightarrow \mathbf{R}$ a solution of the boundary initial value problem

$$\begin{aligned}\partial_t u &= \operatorname{div}(g(|\nabla u|)\nabla u), && \text{in } [0, \infty[\times \Omega, \\ \partial_\nu u &= 0 && \text{on } [0, \infty[\times \partial\Omega, \\ u(0, x) &= u_0(x) && \text{for } x \in \Omega.\end{aligned}$$

Then for all $p \in [2, \infty]$, one has

$$\|u(t, \cdot)\|_p \leq \|u_0\|_p,$$

and moreover,

$$\int_{\Omega} u(t, x) \, dx = \int_{\Omega} u_0(x) \, dx.$$

Proof For $p < \infty$ we set $h(t) = \int_{\Omega} |u(t, x)|^p \, dx$ and differentiate h :

$$\begin{aligned}h'(t) &= \int_{\Omega} \frac{d}{dt} |u(t, x)|^p \, dx \\ &= \int_{\Omega} p|u(t, x)|^{p-2} u(t, x) \partial_t u(t, x) \, dx \\ &= \int_{\Omega} p|u(t, x)|^{p-2} u(t, x) \operatorname{div}(g(|\nabla u|)\nabla u)(t, x) \, dx.\end{aligned}$$

We use the divergence theorem (Theorem 2.81) and get

$$\begin{aligned}&\int_{\Omega} p|u(t, x)|^{p-2} u(t, x) \operatorname{div}(g(|\nabla u|)\nabla u)(t, x) \, dx \\ &= \int_{\partial\Omega} p|u(t, x)|^{p-2} u(t, x) g(|\nabla u(t, x)|) \partial_\nu u(t, x) \, d\mathfrak{H}^{d-1} \\ &\quad - p(p-1) \int_{\Omega} |u(t, x)|^{p-2} g(|\nabla u(t, x)|) |\nabla u(t, x)|^2 \, dx.\end{aligned}$$

The boundary integral vanishes due to the boundary condition, and the other integral is nonnegative. Hence, $h'(t) \leq 0$, i.e., h is decreasing, which proves the claim for $p < \infty$, and also $h(t)^{1/p} = \|u(t, \cdot)\|_p$ is decreasing for all $p \in [2, \infty[$. The case $p = \infty$ now follows by letting $p \rightarrow \infty$.

For the second claim we argue similarly and differentiate the function $\mu(t) = \int_{\Omega} u(t, x) dx$, again using the divergence theorem to get

$$\begin{aligned}\mu'(t) &= \int_{\Omega} \partial_t u(t, x) dx \\ &= \int_{\Omega} 1 \cdot \operatorname{div}(g(|\nabla u|) \nabla u)(t, x) dx \\ &= \int_{\partial\Omega} g(|\nabla u(t, x)|) \partial_{\nu} u(t, x) d\mathfrak{H}^{d-1} - \int_{\Omega} (\nabla 1) \cdot (g(|\nabla u(t, x)|) \nabla u(t, x)) dx.\end{aligned}$$

The claim follows, since both terms are zero. \square

The fact that the Perona-Malik equation reduces the norm $\|\cdot\|_{\infty}$ is also called *maximum principle*.

Now we analyze the Perona-Malik equation in local coordinates. With $\eta = \nabla u / |\nabla u|$ we get from Lemma 5.26

$$\begin{aligned}\operatorname{div}(g(|\nabla u|) \nabla u) &= g'(|\nabla u|) |\nabla u| \partial_{\eta\eta} u + g(|\nabla u|) \Delta u \\ &= (g'(|\nabla u|) |\nabla u| + g(|\nabla u|)) \partial_{\eta\eta} u + g(|\nabla u|) (\Delta u - \partial_{\eta\eta} u),\end{aligned}$$

i.e., we have decomposed the operator $\operatorname{div}(g(|\nabla u|) \nabla u)$ into a part perpendicular to the levelsets and a part tangential to the level sets. The tangential part is $g(|\nabla u|) (\Delta u - \partial_{\eta\eta} u)$, and it has a positive coefficient $g(|\nabla u|)$. The perpendicular part is described with the help of the flux function $f(s) = sg(s)$ as $f'(|\nabla u|) \partial_{\eta\eta} u$. The flux functions for the g 's from (5.14) are

$$f_1(s) = \frac{s}{1 + \frac{s^2}{\lambda^2}}, \quad f_2(s) = s e^{-\frac{s^2}{2\lambda^2}}; \quad (5.16)$$

see Fig. 5.11.

In both cases $f'(s)$ becomes negative for $s > \lambda$. We note that the Perona-Malik equation behaves like *backward diffusion* in the direction η at places with large gradient (i.e. potential places for edges). In the directions tangential to the level

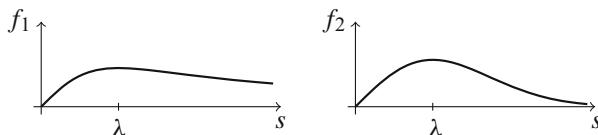


Fig. 5.11 Flux functions f_1 and f_2 from (5.16) to the functions g_1 and g_2 from (5.14)

lines we see that the coefficient $g(|\nabla u|)$ is small for large gradient, leading to slow diffusion. We may conjecture three things:

- The Perona-Malik equation has the ability to make edges steeper (i.e., sharper).
- The Perona-Malik equation is unstable, since it has parts that behave like backward diffusion.
- At steep edges, noise reduction may not be good.

We try to get some rigorous results and ask, what happens to solutions of the Perona-Malik equation at edges.

To answer this question we follow [85] and consider only the one-dimensional equation. We denote by u' the derivative with respect to x and consider

$$\partial_t u = (g(|u'|)u')'. \quad (5.17)$$

By the chain rule we get

$$\partial_t u = (g'(|u'|)|u'| + g(|u'|))u''.$$

Hence, the one-dimensional equation behaves like the equation in higher dimensions in the direction perpendicular to the level lines, and we can use results for the one-dimensional equation to deduce similar properties for the case in higher dimensions. To analyze what happens at edges under the Perona-Malik equation, we define an edge as follows:

Definition 5.30 We say that $u : \mathbf{R} \rightarrow \mathbf{R}$ has an *edge* at x_0 if

1. $|u'(x_0)| > 0$;
2. $u''(x_0) = 0$;
3. $u'(x_0)u'''(x_0) < 0$.

The first condition says that the image is not flat, while the second and the third conditions guarantee a certain kind of inflection point; see Fig. 5.12.

Since in one dimension we have

$$\partial_\eta u = |u'|, \quad \partial_{\eta\eta} u = u'', \quad \partial_{\eta\eta\eta} u = \operatorname{sgn}(u')u''',$$

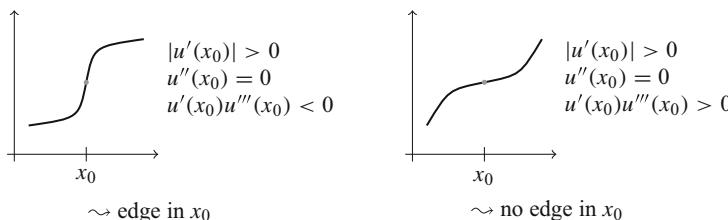


Fig. 5.12 Left: inflection point of the first kind (edge), right: inflection point of the second kind (no edge)

we can characterize an edge by

1. $\partial_\eta u > 0$;
2. $\partial_{\eta\eta} u = 0$;
3. $\partial_{\eta\eta\eta} u < 0$.

Using this notation, we can write the Perona-Malik equation (5.17) as

$$\partial_t u = (g'(\partial_\eta u)\partial_\eta u + g(\partial_\eta u))\partial_{\eta\eta} u = f'(\partial_\eta u)\partial_{\eta\eta} u.$$

In the following we will use the compact notation $u_\eta = \partial_\eta u$, $u_{\eta\eta} = \partial_{\eta\eta} u$, etc. The next theorem shows what happens locally at an edge.

Theorem 5.31 *Let u be a five times continuously differentiable solution of the Perona-Malik equation (5.17). Moreover, let x_0 be an edge of $u(t_0, \cdot)$ where additionally $u_{\eta\eta\eta\eta}(t_0, x_0) = 0$ is satisfied. Then at (t_0, x_0) ,*

1. $\partial_t u_\eta = f'(u_\eta)u_{\eta\eta\eta}$,
2. $\partial_t u_{\eta\eta} = 0$, and
3. $\partial_t u_{\eta\eta\eta} = 3f''(u_\eta)u_{\eta\eta\eta}^2 + f'(u_\eta)u_{\eta\eta\eta\eta\eta}$.

Proof We calculate the derivatives by swapping the order of differentiation:

$$\partial_t u_\eta = \partial_\eta u_t = \partial_\eta(f'(u_\eta)u_{\eta\eta}) = \partial_\eta(f'(u_\eta))u_{\eta\eta} + f'(u_\eta)u_{\eta\eta\eta}.$$

In a similar way we obtain

$$\partial_t u_{\eta\eta} = \partial_{\eta\eta}(f'(u_\eta))u_{\eta\eta} + 2\partial_\eta(f'(u_\eta))u_{\eta\eta\eta} + f'(u_\eta)u_{\eta\eta\eta\eta}$$

and

$$\begin{aligned} \partial_t u_{\eta\eta\eta} &= \partial_{\eta\eta\eta}(f'(u_\eta))u_{\eta\eta} + \partial_{\eta\eta}(f'(u_\eta))u_{\eta\eta\eta} \\ &+ 2(\partial_{\eta\eta}(f'(u_\eta))u_{\eta\eta\eta} + \partial_\eta(f'(u_\eta))u_{\eta\eta\eta\eta}) + \partial_\eta(f'(u_\eta))u_{\eta\eta\eta\eta} + f'(u_\eta)u_{\eta\eta\eta\eta\eta}. \end{aligned}$$

Since there is an edge at x_0 , all terms $u_{\eta\eta\eta}$ and $u_{\eta\eta\eta\eta\eta}$ vanish. With $\partial_\eta(f'(u_\eta)) = f''(u_\eta)u_{\eta\eta}$ and $\partial_{\eta\eta}(f'(u_\eta)) = f'''(u_\eta)u_{\eta\eta}^2 + f''(u_\eta)u_{\eta\eta\eta}$ we obtain the claim. \square

The second assertion of the theorem says that edges remain inflection points. The first assertion indicates whether the edge gets steeper or less steep with increasing t . The third assertion tells us, loosely speaking, whether the inflection point tries to change its type. The specific behavior depends on the functions f and g .

Corollary 5.32 *In addition to the assumptions of Theorem 5.31 assume that the diffusion coefficient g is given by one of the functions (5.14). Moreover, assume that*

$u_{\eta\eta\eta\eta\eta} > 0$ holds at the edge. Then in (t_0, x_0) :

1. $\partial_t u_\eta > 0$ if $u_\eta > \lambda$ and $\partial_t u_\eta < 0$ if $u_\eta < \lambda$;
2. $\partial_t u_{\eta\eta} = 0$;
3. $\partial_t u_{\eta\eta\eta} < 0$ if $\lambda < u_\eta < \sqrt{3}\lambda$.

Proof We consider the function g_2 from (5.14) and f'_2 from (5.16).

Assertion 2 is independent of the function g and follows directly from Theorem 5.31.

Since $u_{\eta\eta\eta} < 0$ holds at an edge, we have by Theorem 5.31, 1, that $\partial_t u_\eta > 0$ if and only if $f'_2(u_\eta) < 0$. From (5.16) one concludes that this holds if and only if $u_\eta > \lambda$. Otherwise, we have $f'_2(u_\eta) > 0$ and thus $\partial_t u_\eta < 0$, which proves assertion 1.

To see assertion 3, we need the second derivative of the flux function:

$$f''_2(s) = -\left(\frac{3s}{\lambda^2} - \frac{s^3}{\lambda^4}\right)e^{-\frac{s^2}{2\lambda^2}}.$$

By Theorem 5.31, 3, one has

$$\partial_t u_{\eta\eta\eta} = 3f''(u_\eta)u_{\eta\eta\eta}^2 + f'(u_\eta)u_{\eta\eta\eta\eta\eta}.$$

Since $u_{\eta\eta\eta\eta\eta} > 0$, we conclude that $f'_2(u_\eta) < 0$ for $u_\eta > \lambda$ and $f''_2(u_\eta) < 0$ for $u_\eta < \sqrt{3}\lambda$, which implies assertion 3.

For g_1 from (5.14) and f'_1 from (5.16) we have that $f'_1(u_\eta) < 0$ if and only if $u_\eta > \lambda$ (which proves assertion 1), and moreover,

$$f''_1(s) = \frac{-(3 - \frac{s^2}{\lambda^2})\frac{2s}{\lambda^2}}{\left(1 + \frac{s^2}{\lambda^2}\right)^3}.$$

Similarly to the case g_2 , we also have $f'_1(u_\eta) < 0$ for $u_\eta > \lambda$ and $f''_1(u_\eta) < 0$ for $u_\eta < \sqrt{3}\lambda$. \square

We may interpret the corollary as follows:

- The second point says that inflection points remain inflection points.
- The first point says that steep edges get steeper and flat edges become flatter. More precisely: Edges no steeper than λ become flatter.
- If an edge is steeper than λ , there are two possibilities: If it is no steeper than $\sqrt{3}\lambda$, the inflection point remains an inflection point of the first kind, i.e. the edge stays an edge. If the edge is steeper than $\sqrt{3}\lambda$, the inflection point may try to change its type ($u_{\eta\eta\eta}$ can grow and become positive). This potentially leads to the so-called *staircasing effect*; see Fig. 5.13.

Using a similar technique, one can derive a condition for a maximum principle of the one-dimensional Perona-Malik equation; see Exercise 5.11.

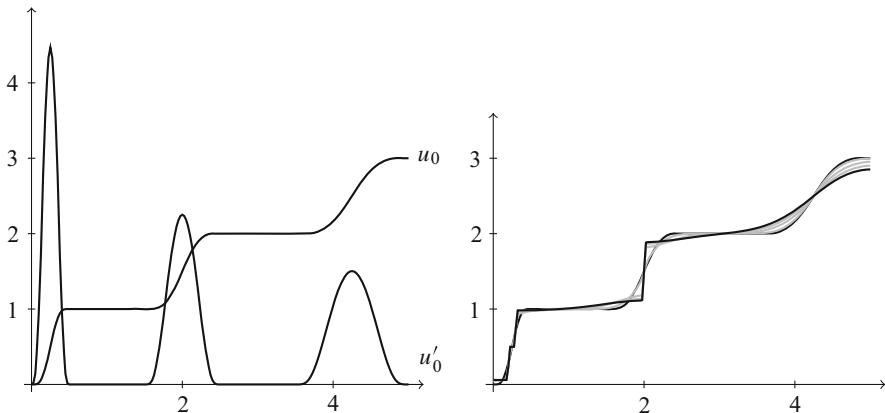


Fig. 5.13 Illustration of the staircasing effect, sharpening and smoothing by the Perona-Malik equation with the function g_2 and $\lambda = 1.6$. Left: Initial value of the Perona-Malik equation and its derivative. Right: solution of the Perona-Malik equation at times $t = 0.5, 2, 3.5, 5$. We see all predicted effects from Corollary 5.32: the first edge is steeper than $\sqrt{3}\lambda$ and indeed we observe staircasing. The middle edge has a slope between λ and $\sqrt{3}\lambda$ and becomes steeper, while the right edge has a slope below λ and is flattened out

Application 5.33 (Perona-Malik as Preprocessing for Edge Detection) The presence of noise severely hinders automatic edge detection. Methods based on derivatives (such as the Canny edge detection from Application 3.23) find many edges in the noise. This can be avoided by a suitable presmoothing. Smoothing by the heat equation is not suitable, since this may move the position of edges quite a bit (see Remark 5.21). A presmoothing by the Perona-Malik equation gives better results; see Fig. 5.14. We also note one drawback of the Perona-Malik equation: there is no smoothing along the edges, and this results in noise at the edges. This reduces the quality of edge detection.

Finally, we turn to the problem that the Perona-Malik equation does not have solutions in general. From an analytical perspective, the problem is that gradients do not stay bounded, and thus the diffusion coefficient may come arbitrarily close to zero. One way to circumvent this difficulty is to slightly change the argument of the function g . It turns out that a slight smoothing is enough. For simplicity, assume that the domain is a square: $\Omega = [0, 1]^2$. For some $\sigma > 0$ we define the Gauss kernel G_σ as in (3.2). For some $u \in L^2(\Omega)$ let $u_\sigma = G_\sigma * u$, where we have extended u by symmetry to the whole of \mathbf{R}^2 (compare Fig. 3.11). We consider the modified Perona-Malik equation

$$\begin{aligned} \partial_t u &= \operatorname{div}(g(|\nabla u_\sigma|)\nabla u) && \text{in } [0, \infty[\times \Omega \\ \partial_\nu u &= 0 && \text{on } [0, \infty[\times \partial\Omega \\ u(0, x) &= u_0(x) && \text{for } x \in \Omega. \end{aligned} \tag{5.18}$$

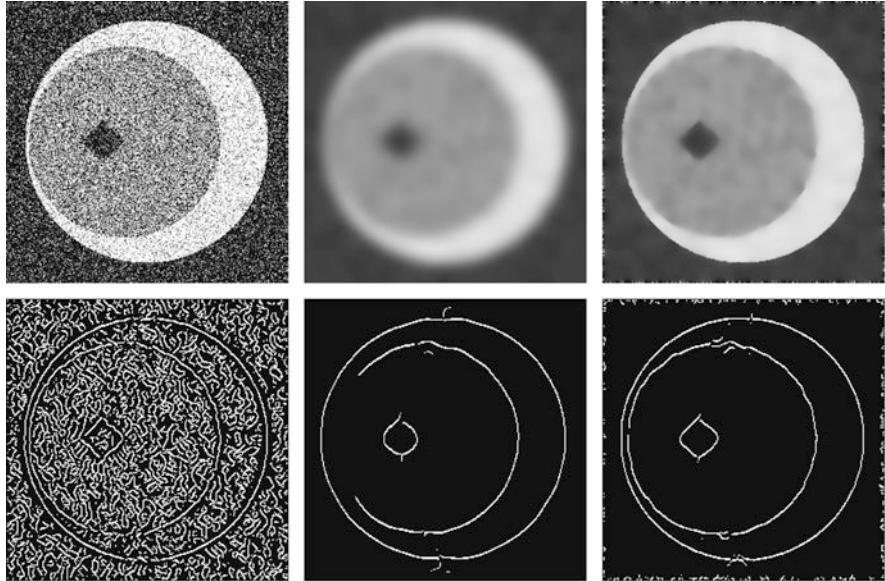


Fig. 5.14 Detection of edges in noisy images. Left column: Noisy image with gray values in $[0, 1]$ and the edges detected by Canny's edge detector from Application 3.23 (parameters: $\sigma = 2$, $\tau = 0.01$). Middle column: Presmoothing by the heat equation, final time $T = 20$. Right column: Presmoothing by the Perona-Malik equation with the function g_1 , final time $T = 20$ and $\lambda = 0.05$

The only difference from the original Perona-Malik equation is the smoothing of u in the argument of g . To show existence of a solution, we need a different notion from that of viscosity solutions, namely the notion of *weak solutions*. This notion is based on an observation that is similar to that in Theorem 5.15.

Theorem 5.34 *Let $A : \Omega \rightarrow \mathbf{R}^{d \times d}$ be differentiable and $T > 0$. Then $u \in \mathcal{C}^2([0, \infty[\times \Omega)$ with $u(t, \cdot) \in L^2(\Omega)$ is a solution of the partial differential equation*

$$\begin{aligned}\partial_t u &= \operatorname{div}(A \nabla u) && \text{in } [0, T] \times \Omega \\ A \nabla u \cdot v &= 0 && \text{on } [0, T] \times \partial\Omega\end{aligned}$$

if and only if for every function $v \in H^1(\Omega)$ and every $t \in [0, 1]$ one has

$$\int_{\Omega} (\partial_t u(t, x)) v(x) dx = - \int_{\Omega} (A(x) \nabla u(t, x)) \cdot \nabla v(x) dx.$$

Proof Let u be a solution of the differential equation with the desired regularity and $v \in H^1(\Omega)$. Multiplying both sides of the differential equation by v , integrating

over Ω , and performing partial integration leads to

$$\begin{aligned} \int_{\Omega} (\partial_t u(t, x)) v(x) dx &= \int_{\Omega} (\operatorname{div}(A \nabla u))(t, x) v(x) dx \\ &= \int_{\partial\Omega} v(x) (A(x) \nabla u(t, x)) \cdot \nu d\mathfrak{H}^{d-1} - \int_{\Omega} (A(x) \nabla u(t, x)) \cdot \nabla v(x) dx. \end{aligned}$$

The boundary integral vanishes because of the boundary condition, and this implies the claim.

Conversely, let the equation for the integral be satisfied for all $v \in H^1(\Omega)$. Similar to the above calculation we get

$$\int_{\Omega} (\partial_t u - \operatorname{div}(A \nabla u))(t, x) v(x) dx = - \int_{\partial\Omega} v(x) (A(x) \nabla u(t, x)) \cdot \nu d\mathfrak{H}^{d-1}.$$

Since v is arbitrary, we can conclude that the integrals on both sides have to vanish. Invoking the fundamental lemma of the calculus of variations (Lemma 2.75) establishes the claim. \square

The characterization of solutions in the above theorem does not use the assumption $u \in \mathcal{C}^2([0, \infty[\times \Omega)$, and also differentiability of A is not needed. The equality of the integrals can be formulated for functions $u \in \mathcal{C}^1([0, T], L^2(\Omega)) \cap L^2(0, T; H^1(\Omega))$. The following reformulation allows us to get rid of the time derivative of u : we define a bilinear form $a : H^1(\Omega) \times H^1(\Omega) \rightarrow \mathbf{R}$ by

$$a(u, v) = \int_{\Omega} (A \nabla u) \cdot v dx.$$

Now we define the notion of weak solution of an initial-boundary value problem:

Definition 5.35 (Weak Solutions) Let $u_0 \in L^2(\Omega)$ and $A \in L^\infty(\Omega, \mathbf{R}^{d \times d})$. A function $u \in L^2(0, T, H^1(\Omega)) \cap \mathcal{C}([0, T], L^2(\Omega))$ is called a *weak solution* of the initial-boundary value problem

$$\begin{aligned} \partial_t u &= \operatorname{div}(A \nabla u) && \text{in } [0, T] \times \Omega, \\ A \nabla u \cdot \nu &= 0 && \text{on } [0, T] \times \partial\Omega, \\ u(0) &= u_0, \end{aligned}$$

if for all $v \in H^1(\Omega)$

$$\frac{d}{dt}(u(t), v) + a(u(t), v) = 0,$$

$$u(0) = u_0.$$

This form of the initial-boundary value problem is called the *weak formulation*.

Remark 5.36 The time derivative in the weak formulation has to be understood in the weak sense as described in Sect. 2.3. In more detail: the first equation of the weak formulation states that for all $v \in H^1(\Omega)$ and $\phi \in \mathcal{D}(]0, 1[)$

$$\int_0^T \left[a(u(t), v)\phi(t) - (u(t), v)\phi'(t) \right] dt = 0.$$

Remark 5.37 The modification proposed in (5.18) can be interpreted in a different way: the diffusion coefficient $g(|\nabla u|)$ should act as an edge detector and slow down diffusion at edges. Our investigation of edges in Application 3.23 showed that some presmoothing makes edge detection more robust. Hence, the proposed diffusion coefficient $g(|\nabla u_\sigma|)$ takes a more robust edge detector than the classical Perona-Malik model. This motivation, as well as the model itself, goes back to [30].

The modified Perona-Malik equation (5.18) does have weak solutions, as has been shown in [30].

Theorem 5.38 Let $u_0 \in L^\infty(\Omega)$, $\sigma > 0$, $T > 0$, and $g : [0, \infty[\rightarrow [0, \infty[$ be infinitely differentiable. Then the equation

$$\begin{aligned} \partial_t u &= \operatorname{div}(g(|\nabla u_\sigma|)\nabla u) && \text{in } [0, \infty[\times \Omega \\ \partial_\nu u &= 0 && \text{on } [0, \infty[\times \partial\Omega \\ u(0, x) &= u_0(x) && \text{for } x \in \Omega \end{aligned}$$

has a unique weak solution $u : [0, T] \rightarrow L^2(\Omega)$. Moreover, u as a mapping from $[0, T]$ to $L^2(\Omega)$ is continuous, and for almost all $t \in [0, T]$, one has $u(t) \in H^1(\Omega)$.

The proof is based on Schauder's fixed point theorem [22] and uses deep results from the theory of linear partial differential equations. The (weak) solutions of the modified Perona-Malik equation have similar properties to the (in general non-existent) solutions of the original Perona-Malik equation. In particular, Theorem 5.29 holds similarly, as you shall show in Exercise 5.12.

Example 5.39 (Denoising with the Modified Perona-Malik Equation) We consider the modified Perona-Malik equation as in Theorem 5.38 and aim to see its performance for denoising of images. As with the standard Perona-Malik equation (5.15), we choose g as a decreasing function from (5.14). Again $\lambda > 0$ acts as a threshold for edges. However, this time we have to take into account that the argument of g is the magnitude of the smoothed gradient. This is, depending on the smoothing parameter σ , smaller than the magnitude of the gradient itself. Thus, the parameter λ should be chosen in dependence on σ . A rule of thumb is, the parameter σ should be adapted to the noise level in u_0 such that $u_0 * G_\sigma$ is roughly noise-free. Then one inspects the magnitude of the gradient of $u_0 * G_\sigma$ and chooses λ such that the dominant edges have a gradient above λ .

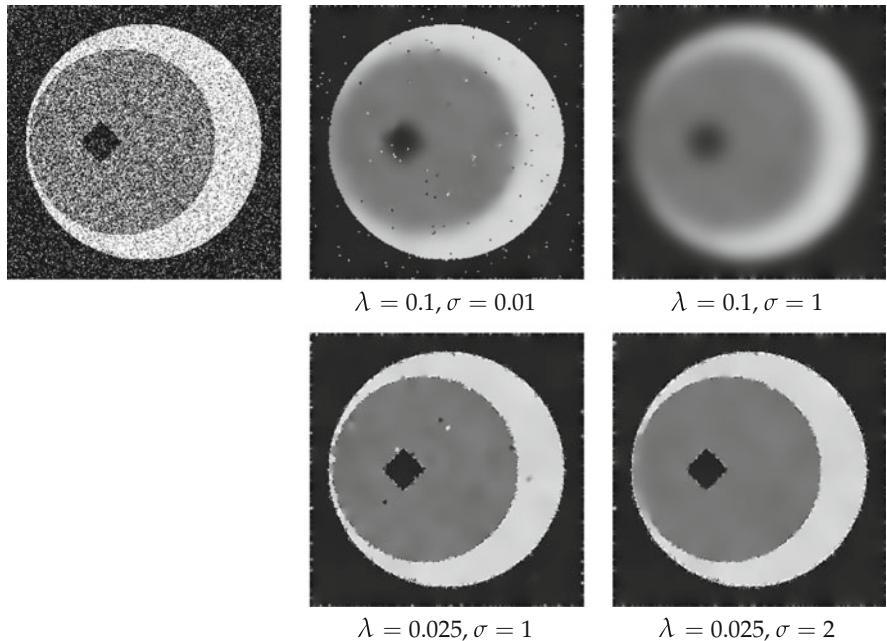


Fig. 5.15 Denoising with the modified Perona-Malik equation

Figure 5.15 shows the denoising capabilities of the modified Perona-Malik equation. The function g_2 from Eq. (5.14) has been used with final time $T = 50$. For very small values of σ (0.01) one does not get good results: for $\lambda = 0.1$, some edges are quite blurred, while the noise has not been removed well. If the value of σ is increased without adjusting λ appropriately, one sees, as expected, that the magnitudes of the gradient at edges fall below the edge threshold due to presmoothing and hence are smoothed away. A suitable adaption at λ to σ leads to good results, as in the case of $\sigma = 1$ and $\sigma = 2$. For larger σ the effect that the Perona-Malik equation cannot remove noise at steep edges is pronounced.

In conclusion, we collect the following insight into the Perona-Malik equation:

- The choice $u * G_\sigma$ as edge detector makes sense, since on the one hand, it gives a more robust edge detector, while on the other hand, it also allows for a rigorous theory for solutions.
- The choice of the parameters λ , σ , and T is of great importance for the result. For the smoothing parameter σ and the edge threshold λ there exist reasonable rules of thumb.
- The modified Perona-Malik equation shows the predicted properties in practice: homogeneous regions are denoised, edges are preserved, and at steep edges, noise is not properly eliminated.

Example 5.40 (Color Images and the Perona-Malik Equation) We use the example of nonlinear diffusion according to Perona-Malik to illustrate some issues that arise in the processing of color images. We consider a color image with three color channels $u_0 : \Omega \rightarrow \mathbf{R}^3$ (cf. Sect. 1.1). Alternatively, we could also consider three separate images or $u_0 : \Omega \times \{1, 2, 3\} \rightarrow \mathbf{R}$, where the image $u_0(\cdot, k)$ is the k th color channel. If we want to apply the Perona-Malik equation to this image, we have several possibilities for choosing the diffusion coefficient. One naive approach would be to apply the Perona-Malik equation to all channels separately:

$$\partial_t u(t, x, k) = \operatorname{div}(g(|\nabla u(\cdot, \cdot, k)|)\nabla u(\cdot, \cdot, k))(t, x), \quad k = 1, 2, 3.$$

This may lead to problems, since it is not clear whether the different color channels have their edges at the same positions. This can be seen clearly in Fig. 5.16. The image there consists of a superposition of slightly shifted, blurred circles in the RGB channels. After the Perona-Malik equation has been applied to all color channels, one can clearly see the shifts. One way to get around this problem is to use the HSV color system. However, there is another problem: edges do not need to have the same slope in the different channels, and this may lead to further errors in the colors. In the HSV system, the V-channel carries the most information, and often it

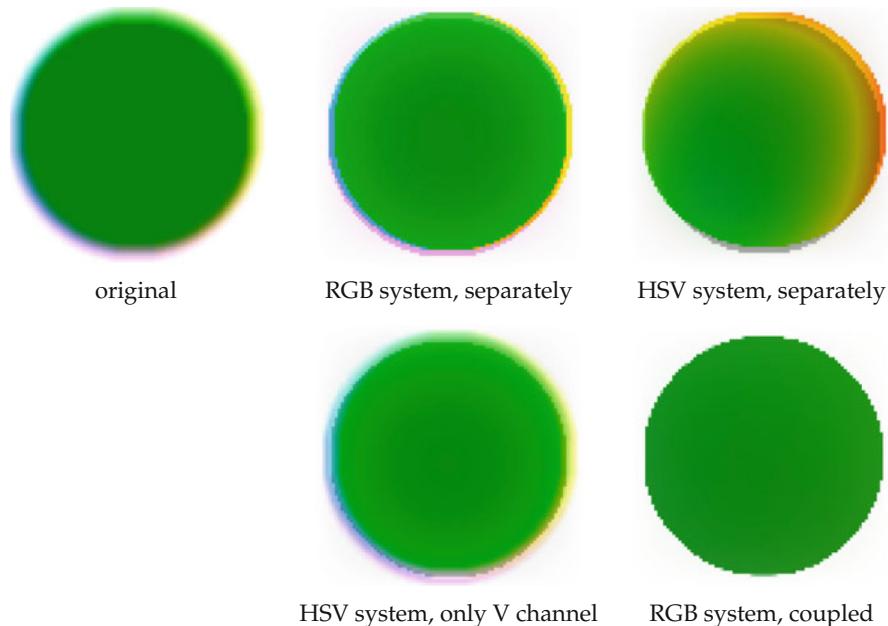


Fig. 5.16 Nonlinear Perona-Malik diffusion for color images. The original image consists of slightly shifted and blurred circles with different intensities in the three color RGB channels. The color system and the choice of the diffusion coefficient play a significant role. The best results are achieved when the diffusion coefficients are coupled among the channels

is enough just to denoise this channel; however, there may still be errors. All these effects can be seen in Fig. 5.16.

Another possibility is to remember the role of the diffusion coefficient as an edge detector. Since being an edge is not a property of a single color channel, edges should be at the same places in all channels. Hence, one should choose the diffusion coefficient to be equal in all channels, for example by using the average of the magnitudes of the gradients:

$$\partial_t u(t, x, k) = \operatorname{div} \left(g \left(\frac{1}{3} \left| \sum_{i=1}^3 \nabla u(\cdot, \cdot, i) \right|^2 \right) \nabla u(\cdot, \cdot, k) \right)(t, x), \quad k = 1, 2, 3.$$

Hence, the diffusion coefficient is coupled among the channels. Usually this gives the best results. The effects can also be seen in real images, e.g., in images where so-called *chromatic aberration* is present. This refers to the effect that is caused by the fact that light rays of different wavelengths are refracted differently. This effect can be observed, for example, in lenses of low quality; see Fig. 5.17.

5.3.2 Anisotropic Diffusion

The Perona-Malik equation has shown good performance for denoising and simultaneous preservation of edges. Smoothing along edges has not been that good, though. This drawback can be overcome by switching to an anisotropic model. The basic idea is to design a diffusion tensor that enables Perona-Malik-like diffusion perpendicular to the edges but linear diffusion along the edges. We restrict ourselves to the two-dimensional case, since edges are curves in this case and there is only one direction along the edges. The development of methods based on anisotropic diffusion goes back to [140].

The diffusion tensor should encode as much local image information as possible. We follow the modified model (5.18) and take ∇u_σ as an edge detector. As preparation for the following, we define the structure tensor:

Definition 5.41 (Structure Tensor) The *structure tensor* for $u : \mathbf{R}^2 \rightarrow \mathbf{R}$ and noise level $\sigma > 0$ is the matrix-valued function $J_0(\nabla u_\sigma) : \mathbf{R}^2 \rightarrow \mathbf{R}^{2 \times 2}$ defined by

$$J_0(\nabla u_\sigma) = \nabla u_\sigma \nabla u_\sigma^\top.$$

It is obvious that the structure tensor does not contain any more information than the smoothed gradient ∇u_σ , namely the information on the local direction of the image structure and the rate of the intensity change. We can find this information in the structure tensor as follows:

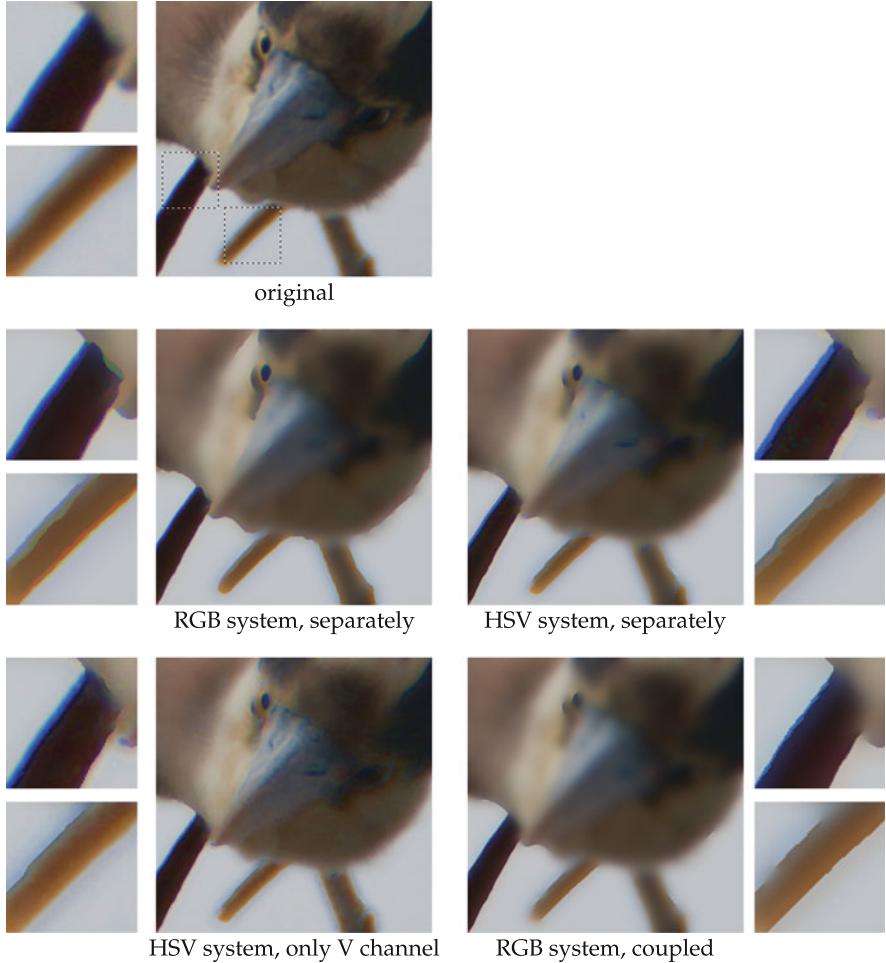


Fig. 5.17 Nonlinear Perona-Malik diffusion for a color image degraded by chromatic aberration. If one treats the color channels separately in either the RGB or HSV system, some color errors along the edges occur. Only denoising of the V channels shows good results; coupling of the color channels gives best results

Lemma 5.42 *The structure tensor has two orthonormal eigenvectors $v_1 \parallel \nabla u_\sigma$ and $v_2 \perp \nabla u_\sigma$. The corresponding eigenvalues are $|\nabla u_\sigma|^2$ and zero.*

Proof For $v_1 = c\nabla u_\sigma$, one has $J_0(\nabla u_\sigma)v_1 = \nabla u_\sigma \nabla u_\sigma^T(c\nabla u_\sigma) = \nabla u_\sigma c|\nabla u_\sigma|^2 = |\nabla u_\sigma|^2 v_1$. Similarly, one sees that $J_0(\nabla u_\sigma)v_2 = 0$. \square

Thus, the directional information is encoded in the eigenvalues of the structure tensor. The eigenvalues correspond, roughly speaking, to the contrast in the

directions of the respective eigenvectors. A second spatial averaging allows one to encode even more information in the structure tensor:

Definition 5.43 The *structure tensor* to $u : \mathbf{R}^2 \rightarrow \mathbf{R}$, noise level $\sigma > 0$, and spatial scale $\rho > 0$ is the matrix-valued function $J_\rho(\nabla u_\sigma) : \mathbf{R}^2 \rightarrow \mathbf{R}^{2 \times 2}$ defined by

$$J_\rho(\nabla u_\sigma) = G_\rho * (\nabla u_\sigma \nabla u_\sigma^\top).$$

The convolution is applied componentwise, i.e. separately for every component of the matrix. Since the smoothed gradient ∇u_σ enters in the structure tensor $J_0(\nabla u_\sigma)$ quadratically, one cannot express the different smoothings by G_σ and G_ρ each in terms of the other. The top left entry of $J_\rho(\nabla u_\sigma)$, for example, has the form $G_\rho * ((G_\sigma * \partial_x u)^2)$. Due to the presence of the square, one cannot express the two convolutions as just one. In other words, the second convolution with G_ρ is not just another averaging of the same type as the convolution with G_σ .

Lemma 5.44 *The structure tensor $J_\rho(\nabla u_\sigma)(x)$ is positive semi-definite for all x .*

Proof Obviously, the matrix $J_0(\nabla u_\sigma)(x)$ is positive semi-definite for every x (the eigenvalues are $|\nabla u_\sigma(x)|^2 \geq 0$ and zero). In particular, for every vector $v \in \mathbf{R}^2$ and every x one has $v^\top J_0(\nabla u_\sigma)(x) v \geq 0$. Hence

$$\begin{aligned} v^\top J_\rho(\nabla u_\sigma)(x) v &= v^\top \int_{\mathbf{R}^2} G_\sigma(x - y) J_0(\nabla u_\sigma)(y) dy v \\ &= \int_{\mathbf{R}^2} \underbrace{G_\sigma(x - y)}_{\geq 0} \underbrace{v^\top J_0(\nabla u_\sigma)(y) v}_{\geq 0} dy \\ &\geq 0. \end{aligned}$$

□

As a consequence of the above lemma, $J_\rho(\nabla u_\sigma)$ also has orthonormal eigenvectors v_1, v_2 and corresponding nonnegative eigenvalues $\mu_1 \geq \mu_2 \geq 0$. We interpret these quantities as follows:

- The eigenvalues μ_1 and μ_2 are the “averaged contrasts” in the directions v_1 and v_2 , respectively.
- The vector v_1 points in the direction of “largest averaged gray value variation.”
- The vector v_2 points in the direction of “average local direction of an edge.” In other words, v_2 is the “averaged direction of coherence.”

Starting from this interpretation we can use the eigenvalues μ_1 and μ_2 to discriminate different regions of an image:

- μ_1, μ_2 small: There is no direction of significant change in gray values. Hence, this is a *flat region*.
- μ_1 large, μ_2 small: There is a large gray value variation in one direction, but not in the orthogonal direction. Hence, this is an *edge*.

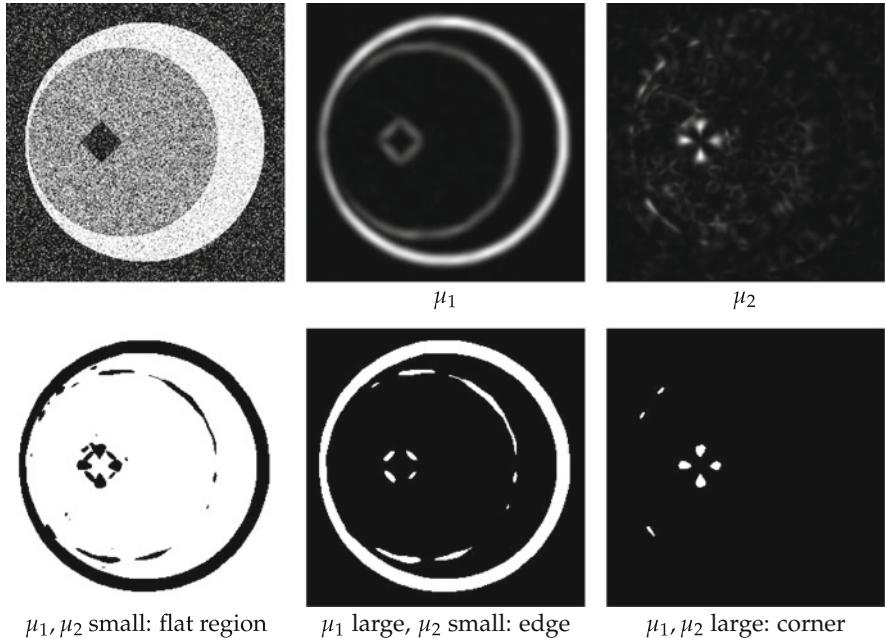


Fig. 5.18 The structure tensor $J_\rho(\nabla u_\sigma)$ encodes information about flat regions, edges, and corners. In the bottom row, the respective regions have been colored in white. In this example the noise level is $\sigma = 4$ and the spatial scale is $\rho = 2$

- μ_1, μ_2 both large: There are two orthogonal directions of significant gray value change, and this is a *corner*.

See Fig. 5.18 for an illustration. We observe that the structure tensor $J_\rho(\nabla u_\sigma)$ indeed contains more information than $J_0(\nabla u_\sigma)$: For the latter, there is always one eigenvalue zero, and thus it cannot see corners. The matrix $J_\rho(\nabla u_\sigma)$ is capable of doing this, since direction information from some neighborhood is used.

Before we develop special methods for anisotropic diffusion, we cite a theorem about the existence of solutions of anisotropic diffusion equations where the diffusion tensor is based on the structure tensor. The theorem is due to Weickert [140] and is a direct generalization of Theorem 5.38.

Theorem 5.45 *Let $u_0 \in L^\infty(\Omega)$, $\rho \geq 0$, $\sigma > 0$, $T > 0$, and let $D : \mathbf{R}^{2 \times 2} \rightarrow \mathbf{R}^{2 \times 2}$ satisfy the following properties:*

- $D \in \mathcal{C}^\infty(\Omega, \mathbf{R}^{2 \times 2})$.
- $D(J)$ is symmetric for every symmetric J .
- *For every bounded function $w \in L^\infty(\Omega, \mathbf{R}^2)$ with $\|w\|_\infty \leq K$ there exists a constant $v(K) > 0$ such that the eigenvalues of $D(J_\rho(w))$ are larger than $v(K)$.*

Then the equation

$$\begin{aligned}\partial_t u &= \operatorname{div} (D(J_\rho(\nabla u_\sigma)) \nabla u) && \text{in } [0, \infty[\times \Omega, \\ \partial_{D(J_\rho(\nabla u_\sigma))v} u &= 0 && \text{on } [0, \infty[\times \partial\Omega, \\ u(0, x) &= u_0(x) && \text{for } x \in \Omega,\end{aligned}$$

has a unique solution $u : [0, T] \rightarrow L^2(\Omega)$. Moreover, u is a continuous mapping from $[0, T]$ to $L^2(\Omega)$ and for almost all $t \in [0, T]$, one has $u(t) \in H^1(\Omega)$.

The structure tensor helps us to motivate the following anisotropic diffusion equations, which were also developed by Weickert [140].

Example 5.46 (Edge-Enhancing Diffusion) In this case we want to have uniform diffusion along the edges and Perona-Malik-like diffusion perpendicular to the edges. With the larger eigenvalue μ_1 of the structure tensor $J_\rho(\nabla u_\sigma)$ we define, using a function g as in (5.14),

$$\lambda_1 = g(\mu_1), \quad \lambda_2 = 1.$$

Using the orthonormal eigenvectors v_1 and v_2 of $J_\rho(\nabla u_\sigma)$, we define a diffusion tensor D by

$$D = (v_1 \ v_2) \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} v_1^T \\ v_2^T \end{pmatrix} = (v_1 \ v_2) \begin{pmatrix} g(\mu_1) & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} v_1^T \\ v_2^T \end{pmatrix}.$$

The eigenvectors of D are obviously the vectors v_1 and v_2 , and the corresponding eigenvalues are λ_1 and λ_2 . Hence, the equation

$$\partial_t u = \operatorname{div} (D(J_\rho(\nabla u_\sigma)) \nabla u)$$

should indeed lead to linear diffusion as in the heat equation along the edges and to Perona-Malik-like diffusion perpendicular to each edge; see Fig. 5.19. For results on existence and uniqueness of solutions we refer to [140].

Example 5.47 (Coherence Enhancing Diffusion) Now we aim to enhance “coherent regions,” i.e. regions where the local structure points in the same direction. To that end, we recall the meaning of the eigenvalues μ_1 and μ_2 of the structure tensor $J_\rho(\nabla u_\sigma)$: they represent the contrast in the orthogonal eigendirections. The local structure is incoherent if both eigenvalues have a similar value. In this case we either have a flat region (if both eigenvalues are small) or some kind of corner (both eigenvalues large). If μ_1 is considerably larger than μ_2 , there is a dominant direction (and this is v_2 , since v_1 is orthogonal to the edges). The idea behind coherence-enhancing diffusion is, to use $|\mu_1 - \mu_2|$ as a measure of local coherence. The diffusion tensor is then basically similar to the previous example: it will have

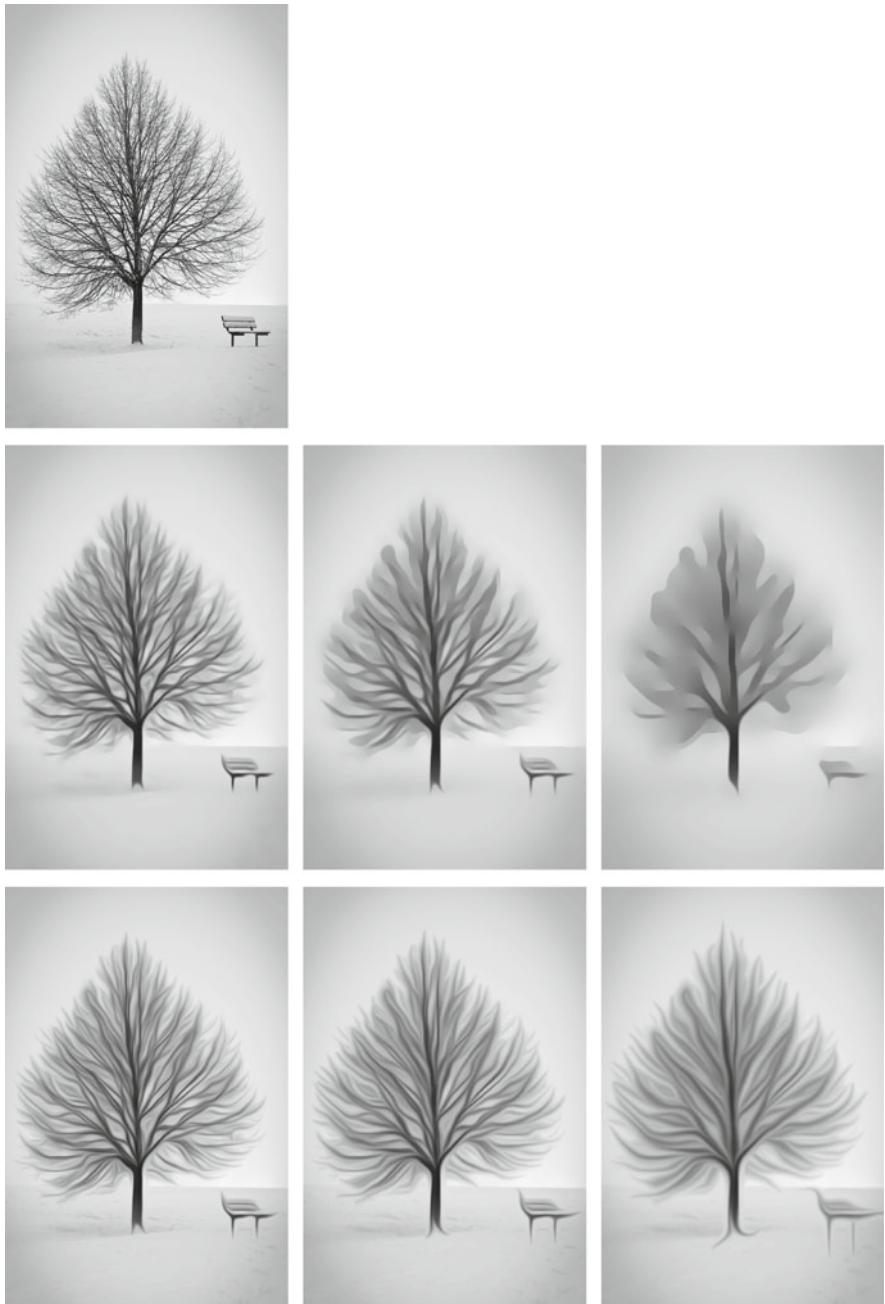


Fig. 5.19 Effect of anisotropic diffusion according to Examples 5.46 and 5.47 based on the structure tensor $J_\rho(\nabla u_\sigma)$ (parameters $\sigma = 0.5$, $\rho = 2$). Top left: original image. Middle row: edge-enhancing diffusion with function g_2 and parameter $\lambda = 0.0005$. Bottom row: coherence-enhancing diffusion with function g_2 and parameters $\lambda = 0.001$, $\alpha = 0.001$. Images are shown at times $t = 25, 100, 500$

the same eigenvectors v_1 and v_2 as the structure tensor, and the eigenvalue for v_2 should become larger for higher local coherence $|\mu_1 - \mu_2|$. With a small parameter $\alpha > 0$ and a function g as in (5.14) we use the following eigenvalues:

$$\lambda_1 = \alpha, \quad \lambda_2 = \alpha + (1 - \alpha)(1 - g(|\mu_1 - \mu_2|)).$$

Similar to Example 5.46, we define the diffusion tensor as

$$D = (v_1 \ v_2) \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} v_1^T \\ v_2^T \end{pmatrix}.$$

The parameter $\alpha > 0$ is needed to ensure that the diffusion tensor is positive definite. As in the previous example we use the model

$$\partial_t u = \operatorname{div}(D(J_\rho(\nabla u_\sigma))\nabla u).$$

The function g is used in a way that the eigenvalue λ_2 is small for low coherence (in the order of α) and close to one for high coherence. In Fig. 5.19 one sees that this equation indeed enhances coherent structures. For further theory we refer to [140].

Application 5.48 (Visualization of Vector Fields) Vector fields appear in several applications, for example vector fields that describe flows as winds in the weather forecast or fluid flows around an object. However, the visualization of vector fields for visual inspection is not easy. Here are some methods: On the one hand, one can visualize a vector field $v : \Omega \rightarrow \mathbf{R}^d$ by plotting small arrows $v(x)$ at some grid points $x \in \Omega$. Another method is to plot so-called integral curves, i.e., curves $\gamma : [0, T] \rightarrow \Omega$ such that the vector field is tangential to the curve, which means $\gamma'(t) = v(\gamma(t))$. The first variant can quickly lead to messy pictures, and the choice of the grid plays a crucial role. for the second variant one has to choose a set of integral curves, and it may happen that these curves accumulate in one region of the image while other regions may be basically empty.

Another method for the visualization of a vector field, building on anisotropic diffusion, has been proposed in [52]. The idea is to design a diffusion tensor that allows diffusion along the vector field, but not orthogonal to the field. The resulting diffusion process is than applied to an image consisting of pure noise. In more detail, the method goes as follows: for a continuous vector field $v : \Omega \rightarrow \mathbf{R}^d$ that is not zero on Ω , there exists a continuous map $B(v)$ that maps a point $x \in \Omega$ to a rotation matrix $B(v)(x)$ that rotates the vector $v(x)$ to the first unit vector e_1 : $B(v)v = |v|e_1$. Using an increasing and positive mapping $\alpha : [0, \infty[\rightarrow [0, \infty[$ and a decreasing map $G : [0, \infty[\rightarrow [0, \infty[$ with $G(r) \rightarrow 0$ for $r \rightarrow \infty$, we define the matrix

$$A(v, r) = B(v)^T \begin{bmatrix} \alpha(|v|) & 0 \\ 0 & G(r) \operatorname{id}_{d-1} \end{bmatrix} B(v).$$

For some initial image $u_0 : \Omega \rightarrow [0, 1]$ and $\sigma > 0$ we consider the following differential equation:

$$\partial_t u = \operatorname{div}(A(v, |\nabla u_\sigma|) \nabla u).$$

Since this equation drives all initial images to a constant image in the limit $t \rightarrow \infty$ (as all the diffusion equations in this chapter do) the authors of [52] proposed to add a source term. For concreteness let $f : [0, 1] \rightarrow \mathbf{R}$ be continuous such that $f(0) = f(1) = 0$, $f < 0$ on $]0, 0.5[$, and $f > 0$ on $]0.5, 1[$. This leads to the modified differential equation

$$\partial_t u = \operatorname{div}(A(v, |\nabla u_\sigma|) \nabla u) + f(u). \quad (5.19)$$

The new term $f(u)$ will push the gray values toward 0 and 1, respectively, and thus leads to higher contrast. This can be accomplished, for example, with the following function:

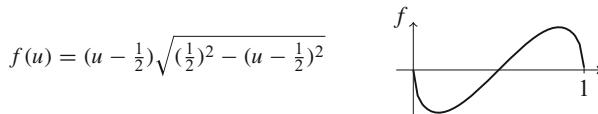


Figure 5.20 shows the effect of Eq. (5.19) with this f and a random initial value. One can easily extract the directions of the vector field, even in regions where the vector field has a small magnitude.

5.4 Numerical Solutions of Partial Differential Equations

To produce images from the equations of the previous sections of this chapter, we have to solve these equations. In general this is not possible analytically, and numerical methods are used. The situation in image processing is a little special in some respects:

- Typically, images are given on a rectangular equidistant grid, and one aims to produce images of a similar shape. Hence, it is natural to use these grids.
- The visual quality of the images is more important than to solve the equations as accurately as possible. Thus, methods with lower order are acceptable if they produce “good images.”
- Some of the partial differential equations we have treated preserve edges or create “kinks.” Two examples are the equations for erosion and dilation (5.12). This poses a special challenge for numerical methods, since the solutions that will be approximated are not differentiable.

We consider an introductory example:

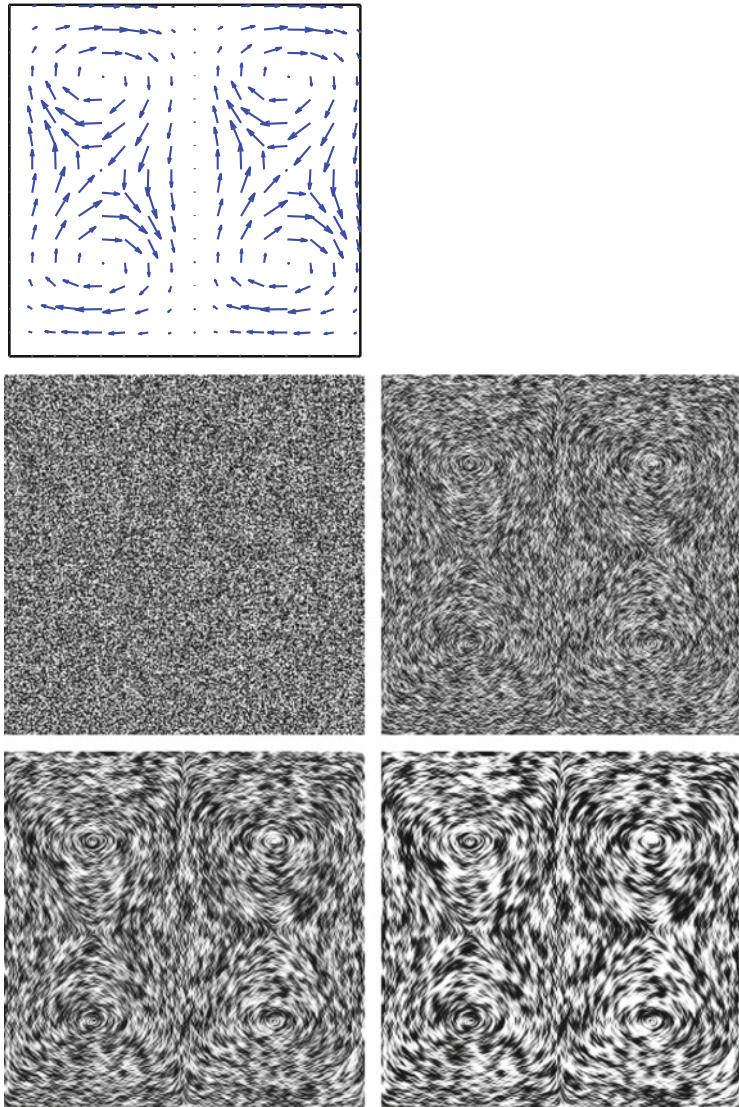
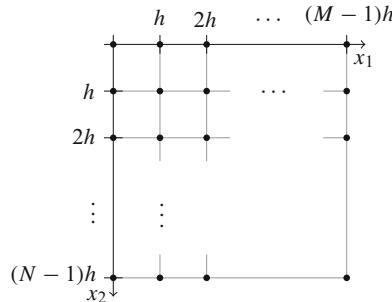


Fig. 5.20 Visualization of vector fields by anisotropic diffusion as in Application 5.48. Top left: The vectorfield. Below: The initial value and different solutions of Eq. (5.19) at different times

Example 5.49 (Discretization of the Heat Equation) We want to solve the following initial-boundary value problem for the heat equation

$$\begin{aligned}\partial_t u &= \Delta u && \text{in } [0, T] \times \Omega, \\ \partial_\nu u &= 0 && \text{on } [0, T] \times \partial\Omega, \\ u(0) &= u^0,\end{aligned}\tag{5.20}$$

on a rectangular domain $\Omega \subset \mathbf{R}^2$. Our initial image u^0 is given in discrete form, i.e., as an $N \times M$ matrix. We assume that the entries of the matrix have been obtained by sampling a continuous image with sampling rate h in the x_1 and x_2 directions, respectively. If we use (by slight abuse of notation) u^0 for both the discrete and the continuous image, we can view the values $u_{i,j}^0$ as $u^0((i-1)h, (j-1)h)$ (we shift by 1 to make the indices i, j start at one and not at zero). Thus, we know the values of u^0 on a rectangular equidistant grid as follows:



For the time variable we proceed similarly and discretize it with step-size τ . Using u for the solution of the initial-boundary value problem (5.20), we want to find $u_{i,j}^n$ as an approximation to $u(n\tau, (i-1)h, (j-1)h)$, i.e., all three equations in (5.20) have to be satisfied. The initial condition $u(0) = u^0$ is expressed by the equation

$$u_{i,j}^0 = u^0((i-1)h, (j-1)h).$$

To satisfy the differential equation $\partial_t u = \Delta u$, we replace the derivatives by difference quotients. We discretized the Laplace operator in Sect. 3.3.3 already:

$$\Delta u(ih, jh) \approx \frac{u((i+1)h, jh) + u((i-1)h, jh) + u(ih, (j+1)h) + u(ih, (j-1)h) - 4u(ih, jh)}{h^2}.$$

The derivative in the direction t is approximated by a forward difference quotient:

$$\partial_t u(n\tau, ih, jh) \approx \frac{u((n+1)\tau, ih, jh) - u(n\tau, ih, jh)}{\tau}.$$

This gives the following equation for the discrete values $u_{i,j}^n$:

$$\frac{u_{i,j}^{n+1} - u_{i,j}^n}{\tau} = \frac{u_{i+1,j}^n + u_{i-1,j}^n + u_{i,j+1}^n + u_{i,j-1}^n - 4u_{i,j}^n}{h^2}. \quad (5.21)$$

There is a problem at the points with $i = 1, N$ or $j = 1, M$: the terms that involve $i = 0, N+1$ or $j = 0, M+1$, respectively, are not defined. To deal with this issue we take the boundary condition $\partial_v u = 0$ into account. In this example, the domain is a rectangle, and the boundary condition has to be enforced for values with $i = 1, N$ and $j = 1, M$. We add the auxiliary points $u_{0,j}^n, u_{N+1,j}^n, u_{i,0}^n$, and $u_{i,M+1}^n$ and replace the derivative by a central difference quotient and get

$$\begin{aligned} \frac{u_{0,j}^n - u_{2,j}^n}{2h} &= 0, & \frac{u_{N-1,j}^n - u_{N+1,j}^n}{2h} &= 0, \\ \frac{u_{i,0}^n - u_{i,2}^n}{2h} &= 0, & \frac{u_{i,M-1}^n - u_{i,M+1}^n}{2h} &= 0. \end{aligned}$$

This leads to the equations

$$\begin{aligned} u_{0,j}^n &= u_{2,j}^n, & u_{N+1,j}^n &= u_{N-1,j}^n \\ u_{i,0}^n &= u_{i,2}^n, & u_{i,M+1}^n &= u_{i,M-1}^n. \end{aligned}$$

Thus, the boundary condition is realized by mirroring the values over the boundary. The discretized equation (5.21) for $i = 1$, for example, has the following form:

$$\frac{u_{1,j}^{n+1} - u_{1,j}^n}{\tau} = \frac{2u_{2,j}^n + u_{1,j+1}^n + u_{1,j-1}^n - 4u_{1,j}^n}{h^2}.$$

We can circumvent the distinction of different cases by the following notation: we solve Eq. (5.21) for $u_{i,j}^{n+1}$ and get

$$u_{i,j}^{n+1} = u_{i,j}^n + \frac{\tau}{h^2}(u_{i+1,j}^n + u_{i-1,j}^n + u_{i,j+1}^n + u_{i,j-1}^n - 4u_{i,j}^n).$$

This can be realized by a discrete convolution as in Sect. 3.3.3:

$$u^{n+1} = u^n + \frac{\tau}{h^2} u^n * \begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix}.$$

One advantage of this formulation is that we can realize the boundary condition by a symmetric extension over the boundary. Starting from the initial value $u_{i,j}^0 = u^0((i-1)h, (j-1)h)$ we can use this to calculate an approximate solution $u_{i,j}^n$ iteratively for every n .

We call the resulting scheme *explicit*, since we can calculate the values u^{n+1} directly from the values u^n . One reason for this is that we discretized the time derivative $\partial_t u$ by a forward difference quotient. If we use a backward difference quotient, we get

$$\frac{u_{i,j}^n - u_{i,j}^{n-1}}{\tau} = \frac{u_{i+1,j}^n + u_{i-1,j}^n + u_{i,j+1}^n + u_{i,j-1}^n - 4u_{i,j}^n}{h^2}$$

or

$$\left(u^n - \frac{\tau}{h^2} u^n * \begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix} \right) = u^{n-1}$$

(again, with a symmetric extension over the boundary to take care of the boundary condition). This is a linear system of equations for u^n , and we call this scheme *implicit*.

This initial example illustrates a simple approach to constructing numerical methods for differential equations:

- Approximate the time derivative by forward or backwards difference quotients.
- Approximate the spatial derivative by suitable difference quotients and use symmetric boundary extension to treat the boundary condition.
- Solve the resulting equation for u^{n+1} .

A little more abstractly: we can solve a partial differential equation of the form

$$\partial_t u(t, x) = \mathcal{L}(u)(t, x)$$

with a differential operator \mathcal{L} that acts on the spatial variable x only by so-called semi-discretization:

- Treat the equation in a suitable space X , i.e., $u : [0, T] \rightarrow X$, such that $\partial_t u = \mathcal{L}(u)$.
- Discretize the operator \mathcal{L} : Choose a spatial discretization of the domain of the x variable and thus, an approximation of the space X . Define an operator L that operates on the discretized space and approximates \mathcal{L} . Thus, this partial differential equation turns into a system of ordinary differential equations

$$\partial_t u = L(u).$$

- Solve the system of ordinary differential equations with some method known from numerical analysis (see, e.g., [135]).

In imaging, one often faces the special case that the domain for the x variable is a rectangle. Moreover, the initial image u^0 is often given on an equidistant grid. This gives a natural discretization of the domain. Hence, many methods in imaging replace the differential operators by difference quotients. This is generally known as the method of finite differences.

The equations from Sects. 5.2 and 5.3 can be roughly divided into two groups: equations of diffusion type (heat equation, nonlinear diffusion) and equations of transport type (erosion, dilation, mean curvature motion). These types have to be treated differently.

5.4.1 Diffusion Equations

In this section we consider equations of diffusion type

$$\partial_t u = \operatorname{div}(A \nabla u),$$

i.e., the differential operator $\mathcal{L}(u) = \operatorname{div}(A \nabla u)$. First we treat the case of isotropic diffusion, i.e., $A : \Omega \rightarrow \mathbf{R}$ is a scalar function. We start with the approximation of the differential operator $\operatorname{div}(A \nabla u) = \partial_{x_1}(A \partial_{x_1} u) + \partial_{x_2}(A \partial_{x_2} u)$ by finite differences. Obviously it is enough to consider the term $\partial_{x_1}(A \partial_{x_1} u)$. At some point (i, j) we proceed as follows:

$$\partial_{x_1}(A \partial_{x_1} u) \approx \frac{1}{h} \left((A \partial_{x_1} u)_{i+\frac{1}{2}, j} - (A \partial_{x_1} u)_{i-\frac{1}{2}, j} \right)$$

with

$$(A \partial_{x_1} u)_{i+\frac{1}{2}, j} = A_{i+\frac{1}{2}, j} \left(\frac{u_{i+1, j} - u_{i, j}}{h} \right), \quad (A \partial_{x_1} u)_{i-\frac{1}{2}, j} = A_{i-\frac{1}{2}, j} \left(\frac{u_{i, j} - u_{i-1, j}}{h} \right).$$

Using a similar approximation for the x_2 direction we get

$$\begin{aligned} \operatorname{div}(A \nabla u) &\approx \frac{1}{h^2} \left(A_{i, j-\frac{1}{2}} u_{i, j-1} + A_{i, j+\frac{1}{2}} u_{i, j+1} + A_{i-\frac{1}{2}, j} u_{i-1, j} + A_{i+\frac{1}{2}, j} u_{i+1, j} \right. \\ &\quad \left. - (A_{i, j-\frac{1}{2}} + A_{i, j+\frac{1}{2}} + A_{i-\frac{1}{2}, j} + A_{i+\frac{1}{2}, j}) u_{i, j} \right). \end{aligned} \tag{5.22}$$

We can arrange this efficiently in matrix notation. To that end, we arrange the matrix $u \in \mathbf{R}^{N \times M}$ in a vector $U \in \mathbf{R}^{NM}$, by stacking the rows into a vector.¹ We define a

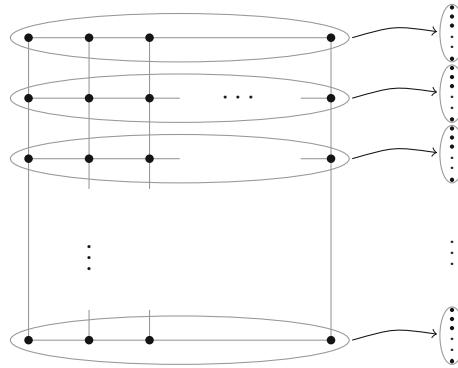
¹Of course, we could also stack the columns into a vector, and indeed, some software packages have this as a default operation. The only difference between these two approaches is the direction of the x_1 and x_2 coordinates.

bijection $\Theta : \{1, \dots, N\} \times \{1, \dots, M\} \rightarrow \{1, \dots, NM\}$ of the index sets by

$$\Theta(i, j) = (i - 1)M + j, \quad \Theta^{-1}(I) = (\lfloor \frac{I}{M} \rfloor + 1, I \bmod M).$$

Thus we get

$$U_{\Theta(i,j)} = u_{i,j}, \quad \text{bzw.} \quad U_I = u_{\Theta^{-1}(I)}.$$



If we denote the right-hand side in (5.22) by $v_{i,j}$ and define $V_{\Theta(i,j)} = v_{i,j}$ and $U_{\Theta(i,j)} = u_{i,j}$, we get $V = \mathbf{A}U$ with a matrix $\mathbf{A} \in \mathbb{R}^{NM \times NM}$ defined by

$$\mathbf{A}_{\Theta(i,j), \Theta(k,l)} = \begin{cases} -(A_{i,j-\frac{1}{2}} + A_{i,j+\frac{1}{2}} + A_{i-\frac{1}{2},j} + A_{i+\frac{1}{2},j}) & \text{if } i = k, j = l, \\ A_{i \pm \frac{1}{2},j} & \text{if } i \pm 1 = k, j = l, \\ A_{i,j \pm \frac{1}{2}} & \text{if } i = k, j \pm 1 = l, \\ 0 & \text{otherwise.} \end{cases} \quad (5.23)$$

We can treat boundary conditions $\partial_\nu u = 0$ similarly to Example 5.49 by introducing auxiliary points and again realize the boundary condition by symmetric extension over the boundary. We obtain the following semi-discretized equation

$$\partial_t U = \frac{1}{h^2} \mathbf{A}U.$$

This is a system of linear ordinary differential equations. Up to now we did not incorporate that the diffusion coefficient A may depend on u (or on the gradient of u , respectively). In this case \mathbf{A} depends on u and we obtain the nonlinear system

$$\partial_t U = \frac{1}{h^2} \mathbf{A}(U)U.$$

Approximating the time derivative by a simple difference, we see three variants:

Explicit:

$$\frac{U^{n+1} - U^n}{\tau} = \frac{1}{h^2} \mathbf{A}(U^n) U^n.$$

Implicit:

$$\frac{U^{n+1} - U^n}{\tau} = \frac{1}{h^2} \mathbf{A}(U^{n+1}) U^{n+1}.$$

Semi-implicit:

$$\frac{U^{n+1} - U^n}{\tau} = \frac{1}{h^2} \mathbf{A}(U^n) U^{n+1}.$$

The implicit variant leads to a nonlinear system of equations, and its solution may pose a significant challenge. Hence, implicit methods are usually not the method of choice. The explicit method can be written as

$$U^{n+1} = (\text{id} + \frac{\tau}{h^2} \mathbf{A}(U^n)) U^n \quad (5.24)$$

and it requires only one discrete convolution per iteration. The semi-implicit method leads to

$$(\text{id} - \frac{\tau}{h^2} \mathbf{A}(U^n)) U^{n+1} = U^n. \quad (5.25)$$

This is a linear system of equations for U^{n+1} , i.e., in every iteration we need to solve one such system.

Now we analyze properties of the explicit and semi-implicit methods.

Theorem 5.50 *Let $A_{i \pm \frac{1}{2}, j} \geq 0$, $A_{i, j \pm \frac{1}{2}} \geq 0$, and $\mathbf{A}(U^n)$ according to Eq. (5.23). For the explicit method (5.24) assume that the step-size restriction*

$$\tau \leq \frac{h^2}{\max_I |\mathbf{A}(U^n)_{I,I}|}$$

holds, while for the semi-implicit method (5.25) no upper bound on τ is assumed. Then the iterates U^n of (5.24) and (5.25), respectively, satisfy the discrete maximum principle, i.e. for all I ,

$$\min_J U_J^0 \leq U_I^n \leq \max_J U_J^0.$$

Moreover, under the same assumptions,

$$\sum_{J=1}^{NM} U_J^n = \sum_{J=1}^{NM} U_J^0,$$

for both the explicit and semi-implicit methods, i.e., the mean gray value is preserved.

Proof First we consider the explicit iteration (5.24) and set $Q(U^n) = \text{id} + \tau/h^2 \mathbf{A}(U^n)$. Then the explicit iteration reads $U^{n+1} = Q(U^n)U^n$. By definition of $\mathbf{A}(U^n)$, we have $\sum_{J=1}^{NM} \mathbf{A}(U^n)_{I,J} = 0$ for all I (note the boundary condition $\partial_\nu u = 0!$), and hence

$$\sum_{J=1}^{NM} Q(U^n)_{I,J} = 1.$$

This immediately implies the preservation of the mean gray value, since

$$\sum_{J=1}^{NM} U_J^{n+1} = \sum_{J=1}^{NM} \sum_{I=1}^{NM} Q(U^n)_{I,J} U_I^n = \sum_{I=1}^{NM} \sum_{J=1}^{NM} Q(U^n)_{I,J} U_I^n = \sum_{I=1}^{NM} U_I^n,$$

which, by recursion, implies the claim.

Moreover, for $I \neq J$ one has $Q(U^n)_{I,J} = \mathbf{A}(U^n)_{I,J} \geq 0$. On the diagonal,

$$Q(U^n)_{I,I} = 1 + \frac{\tau}{h^2} \mathbf{A}(U^n)_{I,I}.$$

The step-size restriction implies $Q(U^n)_{I,I} \geq 0$, which shows that the matrix $Q(U^n)$ has nonnegative components. We deduce that

$$U_I^{n+1} = \sum_{J=1}^{NM} Q(U^n)_{I,J} U_J^n \leq \max_K U_K^n \underbrace{\sum_{J=1}^{NM} Q(U^n)_{I,J}}_{=1} = \max_K U_K^n.$$

Similarly, one shows that

$$U_I^{n+1} \geq \min_K U_K^n,$$

which, again by recursion, implies the maximum principle.

In the semi-implicit case we write $R(U^n) = (\text{id} - \tau/h^2 \mathbf{A}(U^n))$, and hence the iteration reads $U^{n+1} = R(U^n)^{-1}U^n$. For the matrix $\mathbf{A}(U^n)$, we have by

construction

$$\mathbf{A}(U^n)_{I,I} = - \sum_{J \neq I} \mathbf{A}(U^n)_{I,J},$$

and thus

$$R(U^n)_{I,I} = 1 - \frac{\tau}{h^2} \mathbf{A}(U^n)_{I,I} = 1 + \frac{\tau}{h^2} \sum_{J \neq I} \mathbf{A}(U^n)_{I,J} > \frac{\tau}{h^2} \sum_{J \neq I} \mathbf{A}(U^n)_{I,J} = \sum_{J \neq I} |R(U^n)_{I,J}|.$$

The property $R(U^n)_{I,I} > \sum_{J \neq I} |R(U^n)_{I,J}|$ is called “strict diagonal dominance.” This property implies that $R(U^n)$ is invertible and that the inverse matrix $R(U^n)^{-1}$ has nonnegative entries (cf. [72]). Moreover, with $e = (1, \dots, 1)^T \in \mathbf{R}^{NM}$, we have $R(U^n)e = e$, which implies, by invertibility of $R(U^n)$, that $R(U^n)^{-1}e = e$ holds, too. We conclude that

$$\sum_{J=1}^{NM} (R(U^n)^{-1})_{I,J} = 1.$$

Similarly to the explicit case, we deduce the preservation of the mean gray value from the nonnegativity $R(U^n)^{-1}$, and

$$U_I^{n+1} = \sum_{J=1}^{NM} (R(U^n)^{-1})_{I,J} U_J^n \leq \max_K U_K^n \underbrace{\sum_{J=1}^{NM} (R(U^n)^{-1})_{I,J}}_{=1} = \max_K U_K^n$$

implies, by recursion, the full claim. \square

Remark 5.51 (Step-Size Restrictions and the AOS Method) The step-size restriction of the explicit method is a severe limitation. To compute the solution for a large time t , one may need many iterations. Since the matrix \mathbf{A} is sparse, every iteration is comparably cheap (each row has only five non-zero entries). However, if the diffusion coefficient depends on U , \mathbf{A} has to be assembled for every iteration, which also costs some time. The semi-implicit method does not have a step-size restriction, but one needs to solve a linear system in every iteration. Again, by sparseness of \mathbf{A} , this may be done efficiently with an iterative method like the Gauss-Seidel method (see, e.g., [72]). However, for larger steps τ , these methods tend to take longer to converge, and this restricts the advantage of the semi-implicit method. One way to circumvent this difficulty is the method of “additive operator splitting” (AOS, see [140]). Here one treats the differences in the x_1 - and x_2 -directions separately and averages. One need to solve two tridiagonal linear systems per iteration, which can be done in linear time. For details of the implementation we refer to [140].

Example 5.52 (Perona-Malik and Modified Perona-Malik) In the case of Perona-Malik diffusion we have

$$A(u) = g(|\nabla u|) \text{id}.$$

Hence, the entries of A are

$$A_{i \pm \frac{1}{2}, j} = g(|\nabla u|_{i \pm \frac{1}{2}, j}).$$

To calculate the magnitude of the gradient at intermediate coordinates we can use, for example, linear interpolation of the magnitudes of the gradients at neighboring integer places, i.e.,

$$|\nabla u|_{i \pm \frac{1}{2}, j} = \frac{|\nabla u|_{i,j} + |\nabla u|_{i \pm 1, j}}{2}.$$

The gradients at these integer places can be approximated by finite differences. For the modified Perona-Malik equation we have

$$A_{i \pm \frac{1}{2}, j} = g(|\nabla u_\sigma|_{i \pm \frac{1}{2}, j}).$$

Then the entries of A are calculated in exactly the same way, after an initial presmoothing. If the function g is nonnegative, the entries $A_{i \pm \frac{1}{2}, j}$ and $A_{i, j \pm \frac{1}{2}}$ are nonnegative, and Theorem 5.50 applies. This discretization was used to generate the images in Fig. 5.10. Alternative methods for the discretization of isotropic nonlinear diffusion are described, for example, in [142] and [111].

Remark 5.53 (Anisotropic Equations) In the case of anisotropic diffusion with symmetric diffusion tensor

$$A = \begin{bmatrix} B & C \\ C & D \end{bmatrix},$$

there are mixed second derivatives in the divergence. For example, in two dimensions,

$$\operatorname{div}(A \nabla u) = \partial_{x_1} (B \partial_{x_1} u + C \partial_{x_2} u) + \partial_{x_2} (C \partial_{x_1} u + D \partial_{x_2} u).$$

If we form the matrix \mathbf{A} similar to Eqs. (5.22) and (5.23) by finite differences, it is not clear a priori how one can ensure that the entries $A_{i \pm \frac{1}{2}, j}$ and $A_{i, j \pm \frac{1}{2}}$ are nonnegative. In fact, this is nontrivial, and we refer to [140, Section 3.4.2]. One alternative to finite differences is the method of finite elements, and we refer to [52, 115] for details.

5.4.2 Transport Equations

Transport equations are a special challenge. To see why this is so, we begin with a simple one-dimensional example of a transport equation. For $a \neq 0$ we consider

$$\partial_t u + a \partial_x u = 0, \quad t > 0, \quad x \in \mathbf{R},$$

with initial value $u(0, x) = u_0(x)$. It is simple to check that the solution is just the initial value transported with velocity a , i.e.,

$$u(t, x) = u_0(x - at).$$

This has two interesting aspects:

1. The formula for the solution is applicable for general measurable functions u_0 , and no continuity or differentiability is needed whatsoever. Hence, one could also “solve” the equation for this type of initial condition.
2. There are curves (in this case even straight lines) along which the solution is constant. These curves are the solutions of the following ordinary differential equation:

$$X'(t) = a, \quad X(0) = x_0.$$

These curves are called *characteristics*.

The first point somehow explains why methods based on finite differences tend to be problematic for transport equations. The second point can be extended to a simple method, the method of characteristics.

Method of Characteristics

We describe the solution of a transport equation in \mathbf{R}^d :

Lemma 5.54 *Let $a : \mathbf{R}^d \rightarrow \mathbf{R}^d$ be Lipschitz continuous, $u_0 : \mathbf{R}^d \rightarrow \mathbf{R}$ continuous, and u a solution of the Cauchy problem*

$$\partial_t u + a \cdot \nabla u = 0,$$

$$u(0, x) = u_0(x).$$

If X is a solution of the ordinary initial value problem

$$X' = a(X), \quad X(0) = x_0,$$

then $u(t, X(t)) = u_0(x_0)$.

Proof We consider u along the solutions X of the initial value problems and take the derivative with respect to t :

$$\frac{d}{dt}u(t, X(t)) = \partial_t u(t, X(t)) + a(X(t)) \cdot \nabla u(t, X(t)) = 0.$$

Thus, u is constant along X , and by the initial value for X we get at $t = 0$ that

$$u(0, X(0)) = u(0, x_0) = u_0(x_0).$$

□

Also in this case the curves X are called characteristics of the equation. For a given vector field a one can solve the transport equations by calculating the characteristics, which amounts to solving ordinary differential equations.

Application 5.55 (Coordinate Transformations) The coordinate transformation from Example 5.2 has the infinitesimal generator

$$A[u](x) = v(x) \cdot \nabla u(x)$$

(cf. Exercise 5.6). Hence, the scale space is described by the differential equation

$$\partial_t u - v \cdot \nabla u = 0, \quad u(0, x) = u_0(x).$$

This differential equation can be solved by the method of characteristics as follows: for some x_0 we calculate the solution of the ordinary initial value problem

$$X'(t) = v(X(t)), \quad X(0) = x_0,$$

with some suitable routine up to time T . Here one can use, for example, the Runge-Kutta methods, see, e.g., [72]. If v is given only on a discrete set of points, one can use interpolation as in Sect. 3.1.1 to evaluate v at intermediate points. Then one gets $u(T, X(T)) = u_0(x_0)$ (where one may need interpolation again to obtain $u(T, \cdot)$ at the grid points). This method was used to generate the images in Fig. 5.1.

Application 5.56 (Erosion, Dilation, and Mean Curvature Motion) The equations for erosion, dilation, and mean curvature motion can be interpreted as transport equations, cf. Remark 5.25 and Sect. 5.2.2. However, the vector field v depends on u in these cases, i.e.,

$$\partial_t u - v(u) \cdot \nabla u = 0.$$

Hence, the method of characteristics from Application 5.55 cannot be applied in its plain form. One still obtains reasonable results if the function u is kept fixed for the computation of the vector field $v(u)$ for some time. In the example of mean

curvature motion this looks as follows:

- For a given time t_n and corresponding image $u(t_n, x)$ compute the vector field

$$v(u(t_n, x)) = \kappa(t_n, x) \frac{\nabla u(t_n, x)}{|\nabla u(t_n, x)|}.$$

By Exercise 5.9, we have

$$\kappa = \operatorname{div}\left(\frac{\nabla u}{|\nabla u|}\right).$$

Thus we may proceed as follows. First calculate the unit vector field $v(t_n, x) = \frac{\nabla u(t_n, x)}{|\nabla u(t_n, x)|}$, e.g., by finite differences (avoiding division by zero, e.g., by $|\nabla u(t_n, x)| \approx \sqrt{|\nabla u(t_n, x)| + \varepsilon^2}$ with some small $\varepsilon > 0$). Compute $v_{t_n}(x) = (\operatorname{div} v)(t_n, x)v(t_n, x)$, e.g. again by finite differences.

- Solve the equation

$$\partial_t u - v_{t_n} \cdot \nabla u = 0$$

with initial value $u(t_n, x)$ up to time $t_{n+1} = t_n + T$ with T not too large by the method of characteristics and go back to the previous step.

This method was used to produce the images in Fig. 5.8.

Similarly one can apply the method for the equations of erosion and dilation with a circular structure element

$$\partial_t u \pm |\nabla u| = 0;$$

see Fig. 5.21 for the case of dilation. One notes some additional smoothing that results from the interpolation.

In this application we did not treat the nonlinearity in a rigorous way. For a nonlinear transport equation of the form $\partial_t u + \nabla(F(u)) = 0$ one can still define characteristics, but it may occur that two characteristics intersect or are not well defined. The first case leads to so-called “shocks,” while the second case leads to nonunique solutions. Our method does not consider these cases and hence may run into problems.

Finite Difference Methods

The application of finite difference methods to transport equations needs special care. We illustrate this with an introductory example and then state a suitable method.



Fig. 5.21 Solutions of the equation for the dilation by the method of characteristics according to Application 5.56

Example 5.57 (Stability Analysis for the One-Dimensional Case) Again we begin with a one-dimensional example

$$\partial_t u + a \partial_x u = 0, \quad u(0, x) = u_0(x).$$

We use forward differences in the t direction and a central difference quotient in the x direction and get with notation similar to Example 5.49, the explicit scheme

$$u_j^{n+1} = u_j^n + a \frac{\tau}{2h} (u_{j+1}^n - u_{j-1}^n). \quad (5.26)$$

To see that this method is not useful we use the so-called “von Neumann stability analysis.” To that end, we consider the method on a finite interval with periodic boundary conditions, i.e., $j = 1, \dots, M$ and $u_{j+M}^n = u_j^n$. We make a special ansatz

for the solution, namely

$$v_j^n = \xi^n e^{ikj\pi h}, \quad 0 \leq k \leq M = 1/h, \quad \xi \in \mathbf{C} \setminus \{0\}.$$

Plugging this into the scheme gives

$$\xi^{n+1} e^{ikj\pi h} = \xi^n e^{ikj\pi h} + a \frac{\tau}{2h} (\xi^n e^{ik(j+1)\pi h} - \xi^n e^{ik(j-1)\pi h})$$

and after multiplication by $\xi^{-n} e^{-ikj\pi h}$ we obtain the following equation for ξ :

$$\xi = 1 + ia \frac{\tau}{h} \sin(k\pi h).$$

This shows that ξ has a magnitude that is strictly larger than one, and hence for every solution that contains v_j^n , this part will be amplified exponentially in some sense. This contradicts our knowledge that the initial value is only transported and not changed in magnitude and we see that the scheme (5.26) is unstable.

Now we consider a forward difference quotient in the x -direction, i.e., the scheme

$$u_j^{n+1} = u_j^n + a \frac{\tau}{2h} (u_{j+1}^n - u_j^n). \quad (5.27)$$

Similar to the previous calculation, we obtain

$$\xi = 1 + a \frac{\tau}{h} (e^{ikj\pi h} - 1).$$

Now we see that $|\xi| \leq 1$ holds for $0 \leq a \frac{\tau}{h} \leq 1$. Since τ and h are nonnegative, we get stability for $a \geq 0$ only under the condition that

$$a \frac{\tau}{h} \leq 1.$$

Arguing similarly for a backward difference quotient, we see that the condition $-1 \leq a \frac{\tau}{h}$ guarantees stability. Thus, we should use different schemes for different signs of a , i.e., for different transport directions. If a depends on x , we should use different difference quotients for different signs of a , namely

$$u_j^{n+1} = \begin{cases} u_j^n + a_j \frac{\tau}{h} (u_{j+1}^n - u_j^n) & \text{if } a_j \geq 0, \\ u_j^n + a_j \frac{\tau}{h} (u_j^n - u_{j-1}^n) & \text{if } a_j \leq 0, \end{cases}$$

or more compactly,

$$u_j^{n+1} = u_j^n + \frac{\tau}{h} (\max(0, a_j)(u_{j+1}^n - u_j^n) + \min(0, a_j)(u_j^n - u_{j-1}^n)).$$

This scheme is stable under the condition that

$$|a| \frac{\tau}{h} \leq 1.$$

Since one uses the transport direction to adapt the scheme, this method is called an *upwind scheme*. The condition $|a| \frac{\tau}{h} \leq 1$ is called the CFL condition and goes back to Courant et al. [47].

Application 5.58 (Upwind Method in 2D: The Rouy-Tourin Method) We apply the idea of the upwind method to the two-dimensional dilation equation

$$\partial_t u = |\nabla u| = \sqrt{(\partial_{x_1} u)^2 + (\partial_{x_2} u)^2}.$$

Depending on the sign of $\partial_{x_i} u$, we choose either the forward or backward difference, i.e.,

$$\begin{aligned} (\partial_{x_1} u)_{i,j}^2 &\approx \frac{1}{h^2} \max(0, u_{i+1,j} - u_{i,j}, -(u_{i,j} - u_{i-1,j}))^2, \\ (\partial_{x_2} u)_{i,j}^2 &\approx \frac{1}{h^2} \max(0, u_{i,j+1} - u_{i,j}, -(u_{i,j} - u_{i,j-1}))^2. \end{aligned}$$

The resulting method is known as Rouy-Tourin method [120]. Results for this method are shown in Fig. 5.22. Again we note, similar to the method of characteristics from Application 5.56, that a certain blur occurs. This phenomenon is called *numerical viscosity*. Finite difference methods with less numerical viscosity are proposed, for example, in [20].

Remark 5.59 (Upwind Method According to Osher and Sethian) The authors of [106] propose the following different upwind method:

$$\begin{aligned} (\partial_{x_1} u)_{i,j}^2 &\approx \frac{1}{h^2} (\max(0, u_{i+1,j} - u_{i,j})^2 + \max(0, u_{i-1,j} - u_{i,j})^2) \\ (\partial_{x_2} u)_{i,j}^2 &\approx \frac{1}{h^2} (\max(0, u_{i,j+1} - u_{i,j})^2 + \max(0, u_{i,j-1} - u_{i,j})^2). \end{aligned}$$

This method gives results that are quite similar to that of Rouy-Tourin, and hence we do not show extra pictures. Especially, some numerical viscosity can be observed, too.

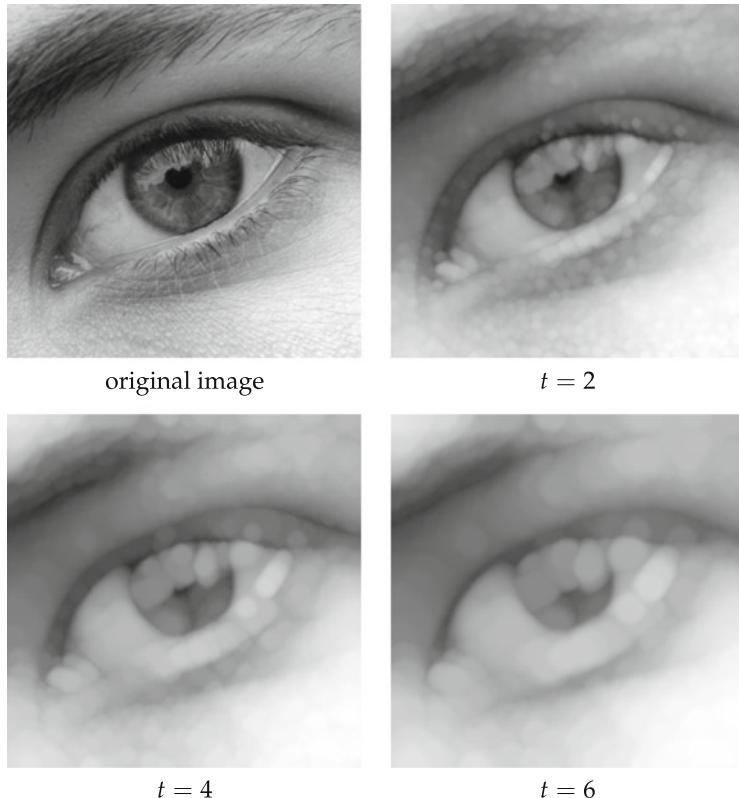


Fig. 5.22 Solution of the dilation equation with the upwind method according to Rouy-Tourin from Application 5.58

5.5 Further Developments

Partial differential equations can be used for many further tasks, e.g., also for inpainting; cf. Sect. 1.2. An idea proposed by Bertalmio [13], is, to “transport” the information of the image into the inpainting domain. Bornemann and März [15] provide the following motivation for this approach: in two dimensions we denote by $\nabla^\perp u$ the gradient of u turned $\frac{\pi}{2}$ to the left,

$$\nabla^\perp u = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \nabla u = \begin{bmatrix} -\partial_{x_2} u \\ \partial_{x_1} u \end{bmatrix}.$$

Now consider the transport equation

$$\partial_t u = -\nabla^\perp (\Delta u) \cdot \nabla u.$$

As illustrated in Application 3.23, the level lines of Δu roughly follow the edges. Hence, the vector $\nabla^\perp(\Delta u)$ is also tangential to the edges, and thus the equation realizes some transport along the edges of the image. This is, roughly speaking, the same one would do by hand to fill up a missing piece of an image: take the edges and extend them into the missing part and then fill up the domain with the right colors. Bornemann and März [15] propose, on the one hand, to calculate the transport direction $\nabla^\perp(\Delta u_\sigma)$ with a presmoothed u_σ and, on the other hand, get further improved results by replacing the transport direction with the eigenvector corresponding to the smaller eigenvalue of the structure tensor $J_\rho(\nabla u_\sigma)$.

Methods that are based on diffusion can be applied to objects different from images; we can also “denoise” surfaces. Here a surface is a manifold, and the Laplace operator has to be replaced by the so-called Laplace-Beltrami operator. Also one can adapt the ideas of anisotropic diffusion to make them work on surfaces, too; see, e.g., [45].

The Perona-Malik equation and its analytical properties are still subject to research. Amann [4] describes a regularization of the Perona-Malik equation that uses a temporal smoothing instead of the spatial smoothing used in the modified model (5.18). This can be interpreted as a continuous analogue of the semi-implicit method (5.25). Chen and Zhang [44] provide a new interpretation of the nonexistence of solutions of the Perona-Malik equation in the context of Young measures. Esedoglu [59] develops a finer analysis of the stability of the discretized Perona-Malik method and proves maximum principle for certain initial values.

5.6 Exercises

Exercise 5.1 (Scale Space Properties of Coordinate Transformations) Show that the coordinate transformations from Example 5.2 satisfy the Axioms [REG], [COMP], [GLSI], [GSI] and [SCALE]. Moreover, show that [TRANS] and [ISO] are not satisfied in general.

Exercise 5.2 (Gray Value Scaling Invariance of Convolution Operators) Show that the multiscale convolution of Example 5.3 does not satisfy the axiom [GSI] of gray value scaling invariance.

Exercise 5.3 (Scale Space Properties of the Moving Average) In the context of Example 5.3 let

$$\varphi(x) = \frac{\Gamma(1+d/2)}{\pi^{d/2}} \chi_{B_1(0)}(x)$$

(cf. Example 2.38) and $\varphi_t(x) = \tau(t)^{-d} \varphi\left(\frac{x}{\tau(t)}\right)$. We consider the scale space

$$\mathcal{T}_t u = \begin{cases} u * \varphi_t & \text{if } t > 0 \\ u & \text{if } t = 0. \end{cases}$$

Which scale space axioms are satisfied, and which are not? Can you show that an infinitesimal generator exists?

Exercise 5.4 (Recursivity of Scaled Dilation) Let $B \subset \mathbf{R}^d$ be nonempty.

1. Show that

$$B \text{ convex} \iff \text{for all } t, s \geq 0, tB + sB = (t + s)B.$$

2. Show that the multiscale dilation from Example 5.4 satisfies the axiom [REC] if B is convex. Which assumptions are needed for the reverse implication?

Exercise 5.5 (Properties of the Infinitesimal Generator) Let the assumptions of Theorem 5.11 be fulfilled. Show the following:

1. If axiom [TRANS] is satisfied in addition, then

$$A[u](x) = F(u(x), \nabla u(x), \nabla^2 u(x)).$$

2. If axiom [GLSI] is satisfied in addition, then

$$A[u](x) = F(x, \nabla u(x), \nabla^2 u(x)).$$

Exercise 5.6 (Infinitesimal Generator of Coordinate Transformations) Let (\mathcal{T}_t) be the multiscale coordinate transformation from Example 5.2, i.e.,

$$(\mathcal{T}_t u)(x) = u(j(t, x)),$$

where $j(\cdot, x)$ denotes the solution of the initial value problem

$$\frac{\partial j}{\partial t}(t, x) = v(j(t, x)), \quad j(0, x) = x.$$

Show that the infinitesimal generator of (\mathcal{T}_t) is given by

$$A[u](x) = v(x) \cdot \nabla u(x).$$

(You may assume that the generator exists.)

Exercise 5.7 (Gradient and Hessian and Gray Value Scalings) Let $h : \mathbf{R} \rightarrow \mathbf{R}$ and $u : \mathbf{R}^d \rightarrow \mathbf{R}$ be twice differentiable. Show that

$$\nabla(h \circ u) = h' \nabla u,$$

$$\nabla^2(h \circ u) = h' \nabla^2 u + h'' \nabla u \otimes \nabla u.$$

Exercise 5.8 (Auxiliary Calculation for Theorem 5.23) Let $X \in S^{d \times d}$ with $x_{d,d} = 0$ and $M = \sum_{i=1}^{d-1} x_{d,i}^2$. Moreover, let $\varepsilon > 0$ and

$$Q = \begin{pmatrix} 1 & & & \\ & \ddots & & \\ & & 1 & \\ & & & 0 \end{pmatrix}, \quad I_\varepsilon = \begin{pmatrix} \varepsilon & & & \\ & \ddots & & \\ & & \varepsilon & \\ & & & \frac{M}{\varepsilon} \end{pmatrix}.$$

Show that

$$Q X Q \preccurlyeq X + I_\varepsilon,$$

$$X \preccurlyeq Q X Q + I_\varepsilon.$$

Exercise 5.9 (Curvature of Level Sets) Let $u : \mathbf{R}^2 \rightarrow \mathbf{R}$ be twice differentiable and let (η, ξ) be the local coordinates.

1. Show that

$$\operatorname{div}\left(\frac{\nabla u}{|\nabla u|}\right) = \frac{\partial_{\xi\xi} u}{|\nabla u|}.$$

2. Let $c : [0, 1] \rightarrow \mathbf{R}^2$ be a twice differentiable curve. The *curvature* of c is, expressed in the coordinate functions $c(s) = (x(s), y(s))^\top$, as follows:

$$\kappa = \frac{x'y'' - x''y'}{\left((x')^2 + (y')^2\right)^{3/2}}.$$

Let $u : \mathbf{R}^2 \rightarrow \mathbf{R}$ be such that the zero level set $\{(x, y) \mid u(x, y) = 0\}$ is parameterized by such a curve c .

Show that on this zero level set at points with $\nabla u \neq 0$ one has

$$\kappa = \operatorname{div}\left(\frac{\nabla u}{|\nabla u|}\right).$$

Exercise 5.10 (Infinitesimal Generator of the Perona-Malik Equation) Show that the Perona-Malik equation has the infinitesimal generator

$$F(p, X) = \frac{g'(|p|)}{|p|} p^\top X p + g(|p|) \operatorname{trace} X$$

(cf. Lemma 5.26).

Exercise 5.11 (Maximum Principle for the One-Dimensional Perona-Malik Equation) We consider the Cauchy problem

$$\begin{aligned} \partial_t u &= \partial_x \left(g((\partial_x u)^2) \partial_x u \right) && \text{in } [0, \infty[\times \mathbf{R} \\ u(0, x) &= u_0(x) && \text{for } x \in \mathbf{R}. \end{aligned} \tag{5.28}$$

Let g be differentiable, u a solution of this Cauchy problem, and (t_0, x_0) a point such that the map $x \mapsto u(t_0, x)$ has a local maximum at x_0 .

1. Under what conditions on g is the map $t \mapsto u(t, x_0)$ decreasing at the point t_0 ?
2. Use the previous result to deduce a condition which implies that for all solutions u of (5.28) and all $t \geq 0, x \in \mathbf{R}$, one has

$$\inf_{x \in \mathbf{R}} u_0(x) \leq u(t, x) \leq \sup_{x \in \mathbf{R}} u_0(x).$$

(In this case one says that the solution satisfies a *maximum principle*.)

Exercise 5.12 (Decrease of Energy and Preservation of the Mean Gray Value for the Modified Perona-Malik Equation) Let $\Omega \subset \mathbf{R}^d$, $u_0 \in L^\infty(\Omega)$, $g : [0, \infty[\rightarrow [0, \infty[$ infinitely differentiable and $u : [0, T] \times \Omega \rightarrow \mathbf{R}$ a solution of the modified Perona-Malik equation, i.e., a solution of the initial-boundary value problem

$$\begin{aligned} \partial_t u &= \operatorname{div}(g(|\nabla u_\sigma|) \nabla u) && \text{in } [0, T] \times \Omega \\ \partial_\nu u &= 0 && \text{on } [0, T] \times \partial\Omega \\ u(0, x) &= u_0(x) && \text{for } x \in \Omega. \end{aligned} \tag{5.29}$$

1. Show that the quantity

$$h(t) = \int_{\Omega} u(t, x)^p \, dx$$

is nonincreasing for all $p \in [2, \infty[$.

2. Show that the quantity

$$\mu(t) = \int_{\Omega} u(t, x) \, dx$$

is constant.

Chapter 6

Variational Methods



6.1 Introduction and Motivation

To motivate the general approach and possibilities of variational methods in mathematical imaging, we begin with several examples. First, consider the problem to remove additive noise from a given image, i.e., of reconstructing u^\dagger from the data

$$u^0 = u^\dagger + \eta,$$

where the noise η is unknown. As we have already seen, there are different approaches to solve the denoising problem, e.g., the application of the moving average, morphological opening, the median filter, and the solution of the Perona-Malik equation.

Since we do not know the noise η , we need to make assumptions on u^\dagger and η and hope that these assumptions are indeed satisfied for the given data u^0 . Let us make some basic observations:

- The noise $\eta = u^0 - u^\dagger$ is a function whose value at every point is independent of the values in the neighborhood of that point. There is no special spatial structure in η .
- The function u^\dagger represents an image that has some spatial structure. Hence, it is possible to make assertions about the behavior of the image in the neighborhood of a point.

In a little bit more abstract terms, the image and the noise will have different characteristics that allow one to discriminate between these two; in this case these characteristics are given by the local behavior. These assumptions, however, do not lead to a mathematical model and of course not to a denoising method. The basic idea of variational methods in mathematical imaging is to express the above assumptions in quantitative expressions. Usually, these expressions say how “well”

a function “fits” the modeling assumption; it should be small for a good fit, and large for a bad fit.

With this in mind, we can reformulate the above points as follows:

- There is a real-valued function Φ that gives, for every “noise” function η , the “size” of the noise. The function Φ should use only point-wise information. “Large” noise, or the presence of spatial structure, should lead to large values.
- There is a real valued function Ψ that says how much an image u looks like a “natural image.” The function should use information from neighborhoods and should be “large” for “unnatural” images and “small” for “natural” images.

These assumptions are based on the hope that the quantification of the local behavior is sufficient to discriminate image information from noise. For suitable functions Φ and Ψ one chooses a weight $\lambda > 0$, and this leads, for every image u (and, consequently, for every noise $\eta = u - u^0$), to the expression

$$\Phi(u^0 - u) + \lambda \Psi(u),$$

which gives a value that says how well both requirements are fulfilled; the smaller, the better. Thus, it is natural to look for an image u that minimizes the expression, i.e., we are looking for u^* for which

$$\Phi(u^0 - u^*) + \lambda \Psi(u^*) = \min_u \Phi(u^0 - u) + \lambda \Psi(u)$$

holds.

Within the model given by Φ , Ψ , and λ , the resulting u^* is optimal and gives the denoised image. A characteristic feature of these methods is the solution of a minimization problem. Since one varies over all u to search for an optimum u^* , these methods are called *variational methods* or *variational problems*. The function to be minimized in a variational problem is called an *objective functional*. Since Φ measures the difference $u^0 - u$, it is often called a *discrepancy functional* or *discrepancy term*. In this context, Ψ is also called a *penalty functional*.

The following example, chosen such that the calculations remain simple, gives an impression as to how the model assumption can be transferred to a functional and what mathematical questions can arise in this context.

Example 6.1 (L^2 - H^1 Denoising) Consider the whole space \mathbf{R}^d and a (complex-valued) noisy function $u^0 \in L^2(\mathbf{R}^d)$. It appears natural to choose for Φ the squared norm

$$\Phi(u) = \frac{1}{2} \int_{\mathbf{R}^d} |u(x)|^2 dx.$$

Indeed, it uses only point-wise information. In contrast, the functional

$$\Psi(u) = \frac{1}{2} \int_{\mathbf{R}^d} |\nabla u(x)|^2 dx,$$

uses the gradient of u and hence uses in some sense also information from a neighborhood. The corresponding variational problem reads

$$\min_{u \in H^1(\mathbf{R}^d)} \frac{1}{2} \int_{\mathbf{R}^d} |u^0(x) - u(x)|^2 dx + \frac{\lambda}{2} \int_{\mathbf{R}^d} |\nabla u(x)|^2 dx. \quad (6.1)$$

Note that the gradient has to be understood in the weak sense, and hence the functional is well defined in the space $H^1(\mathbf{R}^d)$. Something that is not clear a priori is the existence of a minimizer and hence the justification to use a minimum instead of an infimum.

We will treat the question of existence of minimizers in greater detail later in this chapter and content ourselves with a formal solution of the minimization problem: by the Plancherel formula (4.2) and the rules for derivatives from Lemma 4.28, we can reformulate the problem (6.1) as

$$\min_{u \in H^1(\mathbf{R}^d)} \frac{1}{2} \int_{\mathbf{R}^d} |\widehat{u^0}(\xi) - \widehat{u}(\xi)|^2 d\xi + \frac{\lambda}{2} \int_{\mathbf{R}^d} |\xi|^2 |\widehat{u}(\xi)|^2 d\xi.$$

We see that this is now a minimization problem for \widehat{u} in which we aim to minimize an integral that depends only on $\widehat{u}(\xi)$. For this ‘‘point-wise problem,’’ and we will argue in more detail later, the overall minimization is achieved by ‘‘point-wise almost everywhere minimization.’’ The point-wise minimizer u^* satisfies for almost all ξ ,

$$\widehat{u}(\xi) = \arg \min_{z \in \mathbf{C}} \frac{1}{2} |\widehat{u^0}(\xi) - z|^2 + \frac{\lambda}{2} |\xi|^2 |z|^2.$$

Rewriting and with $z = |z| \operatorname{sgn}(z)$, we get

$$\frac{1}{2} |\widehat{u^0}(\xi) - z|^2 + \frac{\lambda}{2} |\xi|^2 |z|^2 = \frac{1}{2} (1 + \lambda |\xi|^2) |z|^2 + \frac{1}{2} |\widehat{u^0}(\xi)|^2 - |z| \operatorname{Re} (\operatorname{sgn}(z) \overline{\widehat{u^0}(\xi)}),$$

and hence the minimization with respect to the argument $\operatorname{sgn}(z)$ yields $\operatorname{sgn}(z) = \operatorname{sgn}(\widehat{u^0}(\xi))$. This leads to

$$\frac{1}{2} |\widehat{u^0}(\xi) - z|^2 + \frac{\lambda}{2} |\xi|^2 |z|^2 = \frac{1}{2} (1 + \lambda |\xi|^2) |z|^2 - |z| |\widehat{u^0}(\xi)| + \frac{1}{2} |\widehat{u^0}(\xi)|^2,$$

which we minimize with respect to the absolute value $|z|$ and obtain $|z| = |\widehat{u^0}(\xi)| / (1 + \lambda |\xi|^2)$. In total we get $z = \widehat{u^0}(\xi) / (1 + \lambda |\xi|^2)$, and hence \widehat{u}^* is unique and given by

$$\widehat{u}^*(\xi) = \frac{\widehat{u^0}(\xi)}{1 + \lambda |\xi|^2} \quad \text{for almost every } \xi \in \mathbf{R}^d.$$

Letting $\widehat{P_\lambda}(\xi) = (2\pi)^{-d/2}/(1 + \lambda|\xi|^2)$, we get with Theorem 4.27 on the convolution and the Fourier transform

$$u^* = u^0 * P_\lambda.$$

Using the $(d/2 - 1)$ th *modified Bessel function of the second kind* $K_{d/2-1}$, we can write P_λ as

$$P_\lambda(x) = \frac{|x|^{1-d/2}}{(2\pi)^{d-1}\lambda^{(d+2)/4}} K_{d/2-1}\left(\frac{2\pi|x|}{\sqrt{\lambda}}\right) \quad (6.2)$$

(cf. Exercise 6.1).

We see that variational denoising with squared L^2 -norm and squared H^1 -seminorm on the whole space leads to a certain class of linear convolution filters. In particular, they give a motivation to use the convolution kernels P_λ .

It is simple to implement the method numerically: instead of the continuous convolution one uses the discrete version and realizes the multiplication in frequency space with the help of the fast Fourier transform (see also Exercise 4.12). The result of this method can be seen in Fig. 6.1. Since the method corresponds to a linear filter, the noise reduction also leads to a loss of sharp edges in the image; cf. Example 4.19.

Minimization of functionals is used not only for denoising. All problems in which the inversion of an operation shall be done implicitly, so-called *inverse problems*, can be formulated in a variational context. This can be done by an adaptation of the discrepancy term.

Example 6.2 (H^1 Deconvolution) In Remark 4.21 we saw that blur can be modeled by the linear operation of convolution and that it can, in principle, be reversed by division in the Fourier-space. However, even for distortions so small that they are not visually detectable, such as a quantization to 256 gray levels, direct reconstruction leads to notably worse reconstruction quality. If the image u^0 even contains noise, or if the Fourier transform of the convolution kernel is close to zero at many places, deconvolution is not possible in this way.

Let us model the deblurring problem as a variational problem. If we assume that the convolution kernel $k \in L^1(\mathbf{R}^d) \cap L^2(\mathbf{R}^d)$ is known and satisfies $\int_{\mathbf{R}^d} k \, dx = 1$, then the data and the noise satisfy the identities

$$u^0 = u^\dagger * k + \eta \quad \text{and} \quad \eta = u^0 - u^\dagger * k, \quad \text{respectively.}$$

Hence, we should replace the term $\Phi(u^0 - u)$ by $\Phi(u^0 - u * k)$ in the respective minimization problem. Choosing Φ and Ψ similar to Example 6.1, we obtain the problem

$$\min_{u \in H^1(\mathbf{R}^d)} \frac{1}{2} \int_{\mathbf{R}^d} |u^0(x) - (u * k)(x)|^2 \, dx + \frac{\lambda}{2} \int_{\mathbf{R}^d} |\nabla u(x)|^2 \, dx. \quad (6.3)$$

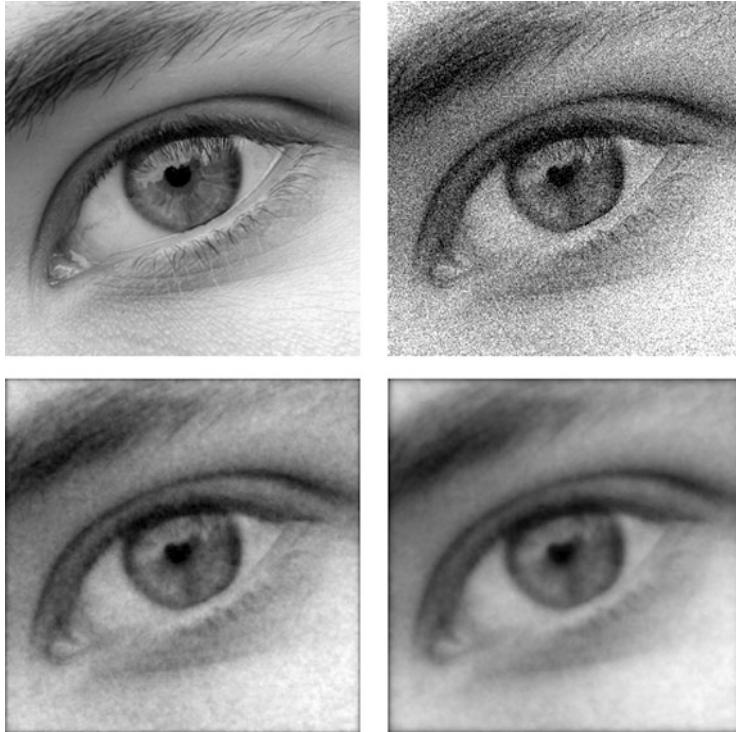


Fig. 6.1 Denoising by solving problem (6.1). Upper left: Original image u^\dagger with 256×256 pixels, upper right: Noisy version u^0 ($\text{PSNR}(u^0, u^\dagger) = 19.98 \text{ dB}$). Bottom row: Denoised images by solving the minimization problem (6.1) u^1 (left, $\text{PSNR}(u^1, u^\dagger) = 26.21 \text{ dB}$) and u^2 (right, $\text{PSNR}(u^2, u^\dagger) = 24.30 \text{ dB}$). The regularization parameters are $\lambda_1 = 25 \times 10^{-5}$ and $\lambda_2 = 75 \times 10^{-5}$, respectively and u^0 has been extended to \mathbf{R}^2 by 0

Similar to Example 6.1 we get, using the convolution theorem this time (Theorem 4.27), that the minimization is equivalent to

$$\min_{u \in H^1(\mathbf{R}^d)} \frac{1}{2} \int_{\mathbf{R}^d} |\widehat{u^0}(\xi) - (2\pi)^{d/2} \widehat{k}(\xi) \widehat{u}(\xi)|^2 d\xi + \frac{\lambda}{2} \int_{\mathbf{R}^d} |\xi|^2 |\widehat{u}(\xi)|^2 d\xi,$$

which can again be solved by point-wise almost everywhere minimization. Some calculations lead to

$$\widehat{u^*}(\xi) = \frac{\widehat{u^0}(\xi) (2\pi)^{d/2} \overline{\widehat{k}(\xi)}}{(2\pi)^d |\widehat{k}(\xi)|^2 + \lambda |\xi|^2} \quad \text{for almost all } \xi \in \mathbf{R}^d,$$

and hence the solution is again obtained by convolution, this time with the kernel

$$u^* = u^0 * k_\lambda, \quad k_\lambda = \mathcal{F}^{-1} \left(\frac{\widehat{k}}{(2\pi)^d |\widehat{k}|^2 + \lambda |\cdot|^2} \right). \quad (6.4)$$

We note that the assumptions $k \in L^1(\mathbf{R}^d) \cap L^2(\mathbf{R}^d)$ and $\int_{\mathbf{R}^d} k \, dx = 1$ guarantee that the denominator is continuous and bounded away from zero, and hence, we have $k_\lambda \in L^2(\mathbf{R}^d)$. For $\lambda \rightarrow 0$ it follows that $(2\pi)^{d/2} \widehat{k}_\lambda \rightarrow (2\pi)^{-d/2} \widehat{k}^{-1}$ point-wise, and hence we can say that the convolution with k_λ is in some sense a regularization of the division by $(2\pi)^{d/2} \widehat{k}$, which would be “exact” deconvolution.

The numerical implementation can be done similarly to Example 6.1. Figures 6.2 and 6.3 show some results for this method. In contrast to Remark 4.21 we used

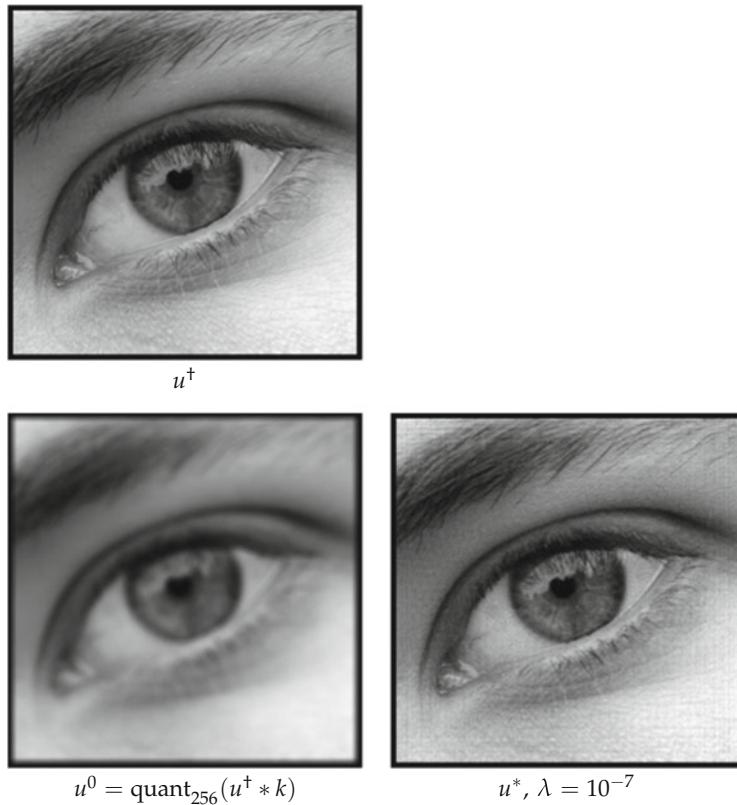


Fig. 6.2 Solutions of the deblurring problem in (6.3). Top left: Original image, extended by zero (264 × 264 pixel). Bottom left: Convolution with an out-of-focus kernel (diameter of 8 pixels) and quantized to 256 gray values (not visually noticeable). Bottom right: Reconstruction with (6.4) ($\text{PSNR}(u^*, u^\dagger) = 32.60 \text{ dB}$)

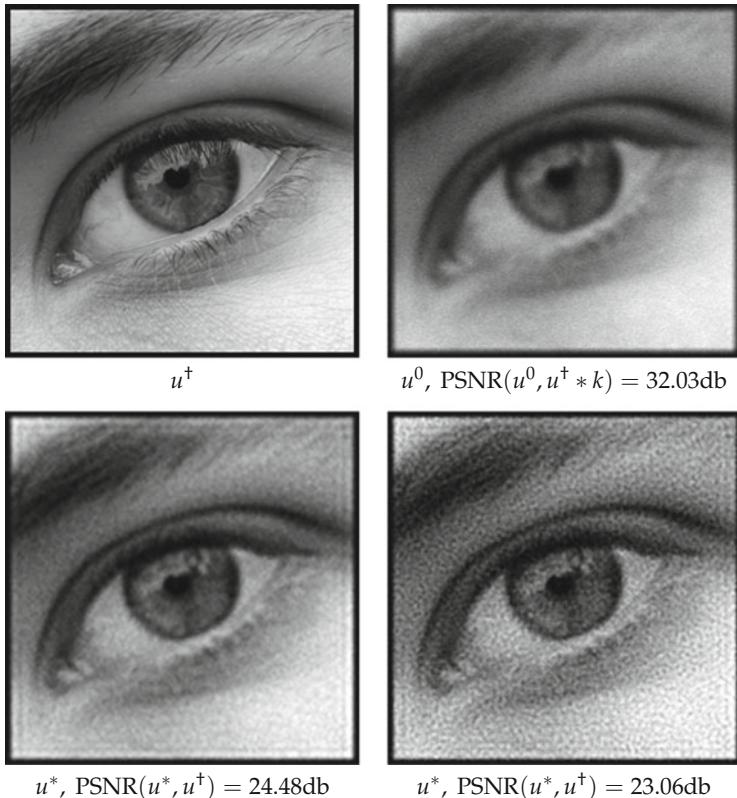


Fig. 6.3 Illustration of the influence of noise in u^0 on the deconvolution according to (6.3). Top left: Original image. Top right: Convolved image corrupted by additive normally distributed noise (visually noticeable). Bottom row: Reconstruction with different parameters ($\lambda = 5 \times 10^{-6}$ left, $\lambda = 10^{-6}$ right). Smaller parameters amplify the artifacts produced by noise

a convolution kernel for which the Fourier transform is equal to zero at many places. Hence, division in the Fourier space is not an option. This problem can be solved by the variational approach (6.3): even after quantization to 256 gray levels we can achieve a satisfactory, albeit not perfect, deblurring (Fig. 6.2; cf. Fig. 4.3). However, the method has its limitations. If we distort the image by additive noise, this will be amplified during the reconstruction (Fig. 6.3). We obtain images that look somehow sharper, but they contain clearly perceivable artifacts. Unfortunately, both “sharpness” and artifacts increase for smaller λ , and the best results (in the sense of visual perception) are obtained with a not-too-small parameter.

This phenomenon is due to the fact that deconvolution is an *ill-posed inverse problem*, i.e., minimal distortions in the data lead to arbitrarily large distortions in the solution. A clever choice of λ can lead to an improvement of the image, but some essential information seems to be lost. To some extent, this information is

amended by the “artificial” term $\lambda\Psi$ in the minimization problem. The penalty term Ψ , however, is part of our model that we have for our original image; hence, the result depends on how well u^\dagger is represented by that model. Two questions come up: what reconstruction quality can be expected, despite the loss of information, and how does the choice of the minimization problem influence this quality?

As a consequence, the theory of inverse problems [58] is closely related to mathematical imaging [126]. There one deals with the question how to overcome the problems induced by ill-posedness in greater detail and also answers, to some extent, the question how to choose the parameter λ . In this book we do not discuss questions of parameter choice and assume that the parameter λ is given.

Remark 6.3 The assumption that the convolution kernel k is known is a quite strong assumption in practice. If a blurred image u^0 is given, k can usually not be deduced from the image. Hence, one faces the problem of reconstructing both u^\dagger and k simultaneously, a problem called *blind deconvolution*. The task is highly ambiguous since one sees by the convolution theorem that for every u^0 there exist many pairs of u and k such that $(2\pi)^{d/2}\widehat{ku} = \widehat{u^0}$. This renders blind deconvolution considerably more difficult and we will restrict ourselves to the, already quite difficult, task of “non-blind” deconvolution. For blind deconvolution we refer to [14, 26, 39, 81].

As a final introductory example we consider a problem that seems to have a different flavor at first sight: inpainting.

Example 6.4 (Harmonic Inpainting) The task to fill in a missing part of an image in a natural way, called *inpainting*, can also be written as a minimization problem. We assume that the “true,” real valued image u^\dagger is given on a domain $\Omega \subset \mathbf{R}^d$, but on a proper subset $\Omega' \subset \Omega$ with $\overline{\Omega'} \subset\subset \Omega$ it is not known. Hence, the given data consists of Ω' and $u^0 = u^\dagger|_{\Omega \setminus \Omega'}$.

Since we have to “invent” suitable data, the model that we have for an image is of great importance. Again we note that the actual brightness value of an image is not so important in comparison to the behavior of the image in a local neighborhood. As before, we take the Sobolev space $H^1(\Omega)$ as a model (this time real-valued) and postulate $u^\dagger \in H^1(\Omega)$. In particular, we assume that the corresponding (semi-)norm measures how well an element $u \in H^1(\Omega)$ resembles a natural image. The task of inpainting is then formulated as the minimization problem

$$\min_{\substack{u \in H^1(\Omega) \\ u=u^0 \text{ on } \Omega \setminus \Omega'}} \frac{1}{2} \int_{\Omega} |\nabla u(x)|^2 \, dx. \quad (6.5)$$

Again, we are interested in a concrete minimizer, but in contrast to the previous examples, the Fourier transform is not helpful here, since the domains Ω and Ω' introduce an inherent dependence of the space that is not reflected in the frequency representation.

We use a different technique. Assume that we know a minimizer $u^* \in H^1(\Omega)$. For every other $u \in H^1(\Omega)$ with $u = u^0$ in $\Omega \setminus \Omega'$ we consider the function

$$F_u(t) = \frac{1}{2} \int_{\Omega} |\nabla(u^* + t(u - u^*))(x)|^2 dx,$$

which has, by minimality of u^* , a minimum at $t = 0$. We take the derivative of F_u with respect to t in $t = 0$ and obtain

$$\begin{aligned} \frac{\partial F_u}{\partial t}(0) &= \lim_{t \rightarrow 0} \frac{F_u(t) - F_u(0)}{t} \\ &= \lim_{t \rightarrow 0} \left[\frac{1}{2t} \int_{\Omega} |\nabla u^*(x)|^2 dx + \int_{\Omega} \nabla u^*(x) \cdot \nabla(u - u^*)(x) dx \right. \\ &\quad \left. + \frac{t}{2} \int_{\Omega} |\nabla(u - u^*)(x)|^2 dx - \frac{1}{2t} \int_{\Omega} |\nabla u^*(x)|^2 dx \right] \\ &= \int_{\Omega} \nabla u^*(x) \cdot \nabla(u - u^*)(x) dx = 0. \end{aligned}$$

The set

$$\{u - u^* \in H^1(\Omega) \mid u = u^0 \text{ in } \Omega \setminus \Omega'\} = \{v \in H^1(\Omega) \mid v = 0 \text{ in } \Omega \setminus \Omega'\}$$

is a subspace of $H^1(\Omega)$, and it can be shown that it is equal to $H_0^1(\Omega')$ (see Exercise 6.2). Hence, u^* has to satisfy

$$\int_{\Omega'} \nabla u^*(x) \cdot \nabla v(x) dx = 0 \text{ for all } v \in H_0^1(\Omega'). \quad (6.6)$$

This is the weak form of the so-called *Euler-Lagrange equation* associated to (6.5). In fact, (6.6) is the weak form of a partial differential equation. If u^* is twice continuously differentiable in Ω' , one has for every $v \in \mathcal{D}(\Omega')$ that

$$\int_{\Omega'} \nabla u^*(x) \cdot \nabla v(x) dx = - \int_{\Omega'} \Delta u^*(x) v(x) dx = 0$$

and by the fundamental lemma of the calculus of variations (Lemma 2.75) we obtain $\Delta u^* = 0$ in Ω' , i.e. the function u^* is *harmonic* there (see [80] for an introduction to the theory of harmonic functions). It happens that u^* is indeed always twice differentiable, since it is *weakly harmonic*, i.e., it satisfies

$$\int_{\Omega'} u^*(x) \Delta v(x) dx = 0 \text{ for all } v \in \mathcal{D}(\Omega').$$

By Weyl's lemma [148, Theorem 18.G] such functions are infinitely differentiable in Ω' . If we further assume that $u^*|_{\Omega'}$ has a trace on $\partial\Omega'$, it has to be the same as the trace of u^0 (which is given only on $\Omega \setminus \Omega'$). This leads to the conditions

$$\Delta u^* = 0 \text{ in } \Omega', \quad u^* = u^0 \text{ on } \partial\Omega',$$

which is the strong form of the Euler-Lagrange equation for (6.5). Hence, the inpainting problem with H^1 norm leads to the solution of the Laplace equation with so-called *Dirichlet boundary values*, and is also called *harmonic inpainting*.

Some properties of u^* are immediate. On the one hand, harmonic functions satisfy the maximum principle, which says that non constant u^* do not have local maxima and minima in Ω' . This says that the solution is unique and that harmonic inpainting cannot introduce new structures. This seems like a reasonable property. On the other hand, u^* satisfies the mean-value property

$$u^*(x) = \frac{1}{\mathcal{L}^d(B_r(0))} \int_{B_r(0)} u^*(x - y) dy$$

as soon as $\overline{B_r(x)} \subset \Omega'$. In other words, u^* is invariant under “out-of-focus” blur (compare Example 3.12). The strength of the averaging, i.e., the radius r depends on Ω' : for “larger” regions we choose r larger. We already know that the moving average blurs edges and regions of high contrast. Since u^* is the result of an average, we cannot expect sharp edges there. Thus, the model behind the variational method (6.5) cannot extend edges into the inpainting region. We may conclude that harmonic inpainting is most useful in filling in homogeneous regions.

A numerical implementation can be done, for example, by finite differences as in Sect. 5.4. Figure 6.4 shows an application of this method. In the left example it looks, at first glance, that homogeneous region can be filled in a plausible way. A closer look shows that indeed edges and regions with high contrast changes are blurred. This effect is much stronger in the right example. The reconstruction shows blurred regions that do not look plausible. One also notices that the blurred regions do not look that much diffuse if the domain of inpainting is small in relation to the image structures. In conclusion, we see a good fit with the theoretical results. It remains open whether a better choice of the model (or minimization problem) leads to a method with comparably good reconstruction but that is able to extend edges.

Remark 6.5 (Compression by Inpainting) Figure 6.4 shows that harmonic inpainting can be used to compress images: Is it possible to choose Ω' in a way such that it is possible to reconstruct a high quality image? If Ω' is chosen such that it contains only smooth regions, the reconstruction by harmonic inpainting looks similar to the original image. This approach is worked out in [64], for example.



Fig. 6.4 Inpainting by minimization of the H^1 norm. Left column: The original image (top, 256×256 pixels) contains some homogeneous regions that should be reconstructed by inpainting (middle, Ω' is the checkerboard pattern). Bottom, result of the minimization of (6.5). Right column: The original image (top, 256×256 pixels) contains some fine structures with high contrast, which has been removed in the middle picture. The bottom shows the result of harmonic inpainting

As we have seen, we can use minimization problems to solve various tasks in mathematical imaging. We aim to provide a general mathematical foundation to treat such problems. As a start, we collect the questions that have arisen so far:

- How do we motivate the choice of the discrepancy and penalty terms and what influence do they have?

In the previous example we had, similarly to linear regression, quadratic discrepancy terms. Also, the gradient entered quadratically into the penalty term. What changes if we use different models or different functionals and which functionals are better suited for imaging?

- How can we ensure existence and uniqueness of minimizers?

Our calculation in the previous examples implicitly assumed the existence of minimizers. Strictly speaking, we still need to prove that the obtained functions indeed minimize the respective functionals (although this seems obvious in the example of point-wise almost everywhere minimization in Fourier space). A theory that provides existence, and, if possible, also uniqueness from general assumptions is desirable.

- Are there general methods to calculate minimizers?

For a general functional it will not be possible to obtain an explicit expression for the solution of a variational problem. How can we still describe the solutions? If the solution cannot be given explicitly, we are in need of numerical approximation schemes that are broadly applicable.

Each of these questions can be answered to a good extent by well-developed mathematical theories. In the following we describe the aspects of these theories that are relevant for mathematical imaging. We aim for great generality, but keep the concrete applications in mind. We will begin with fundamentals of the topic.

- The question of existence of solutions of minimization problem can be answered with techniques from functional analysis. Here one deals with a generalization of the theorem of Weierstrass (also called the extreme value theorem) to infinite dimensional (function) spaces. We give an overview of generalizations that are relevant for our applications.
- An important role for the uniqueness of solutions, but also for their description, is played by convex functionals. *Convex analysis* and the calculus of *subdifferentials* provide a unified theory for convex minimization problems. In the following we will develop the most important and fundamental results of this theory.
- Functions with discontinuities play a prominent role in mathematical imaging, since these discontinuities model edges in images. The space of functions with bounded total variation is the most widely used model in this context. We introduce this Banach space and prove some important results.

Building on the developed theory, we will model and analyze various variational method in the context of mathematical imaging. In particular, we treat the following problems and show how variational methods and the analysis of these can be used to solve these problems:

- Denoising and image decomposition,
- deblurring,
- restoration/inpainting,
- interpolation/zooming.

Moreover, we will discuss numerical methods to solve the respective minimization problems.

6.2 Foundations of the Calculus of Variations and Convex Analysis

6.2.1 The Direct Method

One of the most widely used proof techniques to show the existence of minimizers of some functional is the *direct method in the calculus of variations*. Its line of argumentation is very simple and essentially follows three abstract steps.

Before we treat the method we recall the definition of the *extended real numbers*:

$$\mathbf{R}_\infty =]-\infty, \infty] = \mathbf{R} \cup \{\infty\}.$$

Of course, we set $t \leq \infty$ for all $t \in \mathbf{R}_\infty$ and $t < \infty$ if and only if $t \in \mathbf{R}$. Moreover, we use the formal rules $t + \infty = \infty$ for $t \in \mathbf{R}_\infty$ and $t \cdot \infty = \infty$ if $t > 0$ as well as $0 \cdot \infty = 0$. Subtraction of ∞ as well as multiplication of ∞ by negative numbers are not defined. For mappings $F : X \rightarrow \mathbf{R}_\infty$ let $\text{dom } F = \{u \in X \mid F(u) < \infty\}$ be the *effective domain of definition*. Often we want to exclude the special case $\text{dom } F = \emptyset$, and hence we call F *proper* if F is not constant ∞ .

We further recall the following notions.

Definition 6.6 (Epigraph) The *epigraph* of a functional $F : X \rightarrow \mathbf{R}_\infty$ is the set

$$\text{epi } F = \{(u, t) \in X \times \mathbf{R} \mid F(u) \leq t\}.$$

Definition 6.7 (Sequential Lower Semicontinuity) A functional $F : X \rightarrow \mathbf{R}_\infty$ on a topological space X is *sequential lower semicontinuous* if for every sequence (u^n) and $u \in X$ with $\lim_{n \rightarrow \infty} u^n = u$, one has

$$F(u) \leq \liminf_{n \rightarrow \infty} F(u^n).$$

Remark 6.8 A functional $F : X \rightarrow \mathbf{R}_\infty$ is sequential lower semicontinuous if and only if $\text{epi } F$ is sequentially closed in $X \times \mathbf{R}$:

For every sequence $((u^n, t_n))$ in $\text{epi } F$ with $\lim_{n \rightarrow \infty} (u^n, t_n) = (u, t)$, one has for every n that $F(u^n) \leq t_n$ and thus

$$F(u) \leq \liminf_{n \rightarrow \infty} F(u^n) \leq \liminf_{n \rightarrow \infty} t_n = t \quad \Rightarrow \quad (u, t) \in \text{epi } F.$$

Conversely, for every sequence (u^n) in X with $\lim_{n \rightarrow \infty} u^n = u$ there exists a subsequence (u^{n_k}) such that for $t_n = F(u^n)$, one has $\lim_{k \rightarrow \infty} t_{n_k} = \liminf_{n \rightarrow \infty} F(u^n)$. We conclude that $F(u) \leq \liminf_{n \rightarrow \infty} F(u^n)$.

With these notions in hand, we can describe the direct method as follows:

To Show

The functional $F : X \rightarrow \mathbf{R}_\infty$, defined on a topological space X has a minimizer u^* , i.e.,

$$F(u^*) = \inf_{u \in X} F(u) = \min_{u \in X} F(u).$$

The Direct Method

1. Establish that F is bounded from below, i.e., $\inf_{u \in X} F(u) > -\infty$. By the definition of the infimum this implies the existence of a minimizing sequence (u^n) , i.e. a sequence such that $F(u^n) < \infty$ for all n and $\lim_{n \rightarrow \infty} F(u^n) = \inf_{u \in X} F(u)$.
2. Show that the sequence (u^n) lies in a set that is sequentially compact with respect to the topology on X . This implies the existence of a subsequence of (u^n) , denoted by (u^{n_k}) , and a $u^* \in X$ such that $\lim_{k \rightarrow \infty} u^{n_k} = u^*$, again with respect to the topology in X . This u^* is a “candidate” for a minimizer.
3. Prove sequential lower semicontinuity of F with respect to the topology on X . Applying this to the subsequence from above, we obtain

$$\inf_{u \in X} F(u) \leq F(u^*) \leq \liminf_{k \rightarrow \infty} F(u^{n_k}) = \inf_{u \in X} F(u),$$

which shows that u^* has to be a minimizer.

If one is about to apply this method to a given minimization problem, one needs to decide on one open point, namely the choice of the topology on the set X . While the first step is independent of the topology, it influences the following two steps. However, the second and third steps have somehow competing demands on the topology: in weaker (or coarser) topologies there are more convergent sequences, and consequently, more sequentially compact sets. By contrast, the functional F has to satisfy the \liminf condition in step three for more sequences, and hence there are fewer sequentially lower semicontinuous functionals for weaker topologies.

We exemplify this issue for topologies on Banach spaces using Examples 6.1–6.4, since most applications are in this context. Assume that a given minimization problem is stated on a Banach space X , and one chooses the, somehow natural, strong topology on this space. Then the requirement of compactness is an unfortunately strong restriction.

Remark 6.9 (Compactness in Examples 6.1–6.4) For the functional in (6.1), (6.3), and (6.5) one can show only that every minimizing sequence (u^n) is bounded in $X = H^1(\mathbf{R}^d)$ or $X = H^1(\Omega)$, respectively. For the functionals in (6.1) and (6.3), for example, the form of the objective functional implies that there exists a $C > 0$ such that for all n ,

$$\|\nabla u^n\|_2^2 \leq \frac{2}{\lambda} F(u^n) \leq C,$$

which immediately implies the boundedness of the sequence (u^n) , (translations by constant functions are not possible in these examples). It is easy to see, that also the incorporation of the discrepancy term does not allow stronger claims. Since the space $H^1(\mathbf{R}^d)$ is infinite-dimensional, there is no assertion about (pre-)compactness of the sequence (u_n) possible in this situation. In the case of the functional in (6.5), one directly concludes that $\|\nabla u^n\|_2^2 \leq C$ for a minimizing sequence and continuous the reasoning as above.

What seems like a dead end at first sight has a simple remedy. An infinite dimensional Banach space X has a larger collection of sets that are compact in the weak topology. The situation is especially simple for reflexive Banach spaces: as a consequence of the theorem of Eberlein-Šmulian, every bounded and weakly sequentially closed set is weakly compact. In our examples we can conclude the following:

Remark 6.10 (Weak Sequential Compactness in Examples 6.1–6.4) The spaces $X = H^1(\mathbf{R}^d)$ and $X = H^1(\Omega)$ are Hilbert spaces and hence reflexive. As we have seen in Remark 6.9, every minimizing sequence for (6.1), (6.3), (6.5) is bounded in these spaces. Consequently, we can always extract a subsequence that is at least weakly convergent in X .

Therefore, one often uses the weak topology on reflexive Banach spaces for the second step of the direct method. An important notion in this context, which also gives a simple criterion for boundedness on minimizing sequences, is the following:

Definition 6.11 (Coercivity) Let X be a Banach space. A functional $F : X \rightarrow \mathbf{R}_\infty$ is called *coercive* if

$$F(u^n) \rightarrow \infty \quad \text{for} \quad \|u^n\|_X \rightarrow \infty.$$

Furthermore, it is *strongly coercive* if $F(u^n)/\|u^n\|_X \rightarrow \infty$ for $\|u^n\|_X \rightarrow \infty$.

With this notion we argue as follows:

Lemma 6.12 *Let X be reflexive and $F : X \rightarrow \mathbf{R}_\infty$ proper and coercive. Then every minimizing sequence (u^n) for F has a weakly convergence subsequence.*

Proof Coercivity of F implies that every minimizing sequence (u^n) is bounded (otherwise, there would be a subsequence, which again would be a minimizing sequence, for which the functional value would tend to infinity, and hence the

sequence would not be a minimal sequence). Hence, (u^n) is contained in a closed ball, i.e., a weakly sequentially compact set. \square

Remark 6.13 The following criterion is sufficient for coercivity: there exist an $R \geq 0$ and a $\varphi : [R, \infty[\rightarrow \mathbf{R}_\infty$ with $\lim_{t \rightarrow \infty} \varphi(t) = \infty$ such that

$$F(u) \geq \varphi(\|u\|_X) \quad \text{for} \quad \|u\|_X \geq R.$$

A similar conclusion holds for strong coercivity if $\lim_{t \rightarrow \infty} \varphi(t)/t = \infty$.

It is not difficult to see that coercivity is not necessary for the existence of minimizers (Exercise 6.3). Strong coercivity is a special case of coercivity and may play a role if the sum of two functionals is to be minimized.

To complete the argument of the direct method, F has to be weakly sequentially lower semicontinuous, i.e., sequentially lower semicontinuous with respect to the weak topology. This allows us to perform the third and final step of the direct method. To show weak sequential lower semicontinuity (in the following we say simply weak lower semicontinuity) it is helpful to know some sufficient conditions:

Lemma 6.14 *Let X, Y be Banach spaces and $F : X \rightarrow \mathbf{R}_\infty$ a functional. Then the following hold:*

1. *If F is weakly lower semicontinuous, then so is αF for $\alpha \geq 0$.*
2. *If F and $G : X \rightarrow \mathbf{R}_\infty$ are weakly lower semicontinuous, then so is $F + G$.*
3. *If F is weakly lower semicontinuous and $\varphi : \mathbf{R}_\infty \rightarrow \mathbf{R}_\infty$ is monotonically increasing and lower semicontinuous, then $\varphi \circ F$ is weakly lower semicontinuous.*
4. *If $\Phi : Y \rightarrow X$ is weakly sequentially continuous and F weakly lower semicontinuous, then $F \circ \Phi$ is weakly lower semicontinuous.*
5. *For every nonempty family $F_i : X \rightarrow \mathbf{R}_\infty$, $i \in I$, of weakly lower semicontinuous functionals, the point-wise supremum $\sup_{i \in I} F_i$ is also weakly lower semicontinuous.*
6. *Every functional of the form $L_{x^*, \varphi} = \varphi \circ \langle x^*, \cdot \rangle_{X^* \times X}$ with $x^* \in X^*$ and $\varphi : \mathbf{K} \rightarrow \mathbf{R}_\infty$ lower semicontinuous is weakly lower semicontinuous on X .*

Proof In the following let (u^n) be a sequence in X such that $u^n \rightharpoonup u$ for some $u \in X$.

Assertions 1 and 2: Simple calculations show that

$$\alpha F(u) \leq \alpha \liminf_{n \rightarrow \infty} F(u^n) = \liminf_{n \rightarrow \infty} (\alpha F)(u^n),$$

$$(F + G)(u) \leq \liminf_{n \rightarrow \infty} F(u^n) + \liminf_{n \rightarrow \infty} G(u^n) \leq \liminf_{n \rightarrow \infty} (F + G)(u^n).$$

Assertion 3: Applying monotonicity of φ to the defining property of weak lower semicontinuity gives

$$F(u) \leq \liminf_{n \rightarrow \infty} F(u^n) \quad \Rightarrow \quad \varphi(F(u)) \leq \varphi\left(\liminf_{n \rightarrow \infty} F(u^n)\right).$$

For any subsequence n_k , for which $F(u^{n_k})$ converges, we get by monotonicity and lower semicontinuity of φ

$$\varphi\left(\liminf_{n \rightarrow \infty} F(u^n)\right) \leq \varphi\left(\lim_{k \rightarrow \infty} F(u^{n_k})\right) \leq \liminf_{k \rightarrow \infty} \varphi(F(u^{n_k})).$$

Since we can argue as above for any subsequence, we obtain the claim.

Assertion 4: For $u^n \rightharpoonup u$ in Y one has $\Phi(u^n) \rightharpoonup \Phi(u)$ in X , and thus

$$F(\Phi(u)) \leq \liminf_{n \rightarrow \infty} F(\Phi(u^n)).$$

Assertion 5: For every $n \in N$ and $i \in I$ one has $F_i(u^n) \leq \sup_{i \in I} F_i(u^n)$, and hence we conclude that

$$F_i(u) \leq \liminf_{n \rightarrow \infty} F_i(u^n) \leq \liminf_{n \rightarrow \infty} \sup_{i \in I} F_i(u^n) \quad \Rightarrow \quad \sup_{i \in I} F_i(u) \leq \liminf_{n \rightarrow \infty} \sup_{i \in I} F_i(u^n)$$

by taking the supremum.

Assertion 6: By the definition of weak convergence, $u \mapsto \langle x^*, u \rangle_{X^* \times X}$ is continuous, and the assertion follows from Assertion 4. \square

Corollary 6.15 *For monotonically increasing and lower semicontinuous $\varphi : [0, \infty[\rightarrow \mathbf{R}$ the functional $F(u) = \varphi(\|u\|_X)$ is weakly lower semicontinuous.*

Proof In a Banach space we have

$$\|u\|_X = \sup_{\|x^*\|_{X^*} \leq 1} |\langle x^*, u \rangle| = \sup_{\|x^*\|_{X^*} \leq 1} L_{x^*, |\cdot|}(u),$$

and hence the norm is, by Lemma 6.14, items 5 and 6, weakly lower semicontinuous. The claim follows by item 3 of that lemma. \square

Example 6.16 (Weak Lower Semicontinuity for Examples 6.1–6.4) Now we have all the ingredients to prove weak lower semicontinuity of the functionals in the examples from the beginning of this chapter.

1. We write the functional $F_1(u) = \frac{1}{2} \int_{\mathbf{R}^d} |u^0 - u|^2 dx$ with $\varphi(x) = \frac{1}{2}x^2$ and $\Phi(u) = u - u^0$ as

$$F_1 = \varphi \circ \|\cdot\|_{L^2} \circ \Phi.$$

Here we consider the mapping Φ as a mapping from $H^1(\mathbf{R}^d)$ to $L^2(\mathbf{R}^d)$. Since the embedding $H^1(\mathbf{R}^d) \hookrightarrow L^2(\mathbf{R}^d)$ is linear and continuous (which is simple to see), it is also weakly sequentially continuous (see Remark 2.24), and hence the same holds for Φ , since it involves only an additional translation by u^0 . By Corollary 6.15, the composition $\varphi \circ \|\cdot\|_{L^2}$ is weakly lower semicontinuous, and by Lemma 6.14, item 4, F_1 is so, too. The weak lower semicontinuity of

$F_2(u) = \frac{\lambda}{2} \int_{\mathbf{R}^d} |\nabla u|^2 dx$ is shown similarly: we use $\varphi(x) = \frac{\lambda}{2}x^2$ and $\Phi = \nabla$ and write

$$F_2 = \varphi \circ \|\cdot\|_{L^2} \circ \Phi$$

(where $\nabla : H^1(\mathbf{R}^d) \rightarrow L^2(\mathbf{R}^d, \mathbf{R}^d)$). By Lemma 6.14, item 2, we get that $F = F_1 + F_2$ is weakly lower semicontinuous on $H^1(\mathbf{R}^d)$, which is exactly the functional from (6.1).

2. For (6.3) we work analogously. The only difference is that we use $\Phi(u) = u^0 - u * k$ for the functional F_1 . The weak lower semicontinuity is then a simple consequence of Young's inequality from Theorem 3.13.
3. For Example 6.4, we need to consider the restriction $u = u^0$ on $\Omega \setminus \Omega'$. We treat it as follows:

$$F_1(u) = I_{H_0^1(\Omega')}(u - u^0), \quad I_{H_0^1(\Omega')}(v) = \begin{cases} 0 & \text{if } v \in H_0^1(\Omega') \\ \infty & \text{otherwise.} \end{cases}$$

It is easy to see (Exercise 6.2) that $H_0^1(\Omega')$ is a closed subspace of $H^1(\Omega)$. For every sequence (u^n) in $H_0^1(\Omega') \subset H^1(\Omega)$ with weak limit $u \in H^1(\Omega)$, one has

$$v \in H_0^1(\Omega')^\perp \subset H^1(\Omega) : \quad (u, v) = \lim_{n \rightarrow \infty} (u^n, v) = 0.$$

Hence, we have $u \in H_0^1(\Omega')$, and this shows that $I_{H_0^1(\Omega')}$ is weakly lower semicontinuous. With F_2 from item 1 and $F = F_1 + F_2$ we obtain the weak lower semicontinuity of the functional in (6.5).

This settles the question of existence of minimizing elements for the introductory problems. We note the general procedure in the following theorem:

Theorem 6.17 (The Direct Method in Banach Spaces) *Let X be a reflexive Banach space and let $F : X \rightarrow \mathbf{R}_\infty$ be bounded from below, coercive, and weakly lower semicontinuous. Then the problem*

$$\min_{u \in X} F(u)$$

has a solution in X .

For dual spaces X^* of separable normed spaces (not necessarily reflexive) one can prove a similar claim under the assumption that F is weakly* lower semicontinuous (Exercise 6.4). As we have seen, the notion of weak lower semicontinuity is a central element of the argument. Hence, the question as to which functionals have this property is well investigated within the calculus of variations. However, there is no general answer to this question. The next example highlights some of the difficulties that can arise with weak lower semicontinuity.

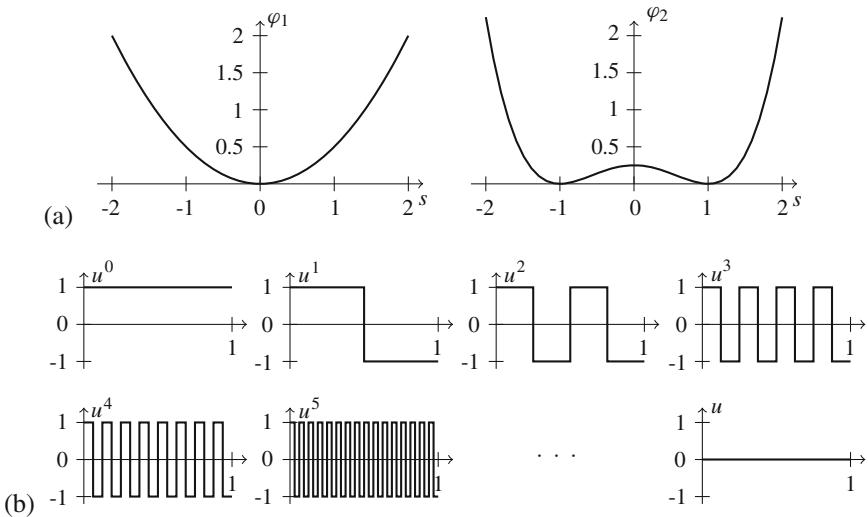


Fig. 6.5 Visualization of the functions from Example 6.18. (a) The pointwise “energy functions” φ_1 and φ_2 . (b) The first elements of the sequence (u^n) and its weak limit u

Example 6.18 (Example/Counterexample for Weak Lower Semicontinuity) Let $X = L^2([0, 1])$, $\varphi_1(x) = \frac{1}{2}x^2$ and $\varphi_2(x) = \frac{1}{4}(x-1)^2(x+1)^2$, depicted in Fig. 6.5a. Consider the functionals

$$F_1(u) = \int_0^1 \varphi_1(u(t)) dt, \quad F_2(u) = \int_0^1 \varphi_2(u(t)) dt.$$

By Corollary 6.15, $F_1 = \varphi_1 \circ \|\cdot\|_2$ is weakly lower semicontinuous on X . However, F_2 is not: the sequence (u^n) given by

$$u^n(t) = v(2^n t), \quad v(t) = \begin{cases} 1 & \text{if } 2k \leq t < 2k + 1 \text{ for some } k \in \mathbf{Z}, \\ -1 & \text{otherwise,} \end{cases}$$

forms a set of mutually orthonormal vectors, and hence by a standard argument in Hilbert space theory (see Sect. 2.1.3), it converges weakly to $u = 0$ (cf. Fig. 6.5b). But

$$\forall n \in \mathbf{N} : \quad F_2(u^n) = 0, \quad F_2(u) = \frac{1}{4},$$

and thus $F_2(u) > \liminf_{n \rightarrow \infty} F_2(u^n)$.

Although the functionals F_1 and F_2 have a similar structure, they differ with respect to weak lower semicontinuity. The reason for this difference lies in the fact

that φ_1 is convex, while φ_2 is not. This is explained within the theory of *convex analysis*, which we treat in the next section.

During the course of the chapter we will come back to the notion of weak lower semicontinuity of functionals. But for now, we end this discussion with a remark.

Remark 6.19 For every functional $F : X \rightarrow \mathbf{R}_\infty$, for which there exists a weakly lower semicontinuous $F_0 : X \rightarrow \mathbf{R}_\infty$ such that $F_0 \leq F$ holds pointwise, one can consider the following construction:

$$\underline{F}(u) = \sup \{G(u) \mid G : X \rightarrow \mathbf{R}_\infty, G \leq F, G \text{ weakly lower semicontinuous}\}.$$

By Lemma 6.14, item 5, this leads to a weakly lower semicontinuous functional with $\underline{F} \leq F$. By construction, it is the largest such functional below F , and it is called the *weak lower semicontinuous envelope* of F .

6.2.2 Convex Analysis

This and the following subsection give an overview of the basic ideas of convex analysis, where we focus on the applications to variational problems in mathematical imaging. The results, more details, and further material can be found in standard references on convex analysis such as [16, 57, 118]. We focus our study of convex analysis on convex functionals and recall their definition.

Definition 6.20 (Convexity of Functionals) A functional $F : X \rightarrow \mathbf{R}_\infty$ on a normed space X is called *convex* if for all $u, v \in X$ and $\lambda \in [0, 1]$, one has

$$F(\lambda u + (1 - \lambda)v) \leq \lambda F(u) + (1 - \lambda)F(v).$$

It is called *strictly convex* if for all $u, v \in X$ with $u \neq v$ and $\lambda \in]0, 1[$, one has

$$F(\lambda u + (1 - \lambda)v) < \lambda F(u) + (1 - \lambda)F(v).$$

We will study convex functionals in depth in the following, and we will see that they have several nice properties. These properties make them particularly well suited for minimization problems. Let us start with fairly obvious constructions and identify some general examples of convex functionals. The method from Lemma 6.14 can be applied to convexity in a similar way.

Lemma 6.21 (Construction of Convex Functionals) *Let X, Y be normed spaces and $F : X \rightarrow \mathbf{R}_\infty$ convex. Then we have the following:*

1. *For $\alpha \geq 0$ the functional αF is convex.*
2. *If $G : X \rightarrow \mathbf{R}_\infty$ is convex, then so is $F + G$.*

3. For $\varphi : \mathbf{R}_\infty \rightarrow \mathbf{R}_\infty$ convex and monotonically increasing on the range of F , the composition $\varphi \circ F$ is also convex.
4. For $\Phi : Y \rightarrow X$ affine linear, i.e., $\Phi(\lambda u + (1 - \lambda)v) = \lambda\Phi(u) + (1 - \lambda)\Phi(v)$ for all $u, v \in Y$ and $\lambda \in [0, 1]$, the composition $F \circ \Phi$ is convex on Y .
5. The pointwise supremum $F(u) = \sup_{i \in I} F_i(u)$ of a family of convex functionals $F_i : X \rightarrow \mathbf{R}_\infty$ with $i \in I$ and $I \neq \emptyset$ is convex.

Proof Assertions 1 and 2: This follows by simple calculations.

Assertion 3: One simply checks that for $u, v \in X$, $\lambda \in [0, 1]$, one has

$$\varphi(F(\lambda u + (1 - \lambda)v)) \leq \varphi(\lambda F(u) + (1 - \lambda)F(v)) \leq \lambda\varphi(F(u)) + (1 - \lambda)\varphi(F(v)).$$

Assertion 4: Affine linear mappings are compatible with convex combinations. Hence,

$$F(\Phi(\lambda u + (1 - \lambda)v)) = F(\lambda\Phi(u) + (1 - \lambda)\Phi(v)) \leq \lambda F(\Phi(u)) + (1 - \lambda)F(\Phi(v))$$

for $u, v \in Y$ and $\lambda \in [0, 1]$.

Assertion 5: Let $u, v \in X$ and $\lambda \in [0, 1]$. For all $i \in I$ we have, by definition of the supremum, that

$$F_i(\lambda u + (1 - \lambda)v) \leq \lambda F_i(u) + (1 - \lambda)F_i(v) \leq \lambda F(u) + (1 - \lambda)F(v),$$

and consequently, the same holds for the supremum $F(\lambda u + (1 - \lambda)v)$ over all $i \in I$.

□

Remark 6.22 Similar claims to those in Lemma 6.21 can be made for strictly convex functionals F :

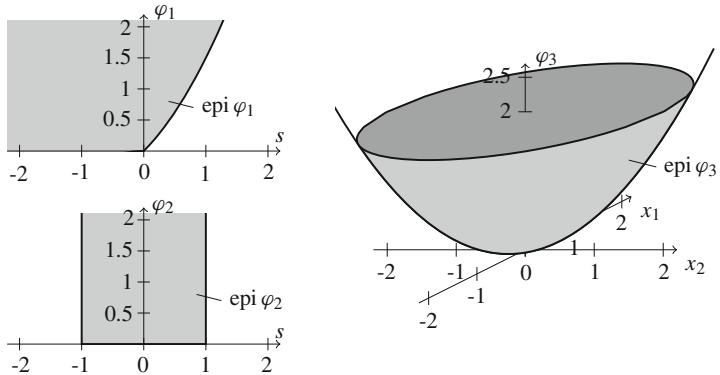
- $\left. \begin{array}{l} \alpha > 0, \quad G \text{ convex}, \\ \varphi \text{ convex, strictly increasing} \end{array} \right\} \Rightarrow \alpha F, \quad F + G, \quad \varphi \circ F \text{ strictly convex},$
- $\left. \begin{array}{l} \Phi \text{ affine linear, injective}, \\ F_1, \dots, F_N \text{ strictly convex} \end{array} \right\} \Rightarrow F \circ \Phi, \quad \max_{i=1, \dots, N} F_i \text{ strictly convex}.$

Besides these facts we give some concrete and some more abstract examples; see also Fig. 6.6.

Example 6.23 (Convex Functionals)

1. Exponentiation

The functions $\varphi : \mathbf{R} \rightarrow \mathbf{R}$ defined by $x \mapsto |x|^p$ are convex for $p \geq 1$ and strictly convex for $p > 1$. Every twice continuously differentiable function $\varphi : \mathbf{R} \rightarrow \mathbf{R}$ with $\varphi''(x) \geq 0$ for all x is convex, and strictly so if strict inequality holds.



$$\varphi_1(s) = \begin{cases} \frac{3}{10}s^2, & s \leq 0, \\ \frac{1}{2}s^2 + s, & s > 0, \end{cases} \quad \varphi_2(s) = \begin{cases} 0, & s \in [-1, 1], \\ \infty, & \text{otherwise}, \end{cases} \quad \varphi_3(x) = \frac{1}{2}(x_1^2 + x_2^2).$$

Fig. 6.6 Examples of convex functionals on \mathbf{R} and \mathbf{R}^2

2. Norms

Every norm $\|\cdot\|_X$ on a normed space is convex, since for all $u, v \in X$ and $\lambda \in \mathbf{K}$,

$$\|\lambda u + (1 - \lambda)v\|_X \leq |\lambda| \|u\|_X + |1 - \lambda| \|v\|_X.$$

For a normed space Y with $Y \subset X$, we can extend the norm $\|\cdot\|_Y$ to X by ∞ and obtain a convex functional on X . With a slight abuse of notation we write for $u \in X$,

$$\|u\|_Y = \begin{cases} \|u\|_Y & \text{if } u \in Y, \\ \infty & \text{otherwise.} \end{cases}$$

We remark that a norm on a nontrivial normed space is never strictly convex (due to positive homogeneity). The situation is different for strictly convex functions of norms: for $\varphi : [0, \infty[\rightarrow \mathbf{R}_\infty$ strictly monotonically increasing and strictly convex, the functional $F(u) = \varphi(\|u\|_X)$ is strictly convex if and only if the norm in X is *strictly convex*, i.e. for all $u, v \in X$ with $\|u\|_X = \|v\|_X = 1$, $u \neq v$, and for all $\lambda \in]0, 1[$, one has $\|\lambda u + (1 - \lambda)v\|_X < 1$.

The norm in a Hilbert space is always strictly convex, since for $\|u\|_X = \|v\|_X = 1$ and $u \neq v$, the function $F_\lambda : \lambda \mapsto \|\lambda u + (1 - \lambda)v\|_X^2$ is twice continuously differentiable with $F''_\lambda(\lambda) = 2\|u - v\|_X^2 > 0$, and hence convex.

3. Indicator functionals

The *indicator functional* of a set $K \subset X$, i.e.,

$$I_K(u) = \begin{cases} 0 & \text{if } u \in K, \\ \infty & \text{otherwise,} \end{cases}$$

is convex if and only if K is convex. Such functionals are used to describe constraints for minimization problems; see Example 6.4. The minimization of F over K is then written as

$$\min_{u \in X} F(u) + I_K(u).$$

4. Functionals in X^*

An element $x^* \in X^*$ and a convex function $\varphi : \mathbf{K} \rightarrow \mathbf{R}_\infty$ always lead to a composition $\varphi \circ \langle x^*, \cdot \rangle_{X^* \times X}$ that is convex.

5. Composition with a linear map $F \circ A$

If $A : \text{dom } A \subset Y \rightarrow X$ is a linear map defined on a subspace $\text{dom } A$ and $F : X \rightarrow \mathbf{R}_\infty$ is convex, then the functional $F \circ A : Y \rightarrow \mathbf{R}_\infty$ with

$$(F \circ A)(y) = \begin{cases} F(Ay) & \text{if } y \in \text{dom } A, \\ \infty & \text{otherwise,} \end{cases}$$

is also convex.

6. Convex functions in integrals

Let $(\Omega, \mathfrak{F}, \mu)$ be a measure space, $X = L^p(\Omega, \mathbf{K}^N)$ for some $N \geq 1$, and $\varphi : \mathbf{K}^N \rightarrow \mathbf{R}_\infty$ convex and lower semicontinuous. Then

$$F(u) = \int_\Omega \varphi(u(x)) \, dx$$

is convex, at least at the points where the integral exists (it may happen that $|\cdot| \circ \varphi \circ u$ is not integrable; the function $\varphi \circ u$ is, due to lower semicontinuity of φ , always measurable). A similar claim holds for strict convexity of φ , since for nonnegative $f \in L^p(\Omega)$, it is always the case that $\int_\Omega f \, dx = 0$ implies that $f = 0$ almost everywhere.

In particular, the norms $\|u\|_p = (\int_\Omega |u(x)|^p \, dx)^{1/p}$ in $L^p(\Omega, \mathbf{K}^N)$ are strictly convex norms for $p \in]1, \infty[$ if the vector norm $|\cdot|$ on \mathbf{K}^N is strictly convex.

Remark 6.24 (Convexity in the Introductory Examples) The functional in (6.1) from Examples 6.1 is strictly convex: We see this by Remark 6.22 or Lemma 6.21 and item 2 of Example 6.23. Similarly one sees strict convexity of the functionals in (6.3) and (6.5) for Examples 6.2–6.4.

Convex functions satisfy some continuity properties. These can be deduced, quite remarkably, only from assumptions on boundedness. Convexity allows us to transfer local properties to global properties.

Theorem 6.25 *If $F : X \rightarrow \mathbf{R}_\infty$ is convex and there exists $u^0 \in X$ such that F is bounded from above in a neighborhood of u^0 , then F is locally Lipschitz continuous at every interior point of $\text{dom } F$.*

Proof We begin with the proof of the following claim: If F bounded from above in a neighborhood of $u^0 \in X$, then it is Lipschitz continuous in a neighborhood of u^0 . By assumption, there exist $\delta_0 > 0$ and $R > 0$, such that $F(u) \leq R$ for $u \in B_{\delta_0}(u^0)$.

Since F is bounded from below on $B_{\delta_0}(u^0)$ we conclude that for $u \in B_{\delta_0}(u^0)$ we have $u^0 = \frac{1}{2}u + \frac{1}{2}(2u^0 - u)$, and consequently

$$F(u^0) \leq \frac{1}{2}F(u) + \frac{1}{2}F(2u^0 - u).$$

Since $2u^0 - u \in B_{\delta_0}(u^0)$, we obtain the lower bound $-L = 2F(u^0) - R$:

$$F(u) \geq 2F(u^0) - F(2u^0 - u) \geq 2F(u^0) - R = -L.$$

For distinct $u, v \in B_{\delta_0/2}(u^0)$ the vector $w = u + \alpha^{-1}(u - v)$ with $\alpha = 2\|u - v\|_X/\delta_0$ is still in $B_{\delta_0}(u^0)$, since

$$\|w - u^0\|_X \leq \|u - u^0\|_X + \alpha^{-1}\|v - u\|_X < \frac{\delta_0}{2} + \frac{\delta_0}{2\|u - v\|_X}\|u - v\|_X = \delta_0.$$

We write $u = \frac{1}{1+\alpha}v + \frac{\alpha}{1+\alpha}w$ as a convex combination, and conclude that

$$\begin{aligned} F(u) - F(v) &\leq \frac{1}{1+\alpha}F(v) + \frac{\alpha}{1+\alpha}F(w) - F(v) = \frac{\alpha}{1+\alpha}(F(w) - F(v)) \\ &\leq \alpha(R + L) = C\|u - v\|_X \end{aligned}$$

(here we used the boundedness of F in $B_{\delta_0}(u^0)$ from above and below and the definition of α). Swapping the roles of u and v in this argument, we obtain the Lipschitz estimate $|F(u) - F(v)| \leq C\|u - v\|_X$ in $B_{\delta_0/2}(u^0)$.

Finally, we show that every interior point u^1 of $\text{dom } F$ has a neighborhood on which F is bounded from above. To that end, let $\lambda \in]0, 1[$ be such that $F(\lambda^{-1}(u^1 - (1-\lambda)u^0)) = S < \infty$. Such a λ exists, since the mapping $\lambda \mapsto \lambda^{-1}(u^1 - (1-\lambda)u^0)$ is continuous at $\lambda = 1$ and has the value u^1 there.

Furthermore, for a given $v \in B_{(1-\lambda)\delta_0}(u^1)$ we choose a vector $u = u_0 + (v - u^1)/(1 - \lambda)$ (which is also in $B_{\delta_0}(u^0)$). This v is a convex combination, since $v = \lambda\lambda^{-1}(u^1 - (1-\lambda)u^0) + (1-\lambda)u$, and we conclude, that

$$F(v) \leq \lambda F(\lambda^{-1}(u^1 - (1-\lambda)u^0)) + (1-\lambda)F(u) \leq S + R.$$

These v form a neighborhood of u^1 on which F is bounded, and hence F is locally Lipschitz continuous. \square

Remark 6.26 The following connection of convex functionals and convex sets, similar to Remark 6.8, is simple to see:

- A functional is convex if and only if its epigraph is convex.
- It is convex and lower semicontinuous if and only if its epigraph is convex and closed.
- Especially, for a convex and lower semicontinuous F and all $t \in \mathbf{R}$, the *sublevel sets* $\{u \in X \mid F(u) \leq t\}$ are convex and closed.

This observation is behind the proof of the following fundamental property of closed sets in Banach spaces:

Lemma 6.27 (Convex, Weakly Closed Sets) *A convex subset $A \subset X$ of a Banach space X is weakly sequentially closed if and only if it is closed.*

Proof Since strong convergence implies weak convergence, we immediately see that weakly sequentially closed sets are also closed. For the reverse implication, we assume that there exist a sequence (u^n) in A and some $u \in X \setminus A$ such that $u^n \rightharpoonup u$ for $n \rightarrow \infty$. Since A is convex and closed, Hahn-Banach separation theorems (Theorem 2.29) imply that there exist an $x^* \in X^*$ and a $\lambda \in \mathbf{R}$ such that $\operatorname{Re} \langle x^*, u^n \rangle_{X^* \times X} \leq \lambda$ and $\operatorname{Re} \langle x^*, u \rangle_{X^* \times X} > \lambda$. However, this implies that $\lim_{n \rightarrow \infty} \langle x^*, u^n \rangle_{X^* \times X} \neq \langle x^*, u \rangle_{X^* \times X}$, which contradicts the weak convergence. Hence, A has to be weakly sequentially closed. \square

This result implies a useful characterization of weak (sequential) lower semicontinuity.

Corollary 6.28 *A convex functional $F : X \rightarrow \mathbf{R}_\infty$ is weakly lower semicontinuous if and only if it is lower semicontinuous.*

Proof This follows by applying Lemma 6.27 to the epigraph in combination with Remark 6.26. \square

We apply this to prove the weak lower semicontinuity of functionals of similar type to those we have already seen in Examples 6.1, 6.18, and 6.23.

Example 6.29 (Convex, Weakly Lower Semicontinuous Functionals)

1. Norms

Let Y be a reflexive Banach space, continuously embedded in X . Then $\|\cdot\|_Y$ is lower semicontinuous on X : Let (u^n) with $u^n \rightarrow u$ in X for $n \rightarrow \infty$. If $\liminf_{n \rightarrow \infty} \|u^n\|_Y = \infty$ the desired inequality is surely satisfied; otherwise, take a subsequence (u^{n_k}) with $\liminf_{n \rightarrow \infty} \|u^n\|_Y = \lim_{k \rightarrow \infty} \|u^{n_k}\|_Y < \infty$. By reflexivity of Y we assume without loss of generality that (u^{n_k}) converges weakly to some $v \in Y$. By the continuous embedding in X we also get $u^{n_k} \rightharpoonup v$ for $k \rightarrow \infty$ in X , and hence $v = u$. Since $\|\cdot\|_Y$ is weakly lower semicontinuous,

we conclude that

$$\|u\|_Y \leq \liminf_{k \rightarrow \infty} \|u^{n_k}\|_Y = \liminf_{n \rightarrow \infty} \|u^n\|_Y.$$

Moreover, $\|\cdot\|_Y$ is convex (cf. Example 6.23), and hence also weakly sequentially lower semicontinuous.

The claim also holds if we assume that Y is the dual space of a separable normed space (by weak* sequential compactness of the unit ball, see Theorem 2.21) and that the embedding into X is weakly*-to-strongly closed.

2. Composition with a linear map $F \circ A$

For reflexive Y , $F : Y \rightarrow \mathbf{R}_\infty$ convex, lower semicontinuous and coercive and a strongly-to-weakly closed linear mapping $A : X \supset \text{dom } A \rightarrow Y$, the composition $F \circ A$ is convex (see Example 6.23) and also lower semicontinuous: if $u^n \rightarrow u$ in X and $\liminf_{n \rightarrow \infty} F(Au^n) < \infty$, then by coercivity, $(\|u^n\|_Y)$ is bounded. Hence, there exists a weakly convergent subsequence $(Au^{n_k}) \rightharpoonup v$ with $v \in Y$, and without loss of generality, we can assume that $\lim_{k \rightarrow \infty} F(Au^{n_k}) = \liminf_{n \rightarrow \infty} F(Au^n)$. Moreover, $u^{n_k} \rightarrow u$, and we conclude that $v = Au$ and thus

$$F(Au) \leq \lim_{k \rightarrow \infty} F(Au^{n_k}) = \liminf_{n \rightarrow \infty} F(Au^n).$$

A similar argument establishes the convexity and lower semicontinuity of $F \circ A$ if Y is the dual space of a separable Banach space, $F : Y \rightarrow \mathbf{R}_\infty$ is convex, weakly* lower semicontinuous and coercive, and A is strong-to-weakly* closed.

3. Indicator functionals

It is straightforward to see that an indicator function I_K is lower semicontinuous, if and only if K is closed. Hence, for convex and closed K , the functional I_K is weakly sequentially lower semicontinuous.

4. Convex functions in integrals

Let $(\Omega, \mathfrak{F}, \mu)$ be a measure space, and for $1 \leq p < \infty$ and $N \geq 1$, let $X = L^p(\Omega, \mathbf{K}^N)$ be the respective Lebesgue space. Moreover, let $\varphi : \mathbf{K}^N \rightarrow \mathbf{R}_\infty$ be convex, lower semicontinuous, and satisfy $\varphi(t) \geq 0$ for all $t \in \mathbf{K}^n$ and $\varphi(0) = 0$ if Ω has infinite measure or let φ be bounded from below if Ω has finite measure. Then

$$F(u) = \int_{\Omega} \varphi(u(x)) \, dx,$$

$F : X \rightarrow \mathbf{R}_\infty$ is well defined and weakly lower semicontinuous: Assume $u^n \rightarrow u$ for some (u^n) and u in $L^p(\Omega, \mathbf{K}^N)$. The theorem of Fischer-Riesz (Theorem 2.48) shows the existence of a subsequence, that we again call (u^n) , which converges pointwise almost everywhere. This leads, for almost all $x \in \Omega$, to the estimate

$$\varphi(u(x)) \leq \liminf_{n \rightarrow \infty} \varphi(u^n(x)),$$

which together with Fatou's lemma (Lemma 2.46) implies the inequality

$$\begin{aligned} F(u) &= \int_{\Omega} \varphi(u(x)) \, dx \leq \int_{\Omega} \liminf_{n \rightarrow \infty} \varphi(u^n(x)) \, dx \\ &\leq \liminf_{n \rightarrow \infty} \int_{\Omega} \varphi(u^n(x)) \, dx = \liminf_{n \rightarrow \infty} F(u^n). \end{aligned}$$

This prove the lower semicontinuity of F and, together with convexity, the weak sequential lower semicontinuity (Corollary 6.28).

Lower continuity of convex functionals implies not only weak lower semicontinuity but also strong continuity in the interior of the effective domain.

Theorem 6.30 *A convex and lower semicontinuous functional $F : X \rightarrow \mathbf{R}_{\infty}$ on a Banach space X is continuous at every interior point of $\text{dom } F$.*

Proof For the nontrivial case it is, by Theorem 6.25, enough to show, that F is bounded from above in a neighborhood. We choose $u^0 \in \text{int}(\text{dom } F)$, $R > F(u^0)$ and define $V = \{u \in X \mid F(u) \leq R\}$, so that the sets

$$V_n = \left\{ u \in X \mid u_0 + \frac{u - u^0}{n} \in V \right\},$$

$n \geq 1$, are a sequence of convex and closed sets (since F is convex and lower semicontinuous; see Remark 6.26). Moreover, $V_{n_0} \subset V_{n_1}$ for $n_0 \leq n_1$, since $u^0 + n_1^{-1}(u - u^0) = u^0 + (n_0/n_1)n_0^{-1}(u - u^0)$ is, by convexity, contained in V if $u^0 + n_0^{-1}(u - u^0) \in V$.

Finally, we see that for all $u \in X$, the convex function $t \mapsto F(u^0 + t(u - u^0))$ is finite in a neighborhood of 0 (otherwise, u^0 would not be an interior point of $\text{dom } F$). Without loss of generality, we assume that F_u is continuous even in this neighborhood, see Theorem 6.25. Hence, there exists $n \geq 1$ such that $u^0 + n^{-1}(u - u^0) \in V$. This shows that $\bigcup_{n \geq 1} V_n = X$.

By the Baire category theorem (Theorem 2.14), one V_n has an interior point, and hence V has an interior point, which implies the boundedness of F in a neighborhood. \square

Now we state the direct method for the minimization of convex functionals.

Theorem 6.31 (The Direct Method for Convex Functionals in a Banach Space) *Let X be a reflexive Banach space and $F : X \rightarrow \mathbf{R}_{\infty}$ a convex, lower semicontinuous, and coercive functional. Then, there is a solution of the minimization problem*

$$\min_{u \in X} F(u).$$

The solution is unique if F is strictly convex.

Proof We use Theorem 6.17. If F is everywhere infinite, there is nothing to prove. So let F be proper, i.e., $\text{epi } F \neq \emptyset$. Since F is by Corollary 6.28 weakly lower semicontinuous, we have to prove boundedness from below. To that end, we use the separation theorem (Theorem 2.29) for the closed set $\text{epi } F$ and the compact set $\{(u^0, F(u^0) - 1)\}$ (which are disjoint by definition). Hence, there exist a pair $(x^*, t^*) \in X^* \times \mathbf{R}$ and some $\lambda \in \mathbf{R}$ such that both

$$\operatorname{Re} \langle x^*, u \rangle + t^* F(u) \geq \lambda \quad \forall u \in X$$

and

$$\operatorname{Re} \langle x^*, u^0 \rangle + t^*(F(u^0) - 1) = \operatorname{Re} \langle x^*, u^0 \rangle + t^* F(u^0) - t^* < \lambda$$

hold. This shows that $\lambda - t^* < \lambda$, i.e., $t^* > 0$. For all $R > 0$ we get, for $u \in X$ with $\|u\|_X \leq R$, the estimate

$$\frac{\lambda - R \|x^*\|_{X^*}}{t^*} \leq \frac{\lambda - \operatorname{Re} \langle x^*, u \rangle}{t^*} \leq F(u).$$

Coercivity of F implies the existence of some $R > 0$ such that $F(u) \geq 0$ for all $\|u\|_X \geq R$. This shows that F is bounded from below, and Theorem 6.17 shows that a minimizer exists.

If we now assume strict convexity of F and let $u^* \neq u^{**}$ be two minimizers of F , we obtain

$$\min_{u \in X} F(u) = \frac{1}{2} F(u^*) + \frac{1}{2} F(u^{**}) > F\left(\frac{u^* + u^{**}}{2}\right) \geq \min_{u \in X} F(u),$$

a contradiction. Hence, there is only one minimizer. \square

Example 6.32 (Tikhonov Functionals) Let X, Y be Banach spaces, X reflexive, $A \in \mathcal{L}(X, Y)$, and $u^0 \in Y$. The linear map A should be a model of some *forward operator*, which maps a image to the data that can be measured. This could be, for example, a convolution operator (Example 6.2) or the identity (Example 6.1). Moreover, u^0 should be the noisy data. Building on the considerations of Sect. 6.1, we use the norm Y to quantify the discrepancy. Moreover, the space X should be a good model for the reconstructed data u . This motivates the choice

$$\Phi(v) = \frac{1}{p} \|v\|_Y^p, \quad \Psi(u) = \frac{1}{q} \|u\|_X^q$$

with $p, q \geq 1$. The corresponding variational problem, in this abstract situation, is $\min_{u \in X} \Phi(Au - u^0) + \lambda \Psi(u)$ or

$$\min_{u \in X} F(u), \quad F(u) = \frac{\|Au - u^0\|_Y^p}{p} + \lambda \frac{\|u\|_X^q}{q} \quad (6.7)$$

with some $\lambda > 0$. Functionals F of this form are also called *Tikhonov functionals*. They play an important role in the theory of ill-posed problems.

Using Lemma 6.21 and the considerations in Example 6.23, it is not hard to show that $F : X \rightarrow \mathbf{R}_\infty$ is convex. The functional is finite on all of X , hence continuous (Theorem 6.25), in particular lower semicontinuous. The term $\frac{\lambda}{q} \|u\|_X^q$ implies the coercivity of F (see Remark 6.13). Hence, we can apply Theorem 6.31 and see that there exists a minimizer $u^* \in X$. Note that the penalty $\lambda \Psi$ is crucial for this claim, since $u \mapsto \Phi(Au - u^0)$ is not coercive in general. If it were, there would exist A^{-1} on $\text{rg}(A)$ and it would be continuous; in the context of inverse problems, $Au = u^0$ would not be ill-posed (see also Exercise 6.6).

For the uniqueness of the minimizer we immediately see two sufficient conditions. On the one hand, strict convexity of $\|\cdot\|_X^q$ ($q > 1$) implies strict convexity of Ψ and also of F . On the other hand, an injective A and strictly convex $\|\cdot\|_Y^p$ ($p > 1$) lead to strict convexity of $u \mapsto \Phi(Au - u^0)$ and also of F . In both cases we obtain the uniqueness of u^* by Theorem 6.31.

If we compare F with the functionals in Examples 6.1–6.4, we note that the functionals there are not Tikhonov functionals in this sense. This is due to the fact that the term Ψ is only a squared seminorm $\|\nabla \cdot\|_2$ on a Sobolev space. It would be desirable to have a generalization of the above result on existence to the case $\Psi(u) = \frac{1}{q}|u|_X^q$ with a suitable seminorm $|\cdot|_X$. Since seminorms vanish on a subspace we need additional assumptions on A to obtain coercivity. One such case is treated in Exercise 6.7.

Finally, we note that in view of item 1 in Example 6.29, one can also use the q th power of a norm in a reflexive Banach space Z , which is continuously embedded in X ; in this case one considers the functional on the space Z . The same can be done for certain seminorms in Z . However, it may be advantageous to consider the problem in X if it comes to optimality conditions, as we will see later.

Once the existence of minimizers of some convex functional is settled, e.g., by the direct method, one aims to calculate one of the minimizers. The following approach is similar to what one might see in any calculus course:

Suppose $u^* \in X$ is a minimizer of $F : X \rightarrow \mathbf{R}$. For every direction $v \in X$ we vary in this direction, i.e., we consider $F_v(t) = F(u^* + tv)$, and we get, that 0 is a minimizer of F_v . If we further assume that F_v is differentiable at 0, we get $F'_v(0) = 0$ and if F is also Gâteaux differentiable at u^* with derivative $DF(u^*) \in X^*$,

$$F'_v(0) = \langle DF(u^*), v \rangle = 0 \quad \forall v \in X \quad \Rightarrow \quad DF(u^*) = 0.$$

In other words, every minimizer of F is necessarily a stationary point. Until now we have not used any convexity in this argument. However, convexity allows us to conclude that this criterion is also sufficient.

Theorem 6.33 Let $F : U \rightarrow \mathbf{R}$ be a functional that is Gâteaux differentiable on an open neighborhood U of a convex set K of a normed space X . Then F is convex in K if and only if

$$F(u) + \langle DF(u), v - u \rangle \leq F(v) \quad \forall u, v \in K. \quad (6.8)$$

If u is an interior point of K , then $w = DF(u)$ is the unique element in X^* for which

$$F(u) + \langle w, v - u \rangle \leq F(v) \quad \forall v \in K. \quad (6.9)$$

Proof Let F be convex. We choose $u, v \in K$ and $t \in]0, 1]$, and obtain $F(tv + (1-t)u) \leq F(u) + t(F(v) - F(u))$, and dividing by t and letting $t \rightarrow 0$, yields

$$\langle DF(u), v - u \rangle = \lim_{t \rightarrow 0} \frac{F(u + t(v - u)) - F(u)}{t} \leq F(v) - F(u)$$

and thus (6.8). Now assume that (6.8) holds. Swapping the roles of u and v and adding the inequalities gives

$$\langle DF(u) - DF(v), u - v \rangle \geq 0 \quad \forall u, v \in K.$$

For fixed $u, v \in K$ we consider the functions $f(t) = F(u + t(v - u))$, defined and continuous on $[0, 1]$. By Gâteaux differentiability and the above inequality we get

$$(f'(t) - f'(s))(t - s) = \langle DF(u + t(v - u)) - DF(u + s(v - u)), (t - s)(v - u) \rangle \geq 0$$

for all $s, t \in]0, 1[$. Hence, f' is monotonically increasing. If we choose some $t \in]0, 1[$, we obtain by the mean value theorem, applied to 0 and t , and t and 1, respectively, the existence of s, s' with $0 < s < t < s' < 1$ such that

$$\frac{f(t) - f(0)}{t - 0} = f'(s) \leq f'(s') = \frac{f(1) - f(t)}{1 - t}.$$

Some rearrangement gives $F(tv + (1-t)u) = f(t) \leq tf(1) + (1-t)f(0) = tF(v) + (1-t)F(u)$. This proves the convexity in all non-trivial cases, since $u, v \in K$ and $t \in]0, 1[$ are arbitrary.

For the last point, let u be an interior point of K and $w \in X^*$ such that (6.9) is satisfied. for all $\bar{v} \in X$ we have $v = u + t\bar{v} \in K$ in some interval $t \in]-\varepsilon, \varepsilon[$. For $t > 0$ we conclude by Gâteaux differentiability, that

$$\langle w, \bar{v} \rangle \leq \lim_{t \rightarrow 0^+} \frac{F(u + t\bar{v}) - F(u)}{t} = \langle DF(u), \bar{v} \rangle$$

and similarly for $t < 0$, we get $\langle w, \bar{v} \rangle = \langle DF(u), \bar{v} \rangle$. This shows that $w = DF(u)$.

□

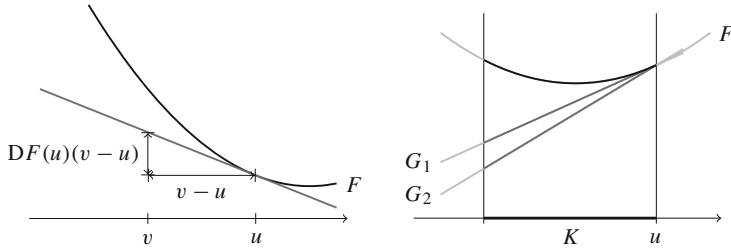


Fig. 6.7 Left: Interpretation of the derivative of a convex function at some point as the slope of the respective affine support function. Right: The characterization does not hold outside of the interior of the domain K . Both $G_1(v) = F(u) + DF(u)(v - u)$ and $G_2(v) = F(u) + w(v - u)$ with $w > DF(u)$ are estimated by F in K

Remark 6.34 For convex Gâteaux differentiable functionals we can characterize the derivative at any interior point u of K by the inequality (6.9). More geometrically, this means that there is an affine linear *support functional* at the point u , that is tight at $F(u)$ and below F on the whole of K . The “slope” of this functional must be equal to $DF(u)$; see Fig. 6.7.

Corollary 6.35 If in Theorem 6.33 the functional F is also convex and if $DF(u^*) = 0$ holds for some $u^* \in K$, then u^* is a minimizer of F in K .

Proof Just plug u^* in (6.8). □

In the special case $K = X$ we obtain that

if $F : X \rightarrow \mathbf{R}$ is convex and Gâteaux differentiable, then u^* is a minimizer if and only if

$$DF(u^*) = 0.$$

Example 6.36 (Euler-Lagrange Equation for Example 6.1) Let us consider again Example 6.1, now with real-valued u^0 and respective real spaces $L^2(\mathbf{R}^d)$ and $H^1(\mathbf{R}^d)$. The functional F from (6.1) is Gâteaux differentiable, and it is easy to compute the derivative as

$$(DF(u))(v) = \int_{\mathbf{R}^d} (u(x) - u^0(x))v(x) \, dx + \lambda \int_{\mathbf{R}^d} \nabla u(x) \cdot \nabla v(x) \, dx.$$

We can find an optimal u^* by solving $DF(u^*) = 0$, or equivalently,

$$\int_{\mathbf{R}^d} u^*(x)v(x) + \lambda \nabla u^*(x) \cdot \nabla v(x) \, dx = \int_{\mathbf{R}^d} u^0(x)v(x) \, dx$$

for all $v \in H^1(\mathbf{R}^d)$. This is a weak formulation (see also Example 6.4) of the equation

$$u^* - \lambda \Delta u^* = u^0 \quad \text{in } \mathbf{R}^d.$$

Hence, the Euler-Lagrange equation of the denoising problem is a partial differential equation. Transformation into frequency space gives

$$(1 + \lambda|\xi|^2)\widehat{u}^* = \widehat{u}^0,$$

which leads to the solution (6.2) we obtained before.

We note that we can use this method also for Example 6.2; we already used a variant of it in Example 6.4. However, the method has its limitations as we already see in the case of convex constraints. In this case we get different conditions for optimality.

Example 6.37 (Minimization with Convex Constraints) Let $F : \mathbf{R} \rightarrow \mathbf{R}$ be convex and continuously differentiable and consider the minimization problem

$$\min_{u \in [-1, 1]} F(u),$$

which has a solution u^* . For $u^* \in]-1, 1[$ we have necessarily $DF(u^*) = 0$, but in the case $u^* = 1$ we get

$$-DF(u^*) = \lim_{t \rightarrow 0} \frac{F(1-t) - F(1)}{t} \geq 0,$$

and similarly, in the case $u^* = -1$ that $DF(u^*) \geq 0$. By (6.8) these conditions are also sufficient for u^* to be a minimizer. Hence, we can characterize optimality of u^* as follows: there exists $\mu^* \in \mathbf{R}$ such that

$$DF(u^*) + \mu^* \operatorname{sgn}(u^*) = 0, \quad |u^*| - 1 \leq 0, \quad \mu^* \geq 0, \quad (|u^*| - 1)\mu^* = 0.$$

The variable μ^* is called the *Lagrange multiplier* for the constraint $|u|^2 - 1 \leq 0$. In the next section we will investigate this notion in more detail.

Since we are interested in minimization problems in (infinite-dimensional) Banach spaces, we ask ourselves how we can transfer the above technique to this setting. For a domain $\Omega \subset \mathbf{R}^d$ and $F : L^2(\Omega) \rightarrow \mathbf{R}$ convex and differentiable on the real Hilbert space $L^2(\Omega)$ and u^* a solution of

$$\min_{u \in L^2(\Omega), \|u\|_2 \leq 1} F(u),$$

we have $DF(u^*) = 0$ if $\|u^*\|_2 < 1$. Otherwise, we vary by $t v$ with $(u^*, v) < 0$, and conclude that $\frac{\partial}{\partial t} \|u + tv\|_2^2 = 2(u, v) + 2t\|v\|_2^2$ and for t small enough, we have $\|u^* + tv\|_2 < 1$. Using that u^* is a minimizer, we obtain $(DF(u^*), v) = \lim_{t \rightarrow 0} \frac{1}{t} (F(u^* + tv) - F(u^*)) \geq 0$. For this to hold for all v with $(v, u^*) < 0$, we need that $DF(u^*) + \mu^* u^* = 0$ for some $\mu^* \geq 0$ (Exercise 6.8). Hence, we can write

the optimality system as

$$DF(u^*) + \mu^* \frac{u^*}{\|u^*\|_2} = 0, \quad \|u^*\|_2 - 1 \leq 0, \quad \mu^* \geq 0, \quad (\|u^*\|_2 - 1)\mu^* = 0.$$

Again, we have a Lagrange multiplier $\mu^* \in \mathbf{R}$.

The situation is different if we consider a pointwise almost everywhere constraint, i.e.,

$$\min_{u \in L^2(\Omega), \|u\|_\infty \leq 1} F(u).$$

If we have a minimizer with $\|u^*\|_\infty < 1$, we can not conclude that $DF(u^*) = 0$, since the set $\{u \in L^2(\Omega) \mid \|u\|_\infty \leq 1\}$ has empty interior (otherwise, the embedding $L^\infty(\Omega) \hookrightarrow L^2(\Omega)$ would be surjective, by the open mapping theorem, Theorem 2.16, which is a contradiction). However, for $t > 0$ small enough we have, $\|u^* + tv\|_\infty < 1$ for all measurable $v = \sigma \chi_{\Omega'}, \sigma \in \{-1, 1\}, \Omega' \subset \Omega$. Thus we have $\int_{\Omega'} |DF(u^*)| dx = 0$ for all measurable $\Omega' \subset \Omega$ and hence $DF(u^*) = 0$.

The case $\|u^*\|_\infty = 1$ is more difficult to analyze. On the one hand, we see, similarly to the above argumentation, that for all measurable subsets $\Omega' \subset \Omega$ with $\|u^*|_{\Omega'}\|_\infty < 1$, one has $DF(u^*)|_{\Omega'} = 0$; and a similar claim holds on the union of all such sets Ω_0 . With $\Omega_+ = \{x \in \Omega \mid u^*(x) = 1\}$, $\Omega_- = \{x \in \Omega \mid u^*(x) = -1\}$ we get, up to a null set, a disjoint partition of Ω in $\Omega_0 \cup \Omega_+ \cup \Omega_-$. Now we conclude for all measurable $\Omega' \subset \Omega_+$ that $\int_{\Omega'} DF(u^*) dx \leq 0$; i.e. $DF(u^*) \leq 0$ almost everywhere in Ω_+ . Similarly, we obtain $DF(u^*) \geq 0$ almost everywhere in Ω_- . In conclusion, we have the following optimality system: there exists $\mu^* \in L^2(\Omega)$, such that

$$DF(u^*) + \mu^* \operatorname{sgn}(u^*) = 0 \quad \text{and} \\ |u^*| - 1 \leq 0, \quad \mu^* \geq 0, \quad (|u^*| - 1)\mu^* = 0 \quad \text{almost everywhere in } \Omega.$$

The Lagrange multiplier is, in contrast to the above cases, an infinite-dimensional object.

As we see, different constraints lead to qualitatively different optimality conditions. Moreover, the derivation of these conditions can be lengthy and depends on the underlying space. Hence, we would like to have methods to treat convex constraints in a unified way. We will see that the subdifferential is well suited for this task. Before we come to its definition, we present the following motivation for this generalized notion of differentiability.

Example 6.38 For $a, b, c \in \mathbf{R}$ with $a > 0, ac - b^2 > 0$, and $f \in \mathbf{R}^2$ we consider the minimization problem

$$\min_{u \in \mathbf{R}^2} F(u), \quad F(u) = \frac{au_1^2 + 2bu_1u_2 + cu_2^2}{2} - f_1u_1 - f_2u_2 + |u_1| + |u_2|.$$

It is easy to see that a unique solution exists. The functional F is convex, continuous, and coercive, but not differentiable. Nonetheless, a suitable case distinction allows us to determine the solution. As an example, we treat the case $a = 1$, $b = 0$, and $c = 1$.

1. If $u_1^* \neq 0$ and $u_2^* \neq 0$, then $DF(u^*) = 0$ has to hold, i.e.,

$$\begin{cases} u_1^* - f_1 + \operatorname{sgn}(u_1^*) = 0, \\ u_2^* + f_2 + \operatorname{sgn}(u_2^*) = 0, \end{cases} \iff \begin{cases} u_1^* + \operatorname{sgn}(u_1^*) = f_1, \\ u_2^* + \operatorname{sgn}(u_2^*) = f_2, \end{cases}$$

and hence $|f_1| > 1$ as well as $|f_2| > 1$ (since we would get a contradiction otherwise). Is it easy to check that the solution is

$$u_1^* = f_1 - \operatorname{sgn}(f_1), \quad u_2^* = f_2 - \operatorname{sgn}(f_2),$$

in this case.

2. If $u_1^* = 0$ and $u_2^* \neq 0$, then F is still differentiable with respect to u_2 , and we obtain $u_2^* + \operatorname{sgn}(u_2^*) = f_2$, and hence $|f_2| > 1$ and $u_2^* = f_2 - \operatorname{sgn}(f_2)$.
3. For $u_1^* \neq 0$ and $u_2^* = 0$ we obtain similarly $|f_1| > 1$ and $u_1^* = f_1 - \operatorname{sgn}(f_1)$.
4. The case $u_1^* = u_2^* = 0$ does not lead to any new conclusion.

All in all, we get

$$u^* = \begin{cases} (0, 0) & \text{if } |f_1| \leq 1, |f_2| \leq 1, \\ (f_1 - \operatorname{sgn}(f_1), 0) & \text{if } |f_1| > 1, |f_2| \leq 1, \\ (0, f_2 - \operatorname{sgn}(f_2)) & \text{if } |f_1| \leq 1, |f_2| > 1, \\ (f_1 - \operatorname{sgn}(f_1), f_2 - \operatorname{sgn}(f_2)) & \text{if } |f_1| > 1, |f_2| > 1, \end{cases}$$

since anything else would contradict the conclusions of the above cases 1–3.

In the case of general a, b, c , the computations get a little bit more involved, and a similar claim holds in higher dimensions. In the case of infinite dimensions with $A \in \mathcal{L}(\ell^2, \ell^2)$ symmetric and positive definite, $f \in \ell^2$, however, we cannot apply the above reasoning to the problem

$$\min_{u \in \ell^2} \frac{(u, Au)}{2} - (f, u) + \sum_{i=1}^{\infty} |u_i|,$$

since the objective functional is nowhere continuous, and consequently not differentiable.

The above example shows again that a unified treatment of minimization problems with nondifferentiable (or even better, noncontinuous) convex objectives is desirable. The subdifferential is an appropriate tool in these situations.

6.2.3 Subdifferential Calculus

Some preparations are in order before we define subgradients and the subdifferential.

Lemma 6.39 *Let X be a complex normed space. Then there exist a real normed space $X_{\mathbf{R}}$ and norm-preserving maps $i_X : X \rightarrow X_{\mathbf{R}}$ and $j_{X^*} : X^* \rightarrow X_{\mathbf{R}}^*$, such that $\langle j_{X^*}x^*, i_X x \rangle = \operatorname{Re} \langle x^*, x \rangle$ for all $x \in X$ and $x^* \in X^*$.*

Proof The complex vector space X turns into a real one $X_{\mathbf{R}}$ by restricting the scalar multiplication to real numbers. Then $\|\cdot\|_{X_{\mathbf{R}}} = \|\cdot\|_X$ is a norm on $X_{\mathbf{R}}$, and hence $i_X = \operatorname{id}$ maps $X \rightarrow X_{\mathbf{R}}$ and preserves the norm. We define $j_{X^*} : X^* \rightarrow X_{\mathbf{R}}^*$ via

$$\langle j_{X^*}x^*, x \rangle_{X_{\mathbf{R}}^* \times X_{\mathbf{R}}} = \operatorname{Re} \langle x^*, i_X^{-1}x \rangle_{X^* \times X} \quad \forall x \in X_{\mathbf{R}}.$$

It remains to show that j_{X^*} preserves the norm. On the one hand, we have

$$\|j_{X^*}x^*\|_{X_{\mathbf{R}}^*} = \sup_{\|x\|_{X_{\mathbf{R}}} \leq 1} |\operatorname{Re} \langle x^*, i_X^{-1}x \rangle| \leq \sup_{\|x\|_X \leq 1} |\langle x^*, x \rangle| = \|x^*\|_{X^*}.$$

On the other hand, we can choose for every sequence (x^n) in X with $\|x^n\|_X \leq 1$ and $|\langle x^*, x^n \rangle| \rightarrow \|x^*\|_{X^*}$ the sequence $\bar{x}^n = i_X(\operatorname{sgn} \langle x^*, x^n \rangle x^n)$ in $X_{\mathbf{R}}$, which also satisfies $\|\bar{x}^n\|_{X_{\mathbf{R}}} \leq 1$, and moreover

$$\begin{aligned} \|j_{X^*}x^*\|_{X_{\mathbf{R}}^*} &\geq \lim_{n \rightarrow \infty} |\langle j_{X^*}x^*, \bar{x}^n \rangle| = \lim_{n \rightarrow \infty} |\operatorname{Re} \langle x^*, \overline{\operatorname{sgn} \langle x^*, x^n \rangle} x^n \rangle| \\ &= \lim_{n \rightarrow \infty} |\langle x^*, x^n \rangle| = \|x^*\|_{X^*}. \end{aligned}$$

Hence $\|j_{X^*}x^*\|_{X_{\mathbf{R}}^*} = \|x^*\|_{X^*}$, as claimed. \square

Remark 6.40

- We see that $j_{X^*}i_{X^*}^{-1} : (X^*)_{\mathbf{R}} \rightarrow X_{\mathbf{R}}^*$ is always an isometric isomorphism, hence we will tacitly identify these spaces.
- Analogously, for $A \in \mathcal{L}(X, Y)$ we can form $A_{\mathbf{R}} = i_Y A i_X^{-1}$ in $\mathcal{L}(X_{\mathbf{R}}, Y_{\mathbf{R}})$. The adjoint is $A_{\mathbf{R}}^* = j_{X^*} A^* j_{Y^*}^{-1}$.
- For (pre-) Hilbert spaces X , this construction leads to the scalar product $(x, y)_{X_{\mathbf{R}}} = \operatorname{Re} (i_X^{-1}x, i_X^{-1}y)_X$ for $x, y \in X_{\mathbf{R}}$.

Due to the above consideration we can restrict ourselves to real Banach and Hilbert spaces in the following and can still treat complex-valued function spaces. As another prerequisite we define a suitable linear arithmetic for set-valued mappings, also called *graphs*.

Definition 6.41 (Set-Valued Mappings, Graphs) Let X, Y be real normed spaces.

1. A *set-valued mapping* $F : X \rightrightarrows Y$ or *graph* is a subset $F \subset X \times Y$. We write $F(x) = \{y \in Y \mid (x, y) \in F\}$ and use $y \in F(x)$ synonymous to $(x, y) \in F$.
2. For every mapping $F : X \rightarrow Y$ we denote its graph also by $F = \{(x, F(x)) \mid x \in X\}$ and use $F(x) = y$ and $F(x) = \{y\}$ interchangeably.
3. For set-valued mappings $F, G : X \rightrightarrows Y$ and $\lambda \in \mathbf{R}$ let

$$(F + G)(x) = \{y_1 + y_2 \mid y_1 \in F(x), y_2 \in G(x)\},$$

$$(\lambda F)(x) = \{\lambda y \mid y \in F(x)\}.$$

4. For a real normed space Z and $G : Y \rightrightarrows Z$, we define the composition of G and F as

$$(G \circ F)(x) = \{z \in G(y) \mid y \in F(x)\}.$$

5. The inversion $F^{-1} : Y \rightrightarrows X$ of $F : X \rightrightarrows Y$ is defined by

$$F^{-1}(y) = \{x \in X \mid y \in F(x)\}.$$

As a motivation for the definition of the subgradient we recall Theorem 6.33. The derivative of a convex functional F , which is also Gâteaux differentiable, is characterized by the inequality (6.9). The idea behind the subgradient is, to omit the assumption of Gâteaux differentiability to obtain a generalized notion of derivative.

Definition 6.42 (Subgradient, Subdifferential) Let X be a real normed space and $F : X \rightarrow \mathbf{R}_\infty$ a convex functional. An element $w \in X^*$ is called a *subgradient* if

$$F(u) + \langle w, v - u \rangle \leq F(v) \quad \forall v \in X. \quad (6.10)$$

The relation $((u, w) \in \partial F \Leftrightarrow (u, w) \text{ satisfies (6.10)})$ defines a graph $\partial F : X \rightrightarrows X^*$, called the *subdifferential* of F . The inequality (6.10) is called the *subgradient inequality*.

Hence, the set $\partial F(u)$ consists of all slopes of affine linear supporting functionals that realize the value $F(u)$ at u and are below F . It may well happen that $\partial F(u)$ is empty or contains more than one element.

The subdifferential provides a handy generalization of Corollary 6.35 for minimization problems with convex functionals.

Theorem 6.43 (Optimality for Convex Minimization Problems) *Let $F : X \rightarrow \mathbf{R}_\infty$ be a convex functional on a real normed space. Then*

$$u^* \in X \text{ solves } \min_{u \in X} F(u) \iff 0 \in \partial F(u^*).$$

Proof An element $u \in X$ is a solution of $\min_{u \in X} F(u)$ if and only if $F(u^*) \leq F(u)$ for all $u \in X$. But $F(u^*) = F(u^*) + \langle 0, u - u^* \rangle$ is, by the definition of subgradients, equivalent to $0 \in \partial F(u^*)$. \square

To apply the above result to concrete problems, we spend some time to analyze further properties of the subdifferential. In general, the subgradient obeys “almost” the same rules as the classical derivative, but some care has to be used. Examples will show that many functionals relevant to our applications that do not have a classical derivative can be treated with this generalized concept.

Example 6.44 (Subdifferential in \mathbf{R}) Let us discuss the subgradients of the convex functions $\varphi_1, \varphi_2 : \mathbf{R} \rightarrow \mathbf{R}_\infty$ given by

$$\varphi_1(t) = \begin{cases} \frac{3}{10}t^2 - \frac{1}{4}t & \text{if } t \leq 0, \\ \frac{1}{2}t^2 + t & \text{if } t > 0, \end{cases} \quad \varphi_2(t) = \begin{cases} 0 & \text{if } t \in [-1, 1], \\ \infty & \text{otherwise.} \end{cases}$$

The first function is differentiable on $\mathbf{R} \setminus \{0\}$. It has a kink at the origin, and it is easy to see that $st \leq \frac{1}{2}t^2 + t$ for all $t \geq 0$ if and only if $s \leq 1$. Similarly, $st \leq \frac{3}{10}t^2 - \frac{1}{4}t$ holds for all $t \leq 0$ if and only if $s \geq -\frac{1}{4}$. Hence, $\partial\varphi_1(0) = [-\frac{1}{4}, 1]$.

The function φ_2 is constant on $]-1, 1[$, hence differentiable there with derivative 0. At the point 1 we note that $s(t-1) \leq 0$ for all $t \in [-1, 1]$ if and only if $s \geq 0$. This shows that $\partial\varphi_2(1) = [0, \infty[$. A similar argument shows that $\partial\varphi_2(-1) =]-\infty, 0]$, and the subgradient is empty at all other points. In conclusion we get

$$\partial\varphi_1(t) = \begin{cases} \{\frac{6}{10}t - \frac{1}{4}\} & \text{if } t < 0, \\ [-\frac{1}{4}, 1] & \text{if } t = 0, \\ \{t+1\} & \text{if } t > 0, \end{cases} \quad \partial\varphi_2(t) = \begin{cases}]-\infty, 0] & \text{if } t = -1, \\ \{0\} & \text{if } t \in]-1, 1[, \\ [0, \infty[& \text{if } t = 1, \\ \emptyset & \text{otherwise} \end{cases}$$

(see also Fig. 6.8).

Before we come to the proofs of the next theorems we note that an element of the subgradient has a geometric interpretation. Every $w \in \partial F(u)$ gives via

$$\{(v, t) \mid \langle w, v \rangle - t = \langle w, u \rangle - F(u)\}$$

a closed hyperplane in $X \times \mathbf{R}$ that separates $\text{epi } F$ and $\{(u, F(u))\}$. It is not “vertical” in the sense that its projection onto X is the whole space. Compare this with the Hahn-Banach theorem: in this generality we get only hyperplanes that separate $\text{epi } F$ and $\{(u, F(u) - \varepsilon)\}$ for every $\varepsilon > 0$. To treat the limiting case, which is what we need for the subgradient, we note the following variant of the separation theorem of Hahn-Banach, sometimes called *Eidelheit’s theorem*.

Lemma 6.45 (Eidelheit’s Theorem) *Let $A, B \subset X$ be nonempty convex subsets of a normed space X . If $\text{int}(A) \neq \emptyset$ and $\text{int}(A) \cap B = \emptyset$, then there exist $x^* \in X^*$,*

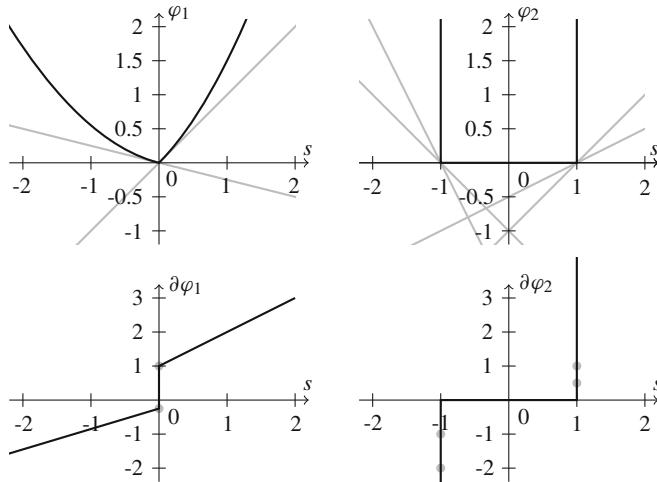


Fig. 6.8 Example of subdifferentials of convex functions. The top row shows the graphs, the bottom row the respective subdifferentials. The gray affine supporting functionals correspond to the gray points in the subdifferential. The function φ_1 is differentiable except at the origin, and there the subgradient is a compact interval; the subgradient of the indicator functional φ_2 is the nonpositive semiaxis at -1 and the nonnegative semiaxis at 1 . Outside of $[-1, 1]$ it is empty

$x^* \neq 0$, and $\lambda \in \mathbf{R}$ such that

$$\operatorname{Re} \langle x^*, x \rangle \leq \lambda \quad \text{for all } x \in A \quad \text{and} \quad \operatorname{Re} \langle x^*, x \rangle \geq \lambda \quad \text{for all } x \in B.$$

Proof First, Theorem 2.29 provides us with $x^* \in X^*$, $x^* \neq 0$, and $\lambda \in \mathbf{R}$ such that

$$\operatorname{Re} \langle x^*, x \rangle \leq \lambda \quad \text{for all } x \in \operatorname{int}(A) \quad \text{and} \quad \operatorname{Re} \langle x^*, x \rangle \geq \lambda \quad \text{for all } x \in B.$$

But since $\overline{\operatorname{int}(A)} = A$ (Exercise 6.9), there exists for every $x \in A$ a sequence (x^n) in $\operatorname{int}(A)$ converging to x , and hence $\operatorname{Re} \langle x^*, x \rangle = \lim_{n \rightarrow \infty} \operatorname{Re} \langle x^*, x^n \rangle \leq \lambda$. \square

Now we collect some fundamental properties of the subdifferential.

Theorem 6.46 *Let $F : X \rightarrow \mathbf{R}_\infty$ be a convex function on a real normed space X . Then the subdifferential ∂F satisfies the following:*

1. *For every u , the set $\partial F(u)$ is a convex weakly* closed subset of X^* .*
2. *If F is also lower semicontinuous, then ∂F is a strongly-weakly* and also weakly-strongly closed subset of $X \times X^*$, i.e., for sequences $((u^n, w^n))$ in ∂F ,*

$$\left. \begin{array}{l} u^n \rightarrow u \quad \text{in } X, \quad w^n \xrightarrow{*} w \quad \text{in } X^* \\ \text{or} \quad u^n \rightharpoonup u \quad \text{in } X, \quad w^n \rightarrow w \quad \text{in } X^* \end{array} \right\} \quad \Rightarrow \quad (u, w) \in \partial F.$$

3. *If F is continuous at u , then $\partial F(u)$ is nonempty and bounded.*

Proof Assertions 1 and 2: The proof is a good exercise (Exercise 6.10).

Assertion 3: Let F be continuous at u . Then there is a $\delta > 0$ such that $|F(v) - F(u)| < 1$ for $v \in B_\delta(u)$. In particular, for some $w \in \partial F(u)$ and all $v \in X$ with $\|v\|_X < 1$, one has the inequality

$$1 > F(u + \delta v) - F(u) \geq \langle w, \delta v \rangle \quad \Rightarrow \quad \langle w, v \rangle < \delta^{-1}.$$

Taking the supremum over $\|v\|_X < 1$, we obtain $\|w\|_{X^*} \leq \delta^{-1}$, and hence $\partial F(u)$ is bounded.

To show that $\partial F(u)$ is not empty, note that $\text{epi } F$ has nonempty interior, since the open set $B_\delta(u) \times]F(u) + 1, \infty[$ is a subset of the epigraph of F . Moreover, $(u, F(u))$ is not in $\text{int}(\text{epi } F)$, since every $(v, t) \in \text{int}(\text{epi}(F))$ satisfies $t > F(v)$. Now we choose $A = \text{epi } F$ and $B = \{(u, F(u))\}$ in Lemma 6.45 to get $0 \neq (w^0, t_0) \in X^* \times \mathbf{R}$,

$$\langle w^0, v \rangle + t_0 t \leq \lambda \quad \forall v \in \text{dom } F, \quad F(v) \leq t \quad \text{and} \quad \langle w^0, u \rangle + t_0 F(u) \geq \lambda.$$

Taking $v = u$, $t = F(u)$ shows that $\lambda = \langle w^0, u \rangle + t_0 F(u)$. Also $t_0 < 0$ holds, since $t_0 > 0$ leads to a contradiction by letting $t \rightarrow \infty$ and for $t_0 = 0$ we would get $\langle w^0, v - u \rangle \leq 0$ for all $v \in B_\delta(u)$, which would imply $w^0 = 0$, which contradicts $(w^0, t_0) \neq 0$. With $w = -t_0^{-1}w^0$ and some rearrangements we finally get

$$F(u) + \langle w, v - u \rangle \leq F(v) \quad \forall v \in \text{dom } F \quad \Rightarrow \quad w \in \partial F(u). \quad \square$$

If F is continuous at u , then one needs to check the subgradient inequality only for v in a dense subset of $\text{dom } F$ to calculate $\partial F(u)$. In the case of finite-dimensional spaces, this even follows, somewhat surprisingly, from convexity alone, even at points where F is not continuous.

Lemma 6.47 *Let $F : \mathbf{R}^N \rightarrow \mathbf{R}_\infty$ be proper and convex, and $V \subset \text{dom } F$ a dense subset of $\text{dom } F$. If for some $u \in \text{dom } F$, $w \in \mathbf{R}^N$ and all $v \in V$ one has*

$$F(u) + w \cdot (v - u) \leq F(v),$$

then $w \in \partial F(u)$.

Proof We show that for every $v^0 \in \text{dom } F$ there exists a sequence (v^n) in V with $\lim_{n \rightarrow \infty} v^n = v^0$ and $\limsup_{n \rightarrow \infty} F(v^n) \leq F(v^0)$. Then the claim follows by taking limits in the subgradient inequality.

Let $v^0 \in \text{dom } F$ and set

$$K = \max \{k \in \mathbf{N} \mid \exists u^1, \dots, u^k \in \text{dom } F, \quad u^1 - v^0, \dots, u^k - v^0 \text{ linearly independent}\}.$$

The case $K = 0$ is trivial; hence we assume $K \geq 1$ and choose $u^1, \dots, u^K \in \text{dom } F$ such that

$$\text{dom } F \subset U = v^0 + \text{span}(u^1 - v^0, \dots, u^K - v^0).$$

Now we consider the sets

$$S_n = \left\{ v^0 + \sum_{i=1}^K \lambda_i (u^i - v^0) \mid \sum_{i=1}^K \lambda_i \leq \frac{1}{n}, \lambda_1, \dots, \lambda_K \geq 0 \right\}$$

for $n \geq 1$ and note that their interior with respect to the relative topology on U is not empty. Hence, for every $n \geq 1$ there exists some $v^n \in S_n \cap V$, since V is also dense in S_n . We have

$$v^n = v^0 + \sum_{i=1}^K \lambda_i^n (u^i - v^0) = \left(1 - \sum_{i=1}^K \lambda_i^n\right) v^0 + \sum_{i=1}^K \lambda_i^n u^i$$

for suitable $\lambda_1^n, \dots, \lambda_K^n \geq 0$ with $\sum_{i=1}^K \lambda_i^n \leq \frac{1}{n}$. Thus, $\lim_{n \rightarrow \infty} v^n = v^0$ and by convexity of F

$$\limsup_{n \rightarrow \infty} F(v^n) \leq \limsup_{n \rightarrow \infty} \left(1 - \sum_{i=1}^K \lambda_i^n\right) F(v^0) + \sum_{i=1}^K \lambda_i^n F(u^i) = F(v^0). \quad \square$$

The following example is a first nontrivial application of the calculus of subdifferentials.

Example 6.48 (Subdifferential of Convex and Closed Constraints) Let K be a nonempty, convex, and closed set in a real normed space X . The subdifferential of the indicator functional ∂I_K at $u \in K$ is characterized by

$$w \in \partial I_K(u) \iff \langle w, v - u \rangle \leq 0 \quad \text{for all } v \in K.$$

It is easy to see that $\partial I_K(u)$ is a *convex cone*: for $w^1, w^2 \in \partial I_K(u)$ we also have $w^1 + w^2 \in \partial I_K(u)$, and for $u \in \partial I_K(u)$, $\alpha \geq 0$, also $\alpha u \in \partial I_K(u)$. This cone is called the *normal cone* (of K at u). Moreover, it is always the case that $0 \in \partial I_K(u)$, i.e., the subgradient is nonempty exactly on K . The special case $K = U + u^0$, with a closed subspace U and $u^0 \in X$, leads to $\partial I_K(u) = U^\perp$ for all $u \in K$.

If $K \neq \emptyset$ satisfies

$$K = \{u \in X \mid G(u) \leq 0, G : X \rightarrow \mathbf{R} \text{ convex and Gâteaux differentiable}\},$$

and if some u with $G(u) < 0$ exists, then we claim that

$$\partial I_K(u) = \begin{cases} \{\mu DG(u) \mid \mu \geq 0, \mu G(u) = 0\} & \text{if } u \in K, \\ \emptyset & \text{otherwise.} \end{cases}$$

For $G(u) < 0$ we need to show that $\partial I_K(u) = \{0\}$. By continuity of G we have for a suitable $\delta > 0$ that $G(u + v) < 0$ for all $\|v\|_X < \delta$, and hence, every $w \in \partial I_K(u)$ satisfies the inequality $\langle w, v \rangle = \langle w, u + v - u \rangle \leq 0$ for all $\|v\|_X < \delta$; thus $w = 0$ has to hold. Consequently, $\{0\}$ is the only element in $\partial G(u)$.

For $G(u) = 0$ the claim is $\partial I_K(u) = \{\mu DG(u) \mid \mu \geq 0\}$. We argue that $DG(u) \neq 0$ has to hold: otherwise, u would be a minimizer of G and the functional could not take on negative values. Now choose $w \in \partial I_K(u)$. For every $v \in X$ with $\langle DG(u), v \rangle = -\alpha_v < 0$, Gâteaux differentiability enables us to find some $t > 0$ such that

$$G(u + tv) - G(u) - \langle DG(u), tv \rangle \leq \frac{\alpha_v}{2}t.$$

This implies $G(u + tv) \leq t(\frac{\alpha_v}{2} + \langle DG(u), v \rangle) = -t\frac{\alpha_v}{2} < 0$ and hence $u + tv \in K$. Plugging this into the subgradient inequality we get

$$\langle w, v \rangle \leq 0 \quad \text{for all } \langle DG(u), v \rangle < 0.$$

Now assume that $w \notin \{\mu DG(u) \mid \mu \geq 0\}$. For $w = \mu DG(u)$ with $\mu < 0$, we get for every v with $\langle DG(u), v \rangle < 0$ the inequality $\langle w, v \rangle > 0$, a contradiction. Hence, we can assume that w and $DG(u)$ are linearly independent. We conclude that the mapping $T : X \rightarrow \mathbf{R}^2$ with $v \mapsto (\langle DG(u), v \rangle, \langle w, v \rangle)$ is surjective: otherwise, there would be a pair $(\alpha, \beta) \neq 0$ with $\alpha \langle DG(u), v \rangle = \beta \langle w, v \rangle$ for all $v \in X$, $DG(u)$, and w would not be linearly independent. This shows the existence of some $v \in X$ with $\langle DG(u), v \rangle < 0$ and $\langle w, v \rangle > 0$, which implies the desired contradiction. Hence, we conclude that $w = \mu DG(u)$ with $\mu \geq 0$.

Finally, every $w = \mu DG(u)$ with $\mu \geq 0$ is an element of $\partial G(u)$, since from Theorem 6.33 we get that for all $v \in K$,

$$\langle w, v - u \rangle = \mu G(u) + \mu \langle DG(u), v - u \rangle \leq \mu G(v) \leq 0.$$

Hence, the subgradient of I_K can be expressed in terms of G and its derivative only. Moreover, it contains at most one direction, a property that is not satisfied in the general case (see Fig. 6.9).

Example 6.49 (Subdifferential of Norm Functionals) Let $\varphi : [0, \infty[\rightarrow \mathbf{R}_\infty$ be a convex monotonically increasing function and set $R = \sup \{t \geq 0 \mid \varphi(t) < \infty\}$, where we allow $R = \infty$. Then the functional $F(u) = \varphi(\|u\|_X)$ is convex on the real normed space X .

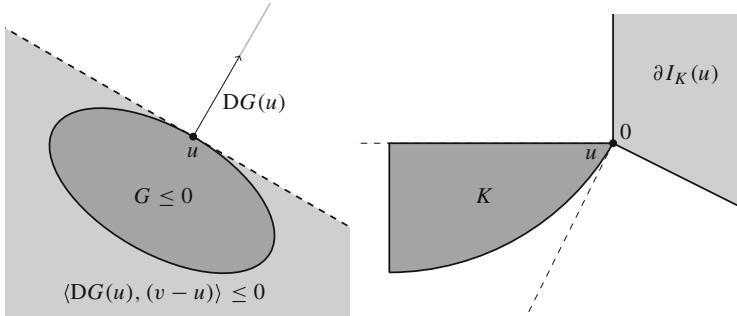


Fig. 6.9 Visualization of the normal cone associated with convex constraints. Left: The normal cone for the set $K = \{G \leq 0\}$ with Gâteaux differentiable G consists of the nonnegative multiples of the derivative $DG(u)$. The plane $\langle DG(u), v - u \rangle = 0$ is “tangential” to u , and K is contained in the corresponding nonpositive halfspace $\langle DG(u), v - u \rangle \leq 0$. Right: An example of a convex set for which the normal cone $\partial I_K(u)$ at some point u contains more than one linearly independent direction

The subgradient of F in u is given by

$$\partial F(u) = \{w \in X^* \mid \langle w, u \rangle = \|w\|_{X^*}\|u\|_X \text{ and } \|w\|_{X^*} \in \partial\varphi(\|u\|_X)\}.$$

To prove this claim, let $w \in \partial F(u)$ for $\|u\| \leq R$. For every vector $v \in X$ with $\|v\|_X = \|u\|_X$, the subgradient inequality (6.10) implies

$$\varphi(\|u\|_X) + \langle w, v - u \rangle \leq \varphi(\|u\|_X) \quad \Rightarrow \quad \langle w, v \rangle \leq \langle w, u \rangle \leq \|w\|_{X^*}\|u\|_X.$$

Taking the supremum over all $\|v\|_X = \|u\|_X$, we obtain $\langle w, u \rangle = \|w\|_{X^*}\|u\|_X$.

For $u = 0$, we get, additionally, by the subgradient inequality that for fixed $t \geq 0$ and all $\|v\|_X = t$, one has

$$\varphi(0) + \langle w, v - 0 \rangle \leq \varphi(\|v\|_X) = \varphi(t) \quad \Rightarrow \quad \varphi(0) + \|w\|_{X^*}t \leq \varphi(t).$$

And since $t \geq 0$ is arbitrary, we get $\|w\|_{X^*} \in \partial\varphi(\|u\|_X)$. (For the latter claim we have implicitly extended φ by $\varphi(t) = \inf_{s \geq 0} \varphi(s)$ for $t < 0$.) For the case $u \neq 0$ we plug $v = t\|u\|_X^{-1}u$ for some $t \geq 0$ in the subgradient inequality,

$$\varphi(t) = \varphi(\|v\|_X) \geq \varphi(\|u\|_X) + \langle w, v - u \rangle = \varphi(\|u\|_X) + \|w\|_{X^*}(t - \|u\|_X),$$

and conclude that $\|w\|_{X^*} \in \partial\varphi(\|u\|_X)$ also in this case.

To prove the reverse inclusion, let $w \in X^*$ be such that $\langle w, u \rangle = \|w\|_{X^*}\|u\|_X$ and $\|w\|_{X^*} \in \partial\varphi(\|u\|_X)$. For all $v \in X$, we have

$$\varphi(\|u\|_X) + \langle w, v - u \rangle \leq \varphi(\|u\|_X) + \|w\|_{X^*}(\|v\|_X - \|u\|_X) \leq \varphi(\|v\|_X),$$

i.e., $w \in \partial F(u)$.

Theorem 6.46 says that $\partial F(u) \neq \emptyset$ for all $\|u\|_X < R$. If, additionally, $\partial\varphi(R) \neq \emptyset$ holds, we also have $\partial F(u) \neq \emptyset$ for all $\|u\|_X \leq R$, since there always exists $w \in X^*$ that satisfies $\langle w, u \rangle = \|w\|_{X^*} \|u\|_X$ if the norm $\|w\|_{X^*}$ is prescribed.

If X is a Hilbert space, we can describe ∂F a little bit more concretely. Using the Riesz map J_X^{-1} we argue as follows: For $u \neq 0$, one has $\langle w, u \rangle = \|w\|_{X^*} \|u\|_X$ if and only if $(u, J_X^{-1}w) = \|J_X^{-1}w\|_X \|u\|_X$, which in turn is equivalent to the existence of some $\lambda \geq 0$ such that $J_X^{-1}w = \lambda u$ holds. Then, the condition $\|w\|_{X^*} \in \partial\varphi(\|u\|_X)$ becomes $\lambda \in \partial\varphi(\|u\|_X)/\|u\|_X$, and thus the subgradient is given by $\lambda J_X u$ with $\lambda \in \partial\varphi(\|u\|_X)/\|u\|_X$. For $u = 0$, it consists exactly of these $J_X v \in X^*$ with $\|v\|_X \in \partial\varphi(0)$. In conclusion, we get

$$\partial F(u) = \begin{cases} \partial\varphi(\|u\|_X) \frac{J_X u}{\|u\|_X} & \text{if } u \neq 0, \\ \partial\varphi(0) J_X \{ \|v\|_X = 1 \} & \text{if } u = 0. \end{cases}$$

Example 6.50 (Subdifferential for Convex Functions of Integrands) As in Example 6.23 let $\varphi : \mathbf{R}^N \rightarrow \mathbf{R}_\infty$, $N \geq 1$ be a convex functional, $p \in [1, \infty[$, and $F : L^p(\Omega, \mathbf{R}^N) \rightarrow \mathbf{R}_\infty$ given by $F(u) = \int_\Omega \varphi(u(x)) dx$. If φ is lower semicontinuous, then the subgradient of F at $u \in L^p(\Omega, \mathbf{R}^N)$ is the set

$$\partial F(u) = \{w \in L^{p^*}(\Omega, \mathbf{R}^N) \mid w(x) \in \partial\varphi(u(x)) \text{ for almost all } x \in \Omega\}.$$

This can be seen as follows: If $w \in L^p(\Omega, \mathbf{R}^N)^* = L^{p^*}(\Omega, \mathbf{R}^N)$ satisfies the condition $w(x) \in \partial\varphi(u(x))$ almost everywhere, we take any $v \in L^p(\Omega, \mathbf{R}^N)$ and plug $v(x)$ for almost every x in the subgradient inequality for φ and get, after integration,

$$\int_\Omega \varphi(u(x)) dx + \langle w, v - u \rangle_{L^p \times L^{p^*}} \leq \int_\Omega \varphi(v(x)) dx,$$

i.e., $w \in \partial F(u)$. For the reverse inclusion, let $w \in \partial F(u)$. Then for every $v \in L^p(\Omega, \mathbf{R}^N)$, we have

$$\int_\Omega \varphi(v(x)) - \varphi(u(x)) - w(x) \cdot (v(x) - u(x)) dx \geq 0.$$

Now choose an at most countable set $V \subset \text{dom } \varphi$, which is dense in $\text{dom } \varphi$. For every $\bar{v} \in V$ and every measurable $A \subset \Omega$ with $\mu(A) < \infty$ we can plug $v_A(x) = \chi_A \bar{v} + \chi_{\Omega \setminus A} u$ in the subgradient inequality and get

$$\int_A \varphi(v(x)) - \varphi(u(x)) - w(x) \cdot (v(x) - u(x)) dx \geq 0$$

and consequently

$$\varphi(u(x)) + w(x) \cdot (\bar{v} - u(x)) \leq \varphi(\bar{v}) \quad \text{for almost every } x \in \Omega.$$

Since V is countable, the union of all sets where the above does not hold is still a nullset, and hence we get that for almost every $x \in \Omega$,

$$\varphi(u(x)) + w(x) \cdot (\bar{v} - u(x)) \leq \varphi(\bar{v}) \quad \text{for all } \bar{v} \in V.$$

By Lemma 6.47 we finally conclude that $w(x) \in \partial\varphi(u(x))$ for almost every $x \in \Omega$.

Now we prove some useful rules for subdifferential calculus. Most rules are straightforward generalizations of the respective rules for the classical derivatives, sometimes with additional assumptions.

In this context we denote translations as follows: $T_{u^0}u = u + u^0$. Since this notion is used in this chapter exclusively, there will be no confusion with the distributions T_{u^0} induced by u^0 (cf. Sect. 2.3). The rules of subgradient calculus are particularly useful to find minimizers (cf. Theorem 6.43). Note that these rules often require additional continuity assumptions.

Theorem 6.51 (Calculus for Subdifferentials) *Let X, Y be real normed spaces, $F, G : X \rightarrow \mathbf{R}_\infty$ proper convex functionals, and $A : Y \rightarrow X$ linear and continuous. The subdifferential obeys the following rules:*

1. $\partial(\lambda F) = \lambda\partial F$ for $\lambda > 0$,
2. $\partial(F \circ T_{u^0})(u) = \partial F(u + u^0)$ for $u^0 \in X$,
3. $\partial(F + G) \supset \partial F + \partial G$ and $\partial(F + G) = \partial F + \partial G$ if F is continuous at some point $u^0 \in \text{dom } F \cap \text{dom } G$,
4. $\partial(F \circ A) \supset A^* \circ \partial F \circ A$ and $\partial(F \circ A) = A^* \circ \partial F \circ A$ if F is continuous at some point $u^0 \in \text{rg}(A) \cap \text{dom } F$.

Proof Assertions 1 and 2: It is simple to check the rules by direct application of the definition.

Assertion 3: The inclusion is immediate: for $u \in X$, $w^1 \in \partial F(u)$ and $w^2 \in \partial G(u)$ the subgradient inequality (6.10) implies

$$\begin{aligned} (F(u) + G(u)) + \langle w^1 + w^2, v - u \rangle \\ = F(u) + \langle w^1, v - u \rangle + G(u) + \langle w^2, v - u \rangle \leq F(v) + G(v) \end{aligned}$$

for all $v \in X$.

For the reverse inclusion, let $w \in \partial(F + G)(u)$, which implies $u \in \text{dom } F \cap \text{dom } G$ and hence

$$F(v) - F(u) - \langle w, v - u \rangle \geq G(u) - G(v) \quad \text{for all } v \in \text{dom } F \cap \text{dom } G. \quad (6.11)$$

With $\bar{F}(v) = F(v) - \langle w, v \rangle$ the inequality becomes $\bar{F}(v) - \bar{F}(u) \geq G(u) - G(v)$. Now we aim to find a suitable linear functional that “fits” between this inequality, i.e., some $w^2 \in X^*$ for which

$$\begin{aligned}\bar{F}(v) - \bar{F}(u) &\geq \langle w^2, u - v \rangle \quad \text{for all } v \in \text{dom } F, \\ \langle w^2, u - v \rangle &\geq G(u) - G(v) \quad \text{for all } v \in \text{dom } G.\end{aligned}\tag{6.12}$$

We achieve this by a suitable separation of the sets

$$K_1 = \{(v, t - \bar{F}(u)) \in X \times \mathbf{R} \mid \bar{F}(v) \leq t\}, \quad K_2 = \{(v, G(u) - t) \in X \times \mathbf{R} \mid G(v) \leq t\}.$$

We note that K_1, K_2 are nonempty convex sets, and moreover, $\text{int}(K_1)$ is not empty (the latter due to continuity of \bar{F} in u^0). Also we note that $\text{int}(K_1) \cap K_2 = \emptyset$, since $(v, t) \in \text{int}(K_1)$ implies $t > \bar{F}(v) - \bar{F}(u)$ while $(v, t) \in K_2$ means that $G(u) - G(v) \geq t$. If both were satisfied we would get a contradiction to (6.11). Lemma 6.45 implies that there exist $0 \neq (w^0, t_0) \in X^* \times \mathbf{R}$ and $\lambda \in \mathbf{R}$ such that

$$\begin{aligned}\langle w^0, v \rangle + t_0(t - \bar{F}(u)) &\leq \lambda \quad \forall v \in \text{dom } F, \bar{F}(v) \leq t, \\ \langle w^0, v \rangle + t_0(G(u) - t) &\geq \lambda \quad \forall v \in \text{dom } G, G(v) \leq t.\end{aligned}$$

Now we show that $t_0 < 0$. The case $t_0 > 0$ leads to a contradiction by letting $v = u$, $t > \bar{F}(u)$ and $t \rightarrow \infty$. In the case $t_0 = 0$, we would get $\langle w^0, v \rangle \leq \lambda$ for all $v \in \text{dom } F$ and especially $\langle w^0, u^0 \rangle < \lambda$, since u^0 is in the interior of $\text{dom } F$. However, since $u^0 \in \text{dom } G$, we also get $\langle w^0, u^0 \rangle \geq \lambda$, which is again a contradiction.

With $t = \bar{F}(v)$ and $t = G(v)$, respectively, we obtain

$$\begin{aligned}\bar{F}(v) - \bar{F}(u) &\geq -\langle t_0^{-1}w^0, v \rangle + t_0^{-1}\lambda \quad \forall v \in \text{dom } F, \\ G(u) - G(v) &\leq -\langle t_0^{-1}w^0, v \rangle + t_0^{-1}\lambda \quad \forall v \in \text{dom } G.\end{aligned}$$

Letting $v = u$ in both inequalities and $w^2 = t_0^{-1}w^0$, we see that $\lambda = \langle w^0, u \rangle$, which implies (6.12). On the one hand, this implies $w^2 \in \partial G(u)$, and on the other hand, by definition of \bar{F} , we get $w^1 = w - w^2 \in \partial F(u)$. Altogether, we get $w \in \partial F(u) + \partial G(u)$ and the proof is complete.

Assertion 4: Again, it is simple to see the inclusion: for $w = A^*\bar{w}$ with $\bar{w} \in \partial F(Au)$, we get

$$F(Au) + \langle A^*\bar{w}, v - u \rangle = F(Au) + \langle \bar{w}, Av - Au \rangle \leq F(Av)$$

for all $v \in Y$. This shows that $w \in \partial(F \circ A)(u)$.

The proof of the reverse inclusion proceeds analogously to item 3. Let $w \in \partial(F \circ A)(u)$, i.e.,

$$F(Au) + \langle w, v - u \rangle \leq F(Av) \quad \text{for all } v \in Y.\tag{6.13}$$

We aim to introduce a separating linear functional into this inequality that amounts to a separation of the nonempty convex sets

$$K_1 = \text{epi } F, \quad K_2 = \{(Av, F(Au) + \langle w, v - u \rangle) \in X \times \mathbf{R} \mid v \in Y\}.$$

In analogy to item 3 we note that $\text{int}(K_1)$ is nonempty and also $\text{int}(K_1) \cap K_2 = \emptyset$. Again, by Lemma 6.45 there is some $0 \neq (w^0, t_0) \in X^* \times \mathbf{R}$ such that

$$\begin{aligned} \langle w^0, \bar{v} \rangle + t_0 t &\leq \lambda \quad \forall \bar{v} \in \text{dom } F, t \geq F(\bar{v}), \\ \langle w^0, Av \rangle + t_0(F(Au) + \langle w, v - u \rangle) &\geq \lambda \quad \forall v \in Y. \end{aligned} \tag{6.14}$$

The case $t_0 > 0$ cannot occur, and $t_0 \neq 0$ follows from the continuity of F at u^0 and $u^0 \in \text{dom } F \cap \text{rg}(A)$. If we set $\bar{v} = Au$, $t = F(\bar{v})$ and $v = u$, we also conclude that $\lambda = \langle w^0, Au \rangle + t_0 F(Au)$. By the second inequality in (6.14) we get

$$\langle w^0, Av - Au \rangle + t_0 \langle w, v - u \rangle \geq 0 \quad \forall v \in Y$$

and hence $w = -t_0^{-1} A^* w^0$. Setting $\bar{w} = -t_0^{-1} w^0$, the first inequality in (6.14) implies

$$\langle \bar{w}, \bar{v} - Au \rangle + F(Au) \leq F(\bar{v}) \quad \forall \bar{v} \in \text{dom } F,$$

i.e., $\bar{w} \in \partial F(Au)$. This shows that $\partial(F \circ A)(u) \subset (A^* \circ \partial F \circ A)(u)$. \square

Corollary 6.52 *Let F be convex and continuous at u^0 and $\partial F(u^0) = \{w\}$. Then F is Gâteaux-differentiable at u^0 with $DF(u^0) = w$.*

Proof Without loss of generality, we can assume that $u^0 = 0$. For all $v \in X$ we define the convex function $t \mapsto F_v(t) = F(tv)$. By Theorem 6.51 we have $\partial F_v(0) = \{\langle w, v \rangle\}$, which implies, for $t > 0$,

$$0 \leq \frac{F_v(t) - F_v(0)}{t} - \langle w, v \rangle.$$

On the other hand, for every $\varepsilon > 0$, there exists $t_\varepsilon > 0$ such that $F_v(t_\varepsilon) < F_v(0) + t_\varepsilon \langle w, v \rangle + t_\varepsilon \varepsilon$ (note that $\langle w, v \rangle + \varepsilon \notin \partial F_v(0)$). By convexity of F_v , for every $t \in [0, t_\varepsilon]$ one has

$$F_v(t) \leq \frac{t}{t_\varepsilon} F_v(t_\varepsilon) + \frac{t_\varepsilon - t}{t_\varepsilon} F_v(0) \leq F(0) + t \langle w, v \rangle + t \varepsilon,$$

and hence

$$\frac{F_v(t) - F_v(0)}{t} - \langle w, v \rangle \leq \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, it follows that $\lim_{t \rightarrow 0^+} \frac{1}{t} (F_v(t) - F_v(0)) = \langle w, v \rangle$, which proves Gâteaux-differentiability and $DF(u^0) = w$. \square

Remark 6.53

- The assertion in item 4 of Theorem 6.51 remains valid in other situations as well. If, for example, $\text{rg}(A) = X$ and F is convex, then $\partial(F \circ A) = A^* \circ \partial F \circ A$ without additional assumptions on continuity (cf. Exercise 6.11).

As another example, in the case of a densely defined linear mapping $A : \text{dom } A \subset Y \rightarrow X$, $\overline{\text{dom } A} = Y$, we obtain the same formula for $\partial(F \circ A)$ (cf. Example 6.23) when the adjoint has to be understood in the sense of Definition 2.25 (Exercise 6.12).

- One can also generalize the continuity assumptions. Loosely speaking, the sum rule holds if continuity of F and G at one point holds relatively to some subspaces whose sum is the whole of X and on which one can project continuously. Analogously, the continuity of F with respect to a subspace that contains the complement of $\text{rg}(A)$ is sufficient for the chain rule to hold (cf. Exercises 6.13–6.16).

On suitable spaces, there are considerably more “subdifferentiable” convex functions than Gâteaux-differentiable or continuous ones. We prove this claim using the sum rule.

Theorem 6.54 (Existence of Nonempty Subdifferentials) *Let $F : X \rightarrow \mathbf{R}_\infty$ be proper, convex and lower semicontinuous on a reflexive Banach space X . Then $\partial F \neq \emptyset$.*

Proof Choose some $u^0 \in X$ with $F(u^0) < \infty$ and $t_0 < F(u^0)$ and consider the minimization problem

$$\min_{(v,t) \in X \times \mathbf{R}} \|v - u^0\|_X^2 + (t - t_0)^2 + I_{\text{epi } F}((v, t)). \quad (6.15)$$

Since F is convex and lower semicontinuous, $I_{\text{epi } F}$ is convex and lower semicontinuous (Remark 6.26 and Example 6.29). The functional $\|(v, t)\| = (\|v\|_X^2 + t^2)^{1/2}$ is a norm on $X \times \mathbf{R}$, and hence convex, continuous, and coercive (Example 6.23 and Remark 6.13), and obviously, the same holds for the functional G defined by $G((v, t)) = \|(v - u^0, t - t_0)\|^2$. Hence, problem (6.15) satisfies all assumptions of Theorem 6.31 (see also Lemmas 6.14 and 6.21) and consequently, it has a minimizer $(u, \tau) \in \text{epi } F$.

We prove that $\tau \neq t_0$. If $\tau = t_0$ held, we would get $u \neq u^0$ and also that the segment joining (u, t_0) and $(u^0, F(u^0))$ was contained in $\text{epi } F$. For $\lambda \in [0, 1]$ we would get

$$G(u + \lambda(u^0 - u), t_0 + \lambda(F(u^0) - t_0)) = (\lambda - 1)^2 \|u - u^0\|_X^2 + \lambda^2 (F(u^0) - t_0)^2.$$

Setting $a = \|u - u^0\|_X^2$ and $b = (F(u^0) - t_0)^2$, we calculate that the right-hand side is minimal for $\lambda = a/(a + b) \in]0, 1[$, which leads to

$$G(u + \lambda(u^0 - u), t_0 + \lambda(F(u^0) - t_0)) = \frac{ab^2 + a^2b}{(a + b)^2} < \frac{a(b^2 + 2ab + a^2)}{(a + b)^2} = \|u - u^0\|_X^2,$$

since $\|u - u^0\|_X^2 > 0$. Then (u, τ) would not be a minimizer, which is a contradiction.

By the sum rule (Theorem 6.51, item 3) we write the optimality condition as

$$0 \in \partial(I_{\text{epi } F} + G)(u, \tau) = \partial I_{\text{epi } F}(u, \tau) + \partial G(u, \tau).$$

In particular, there exists $(w, s) \in X^* \times \mathbf{R}$ such that

$$\langle w, v - u \rangle + s(t - \tau) \leq 0 \quad \forall (v, t) \in \text{epi } F, \quad (6.16)$$

and

$$\langle w, v - u \rangle + s(t - \tau) \geq \|u - u^0\|_X^2 + (\tau - t_0)^2 - \|v - u^0\|_X^2 - (t - t_0)^2 \quad (6.17)$$

for all $(v, t) \in X \times \mathbf{R}$. One has $s \leq 0$, since for $s > 0$ we obtain a contradiction to (6.16) by letting $v = u$, $t > \tau$ and $t \rightarrow \infty$. The case $s = 0$ can also not occur: We choose $v = u$ and $t = t_0$, in (6.17), and since $\tau \neq t_0$ we obtain

$$0 = \langle w, u - u \rangle \geq \|u - u^0\|_X^2 + (\tau - t_0)^2 - \|u - u^0\|_X^2 > 0,$$

a contradiction. Hence, $s < 0$ and (6.16) implies with $v = u$ and $t = F(u)$ the inequality $\tau \leq F(u)$, moreover, since $(u, \tau) \in \text{epi } F$, we even get $\tau = F(u)$. For $v \in \text{dom } F$ and $t = F(v)$ we finally get

$$\langle w, v - u \rangle + s(F(v) - F(u)) \leq 0 \quad \forall v \in \text{dom } F,$$

which we rewrite as $F(u) + \langle -s^{-1}w, v - u \rangle \leq F(v)$ for all $v \in X$, showing that $-s^{-1}w \in \partial F(u)$. \square

The obtained results show their potential when applied to concrete minimization problems

Example 6.55

1. Minimization under constraints

Let $F : X \rightarrow \mathbf{R}_\infty$ be a convex and Gâteaux-differentiable functional and K a nonempty, convex, and closed subset of X . By Theorem 6.43, the minimizers u^* of

$$\min_{u \in K} F(u)$$

are exactly the u^* for which $0 \in \partial(F + I_K)(u^*)$. By Theorem 6.51 we can write $\partial(F + I_K)(u^*) = \partial F(u^*) + \partial I_K(u^*) = DF(u^*) + \partial I_K(u^*)$. Using the result of Example 6.48 we get the optimality condition

$$u^* \in K : \quad DF(u^*) + w^* = 0 \text{ with } \langle w^*, v - u^* \rangle \leq 0 \text{ for all } v \in K. \quad (6.18)$$

For the special case in which $K \neq \emptyset$ is given as

$$K = \bigcap_{m=1}^M K_m, \quad K_m = \{u \in X \mid G_m(u) \leq 0\}$$

with convex and Gâteaux-differentiable $G_m : X \rightarrow \mathbf{R}$, where there exists some $u \in X$ with $G_m(u) < 0$ for all $m = 1, \dots, M$, we get, by the characterization in Example 6.48 and the sum rule for subgradients, that

$$\begin{aligned} \partial I_K(u) &= \sum_{m=1}^M \partial I_{K_m}(u) \\ &= \begin{cases} \left\{ \sum_{m=1}^M \mu_m DG_m(u) \mid \mu_m \geq 0, \mu_m G_m(u) = 0 \right\} & \text{if } u \in K, \\ \emptyset & \text{otherwise.} \end{cases} \end{aligned}$$

The optimality condition becomes

$$u^* \in K : \quad DF(u^*) + \sum_{m=1}^M \mu_m^* DG_m(u^*) = 0, \quad \mu_m^* \geq 0, \quad \mu_m^* G_m(u^*) = 0 \quad (6.19)$$

for $m = 1, \dots, M$.

In this context, one calls the variables $\mu_m^* \geq 0$ the *Lagrange multipliers* for the constraints $\{G_m \leq 0\}$. These have to exist for every minimizer u^* .

The subdifferential calculus provides an alternative approach to optimality for general convex constraints. The w^* in (6.18) corresponds to the linear combination of the derivatives $DG_m(u^*)$ in (6.19), and the existence of Lagrange multipliers μ_m^* is abstracted by the condition that w^* is in the normal cone of K .

2. Tikhonov functionals

We consider an example similar to Example 6.32. Let X be a real Banach space, Y a Hilbert space, $A \in \mathcal{L}(X, Y)$, $u^0 \in Y$, and $p \in]1, \infty[$. Moreover, let $Z \hookrightarrow X$ be a real Hilbert space that is densely and continuously embedded in X , $\lambda > 0$, and $q \in]1, \infty[$. We aim at optimality conditions for the minimization of the Tikhonov functional (6.7). We begin with the functional $\Phi(v) = \frac{1}{p} \|v\|_Y^p$,

which is continuous, and by Example 6.49 the subgradient is

$$\partial\Phi(v) = (J_Y v)\|v\|_Y^{p-2}.$$

Using the rules from Theorem 6.51, we obtain for $u \in X$,

$$\partial(\Phi \circ T_{-u^0} \circ A)(u) = A^* J_Y(Au - u^0)\|Au - u^0\|_Y^{p-2}.$$

The objective function F from (6.7) is continuous, and hence we can apply the sum rule for subgradients to obtain

$$\partial F(u) = A^* J_Y(Au - u^0)\|Au - u^0\|_Y^{p-2} + \lambda \partial\Psi(u).$$

To calculate $\partial\Psi$, we note that we can write Ψ as a concatenation of $\frac{1}{q}\|\cdot\|_Z^q$ and the inverse embedding i^{-1} from X to Z with domain of definition $\text{dom } i^{-1} = Z$. By construction, i^{-1} is a closed and densely defined mapping. By continuity of the norm in Z , the respective chain rule (Remark 6.53) holds, hence $\partial\Psi = (i^{-1})^* \circ \partial\frac{1}{q}\|\cdot\|_Z^q \circ i^{-1}$. The space X^* is densely and continuously embedded in Z^* by $i^* : X^* \hookrightarrow Z^*$, and $(i^{-1})^*$ is a closed mapping from Z^* to X^* , and it is simple to see that it is equal to the inverse of the adjoint, i.e., to $(i^*)^{-1}$ (Exercise 6.17). We obtain

$$\partial\Psi(u) = \begin{cases} (J_Z u)\|u\|_Z^{q-2} & \text{if } u \in Z \text{ and } J_Z u \in X^*, \\ \emptyset & \text{otherwise.} \end{cases}$$

For the minimizer u^* of the Tikhonov functional F , it must be the case that $u^* \in Z$, $(J_Z u^*) \in X^*$, and

$$A^* J_Y(Au^* - u^0)\|Au^* - u^0\|_Y^{p-2} + \lambda(J_Z u^*)\|u^*\|_Z^{q-2} = 0.$$

Note that by construction u^* , belongs to the “better” space $\{u \in Z \mid J_Z u \in X^*\}$. This is a consequence of the penalty term Ψ and basically independent of the choice of Φ .

As announced, we can treat the minimization problems from Examples 6.37 and 6.38 in a unified way (Exercise 6.18). Moreover, we can treat more general problems than these in Examples 6.1–6.4; see Sect. 6.3.

Besides the subdifferential, there is another useful notion to treat minimization problems, which we introduce in the next section.

6.2.4 Fenchel Duality

One important concept for the treatment of convex minimization problems is *Fenchel duality*. It is related to the notion of the *dual problem* of a minimization problem. The following example gives some motivation.

Example 6.56 Let X be a real Hilbert space and Y a real, reflexive Banach space that is densely and continuously embedded in X , i.e. the embedding map $j : Y \hookrightarrow X$ is continuous and has dense range. We consider the strictly convex minimization problem

$$\min_{u \in X} \frac{\|u - u^0\|_X^2}{2} + \lambda \|u\|_Y \quad (6.20)$$

for some $\lambda > 0$. This could, for example, model a denoising problem (see Example 6.1, where $X = L^2(\mathbf{R}^d)$, but the penalty is a squared seminorm in $H^1(\mathbf{R}^d)$). To reformulate the problem, we identify $X = X^*$ and write $Y \subset X = X^* \subset Y^*$. Every $u \in Y$ is mapped via $w = j^*ju$ to some $w \in Y^*$, where $j^* : X^* \rightarrow Y^*$ denotes the adjoint of the continuous embedding. It is easy to see that $X = X^* \hookrightarrow Y^*$ densely, and hence

$$\lambda \|u\|_Y = \sup \{ \langle w, u \rangle \mid w \in Y^*, \|w\|_{Y^*} \leq \lambda \} = \sup \{ (w, u) \mid w \in X, \|w\|_{Y^*} \leq \lambda \}.$$

Consequently, we can write (6.20) as

$$\min_{u \in X} \sup_{\|w\|_{Y^*} \leq \lambda} \frac{\|u - u^0\|_X^2}{2} + (w, u).$$

Now assume that we can swap the minimum and the supremum (in general one has only “ $\sup \inf \leq \inf \sup$ ”, see Exercise 6.19) to obtain

$$\inf_{u \in X} \sup_{\|w\|_{Y^*} \leq \lambda} \frac{\|u - u^0\|_X^2}{2} + (w, u) = \sup_{\|w\|_{Y^*} \leq \lambda} \inf_{u \in X} \frac{\|u - u^0\|_X^2}{2} + (w, u).$$

The functional on the right-hand side is rewritten as

$$\frac{\|u - u^0\|_X^2}{2} + (w, u) = \frac{\|u - (u^0 - w)\|_X^2}{2} + \frac{\|u^0\|_X^2}{2} - \frac{\|u^0 - w\|_X^2}{2},$$

and hence is minimal for $u = u^0 - w$. Plugging this into the functional, we obtain

$$\min_{u \in X} \frac{\|u - u^0\|_X^2}{2} + \lambda \|u\|_Y = \max_{\|w\|_{Y^*} \leq \lambda} \frac{\|u^0\|_X^2}{2} - \frac{\|u^0 - w\|_X^2}{2}. \quad (6.21)$$

The maximization problem on the right-hand side is the dual problem to (6.20). Obviously, it is equivalent (in the sense that the solutions are the same) to the projection problem

$$\min_{w \in X} \frac{\|u^0 - w\|_X^2}{2} + I_{\{\|w\|_{Y^*} \leq \lambda\}}(w).$$

The latter has the unique solution $w^* = P_{\{\|w\|_{Y^*} \leq \lambda\}}(u^0)$, since projections onto nonempty, convex and closed sets in Hilbert spaces are well defined (note that $\{\|w\|_{Y^*} \leq \lambda\}$ is closed by the continuous embedding $j^*: X^* \rightarrow Y^*$).

For $\|w\|_{Y^*} \leq \lambda$ and $u \in X$, one has

$$\begin{aligned} \frac{\|u^0\|_X^2}{2} - \frac{\|u^0 - w\|_X^2}{2} &\leq \max_{\|w\|_{Y^*} \leq \lambda} \min_{u \in X} \frac{\|u - u^0\|_X^2}{2} + (w, u) \\ &= \min_{u \in X} \max_{\|w\|_{Y^*} \leq \lambda} \frac{\|u - u^0\|_X^2}{2} + (w, u) \leq \frac{\|u - u^0\|_X^2}{2} + \lambda \|u\|_Y, \end{aligned}$$

and we conclude that

$$0 \leq \frac{\|u^0 - w\|_X^2}{2} - \frac{\|u^0\|_X^2}{2} + \frac{\|u - u^0\|_X^2}{2} + \lambda \|u\|_Y \quad \text{for all } u \in X, \|w\|_Y \leq \lambda,$$

where equality holds if and only if one plugs in the respective optimal solutions u^* and w^* of the primal problem (6.20) and the dual problem (6.21). We rewrite the last inequality as

$$\begin{aligned} 0 &\leq \frac{\|u^0 - w\|_X^2}{2} - \frac{\|u^0\|_X^2}{2} + \frac{\|u - u^0\|_X^2}{2} + (w, u) - (w, u) + \lambda \|u\|_Y \\ &= \frac{\|u^0 - w - u\|_X^2}{2} + (\lambda \|u\|_Y - (w, u)). \end{aligned}$$

Both summands on the right-hand side are nonnegative for all $u \in X$ and $\|w\|_{Y^*} \leq \lambda$, and in particular both equal zero for a pair (u^*, w^*) of primal and dual solutions. This is equivalent to

$$\left(\frac{\|u^0 - w^* - u^*\|_X^2}{2} = 0 \Leftrightarrow u^* = u^0 - w^* \right) \quad \text{and} \quad (w^*, u^*) = \lambda \|u^*\|_Y.$$

If we know the dual solution w^* , we get $u^* = u^0 - w^*$ as primal solution. In conclusion, we have found a way to solve (6.20) by a projection:

$$u^* \text{ minimizes (6.20)} \quad \Leftrightarrow \quad u^* = u^0 - P_{\{\|w\|_{Y^*} \leq \lambda\}}(u^0).$$

The theory of Fenchel duality is a systematic treatment of the techniques we used in the above example. This leads to the definition of the *dual functional* and thus to the dual problem. The argument crucially relies on the fact that the supremum and infimum can be swapped, which is equivalent to the claim that the infimum for the primal problem equals the supremum for the dual problem. In the following we will derive sufficient conditions for this to hold. Finally, we will be interested in the relation of primal and dual solutions, which allows us, for example, to obtain a primal solution from a dual solution.

We begin with the definition of the dual functional. We aim to write a suitable convex functional as the supremum. The following result is the basis of the construction.

Lemma 6.57 *Every convex and lower semicontinuous functional $F : X \rightarrow \mathbf{R}_\infty$ on a real Banach space X is the pointwise supremum of a nonempty family of affine linear functionals, i.e., there exists a subset $K_0 \subset X^* \times \mathbf{R}$, $K_0 \neq \emptyset$, such that*

$$F = \sup_{(w,s) \in K_0} \langle w, \cdot \rangle_{X^* \times X} + s$$

pointwise.

Proof For $F \equiv \infty$, we just set $K_0 = X^* \times \mathbf{R}$. Otherwise, we construct in the following for every pair $(u, t) \in X \times \mathbf{R}$ with $t < F(u)$ a pair $(w, s) \in X^* \times \mathbf{R}$ such that

$$\langle w, u \rangle_{X^* \times X} + s > t \quad \text{and} \quad \langle w, v \rangle_{X^* \times X} + s < F(v) \quad \text{for all } v \in X.$$

If we set K_0 as the collection of all these (w, s) , we obtain the claim by

$$\sup_{(w,s) \in K_0} \langle w, v \rangle_{X^* \times X} + s \leq F(v) \quad \text{for all } v \in X$$

and the observation that for every fixed $u \in X$ and all $t < F(u)$, one has

$$\sup_{(w,s) \in K_0} \langle w, u \rangle_{X^* \times X} + s > t \quad \Rightarrow \quad \sup_{(w,s) \in K_0} \langle w, u \rangle_{X^* \times X} + s \geq \sup_{t < F(u)} t = F(u).$$

So let (u, t) with $t < F(u) < \infty$ be given (such a pair exists in this case). Analogously to Theorem 6.31 we separate $\{(u, t)\}$ from the closed, convex, and nonempty set $\text{epi } F$ by some $(w^0, s_0) \in X^* \times \mathbf{R}$, i.e., for some $\lambda \in \mathbf{R}$ and $\varepsilon > 0$,

$$\langle w^0, v \rangle + s_0 \tau \geq \lambda + \varepsilon \quad \text{for all } v \in \text{dom } F \text{ and } \tau \geq F(v)$$

and $\langle w^0, u \rangle + s_0 t \leq \lambda - \varepsilon$. Using $v = u$ and $\tau = F(u)$ in the above inequality, we see that

$$\langle w^0, u \rangle + s_0 F(u) > \lambda > \langle w^0, u \rangle + s_0 t \quad \Rightarrow \quad s_0(F(u) - t) > 0 \quad \Rightarrow \quad s_0 > 0.$$

Hence, $w = -s_0^{-1}w^0$ and $s = s_0^{-1}\lambda$ with $\tau = F(v)$ gives the desired inequalities $\langle w, u \rangle + s > t$ and $\langle w, v \rangle + s < F(v)$ for all $v \in X$. In particular, K_0 is not empty.

Now we treat the case $t < F(u) = \infty$, where we can also find some $(w^0, s_0) \in X^* \times \mathbf{R}$ and $\lambda \in \mathbf{R}$, $\varepsilon > 0$ with the above properties. If there exists $v \in \text{dom } F$ with $\langle w^0, u - v \rangle > -2\varepsilon$, we plug in v and $\tau > \max(t, F(v))$ and obtain

$$\langle w^0, v \rangle + s_0\tau - \varepsilon \geq \lambda \geq \langle w^0, u \rangle + s_0t + \varepsilon \quad \Rightarrow \quad s_0(\tau - t) \geq \langle w^0, u - v \rangle + 2\varepsilon > 0$$

and thus $s_0 > 0$; moreover, we can choose (w, s) as above. If such a v does not exist, then $\langle w^0, v - u \rangle + 2\varepsilon \leq 0$ for all $v \in \text{dom } F$. By the above consideration there exists a pair $(w^*, s^*) \in X^* \times \mathbf{R}$ with $\langle w^*, \cdot \rangle + s^* < F$. For this, one has with $c > 0$ that

$$\langle w^*, v \rangle + s^* + c(\langle w^0, v - u \rangle + 2\varepsilon) < F(v) \quad \text{for all } v \in X.$$

The choice $c > \max(0, (2\varepsilon)^{-1}(t - s^* - \langle w^*, u \rangle))$ and $w = w^* + cw^0$, $s = s^* + c(2\varepsilon - \langle w^*, u \rangle)$ gives the desired inequality. \square

Note that we already identified proper, convex, and lower semicontinuous functionals as interesting objects for minimization problems (see, for example, Theorem 6.31). Motivated by Lemma 6.57, we study pointwise suprema of continuous affine linear functionals in more detail. To begin with, it is clear that a set K_0 for which $F = \sup_{(w,s) \in K_0} \langle w, \cdot \rangle + s$ holds is not uniquely determined. However, there is always a distinguished set like this.

Corollary 6.58 *Every convex and lower semicontinuous functional $F : X \rightarrow \mathbf{R}_\infty$ on a real Banach space can be written as*

$$F = \sup_{(w,s) \in \text{spt}(F)} \langle w, \cdot \rangle + s, \quad \text{spt}(F) = \{(w, s) \in X^* \times \mathbf{R} \mid \langle w, \cdot \rangle + s \leq F\},$$

where $\text{spt}(F)$ is the set of affine linear supporting functionals for F .

Proof The set $\text{spt}(F)$ contains the set K_0 we constructed in Lemma 6.57, and hence $\sup_{(w,s) \in \text{spt}(F)} \langle w, \cdot \rangle + s \geq F$. The reverse inequality holds by construction, hence we obtain equality. \square

The set $\text{spt}(F)$ has some notable properties.

Lemma 6.59 *For every functional $F : X \rightarrow \mathbf{R}_\infty$ on a real Banach space X , the set $\text{spt}(F)$ of affine linear supporting functionals is convex and closed, and for $(w^0, s_0) \in \text{spt}(F)$ and $s \leq s_0$, one has also $(w^0, s) \in \text{spt}(F)$. For fixed $w \in X^*$ one has that $\{s \in \mathbf{R} \mid (w, s) \in \text{spt}(F)\}$ is unbounded from above if and only if F is not proper.*

Analogous claims hold for functionals on the dual space, i.e., for $G : X^* \rightarrow \mathbf{R}_\infty$ with

$$\text{spt}(G) = \{(u, t) \in X \times \mathbf{R} \mid \langle \cdot, u \rangle_{X^* \times X} + t \leq G\}.$$

Proof All claims follow by direct computations (Exercise 6.20). \square

We now see that the set $\text{spt}(F)$ for a proper F can be seen as the negative epigraph of a convex and lower semicontinuous functional $F^* : X^* \rightarrow \mathbf{R}_\infty$; one simply sets $-F^*(w) = \sup \{s \in \mathbf{R} \mid (w, s) \in \text{spt}(F)\}$. This condition $(w, s) \in \text{spt}(F)$ is equivalent to

$$\langle w, u \rangle - F(u) \leq -s \quad \text{for all } u \in X \quad \Leftrightarrow \quad s \leq -(\sup_{u \in X} \langle w, u \rangle - F(u)),$$

and hence $F^*(w) = \sup_{u \in X} \langle w, u \rangle - F(u)$. This motivates the following definition.

Definition 6.60 (Dual Functional) Let $F : X \rightarrow \mathbf{R}_\infty$ be a proper functional on a real Banach space X . Then

$$F^* : X^* \rightarrow \mathbf{R}_\infty, \quad F^*(w) = \sup_{u \in X} \langle w, u \rangle_{X^* \times X} - F(u)$$

defines a functional, called the *dual functional* or *Fenchel conjugate* of F .

Moreover, for a proper $G : X^* \rightarrow \mathbf{R}_\infty$ we define

$$G^* : X \rightarrow \mathbf{R}_\infty, \quad G^*(u) = \sup_{w \in X^*} \langle w, u \rangle_{X^* \times X} - G(w).$$

In particular, we call $F^{**} : X \rightarrow \mathbf{R}_\infty$ and $G^{**} : X^* \rightarrow \mathbf{R}_\infty$ the *bidual functionals* for F and G , respectively.

Moreover, the pointwise suprema of continuous affine linear functionals are denoted by

$$\Gamma_0(X) = \left\{ F : X \rightarrow \mathbf{R}_\infty \mid F = \sup_{(w,s) \in K_0} \langle w, \cdot \rangle_{X^* \times X} + s \neq \infty \text{ for some } \emptyset \neq K_0 \subset X^* \times \mathbf{R} \right\},$$

$$\Gamma_0(X^*) = \left\{ G : X^* \rightarrow \mathbf{R}_\infty \mid G = \sup_{(u,t) \in K_0} \langle \cdot, u \rangle_{X^* \times X} + t \neq \infty \text{ for some } \emptyset \neq K_0 \subset X \times \mathbf{R} \right\},$$

where ∞ denotes the functional that is constant infinity.

This definition is fundamentally related to the notion of subgradients.

Remark 6.61 (Fenchel Inequality)

- For $F : X \rightarrow \mathbf{R}_\infty$ proper and $(u, w) \in X \times X^*$ with $F(u) < \infty$ and $F^*(w) < \infty$, one has $\langle w, u \rangle - F(u) \leq F^*(w)$. It is easy to see that then (and also in all other cases),

$$\langle w, u \rangle \leq F(u) + F^*(w) \quad \text{for all } u \in X, w \in X^*. \quad (6.22)$$

This inequality is called *Fenchel inequality*.

- If in the same situation, one has $\langle w, u \rangle = F(u) + F^*(w)$, then $F(u) < \infty$ and by definition $\langle w, u \rangle - F(u) \geq \langle w, v \rangle - F(v)$ for all $v \in \text{dom } F$. Put differently, for all $v \in X$, one has $F(u) + \langle w, v - u \rangle \leq F(v)$, i.e., $w \in \partial F(u)$. Conversely, for $w \in \partial F(u)$, one has $\langle w, u \rangle \geq F(u) + F^*(w)$, which implies equality (by Fenchel's inequality).

In conclusion, the subdifferential consists of pairs $(u, w) \in X \times X^*$ for which Fenchel's inequality is sharp:

$$(u, w) \in \partial F \quad \Leftrightarrow \quad \langle w, u \rangle = F(u) + F^*(w).$$

Let us analyze conjugation as a mapping that maps functionals to functionals and collect some fundamental properties.

Remark 6.62 (Conjugation as a Mapping)

- The conjugation of functionals in X or X^* , respectively, maps onto $\Gamma_0(X^*) \cup \{\infty\}$ or $\Gamma_0(X) \cup \{\infty\}$, respectively. The image is constant ∞ if there is no continuous affine linear functional below F .
- If we form the pointwise supremum over all continuous affine linear maps associated with $(w, s) \in \text{spt}(F)$, we obtain, if $\text{spt}(F) \neq \emptyset$,

$$\sup_{(w,s) \in \text{spt}(F)} \langle w, \cdot \rangle + s = \sup_{w \in X^*} \langle w, \cdot \rangle + \sup_{\substack{s \in \mathbf{R}, \\ (w,s) \in \text{spt}(F)}} s = \sup_{w \in X^*} \langle w, \cdot \rangle - F^*(w) = F^{**},$$

and by definition we have $F^{**} \leq F$. In other words, F^{**} is the largest functional in $\Gamma_0(X)$ that is below F . A similar claim holds for functionals $G : X^* \rightarrow \mathbf{R}_\infty$.

- Every functional in $\Gamma_0(X)$ or $\Gamma_0(X^*)$, respectively, is, as a pointwise supremum of convex and lower semicontinuous functions, again convex and lower semicontinuous (see Lemmas 6.14 and 6.21).

Conversely, by Lemma 6.57 one ha

$$\Gamma_0(X) = \{F : X \rightarrow \mathbf{R}_\infty \mid F \text{ proper, convex and lower semicontinuous}\}$$

and if X is reflexive, the analogous statement for $\Gamma_0(X^*)$ holds, too.

With the exception of the constant ∞ , the sets $\Gamma_0(X)$ and $\Gamma_0(X^*)$ are exactly the images of Fenchel conjugation. On these sets, Fenchel conjugation is even invertible.

Lemma 6.63 *Let X be a real Banach space. The Fenchel conjugation ${}^* : \Gamma_0(X) \rightarrow \Gamma_0(X^*)$ is invertible with inverse ${}^* : \Gamma_0(X^*) \rightarrow \Gamma_0(X)$.*

Proof First note that the conjugation on $\Gamma_0(X)$ and $\Gamma_0(X^*)$, respectively, maps by definition to the respective sets (Remark 6.62). For $F \in \Gamma_0(X)$, there exists $\emptyset \neq K_0 \subset X^* \times \mathbf{R}$ with $F = \sup_{(w,s) \in K_0} \langle w, \cdot \rangle + s$. By definition, K_0 is contained in $\text{spt}(F)$, and thus, by Remark 6.62

$$F = \sup_{(w,s) \in K_0} \langle w, \cdot \rangle + s \leq \sup_{(w,s) \in \text{spt}(F)} \langle w, \cdot \rangle + s = F^{**},$$

which shows that $F = F^{**}$. Hence, conjugation on X^* is a left inverse to conjugation on X . A similar argument for $G \in \Gamma_0(X^*)$ leads to $G = G^{**}$, and hence conjugation on X^* is also a right inverse to conjugation on X . \square

Example 6.64 (Fenchel Conjugates)

1. Closed λ -balls and norm functionals

For $\lambda > 0$ and $F = I_{\{\|u\|_X \leq \lambda\}}$ the conjugate is

$$F^*(w) = \sup_{u \in X} \langle w, u \rangle - I_{\{\|u\|_X \leq \lambda\}}(u) = \sup_{\|u\|_X \leq \lambda} \langle w, u \rangle = \lambda \|w\|_{X^*}.$$

A similar claim is true for the conjugate of $G = I_{\{\|w\|_{X^*} \leq \lambda\}}$. The latter situation occurred in Example 6.56.

More generally, for $F(u) = \varphi(\|u\|_X)$ with a proper and even $\varphi : \mathbf{R} \rightarrow \mathbf{R}_\infty$ (i.e., $\varphi(x) = \varphi(-x)$), one has

$$\begin{aligned} F^*(w) &= \sup_{u \in X} \langle w, u \rangle - \varphi(\|u\|_X) \\ &= \sup_{t \geq 0} \sup_{\|u\|_X = t} \langle w, u \rangle - \varphi(t) \\ &= \sup_{t \geq 0} \|w\|_{X^*} t - \varphi(t) \\ &= \sup_{t \in \mathbf{R}} \|w\|_{X^*} t - \varphi(t) = \varphi^*(\|w\|_{X^*}). \end{aligned}$$

2. Powers and positively homogeneous functionals

The real function $\varphi(t) = \frac{1}{p}|t|^p$ with $1 < p < \infty$ has the conjugate $\psi(s) = \frac{1}{p^*}|s|^{p^*}$ with $\frac{1}{p} + \frac{1}{p^*} = 1$. One the one hand, this leads to Young's inequality for products

$$st \leq \frac{1}{p}|t|^p + \frac{1}{p^*}|s|^{p^*} \quad \Rightarrow \quad \varphi^*(s) \leq \frac{1}{p^*}|s|^{p^*} = \psi(s),$$

and on the other hand, we obtain for $t = \operatorname{sgn}(s)|s|^{\frac{1}{p-1}}$ that

$$st - \frac{1}{p}|t|^p = \left(1 - \frac{1}{p}\right)|s|^{\frac{p}{p-1}} = \frac{1}{p^*}|s|^{p^*}.$$

The special cases $p \in \{1, \infty\}$ follow from the first item:

$$\varphi(t) = |t| \Rightarrow \varphi^*(s) = I_{[-1,1]}(s), \quad \varphi(t) = I_{[-1,1]}(t) \Rightarrow \varphi^*(s) = |s|.$$

In general, a positively p -homogeneous function has a positively p^* -homogeneous functional as conjugate; see Exercise 6.21.

3. Convex cones and subspaces

An important special case is that of conjugates of indicator functionals of closed convex cones, i.e., $F = I_K$ with $K \subset X$ closed, nonempty, and with the property that for all $u^1, u^2 \in K$, also $u^1 + u^2 \in K$, and for all $u \in K, \alpha \geq 0$, also $\alpha u \in K$. For $w \in X^*$, one has

$$\begin{aligned} F^*(w) &= \sup_{u \in K} \langle w, u \rangle = \begin{cases} 0 & \text{if } \langle w, u \rangle \leq 0 \text{ for all } u \in K \\ \infty & \text{otherwise,} \end{cases} \\ &= I_{K^\perp}(w) \end{aligned}$$

with $K^\perp = \{w \in X^* \mid \langle w, u \rangle \leq 0 \text{ for all } u \in K\}$. We check that for two elements w^1, w^2 in K^\perp one has

$$\langle w^1 + w^2, u \rangle = \langle w^1, u \rangle + \langle w^2, u \rangle \leq 0 \quad \text{for all } u \in K,$$

and hence $w^1 + w^2 \in K^\perp$. Similarly we obtain for $w \in K^\perp$ and $\alpha \geq 0$ the inequality $\langle \alpha w, u \rangle = \alpha \langle w, u \rangle \leq 0$ for all $u \in K$. Hence, the set K^\perp is again a closed and convex cone, called the *dual cone*. In this sense, K^\perp is dual to K .

Of course, closed subspaces $U \subset X$ are closed convex cones. We note that

$$\langle w, u \rangle \leq 0 \quad \text{for all } u \in U \quad \Leftrightarrow \quad \langle w, u \rangle = 0 \quad \text{for all } u \in U,$$

and hence the set U^\perp is a closed subspace of X^* , called the *annihilator* of U . In a Hilbert space, the annihilator U^\perp is the orthogonal complement if X and X^* are identified by the Riesz map. Hence, Fenchel duality contains the usual duality of closed spaces as a special case.

There is also a geometric interpretation of Fenchel conjugation. For a fixed “slope” $w \in X^*$, one has, as we already observed,

$$-F^*(w) = \sup \{s \in \mathbf{R} \mid \langle w, \cdot \rangle + s \leq F\}.$$

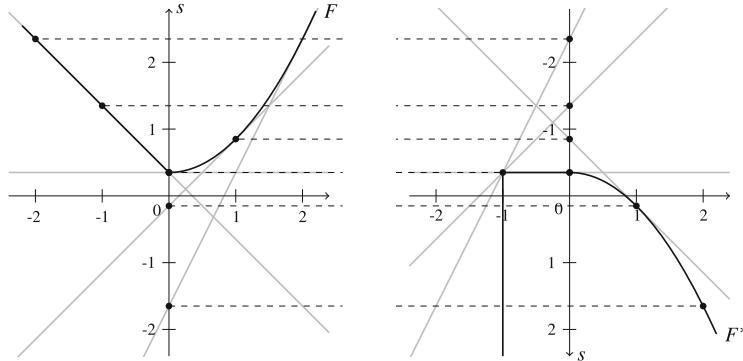


Fig. 6.10 Graphical visualization of Fenchel duality in dimension one. Left: A convex (and continuous) function F . Right: Its conjugate F^* (with inverted s -axis). The plotted lines are maximal affine functionals below the graphs, their intersections with the respective s -axis correspond to the negative values of F^* and F , respectively (suggested by dashed lines). For some slopes (smaller than -1) there are no affine linear functionals below F , and consequently, F^* equals ∞ there

In other words, $m_w = \langle w, \cdot \rangle - F^*(w)$ is, for given $w \in X^*$, the largest affine linear functional below F . We have $m_w(0) = -F^*(w)$, i.e. the intersection of the graph of m_w and the s -axis is the negative value of the dual functional at w . This allows us to construct the conjugate in a graphical way, see Fig. 6.10.

We collect some obvious rules for Fenchel conjugation.

Lemma 6.65 (Calculus for Fenchel Conjugation) *Let $F_1 : X \rightarrow \mathbf{R}_\infty$ be a proper functional on a real Banach space X .*

1. *For $\lambda \in \mathbf{R}$ and $F_2 = F_1 + \lambda$, we have $F_2^* = F_1^* - \lambda$,*
2. *for $\lambda > 0$ and $F_2 = \lambda F_1$, we have $F_2^* = \lambda F_1^* \circ \lambda^{-1} \text{id}$,*
3. *for $w^0 \in X$, $w^0 \in X^*$, and $F_2 = F_1 \circ T_{w^0} + \langle w^0, \cdot \rangle$ we have $F_2^* = (F_1^* - \langle \cdot, w^0 \rangle) \circ T_{-w^0}$,*
4. *for a real Banach space Y and $K \in \mathcal{L}(Y, X)$ continuously invertible, we have for $F_2 = F_1 \circ K$ that $F_2^* = F_1^* \circ (K^{-1})^*$.*

Proof Assertion 1: For $w \in X^*$ we get by definition that

$$F_2^*(w) = \sup_{u \in X} \{ \langle w, u \rangle - F_1(u) - \lambda \} = \sup_{u \in X} \{ \langle w, u \rangle - F_1(u) \} - \lambda = F_1^*(w) - \lambda.$$

Assertion 2: We use that multiplication by positive constants can be interchanged with taking suprema and get

$$F_2^*(w) = \sup_{u \in X} \{ \lambda \langle \lambda^{-1} w, u \rangle - \lambda F_1(u) \} = \lambda \sup_{u \in X} \{ \langle \lambda^{-1} w, u \rangle - F_1(u) \} = \lambda F_1^*(\lambda^{-1} w).$$

Assertion 3: We observe that

$$\begin{aligned}
 \sup_{u \in X} & \left\{ \langle w, u \rangle - F_1(u + u^0) - \langle w^0, u \rangle \right\} \\
 &= \sup_{u \in X} \left\{ \langle w - w^0, u + u^0 \rangle - F_1(u + u^0) \right\} - \langle w - w^0, u^0 \rangle \\
 &= \sup_{\tilde{u}=u+u^0, u \in X} \left\{ \langle w - w^0, \tilde{u} \rangle - F_1(\tilde{u}) \right\} - \langle w - w^0, u^0 \rangle \\
 &= F_1^*(w - w^0) - \langle w - w^0, u^0 \rangle.
 \end{aligned}$$

Assertion 4: Since K is invertible, $\text{rg}(K) = X$, and hence for $\omega \in Y^*$,

$$\begin{aligned}
 F_2^*(\omega) &= \sup_{v \in Y} \left\{ \langle \omega, K^{-1}Kv \rangle - F_1(Kv) \right\} \\
 &= \sup_{\substack{u=Kv, \\ v \in Y}} \left\{ \langle (K^{-1})^*\omega, u \rangle - F_1(u) \right\} = F_1^*((K^{-1})^*\omega). \quad \square
 \end{aligned}$$

In view of Lemmas 6.14 and 6.21, we may ask how conjugation acts for pointwise suprema and sums. Similarly for the calculus of the subdifferential (Theorem 6.51), this question is a little delicate. Let us first look at pointwise suprema. Let $\{F_i\}$, $i \in I \neq \emptyset$ be a family of proper functionals $F_i : X \rightarrow \mathbf{R}_\infty$ with $\bigcap_{i \in I} \text{dom } F_i \neq \emptyset$. For $w \in X^*$ we deduce:

$$\begin{aligned}
 \left(\sup_{i \in I} F_i \right)^*(w) &= \sup_{u \in X} \left\{ \langle w, u \rangle - \sup_{i \in I} F_i(u) \right\} = \sup_{u \in X} \inf_{i \in I} \{ \langle w, u \rangle - F_i(u) \} \\
 &\leq \inf_{i \in I} \sup_{u \in X} \{ \langle w, u \rangle - F_i(u) \} = \inf_{i \in I} F_i^*(w).
 \end{aligned}$$

It is natural to ask whether equality holds, i.e., whether infimum and supremum can be swapped. Unfortunately, this is not true in general: we know that the F_i^* are convex and lower semicontinuous, but these properties are not preserved by pointwise infima, i.e. $\inf_{i \in I} F_i^*$ is in general neither convex nor lower semicontinuous. Hence, this functional is not even a conjugate in general. However, it still contains enough information to extract the desired conjugate:

Theorem 6.66 (Conjugation of Suprema) *Let $I \neq \emptyset$ and $F_i : X \rightarrow \mathbf{R}_\infty$, $i \in I$ and $\sup_{i \in I} F_i$ be proper on a real Banach space X . Then*

$$\left(\sup_{i \in I} F_i \right)^* = \left(\inf_{i \in I} F_i^* \right)^{**}.$$

You should do the *proof* in Exercise 6.23.

Now let us come to the conjugation of sums. Let $F_1, F_2 : X \rightarrow \mathbf{R}_\infty$ be proper functionals with $\text{dom } F_1 \cap \text{dom } F_2 \neq \emptyset$. To calculate the conjugate of $F_1 + F_2$ at

$w \in X^*$ we split $w = w^1 + w^2$ with $w^1, w^2 \in X^*$ and note that

$$\begin{aligned} \sup_{u \in X} \{\langle w, u \rangle - F_1(u) - F_2(u)\} &\leq \sup_{u \in X} \{\langle w^1, u \rangle - F_1(u)\} + \sup_{u \in X} \{\langle w^2, u \rangle - F_2(u)\} \\ &= F_1^*(w^1) + F_2^*(w^2). \end{aligned}$$

Moreover, this holds for all decompositions $w = w^1 + w^2$, and thus

$$(F_1 + F_2)^*(w) \leq \inf_{w=w^1+w^2} F_1^*(w^1) + F_2^*(w^2) = (F_1^* \triangle F_2^*)(w). \quad (6.23)$$

The operation on the right-hand side,

$$\Delta: (F, G) \mapsto F \Delta G, \quad (F \Delta G)(w) = \inf_{w=w^1+w^2} F(w^1) + G(w^2)$$

for $F, G : X \rightarrow \mathbf{R}_\infty$ is called *infimal convolution*. In some cases equality holds in (6.23). A thorough discussion of the needed arguments and the connection with the sum rule for subgradients can be found in Exercise 6.24. We only state the main result here.

Theorem 6.67 (Conjugation of Sums) *Let $F_1, F_2 : X \rightarrow \mathbf{R}_\infty$ be proper, convex, and lower semicontinuous functionals on a reflexive and real Banach space X . Furthermore, let there exist $u^0 \in \text{dom } F_1 \cap \text{dom } F_2$ such that F_1 is continuous at u^0 . Then*

$$(F_1 + F_2)^* = F_1^* \Delta F_2^*.$$

Now we apply Fenchel duality to convex minimization problems. We will treat the following situation, which is general enough for our purposes:

$$\text{Primal problem:} \quad \min_{u \in X} F_1(u) + F_2(Au) \quad (6.24)$$

with $F_1 : X \rightarrow \mathbf{R}_\infty$ proper, convex, and lower semicontinuous on the real Banach space X , $A \in \mathcal{L}(X, Y)$ and $F_2 : Y \rightarrow \mathbf{R}_\infty$ also proper, convex, and lower semicontinuous on the real Banach space Y . We write F_2 as a suitable supremum, and the minimum from above, written as an infimum, becomes

$$\inf_{u \in X} \sup_{w \in Y^*} \langle w, Au \rangle + F_1(u) - F_2^*(w).$$

If we assume that we can swap infimum and supremum and that the supremum is actually assumed, then this turns into

$$\sup_{w \in Y^*} \inf_{u \in X} \langle -A^*w, -u \rangle + F_1(u) - F_2^*(w) = \max_{w \in Y^*} -F_1^*(-A^*w) - F_2^*(w).$$

We have just derived the dual optimization problem, which is

$$\text{Dual problem: } \max_{w \in Y^*} -F_1^*(-A^*w) - F_2^*(w). \quad (6.25)$$

It remains to clarify whether infimum and supremum can be swapped and whether the supremum is realized. In other words, we have to show that the minimum (6.24) equals the maximum in (6.25). The following theorem gives a sufficient criterion for this to hold.

Theorem 6.68 (Fenchel-Rockafellar Duality) *Let $F_1 : X \rightarrow \mathbf{R}_\infty$, $F_2 : Y \rightarrow \mathbf{R}_\infty$ be proper, convex, and lower semicontinuous on the real Banach spaces X and Y , respectively. Further, let $A : X \rightarrow Y$ be linear and continuous, and suppose that the minimization problem*

$$\min_{u \in X} F_1(u) + F_2(Au)$$

has a solution $u^ \in X$. If there exists some $u^0 \in X$ such that $F_1(u^0) < \infty$, $F_2(Au^0) < \infty$ and F_2 is continuous at Au^0 , then*

$$\max_{w \in Y^*} -F_1^*(-A^*w) - F_2^*(w) = \min_{u \in X} F_1(u) + F_2(Au).$$

In particular, the maximum is realized at some $w^ \in Y$.*

Proof Since u^* is a solution of the minimization problem, we immediately get $0 \in \partial(F_1 + F_2 \circ A)(u^*)$. Now we want to use the subdifferential calculus (Theorem 6.51) and note that $F_2 \circ A$ is continuous at u^0 by assumption. Hence,

$$\partial(F_1 + F_2 \circ A) = \partial F_1 + \partial(F_2 \circ A) = \partial F_1 + A^* \circ \partial F_2 \circ A.$$

This means that there exists $w^* \in Y^*$ such that $-A^*w^* \in \partial F_1(u^*)$ and $w^* \in \partial F_2(Au^*)$. Now we reformulate the subgradient inequality for $-A^*w^* \in \partial F_1(u^*)$:

$$\begin{aligned} & F_1(u^*) + \langle -A^*w^*, v - u^* \rangle \leq F_1(v) && \forall v \in X \\ \Leftrightarrow & F_1(u^*) - \langle -A^*w^*, u^* \rangle \leq F_1(v) - \langle -A^*w^*, v \rangle && \forall v \in X \\ \Leftrightarrow & \langle -A^*w^*, u^* \rangle - F_1(u^*) \geq \sup_{v \in X} \langle -A^*w^*, v \rangle - F_1(v) \\ & = F_1^*(-A^*w^*). \end{aligned}$$

Similarly we obtain $\langle w^*, Au^* \rangle - F_2(Au^*) \geq F_2^*(w^*)$. Adding these inequalities, we get

$$\begin{aligned} F_1^*(-A^*w^*) + F_2^*(w^*) & \leq \langle -A^*w^*, u^* \rangle - F_1(u^*) + \langle w^*, Au^* \rangle - F_2(Au^*) \\ & = -F_1(u^*) - F_2(Au^*) \end{aligned}$$

and hence

$$F_1(u^*) + F_2(Au^*) \leq -F_1^*(-A^*w^*) - F_2^*(w^*).$$

On the other hand, we have by Remark 6.62 (see also Exercise 6.19) that

$$\begin{aligned} F_1(u^*) + F_2(Au^*) &= \inf_{u \in X} \sup_{w \in Y^*} F_1(u) + \langle w, Au \rangle - F_2^*(w) \\ &\geq \sup_{w \in Y^*} \inf_{u \in X} F_1(u) - \langle -A^*w, u \rangle - F_2(w) \\ &= \sup_{w \in Y^*} -F_1^*(-A^*w) - F_2^*(w). \end{aligned}$$

We conclude that

$$\begin{aligned} \sup_{w \in Y^*} -F_1^*(-A^*w) - F_2^*(w) &\leq \inf_{u \in X} F_1(u) + F_2(Au) \\ &\leq -F_1^*(-A^*w^*) - F_2^*(w^*) \leq \sup_{w \in Y^*} -F_1^*(-A^*w) - F_2^*(w), \end{aligned}$$

which is the desired equality for the supremum. We also see that it is assumed at w^* . \square

Remark 6.69 The assumption that there exists some $u^0 \in X$ such that $F_1(u^0) < \infty$, $F_2(Au^0) < \infty$ and that F_2 is continuous at Au^0 is used only to apply the sum rule and the chain rule for subdifferentials. Hence, we can replace it with the assumption that $\partial(F_1 + \partial(F_2 \circ A)) = \partial F_1 + A^* \circ \partial F_2 \circ A$. See Exercises 6.11–6.15 for more general sufficient conditions for this to hold.

The previous proof hinges on the applicability of rules for subdifferential calculus and hence fundamentally relies on the separation of suitable convex sets. A closer inspection of the proof reveals the following primal-dual optimality system:

Corollary 6.70 (Fenchel-Rockafellar Optimality System) *If for proper, convex, and lower semicontinuous functionals $F_1 : X \rightarrow \mathbf{R}_\infty$ and $F_2 : Y \rightarrow \mathbf{R}_\infty$ it is the case that*

$$\max_{w \in Y^*} -F_1^*(-A^*w) - F_2^*(w) = \min_{u \in X} F_1(u) + F_2(Au), \quad (6.26)$$

then a pair $(u^, w^*) \in X \times Y^*$ is a solution of the primal-dual problem if and only if*

$$-A^*w^* \in \partial F_1(u^*), \quad w^* \in \partial F_2(Au^*). \quad (6.27)$$

Proof It is clear that $(u^*, w^*) \in X \times Y^*$ is an optimal pair if and only if

$$-F_1^*(-A^*w^*) - F_2^*(w^*) = F_1(u^*) + F_2(Au^*).$$

This is equivalent to

$$\langle -A^*w^*, u^* \rangle + \langle w^*, Au^* \rangle = F_1(u^*) + F_1^*(-A^*w^*) + F_2(Au^*) + F_2^*(w^*),$$

and since by Fenchel's inequality (6.22) one has $\langle -A^*w^*, u^* \rangle \leq F_1(u^*) + F_1^*(-A^*w^*)$ and $\langle w^*, Au^* \rangle \leq F_2(Au^*) + F_2^*(w^*)$, this is equivalent to

$$\langle -A^*w^*, u^* \rangle = F_1(u^*) + F_1^*(-A^*w^*) \quad \text{and} \quad \langle w^*, Au^* \rangle = F_2(Au^*) + F_2^*(w^*).$$

These equalities characterize the inclusions $-A^*w^* \in \partial F_1(u^*)$ and $w^* \in \partial F_2(Au^*)$; see Remark 6.61. \square

Example 6.71 We analyze the special case of Tikhonov functionals from Example 6.32. Let X be a reflexive Banach space, Y a Hilbert space, and $A \in \mathcal{L}(X, Y)$ a given forward operator. We define

$$F_1(u) = \lambda \|u\|_X, \quad F_2(v) = \frac{1}{2} \|v - u^0\|_Y^2,$$

and obtain as primal problem (6.24)

$$\min_{u \in X} \frac{\|Au - u^0\|_Y^2}{2} + \lambda \|u\|_X,$$

the minimization of the Tikhonov functional. In Example 6.32 we showed the existence of a minimizer. Moreover, F_2 is continuous everywhere, and we can apply Theorem 6.68. To that end, we use Example 6.64 and Lemma 6.65 to get

$$F_1^*(\omega) = \begin{cases} 0 & \text{if } \|\omega\|_{X^*} \leq \lambda, \\ \infty & \text{otherwise,} \end{cases} \quad F_2^*(w) = \frac{\|w\|_Y^2}{2} + (w, u^0) = \frac{\|w + u^0\|_Y^2}{2} - \frac{\|u^0\|_Y^2}{2}.$$

We identify $Y = Y^*$ and also $A^* : Y \rightarrow X^*$. The respective dual problem is a constrained maximization problem, namely

$$\max_{\| -A^*w \|_{X^*} \leq \lambda} -\frac{\|w + u^0\|_Y^2}{2} + \frac{\|u^0\|_Y^2}{2}.$$

Substituting $\bar{w} = -w$, flipping the sign, and dropping terms independent of \bar{w} , we see that the problem is equivalent to a projection problem in a Hilbert space:

$$\bar{w}^* = \arg \min_{\|A^*\bar{w}\|_{X^*} \leq \lambda} \frac{\|u^0 - \bar{w}\|_Y^2}{2} \quad \Leftrightarrow \quad \bar{w}^* = P_{\{\|A^*\bar{w}\|_{X^*} \leq \lambda\}}(u^0).$$

The optimal $w^* = -\bar{w}^*$ and every solution u^* of the primal problem satisfy (6.27), and in particular we have

$$w^* \in \partial F_2(Au^*) \Leftrightarrow w^* = Au^* - u^0 \Leftrightarrow Au^* = u^0 - \bar{w}^*,$$

and we see that $u^0 - \bar{w}^*$ lies in the image of A even if this image is not closed. If A is injective, we can apply its inverse (which is not necessarily continuous) and obtain

$$u^* = A^{-1}(u^0 - P_{\{\|A^*\bar{w}\|_{X^*} \leq \lambda\}}(u^0)).$$

This is a formula for the solution of the minimization problem (however, often of limited use in practice), and also we have deduced that the result from Example 6.56 in the case $A = I$ was correct.

At the end of this section we give a more geometric interpretation of the solution of the primal-dual problem.

Remark 6.72 (Primal-Dual Solutions and Saddle Points) We can interpret the simultaneous solution of the primal and dual problem as follows. We define the *Lagrange functional* $L : \text{dom } F_1 \times \text{dom } F_2^* \rightarrow \mathbf{R}$ by

$$L(u, w) = \langle w, Au \rangle + F_1(u) - F_2^*(w),$$

and observe that every optimal pair $(u^*, w^*) \in X \times Y^*$, in the situation of (6.26), has to satisfy the inequalities

$$\begin{aligned} L(u^*, w^*) &\leq \sup_{w \in Y^*} L(u^*, w) = \min_{u \in X} F_1(u) + F_2(Au) \\ &= \max_{w \in Y^*} -F_1^*(-A^*w) - F_2^*(w) = \inf_{u \in X} L(u, w^*) \leq L(u^*, w^*) \end{aligned}$$

and hence is a solution of the *saddle point problem*

$$L(u^*, w) \leq L(u^*, w^*) \leq L(u, w^*) \quad \text{for all } (u, w) \in \text{dom } F_1 \times \text{dom } F_2^*.$$

We say that a solution (u^*, w^*) is a *saddle point* of L . Conversely, every saddle point (u^*, w^*) satisfies

$$\begin{aligned} L(u^*, w^*) &= \sup_{w \in Y^*} L(u^*, w) = F_1(u^*) + F_2(Au^*) \\ &= \inf_{u \in X} L(u, w^*) = -F_1^*(-A^*w^*) - F_2^*(w^*) \end{aligned}$$

which means that (u^*, w^*) is a solution of the primal-dual problem. Hence, the saddle points of L are exactly the primal-dual solutions.

This fact will be useful in deriving so-called *primal-dual algorithms* for the numerical solution of the primal problem (6.24). The basic idea of these methods is to find minimizers of the Lagrange functional in the primal direction and respective maximizers in the dual direction. In doing so, one can leverage the fact that L has a simpler structure than the primal and dual problems. More details on that can be found in Sect. 6.4.

6.3 Minimization in Sobolev Spaces and BV

Now we pursue the goal to apply the theory we developed in the previous sections to convex variational problems in imaging. As described in the introduction and in Examples 6.1–6.4, it is important to choose a good model for images; in our terminology, to choose the penalty Ψ appropriately. Starting from the space $H^1(\Omega)$ we will consider Sobolev spaces with more general exponents. However, we will see that these spaces are not satisfactory for many imaging tasks, since they do not allow a proper treatment of discontinuities, i.e. jumps, in the gray values as they occur on object borders. This matter is notably different in the space of *functions with bounded total variation*, and consequently, this space plays an important role in mathematical imaging. We will develop the basic theory for this Banach space in the context of convex analysis and apply it to some concrete problems.

6.3.1 Functionals with Sobolev Penalty

We begin with the analysis of problems with functions of the Sobolev seminorm in $H^{m,p}(\Omega)$. To that end, we begin with some important notions and results from the theory of Sobolev spaces.

Lemma 6.73 *Let $\Omega \subset \mathbf{R}^d$ be a domain, $\alpha \in \mathbf{N}^d$ a multiindex, and $1 \leq p, q < \infty$. Then the weak partial derivative $\frac{\partial^\alpha}{\partial x^\alpha}$*

$$\frac{\partial^\alpha}{\partial x^\alpha} : \text{dom } \frac{\partial^\alpha}{\partial x^\alpha} = \left\{ u \in L^p(\Omega) \mid \frac{\partial^\alpha u}{\partial x^\alpha} \in L^q(\Omega) \right\} \rightarrow L^q(\Omega),$$

defines a densely defined and closed linear mapping $L^p(\Omega) \rightarrow L^q(\Omega)$.

Moreover, it is weak-to-weak closed.

Proof First we prove the weak-to-weak closedness, which by linearity implies (strong) closedness. Let (u^n) be a sequence in $L^p(\Omega)$ with $u^n \rightharpoonup u$ for some $u \in L^p(\Omega)$ and $\frac{\partial^\alpha}{\partial x^\alpha} u^n \rightharpoonup v$ for some $v \in L^q(\Omega)$. Choose some test function $\varphi \in \mathcal{D}(\Omega)$. These and the derivative $\frac{\partial^\alpha \varphi}{\partial x^\alpha}$ are guaranteed to lie in the respective dual

spaces $L^{p^*}(\Omega)$ and $L^{q^*}(\Omega)$, and hence, by definition of the weak derivative,

$$\int_{\Omega} u \frac{\partial^{\alpha} \varphi}{\partial x^{\alpha}} dx = \lim_{n \rightarrow \infty} \int_{\Omega} u^n \frac{\partial^{\alpha} \varphi}{\partial x^{\alpha}} dx = \lim_{n \rightarrow \infty} (-1)^{|\alpha|} \int_{\Omega} \frac{\partial^{\alpha} u^n}{\partial x^{\alpha}} \varphi dx = (-1)^{|\alpha|} \int_{\Omega} v \varphi dx.$$

Since this holds for every test function, we see that $v = \frac{\partial^{\alpha}}{\partial x^{\alpha}} u$ as desired.

We show that the mapping $\frac{\partial^{\alpha}}{\partial x^{\alpha}}$ is densely defined: every $u \in \mathcal{D}(\Omega)$ also lies in $L^p(\Omega)$ and has a continuous (strong) derivative of order α and in particular, $\frac{\partial^{\alpha}}{\partial x^{\alpha}} u \in L^q(\Omega)$. This shows that $\mathcal{D}(\Omega) \subset \text{dom } \frac{\partial^{\alpha}}{\partial x^{\alpha}}$, and using Lemma 3.16 we obtain the claim. \square

In Sect. 3.3 on linear filters we have seen that smooth functions are dense in Sobolev spaces with $\Omega = \mathbf{R}^d$. The density of $\mathcal{C}^\infty(\overline{\Omega})$ for bounded Ω , however, needs some regularity of the boundary of Ω .

Theorem 6.74 *Let $\Omega \subset \mathbf{R}^d$ be a bounded Lipschitz domain, $m \geq 1$, and $1 \leq p < \infty$. Then there exists a sequence of linear operators (\mathcal{M}_n) in $\mathcal{L}(H^{m,p}(\Omega), H^{m,p}(\Omega))$ such that $\text{rg}(\mathcal{M}_n) \subset \mathcal{C}^\infty(\overline{\Omega})$ for all $n \geq 1$ and the property that for every $u \in H^{m,p}(\Omega)$,*

$$\lim_{n \rightarrow \infty} \mathcal{M}_n u = u \text{ in } H^{m,p}(\Omega).$$

The mappings can be chosen independently of m and p .

As a consequence, $\mathcal{C}^\infty(\overline{\Omega})$ is dense in $H^{m,p}(\Omega)$.

Proof A bounded Lipschitz domain satisfies the so-called *segment condition* (see [2]), i.e., for every $x \in \partial\Omega$ there exist an open neighborhood U_x and some $\eta_x \in \mathbf{R}^d$, $\eta_x \neq 0$, such that for every $y \in \overline{\Omega} \cap U_x$ and $t \in]0, 1[$, one has $y + t\eta_x \in \Omega$.

Now cover $\partial\Omega$ with finitely many such U_x and name them U_1, \dots, U_K (these exist due to the compactness of $\partial\Omega$). Moreover, cover the rest of Ω with another open set U_0 , i.e., $\overline{\Omega} \setminus \bigcup_{k=1}^K U_k \subset\subset U_0$ with $\overline{U_0} \subset\subset \Omega$. The U_0, \dots, U_K cover $\overline{\Omega}$, and for $k = 0, \dots, K$ one can choose open V_k with $\overline{V_k} \subset U_k$, such that the V_0, \dots, V_K still cover $\overline{\Omega}$. We construct a partition of unity that is subordinated to this cover, i.e., $\varphi_k \in \mathcal{D}(\mathbf{R}^d)$, $\text{supp } \varphi_k \subset\subset V_k$, and $\sum_{k=1}^K \varphi_k(x) = 1$ for $x \in \overline{\Omega}$, and is an element of $[0, 1]$ otherwise.

We note that the product $\varphi_k u$ is contained in $H^{m,p}(\Omega)$ and we have

$$\partial^{\alpha}(\varphi_k u) = \sum_{\beta \leq \alpha} \binom{\alpha}{\beta} (\partial^{\alpha} \varphi_k)(\partial^{\beta-\alpha} u) \quad (6.28)$$

for every multiindex $|\alpha| \leq m$. The following argument for $m = 1$ can be easily extended by induction for the full proof. For $i = 1, \dots, d$ and $v \in \mathcal{D}(\Omega)$, one has $\varphi_k \frac{\partial v}{\partial x_i} \in \mathcal{D}(\Omega)$ and

$$\varphi_k \frac{\partial v}{\partial x_i} = \frac{\partial}{\partial x_i}(\varphi_k v) - \frac{\partial \varphi_k}{\partial x_i} v.$$

Hence,

$$\int_{\Omega} u \varphi_k \frac{\partial v}{\partial x_i} dx = \int_{\Omega} u \frac{\partial}{\partial x_i} (\varphi_k v) - u \frac{\partial \varphi_k}{\partial x_i} v dx = - \int_{\Omega} \left(\frac{\partial u}{\partial x_i} \varphi_k + u \frac{\partial \varphi_k}{\partial x_i} \right) v dx,$$

and this means that the weak derivative $\frac{\partial}{\partial x_i} (\varphi_k v) = \frac{\partial u}{\partial x_i} \varphi_k + u \frac{\partial \varphi_k}{\partial x_i}$ is in $L^p(\Omega)$. The representation (6.28) implies that

$$\|\partial^\alpha (\varphi_k u)\|_p \leq \sum_{\beta \leq \alpha} \binom{\alpha}{\beta} \|\partial^\alpha \varphi_k\|_\infty \|\partial^{\beta-\alpha} u\|_p \leq C \|u\|_{m,p},$$

and hence $u \mapsto \varphi_k u$ is a linear and continuous map from $H^{m,p}(\Omega)$ to itself.

In the following we write $u_k = \varphi_k u$ and set $\eta_0 = 0$. Our aim is to translate u_k in the direction of η_k from the segment condition and to convolve the translated function such that u_k is evaluated only on a compact subset of Ω . To that end, choose (t_n) in $]0, 1]$ with $t_n \rightarrow 0$ such that $\overline{V_k} - t_n \eta_k \subset\subset U_k$ holds for all k and n . Let $\psi \in \mathcal{D}(\mathbf{R}^d)$ be a mollifier and choose, for every n , an $\varepsilon_n > 0$ such that the support of the scaled function $\psi^n = \psi_{\varepsilon_n}$ satisfies

$$(\overline{\Omega - \text{supp } \psi^n} + t_n \eta_k) \cap \overline{V_k} \subset\subset \Omega$$

for all $k = 0, \dots, K$. This is possible, since by the segment condition and the choice of t_n , we have

$$(\overline{\Omega} + t_n \eta_k) \cap \overline{V_k} = (\overline{\Omega} \cap (\overline{V_k} - t_n \eta_k)) + t_n \eta_k \subset\subset \Omega.$$

The operation consisting of smooth cutoff, translation, and convolution is then expressed by $u \mapsto (T_{t_n \eta_k} u_k) * \psi^n$ and reads

$$(T_{t_n \eta_k} u_k) * \psi^n(x) = \int_{\mathbf{R}^d} \psi^n(y + t_n \eta_k) u_k(x - y) dy. \quad (6.29)$$

For $x \in \Omega$ we need to evaluate u only at the points $x - y \in V_k$ with $y + t_n \eta_k \in \text{supp } \psi^n$; these points are always contained in $(\overline{\Omega - \text{supp } \psi^n} + t_n \eta_k) \cap \overline{V_k}$ and hence are compactly contained in Ω . Now, Theorems 3.15 and 3.13 show that

$$\|\partial^\alpha ((T_{t_n \eta_k} u_k) * \psi^n)\|_p \leq \|T_{t_n \eta_k} \psi^n\|_1 \|\partial^\alpha u_k\|_p \leq C \|\partial^\alpha u\|_p$$

for all multiindices with $|\alpha| \leq m$. We define \mathcal{M}_n by

$$\mathcal{M}_n u = \sum_{k=0}^K (T_{t_n \eta_k} (\varphi_k u)) * \psi^n, \quad (6.30)$$

which is, by the above consideration, in $\mathcal{L}(H^{m,p}(\Omega), H^{m,p}(\Omega))$. Since ψ is a mollifier, we also have $\mathcal{M}_n u \in \mathcal{C}^\infty(\overline{\Omega})$ for all n and $u \in H^{m,p}(\Omega)$. Note that the construction of M_n is indeed independent of m and p . It remains to show that $\|\mathcal{M}_n u - u\|_{m,p} \rightarrow 0$ for $n \rightarrow \infty$.

To that end, let $\varepsilon > 0$. Since $\varphi_0, \dots, \varphi_K$ is a partition of unity, it follows that

$$\|\mathcal{M}_n u - u\|_{m,p} \leq \sum_{k=0}^K \| (T_{t_n \eta_k} u_k) * \psi^n - u_k \|_{m,p}.$$

For fixed k one has $\partial^\alpha (T_{t_n \eta_k} u_k - u_k) = T_{t_n \eta_k} \partial^\alpha u_k - \partial^\alpha u_k$, and by continuity of translation in $L^p(\Omega)$, we obtain, for n large enough,

$$\|\partial^\alpha (T_{t_n \eta_k} u_k - u_k)\|_p < \frac{\varepsilon}{2M(K+1)}$$

for all multiindices with $|\alpha| \leq m$, and we denote the number of these multiindices by M . With $v_{k,n} = T_{t_n \eta_k} u_k$ we get, by Theorem 3.15, the property that translation and the weak derivative commute, and by Lemma 3.16 that for n large enough,

$$\begin{aligned} \|\partial^\alpha (v_{k,n} - v_{k,n} * \psi^n)\|_p &= \|\partial^\alpha v_{k,n} - (\partial^\alpha v_{k,n}) * \psi^n\|_p \\ &\leq \|\partial^\alpha (u_k - u_k * \psi^n)\|_p < \frac{\varepsilon}{2(K+1)M} \end{aligned}$$

for all multiindices up to order m . Altogether, this gives

$$\|(T_{t_n \eta_k} u_k) * \psi^n - u_k\|_{m,p} \leq \|v_{k,n} * \psi^n - v_{k,n}\|_{m,p} + \|v_{k,n} - u_k\|_{m,p} < \frac{\varepsilon}{K+1},$$

and by summation over k , we finally get $\|\mathcal{M}_n u - u\|_{m,p} < \varepsilon$.

The density is a direct consequence of the above. \square

A direct application of the density result is the chain rule for the weak derivative in the respective spaces.

Lemma 6.75 (Chain Rule for Sobolev Functions) *Let $\varphi : \mathbf{R} \rightarrow \mathbf{R}$ be continuously differentiable with $\|\varphi'\|_\infty \leq L$ for some $L > 0$. Moreover, let Ω be a bounded Lipschitz domain and $1 \leq p < \infty$. Then $u \mapsto \varphi \circ u$ maps the space $H^{1,p}(\Omega)$ into itself, and*

$$\nabla(\varphi \circ u) = \varphi'(u) \nabla u \quad \text{in } L^p(\Omega).$$

The claim is still true for the functions $\varphi(t) = \min(a, t)$ and $\varphi(t) = \max(a, t)$ with some $a \in \mathbf{R}$ (by abuse of notation we set $\varphi'(a) = 0$ here).

Proof For $u \in \mathcal{C}^\infty(\overline{\Omega})$, the result follows from the usual chain rule. For general $u \in H^{1,p}(\Omega)$ we choose a sequence (u^n) in $\mathcal{C}^\infty(\overline{\Omega})$ that converges to u in the

Sobolev norm. By Lipschitz continuity of φ we obtain

$$\begin{aligned}\|\varphi \circ u^n - \varphi \circ u\|_p^p &= \int_{\Omega} |\varphi(u^n(x)) - \varphi(u(x))|^p dx \\ &\leq L^p \int_{\Omega} |u^n(x) - u(x)|^p dx = L^p \|u^n - u\|_p^p\end{aligned}$$

and hence $\varphi \circ u^n \rightarrow \varphi \circ u$ in $L^p(\Omega)$.

By the theorem of Fischer-Riesz (Theorem 2.48) we also get a subsequence (still indexed by n) that converges pointwise almost everywhere to u , and hence $\lim_{n \rightarrow \infty} \varphi'(u^n(x)) = \varphi'(u(x))$ almost everywhere. For $w \in L^{p^*}(\Omega, \mathbf{R}^d)$ we get by Lebesgue's dominated convergence theorem (Theorem 2.47) $\lim_{n \rightarrow \infty} w\varphi'(u^n) = w\varphi'(u)$ in $L^{p^*}(\Omega, \mathbf{R}^d)$, and by continuity of the dual pairing,

$$\lim_{n \rightarrow \infty} \int_{\Omega} w \cdot \varphi'(u^n) \nabla u^n dx = \lim_{n \rightarrow \infty} \int_{\Omega} w \varphi'(u^n) \cdot \nabla u^n dx = \int_{\Omega} w \varphi'(u) \cdot \nabla u dx.$$

Hence, $\varphi'(u^n) \nabla u^n \rightharpoonup \varphi'(u) \nabla u$ in $L^p(\Omega, \mathbf{R}^d)$, and the claim follows from the strong-to-weak closedness of the weak derivative (Lemma 6.73).

Now let $\varphi(t) = \min(a, t)$ and choose, for $\varepsilon > 0$,

$$\varphi_{\varepsilon}(t) = \begin{cases} \sqrt{(t-a)^2 + \varepsilon^2} - \varepsilon + a & \text{if } t > a, \\ a & \text{otherwise,} \end{cases}$$

such that $\nabla(\varphi_{\varepsilon} \circ u) = \varphi'_{\varepsilon}(u) \nabla u$ holds with

$$\varphi'_{\varepsilon}(t) = \begin{cases} \frac{t-a}{\sqrt{(t-a)^2 + \varepsilon^2}} & \text{if } t > a, \\ 0 & \text{otherwise.} \end{cases}$$

We have pointwise convergence $\varphi_{\varepsilon} \rightarrow \varphi$ and hence $\varphi_{\varepsilon} \circ u \rightarrow \varphi \circ u$ almost everywhere. Moreover, we have $a \leq \varphi_{\varepsilon}(t) \leq \min(a, t)$, and by Lebesgue's dominated convergence theorem, $\varphi_{\varepsilon} \circ u \rightarrow \varphi \circ u$ in $L^p(\Omega)$. Similarly we have $\varphi'_{\varepsilon} \rightarrow \chi_{[a, \infty[} = \varphi'$ pointwise and by $\varphi'_{\varepsilon}(t) \in [0, 1]$ (independent of ε) also $\varphi'_{\varepsilon}(u) \nabla u \rightarrow \varphi'(u) \nabla u$ in $L^p(\Omega)$. Again, the claim follows by the closedness of the weak derivative.

The claim for $\varphi(t) = \max(a, t)$ is similar. \square

For the treatment of Sobolev functions it is extremely helpful to know about continuous embeddings into spaces with higher order of integrability but smaller order of differentiability.

Theorem 6.76 (Embedding Between Sobolev Spaces) *Let $\Omega \subset \mathbf{R}^d$ be a bounded Lipschitz domain, $j, m \in \mathbf{N}$, and $p \in [1, \infty]$. Then:*

1. *For $mp < d$, $H^{m+j,p}(\Omega) \hookrightarrow H^{j,q}(\Omega)$ for all $1 \leq q \leq pd/(d-mp)$.*
2. *For $mp = d$, $H^{m+j,p}(\Omega) \hookrightarrow H^{j,q}(\Omega)$ for all $1 \leq q < \infty$.*
3. *For $mp > d$, $H^{m+j,p}(\Omega) \hookrightarrow \mathcal{C}^j(\overline{\Omega})$.*

All embeddings are compact except the upper endpoint case in statement 1 (i.e., $mp < d$ and $q = pd/(d-mp)$).

The above result can be deduced, for example, from Theorems 4.12 and 6.3 in [2]. Note that elements in Sobolev spaces are equivalence classes, and hence the embedding into $\mathcal{C}^j(\overline{\Omega})$ has to be understood in the sense of representatives. Here is another spatial case that can also be found in [2].

Remark 6.77 For $p = d = 1$ and $m, j \in \mathbf{N}$, it is even the case that $H^{m+j,p}(\Omega) \hookrightarrow H^{j,q}(\Omega)$ for all $1 \leq q \leq \infty$.

Now we focus on the Sobolev penalty

$$\Psi(u) = \varphi(\|\nabla^m u\|_p), \quad \|\nabla^m u\|_p^p = \int_{\Omega} \left(\sum_{|\alpha|=m} \binom{m}{\alpha} \left| \frac{\partial^m u}{\partial x^\alpha}(x) \right|^2 \right)^{p/2} dx, \quad (6.31)$$

where, of course, the derivatives are weak derivatives. In the following we also understand $\nabla^m u$ as an \mathbf{R}^{d^m} -valued mapping, i.e., $\nabla^m u(x)$ is a $d \times d \times \cdots \times d$ tuple with

$$(\nabla^m u)_{i_1, i_2, \dots, i_m} = \frac{\partial}{\partial x_{i_1}} \frac{\partial}{\partial x_{i_2}} \cdots \frac{\partial}{\partial x_{i_m}} u.$$

By the symmetry of higher derivatives, one has for every permutation $\pi : \{1, \dots, m\} \rightarrow \{1, \dots, m\}$ that

$$(\nabla^m u)_{i_{\pi(1)}, \dots, i_{\pi(m)}} = (\nabla^m u)_{i_1, \dots, i_m}.$$

To account for this symmetry we denote the coefficients of a symmetric $\xi \in \mathbf{R}^{d^m}$ also by (ξ_α) , and in particular, we can express the Euclidean norm for symmetric $\xi \in \mathbf{R}^{d^m}$ by

$$|\xi| = \left(\sum_{i_1, \dots, i_m=1}^d |\xi_{i_1, \dots, i_m}|^2 \right)^{1/2} = \left(\sum_{|\alpha|=m} \binom{m}{\alpha} |\xi_\alpha|^2 \right)^{1/2}. \quad (6.32)$$

Remark 6.78 The Sobolev seminorm in (6.31) is slightly different from the usual definition $\sum_{|\alpha|=m} \|\partial^\alpha u\|_p$, but both are equivalent. We choose the form (6.31) to ensure that the norm is both translation-invariant and also *rotation*-invariant; see Exercise 6.25.

To guarantee the existence of a solution of minimization problems with Ψ as penalty, it is helpful to analyze the coercivity of the Sobolev seminorm (since the discrepancy function is not expected to be coercive for ill-posed problems; see again Exercise 6.6). While coercivity is obvious for norms, the same question for seminorms is a bit more subtle; for Tikhonov functionals we recall Exercise 6.7. We consider a slightly more general situation, but make use of the fact that $\|\nabla^m \cdot\|_p$ is an *admissible seminorm* on $H^{m,p}(\Omega)$, i.e., that there exist a linear and continuous mapping $P_m : H^{m,p}(\Omega) \rightarrow H^{m,p}(\Omega)$ and constants $0 < c \leq C < \infty$ such that

$$c \|P_m u\|_{m,p} \leq \|\nabla^m u\|_p \leq C \|P_m u\|_{m,p} \quad \forall u \in H^{m,p}(\Omega). \quad (6.33)$$

To construct these mappings P_m we start with the calculation of the kernel of the seminorms $\|\nabla^m \cdot\|_p$.

Lemma 6.79 *Let Ω be a bounded domain, $m \geq 1$, and $1 \leq p, q < \infty$. Then for $u \in L^q(\Omega)$ and $\nabla^m u \in L^p(\Omega, \mathbf{R}^{d^m})$, one has that $\|\nabla^m u\|_p = 0$ if and only if*

$$u \in \Pi^m(\Omega) = \{u : \Omega \rightarrow \mathbf{R} \mid u \text{ is a polynomial of degree } < m\}.$$

Proof Let u be a polynomial of degree less than m ; then $\nabla^m u = 0$ and consequently $\|\nabla^m u\|_p = 0$.

To prove the reverse implication, we begin with two remarks: On the one hand, the implication is clear for m -times continuously differentiable u . On the other hand it is enough to prove that $u \in \Pi^m(\Omega')$ for every open subset Ω' , $\overline{\Omega'} \subset\subset \Omega$. Now we choose for $u \in L^q(\Omega)$ with $\nabla^m u = 0$ (in the weak sense) such an Ω' , a mollifier η in $D(B_1(0))$, and $\varepsilon_0 > 0$ small enough that $\overline{\Omega' + B_\varepsilon(0)} \subset\subset \Omega$ for all $0 < \varepsilon < \varepsilon_0$. The smoothed $u^\varepsilon = u * \eta^\varepsilon$, $\eta^\varepsilon(x) = \varepsilon^{-d} \eta(x/\varepsilon)$ then satisfy $\nabla^m u^\varepsilon = \nabla^m u * \eta^\varepsilon = 0$, at least in Ω' (see Theorem 3.15). But u^ε is m -times continuously differentiable there, and hence we have $u^\varepsilon \in \Pi^m(\Omega')$. Moreover, $u^\varepsilon \rightarrow u$ in $L^q(\Omega')$, and since $\Pi^m(\Omega')$ is finite dimensional, in particular closed, we obtain $u \in \Pi^m(\Omega')$ as desired. \square

A mapping P_m suitable for (6.33) has to send polynomials of degree less than m to zero. We construct such a mapping and also formalize the space of polynomials of a fixed degree that we already used in Lemma 6.79.

Definition 6.80 (Projection onto the Space of Polynomials) Let $\Omega \subset \mathbf{R}^d$ be a bounded domain and $m \in \mathbf{N}$ with $m \geq 1$. The space of *polynomials of degree up to $m - 1$* is

$$\Pi^m(\Omega) = \text{span}(\{x^\alpha : \Omega \rightarrow \mathbf{R} \mid \alpha \in \mathbf{N}^d, |\alpha| < m\}).$$

Moreover, let $q \in [1, \infty]$. The mapping $Q_m : L^q(\Omega) \rightarrow L^q(\Omega)$, defined by

$$Q_m u = v \quad \Leftrightarrow \quad v \in \Pi^m \quad \text{and} \quad \int_{\Omega} v(x) x^\alpha \, dx = \int_{\Omega} u(x) x^\alpha \, dx \quad \forall |\alpha| < m$$

is called the *projection onto Π^m* ; the mapping $P_m : L^q(\Omega) \rightarrow L^q(\Omega)$ defined by $P_m = \text{id} - Q_m$ is the *projection onto the complement of Π^m* .

It is clear that the set of monomials $\{x \mapsto x^\alpha \mid |\alpha| < m\}$ is a basis for $\Pi^m(\Omega)$. Hence, the projection is well defined.

One easily sees that $Q_m^2 = Q_m$. The map $P_m = \text{id} - Q_m$ is also a projection with $\ker(P_m) = \Pi^m$ and should be a map suitable for (6.33) to hold. The upper estimate is clear: since $Q_m u \in \Pi^m$ holds and the seminorm can be estimated by the Sobolev norm, it follows that

$$\|\nabla^m u\|_p = \|\nabla^m(u - Q_m u)\|_p \leq C \|P_m u\|_{m,p}$$

for all $u \in H^{m,p}(\Omega)$. The following lemma establishes the other inequality:

Lemma 6.81 *Let $\Omega \subset \mathbf{R}^d$ be a bounded Lipschitz domain, $1 < p < \infty$, $m \geq 1$. There exists a constant $C > 0$ such that for all $u \in H^{m,p}(\Omega)$ with $Q_m u = 0$ one has*

$$\|u\|_{m-1,p} \leq C \|\nabla^m u\|_p. \quad (6.34)$$

Proof Let us assume that the inequality is wrong, i.e., that there exists a sequence (u^n) in $H^{m,p}(\Omega)$ with $Q_m u^n = 0$, $\|u^n\|_{m-1,p} = 1$, such that $\|\nabla^m u^n\|_p \leq \frac{1}{n}$ for all n , i.e., $\nabla^m u^n \rightarrow 0$ for $n \rightarrow \infty$. Since $H^{m,p}(\Omega)$ is reflexive and u^n is bounded in the respective norm, we can also assume that $u^n \rightharpoonup u$ for some $u \in H^{m,p}(\Omega)$ with $Q_m u = 0$. By the weak closedness of ∇^m we see that also $\nabla^m u = 0$ has to hold. By Lemma 6.79 we have $u \in \Pi^m$ and consequently $u = 0$, since $Q_m u = 0$.

By the compact embedding into $H^{m-1,p}(\Omega)$ (see Theorem 6.76) we obtain $\|P_m u^n\|_{m-1,p} \rightarrow 0$ for $n \rightarrow \infty$. This is a contradiction to $\|P_m u^n\|_{m-1,p} = 1$. \square

Corollary 6.82 (Poincaré-Wirtinger Inequality) *In the above situation, for $k = 0, \dots, m$ one has*

$$\|P_m u\|_{k,p} \leq C_k \|\nabla^m u\|_p \quad \forall u \in H^{m,p}(\Omega). \quad (6.35)$$

Proof First, let $k = m$. We plug $P_m u$ into (6.34), and get $\|P_m u\|_{m-1,p} \leq C \|\nabla^m P_m u\|_p = C \|\nabla^m u\|_p$. Adding $\|\nabla^m P_m u\|_p = \|\nabla^m u\|_p$ on both sides and using the fact that $\|\cdot\|_{m-1,p} + \|\nabla^m \cdot\|_p$ is equivalent to the norm in $H^{m,p}(\Omega)$ yields the claim.

The case $k < m$ follows from the estimate $\|u\|_{k,p} \leq \|u\|_{m,p}$ for all $u \in H^{m,p}(\Omega)$. \square

Now we prove the desired properties of the Sobolev penalty.

Lemma 6.83 *Let $\Omega \subset \mathbf{R}^d$ be a bounded Lipschitz domain and let $\varphi : [0, \infty[\rightarrow \mathbf{R}_\infty$ be proper, convex, lower semicontinuous, and non-decreasing. Further, let $1 < p < \infty$ and $m \geq 0$. Then the functional*

$$\Psi(u) = \begin{cases} \varphi(\|\nabla^m u\|_p) & \text{if } u \in H^{m,p}(\Omega), \\ \infty & \text{otherwise,} \end{cases}$$

is proper, convex, and (weakly) lower semicontinuous on every $L^q(\Omega)$, $1 \leq q < \infty$.

If, moreover, $q \leq pd/(d - mp)$, $mp < d$, and φ is coercive, then Ψ is also coercive in the sense that for every sequence $u^n \in L^q(\Omega)$, one has

$$\|P_m u\|_q \rightarrow \infty \quad \Rightarrow \quad \Psi(u) \rightarrow \infty.$$

If φ is strongly coercive, then $\|P_m u\|_q \rightarrow \infty \Rightarrow \Psi(u)/\|P_m u\|_q \rightarrow \infty$.

Proof Since φ is proper and non-decreasing, it follows that $\varphi(0) < \infty$. Hence, Ψ is proper, since $\Psi(0) = \varphi(0) < \infty$.

The mapping $\xi \mapsto |\xi|$ is by (6.32) a norm on the finite dimensional space \mathbf{R}^{d^m} . Hence, $v \mapsto \|v\|_p = (\int_{\Omega} |v|^p dx)^{1/p}$ is a Lebesgue norm on $L^p(\Omega, \mathbf{R}^{d^m})$, and the functional can be written as $\varphi \circ \|\cdot\|_p \circ \nabla^m$. We consider the linear mapping

$$\nabla^m : \text{dom } \nabla^m \rightarrow L^p(\Omega, \mathbf{R}^{d^m}), \quad \text{dom } \nabla^m = L^q(\Omega) \cap H^{m,p}(\Omega).$$

By Lemma 3.16, ∇^m is densely defined, and we aim to show that it is also strongly-to-weakly closed. To that end, let (u^n) be a sequence in $L^q(\Omega) \cap H^{m,p}(\Omega)$, converging in $L^q(\Omega)$ to some $u \in L^q(\Omega)$ that also satisfies $\nabla^m u^n \rightharpoonup v$ in $L^p(\Omega, \mathbf{R}^{d^m})$. By Lemma 6.73 we get that $v = \nabla^m u$, but it remains to show that $u \in H^{m,p}(\Omega)$.

Here we apply (6.35) and conclude that the sequence $(P_m u^n)$ in $H^{m,p}(\Omega)$ is bounded. By reflexivity we can assume, by moving to a subsequence, that $P_m u^n \rightharpoonup w$ for some $w \in H^{m,p}(\Omega)$, and by the compact embedding from Theorem 6.76 we get $P_m u^n \rightarrow w$ and, as a consequence of the embedding of the Lebesgue spaces, $u^n \rightarrow u$ in $L^1(\Omega)$. This shows strong convergence of the sequence $Q_m u^n = u^n - P_m u^n \rightarrow u - w$ in the finite dimensional space Π^m . We can view this space as a subspace of $H^{m,p}(\Omega)$ and hence, we have $Q_m u^n \rightarrow u - w$ in $H^{m,p}(\Omega)$ by equivalence of norms. This gives $u^n = P_m u^n + Q_m u^n \rightharpoonup w + u - w = u$ in $H^{m,p}(\Omega)$; in particular, u is contained in this Sobolev space.

Since ∇^m is strongly-to-weakly closed, we get from Example 6.23 and Lemma 6.21 the convexity and by Example 6.29 and Lemmas 6.28 and 6.14 the (weak) lower semicontinuity of Ψ .

To prove coercivity, let $q \leq pd/(d - mp)$ if $mp < d$ and let (u^n) be a sequence in $L^q(\Omega)$ with $\|P_m u^n\|_q \rightarrow \infty$. Now assume that there exists a $L > 0$ such that $\|\nabla^m u^n\|_p \leq L$ for infinitely many n . For these n , one has $u^n \in H^{m,p}(\Omega)$ and by the Poincaré-Wirtinger inequality (6.35) and the continuous embedding of $H^{m,p}(\Omega)$ into $L^q(\Omega)$ by Theorem 6.76, we obtain

$$\|P_m u^n\|_q \leq C \|P_m u^n\|_{m,p} \leq C \|\nabla^m u^n\|_p \leq CL, \quad C > 0,$$

a contradiction. Hence, $\|P_m u^n\|_q \rightarrow \infty$ implies $\|\nabla^m u^n\|_p \rightarrow \infty$, and by the coercivity of φ we also get $\Psi(u^n) \rightarrow \infty$. If φ is strongly convex, then the inequality $\|P_m u^n\|_q^{-1} \geq C^{-1} \|\nabla^m u^n\|_p^{-1}$ shows the assertion. \square

Now we have all the ingredients for proving the existence of solutions of minimization problems with Sobolev seminorms as penalty.

Theorem 6.84 (Existence of Solutions with Sobolev Penalty) *Let $\Omega \subset \mathbf{R}^d$ be a bounded Lipschitz domain, $m \in \mathbf{N}$, $m \geq 1$, $1 < p, q < \infty$ with $q \leq pd/(d - mp)$ if $mp < d$. Moreover, let $\Phi : L^q(\Omega) \rightarrow \mathbf{R}_\infty$ be proper on $H^{m,p}(\Omega)$, convex, lower semicontinuous on $L^q(\Omega)$, and coercive on Π^m in the sense that*

$$(\|P_m u\|_q) \text{ bounded and } \|Q_m u\|_q \rightarrow \infty \quad \Rightarrow \quad \Phi(u) \rightarrow \infty.$$

Moreover, set

$$\Psi(u) = \begin{cases} \varphi(\|\nabla^m u\|_p) & \text{if } u \in H^{m,p}(\Omega), \\ \infty & \text{otherwise} \end{cases}$$

where $\varphi : [0, \infty[\rightarrow \mathbf{R}_\infty$ is proper, convex, lower semicontinuous and non-decreasing, and strongly coercive. Then there exists for every $\lambda > 0$ a solution u^* of the minimization problem

$$\min_{u \in L^q(\Omega)} \Phi(u) + \lambda \Psi(u).$$

If φ is strictly convex, any two solutions u^* and u^{**} differ only in Π^m .

Proof By assumption and using Lemma 6.83 we see that $F = \Phi + \lambda \Psi$ is proper, convex, and lower semicontinuous on the reflexive Banach space $L^q(\Omega)$. To apply Theorem 6.31, we need only to show coercivity.

First, note that Φ is bounded from below by an affine linear functional, and by Theorem 6.54 we can even choose $u^0 \in L^q(\Omega)$, $w^0 \in L^{q^*}(\Omega)$ such that $\Phi(u^0) + \langle w^0, u - u^0 \rangle \leq \Phi(u)$ for all $u \in L^q(\Omega)$. In particular, we obtain the boundedness of Φ from below on bounded sets. Now assume that $\|u^n\|_q \rightarrow \infty$ for a sequence (u^n) in $L^q(\Omega)$. For an arbitrary subsequence (u^{n_k}) consider the sequences $(P_m u^{n_k})$ and $(Q_m u^{n_k})$. We distinguish two cases. First, if $\|P_m u^{n_k}\|_q$ is bounded, $\|Q_m u^{n_k}\|_q$ has to be unbounded, and by moving to a subsequence we get $\|Q_m u^{n_k}\|_q \rightarrow \infty$. By assumption we get $\Phi(u^{n_k}) \rightarrow \infty$ and, since $\Psi(u^{n_k}) \geq 0$, also $F(u^{n_k}) \rightarrow \infty$.

Second, if $\|P_m u^{n_k}\|_q$ is unbounded, we get by Lemma 6.83 (again moving to a subsequence if necessary) $\Psi(u^{n_k}) \rightarrow \infty$. If, moreover, $\|Q_m u^{n_k}\|_q$ is bounded, then

$$\begin{aligned} \Phi(u^{n_k}) &\geq \Phi(u^0) + \langle w^0, u^{n_k} - u^0 \rangle \\ &\geq \Phi(u^0) - \|w^0\|_{q^*} \|u^0\|_q - \|w^0\|_{q^*} \|P_m u^{n_k}\|_q - \|w^0\|_{q^*} \|Q_m u^{n_k}\|_q \\ &\geq C - \|w^0\|_{q^*} \|P_m u^{n_k}\|_q \end{aligned}$$

with some $C \in \mathbf{R}$ independent of k . Hence

$$F(u^{n_k}) \geq C + \|P_m u^{n_k}\|_q \left(\frac{\lambda \Psi(u^{n_k})}{\|P_m u^{n_k}\|_q} - \|w^0\|_{q^*} \right) \rightarrow \infty,$$

since the term in parentheses goes to infinity by the strong coercivity of Ψ (again, cf. Lemma 6.83). In the case that $\|Q_m u^{n_k}\|_q$ is unbounded, we obtain (again moving to a subsequence if necessary) $\Phi(u^{n_k}) \rightarrow \infty$ and hence $F(u^{n_k}) \rightarrow \infty$.

Since the above reasoning holds for every subsequence, we see that for the whole sequence we must have $F(u^n) \rightarrow \infty$, i.e., F is coercive. By Theorem 6.31 there exists a minimizer $u^* \in L^q(\Omega)$.

Finally, let u^* and u^{**} be minimizers with $u^* - u^{**} \notin \Pi^m$. Then $\nabla^m u^* \neq \nabla^m u^{**}$, and since $\|\cdot\|_p$ on $L^p(\Omega, \mathbf{R}^{d^m})$ is based on a Euclidean norm on \mathbf{R}^{d^m} (see Definition (6.31) and explanations in Example 6.23 for norms and convex integrands), for strongly convex φ , we have

$$\varphi\left(\left\|\frac{\nabla^m(u^* + u^{**})}{2}\right\|_p\right) < \frac{1}{2}\varphi(\|\nabla^m u^*\|_p) + \frac{1}{2}\varphi(\|\nabla^m u^{**}\|_p)$$

and hence $F\left(\frac{1}{2}(u^* + u^{**})\right) < \frac{1}{2}F(u^*) + \frac{1}{2}F(u^{**})$, a contradiction. We conclude that $u^* - u^{**} \in \Pi^m$. \square

Remark 6.85

- One can omit the assumption $q \leq pd/(d - mp)$ for $mp < d$ if Φ is coercive on the whole space $L^q(\Omega)$.
- Strong coercivity φ can be replaced by mere coercivity if Φ is bounded from below.
- If Φ is strictly convex, minimizers are unique without further assumptions on φ .

We can apply the above existence result to Tikhonov functionals that are associated with the inversion of linear and continuous mappings.

Theorem 6.86 (Tikhonov Functionals with Sobolev Penalty) *Let Ω, d, m, p be as in Theorem 6.84, $q \in]1, \infty[$, and Y a Banach space and $A \in \mathcal{L}(L^q(\Omega), Y)$. If one of the conditions*

1. $q \leq pd/(d - mp)$ for $mp < d$ and A injective on Π^m
2. A injective and $\text{rg}(A)$ closed

is satisfied, then there exists for every $u^0 \in Y, r \in [1, \infty[,$ and $\lambda > 0$ a solution for the minimization problem

$$\min_{u \in L^q(\Omega)} \frac{\|Au - u^0\|_Y^r}{r} + \lambda \frac{\|\nabla^m u\|_p^p}{p}. \quad (6.36)$$

In the case $r > 1$ and strictly convex norm in Y , the solution is unique.

Proof To apply Theorem 6.84 in the first case, it suffices to show coercivity in the sense that $\|P_m u^n\|_q$ bounded and $\|Q_m u\|_q \rightarrow \infty \Rightarrow \frac{1}{r} \|Au - u^0\|_Y^r \rightarrow \infty$. All other conditions are satisfied by the assumptions or are simple consequences of them (see also Example 6.32).

Now consider A restricted to the finite-dimensional space Π^m and note that this restriction is, by assumption, injective and hence boundedly invertible on $\text{rg}(A|_{\Pi^m})$. Hence, there is a $C > 0$ such that $\|u\|_q \leq C \|Au\|_Y$ for all $u \in \Pi^m$. Now let (u^n) be a sequence in $L^q(\Omega)$ with $\|P_m u^n\|_q$ bounded and $\|Q_m u^n\|_q \rightarrow \infty$. Then $(\|AP_m u^n - u^0\|_Y)$ is also bounded (by some $L > 0$), and since Q_m projects onto Π^m , one has

$$\|Au^n - u^0\|_Y \geq \|AQ_m u^n\|_Y - \|AP_m u^n - u^0\|_Y \geq C^{-1} \|Q_m u^n\|_q - L.$$

For large n , the right-hand side is nonnegative, and thus

$$\Phi(u^n) \geq \frac{(C^{-1} \|Q_m u^n\|_q - L)^r}{r} \rightarrow \infty.$$

Hence, Φ has all properties needed in Theorem 6.84, which shows the existence of a minimizer.

In the second case, $\Phi(u) = \frac{1}{r} \|Au - u^0\|_Y^r$ is coercive on $L^q(\Omega)$; this follows from the assertion in Exercise 6.6. By Remark 6.85 there exists a minimizer in this situation.

For the uniqueness in the case $r > 1$ and $\|\cdot\|_Y$ strictly convex note that $\varphi(t) = \frac{t^r}{r}$ is strictly convex and that minimizers u^* and u^{**} differ only in Π^m . In particular, $P_m u^* = P_m u^{**}$. Now assume that $u^* \neq u^{**}$. Then it holds that

$$Au^* = P_m u^* + Q_m u^* \neq P_m u^* + Q_m u^{**} = P_m u^{**} + Q_m u^{**} = Au^{**}$$

by the injectivity of A on Π_m . We obtain

$$\frac{\frac{1}{2}(Au^* - u^0) + \frac{1}{2}(Au^{**} - u^0)\|_Y^r}{r} < \frac{\|Au^* - u^0\|_Y^r}{2r} + \frac{\|Au^{**} - u^0\|_Y^r}{2r},$$

which contradicts the minimizing property of u^* and u^{**} . \square

Remark 6.87 Injectivity (and hence invertibility) of A on Π^m gives the coercivity of the objective functional in the directions that do not change the Sobolev seminorm. So in some sense, one can say that a solution u^* gets its components in Π^m by inversion directly from u^0 . In the case that Y is a Hilbert space, one can show this rigorously; see Exercise 6.26.

Before we turn to concrete applications, we analyze the subgradient of the penalty $\frac{1}{p} \|\nabla^m u\|_p^p$. To that end, we derive the adjoint operator for $\nabla^m : L^q(\Omega) \rightarrow L^p(\Omega, \mathbf{R}^{d^m})$. Since $\nabla^m u$ has values in \mathbf{R}^{d^m} , the adjoint depends on the inner product

in that space. We use the inner product coming from the norm (6.32), namely

$$a \cdot b = \sum_{i_1, \dots, i_m=1}^d a_{i_1, \dots, i_m} b_{i_1, \dots, i_m} \text{ for } a, b \in \mathbf{R}^{d^m}.$$

What is $(\nabla)^* w$ for $w \in \mathcal{D}(\Omega, \mathbf{R}^{d^m})$? We test with $u \in \mathcal{C}^\infty(\overline{\Omega}) \subset \text{dom } \nabla^m$, and get by integration by parts that

$$\begin{aligned} \int_{\Omega} w \cdot \nabla^m u \, dx &= \int_{\Omega} \sum_{i_1, \dots, i_m=1}^d w_{i_1, \dots, i_m} \partial^{i_1} \cdots \partial^{i_m} u \, dx \\ &= (-1)^m \int_{\Omega} \left(\sum_{i_1, \dots, i_m=1}^d \partial^{i_1} \cdots \partial^{i_m} w_{i_1, \dots, i_m} \right) u \, dx. \end{aligned}$$

Since $\mathcal{C}^\infty(\overline{\Omega})$ is dense in $L^p(\Omega)$, we see that the adjoint is the differential operator on the right-hand side. In the case $m = 1$ this amounts to $\nabla^* = -\text{div}$, and hence we write

$$(\nabla^m)^* = (-1)^m \text{div}^m = (-1)^m \sum_{i_1, \dots, i_m=1}^d \partial^{i_1} \cdots \partial^{i_m}.$$

Every element in $\text{dom } \nabla^m$ can be approximated by some element in $\mathcal{C}^\infty(\overline{\Omega})$ in the sense of the norm $\|\cdot\|_q + \|\nabla^m \cdot\|_p$ (see Theorem 6.74), and hence $w \in \text{dom } (\nabla^m)^*$ and we get $\mathcal{D}(\Omega, \mathbf{R}^{d^m}) \subset \text{dom } (\nabla^m)^*$.

The above formulation motivated the definition of the *weak divergence*: An element $v \in L^1_{\text{loc}}(\Omega)$ is called the m th weak divergence of a vector field $w \in L^1_{\text{loc}}(\Omega, \mathbf{R}^{d^m})$ if for every $u \in \mathcal{D}(\Omega, \mathbf{R}^{d^m})$, one has

$$\int_{\Omega} w \cdot \nabla^m u \, dx = (-1)^m \int_{\Omega} vu \, dx.$$

We write $v = \text{div}^m w$ if the weak divergence exists. Similarly to Lemma 6.73 one can show that this defines a closed operator between $L^{p^*}(\Omega, \mathbf{R}^{d^m})$ and $L^{q^*}(\Omega)$. Since $(\nabla^m)^*$ is also closed, we obtain for (w^n) in $\mathcal{D}(\Omega, \mathbf{R}^{d^m})$ with $\lim_{n \rightarrow \infty} w^n = w$ in $L^{p^*}(\Omega, \mathbf{R}^{d^m})$ and $\lim_{n \rightarrow \infty} (-1)^m \text{div}^m w^n = v$ in $L^{q^*}(\Omega)$ also $(\nabla^m)^* w = v = (-1)^m \text{div}^m w$ with the weak m th divergence. We have shown that for

$$\begin{aligned} \mathcal{D}_{\text{div}}^m &= \left\{ w \in L^{p^*}(\Omega, \mathbf{R}^{d^m}) \mid \exists \text{ div}^m w \in L^{q^*}(\Omega) \text{ and sequence } (w^n) \text{ in } \mathcal{D}(\Omega, \mathbf{R}^{d^m}) \right. \\ &\quad \left. \text{with } \lim_{n \rightarrow \infty} \|w^n - w\|_{p^*} + \|\text{div}^m(w^n - w)\|_{q^*} = 0 \right\}, \end{aligned} \quad (6.37)$$

one has $\mathcal{D}_{\text{div}}^m \subset \text{dom}(\nabla^m)^*$. Note that this space depends on p and q , but this dependence is not reflected in the notation. In the case $m = 1$ one can show by simple means that this space coincides with $\text{dom} \nabla^*$.

Theorem 6.88 (Characterization of ∇^*) *For a bounded Lipschitz domain Ω , $1 < p \leq q < \infty$, and the linear operator ∇ between $L^q(\Omega)$ and $L^p(\Omega, \mathbf{R}^d)$ with domain $H^{1,p}(\Omega)$ one has*

$$\nabla^* = -\operatorname{div} \text{ with } \text{dom} \nabla^* = \mathcal{D}_{\text{div}}^1 \text{ as in (6.37).}$$

Proof Let $w \in \text{dom} \nabla^* \subset L^{p^*}(\Omega, \mathbf{R}^d)$. We apply the $L^p - L^{p^*}$ -adjoints of the approximating operators \mathcal{M}_n from (6.30) in Theorem 6.74 to every component of w (and denote them by \mathcal{M}_n^*):

$$w^n = \mathcal{M}_n^* w = \sum_{k=0}^K \varphi_k (T_{t_n \eta_k} (w * \bar{\psi}^n)),$$

where $\bar{\psi}^n = D_{-\operatorname{id}} \psi^n$. Every w^n is infinitely differentiable, so we can consider their supports. Let

$$\Omega_n = \bigcup_{k=0}^K (\overline{\Omega - \text{supp } \psi^n} + t_n \eta_k) \cap \overline{V_k},$$

which satisfies $\Omega_n \subset\subset \Omega$ by construction. Applied to $u \in \mathcal{D}(\Omega \setminus \Omega_n)$ (with extension by zero), we obtain for every k that

$$(T_{t_n \eta_k} (\varphi_k u)) * \psi^n = 0$$

by (6.29) and the fact that u vanished on every $(\overline{\Omega - \text{supp } \psi^n} + t_n \eta_k) \cap \overline{V_k}$. This shows that $\mathcal{M}_n u = 0$. Now we test every w^n with the above u and get

$$\int_{\Omega} (\mathcal{M}_n^* w) u \, dx = \int_{\Omega} w \mathcal{M}_n u \, dx = 0,$$

and by the fundamental lemma of the calculus of variations (Lemma 2.75) we also get $w^n = 0$ on $\Omega \setminus \Omega_n$. This shows that $w^n \in \mathcal{D}(\Omega, \mathbf{R}^d)$.

The sequence w^n converges weakly in $L^{p^*}(\Omega, \mathbf{R}^d)$ to w : for $u \in L^p(\Omega, \mathbf{R}^d)$, one has $\mathcal{M}_n u \rightarrow u$ in $L^p(\Omega, \mathbf{R}^d)$ and hence $\langle w^n, u \rangle = \langle w, \mathcal{M}_n u \rangle \rightarrow \langle w, u \rangle$. Moreover, for $u \in \text{dom} \nabla$, one has

$$\mathcal{M}_n \nabla u = \nabla(\mathcal{M}_n u) - \underbrace{\sum_{k=0}^K (T_{t_n \eta_k} (u \nabla \varphi_k)) * \psi^n}_{=\mathcal{N}_n u} = \nabla(\mathcal{M}_n u) - \mathcal{N}_n u$$

and this shows that $\nabla^* w^n = -\operatorname{div} w^n = -\mathcal{M}_n^*(\operatorname{div} w) - \mathcal{N}_n^* w$, since

$$\begin{aligned} -\langle \operatorname{div} w^n, u \rangle &= \langle \mathcal{M}_n^* w, \nabla u \rangle = \langle w, \mathcal{M}_n \nabla u \rangle = \langle w, \nabla(\mathcal{M}_n u) \rangle - \langle w, \mathcal{N}_n u \rangle \\ &= -\langle \mathcal{M}_n^*(\operatorname{div} w), u \rangle - \langle \mathcal{N}_n^* w, u \rangle. \end{aligned}$$

Similar to Theorem 6.74, one sees that for every $u \in L^q(\Omega)$,

$$\lim_{n \rightarrow \infty} \mathcal{N}_n u = \sum_{k=0}^K u \nabla \varphi_k = 0 \quad \text{in} \quad L^q(\Omega, \mathbf{R}^d)$$

holds, since (φ_k) is a partition of unity. This implies that for every $u \in L^q(\Omega)$,

$$\begin{aligned} |-\langle \operatorname{div} w^n, u \rangle| &\leq |-\langle \operatorname{div} w, \mathcal{M}_n u \rangle| + |\langle w, \mathcal{N}_n u \rangle| \\ &\leq \|\operatorname{div} w\|_{q^*} \sup_n \|\mathcal{M}_n u\|_q + \|w\|_{p^*} \sup_n \|\mathcal{N}_n u\|_p \leq C_u, \end{aligned}$$

since $\|\mathcal{N}_n u\|_p \leq C \|\mathcal{N}_n u\|_q$ because we have $p \leq q$. By the uniform boundedness principle (Theorem 2.15) we get that $(-\operatorname{div} w^n)$ is bounded in $L^{q^*}(\Omega, \mathbf{R}^d)$, and hence there exists a weakly convergent subsequence. By the weak closedness of $-\operatorname{div}$, the weak limit of every weakly convergent subsequence has to coincide with $-\operatorname{div} w$; consequently, the whole sequence converges, i.e., $-\operatorname{div} w^n \rightharpoonup -\operatorname{div} w$ as $n \rightarrow \infty$.

We have shown that

$$\begin{aligned} \operatorname{dom} \nabla^* &= \left\{ w \in L^{p^*}(\Omega, \mathbf{R}^d) \mid \exists \operatorname{div} w \in L^{q^*}(\Omega) \text{ and sequence } (w^n) \text{ in } \mathcal{D}(\Omega, \mathbf{R}^d) \right. \\ &\quad \left. \text{with } w^n \rightharpoonup w \text{ in } L^{p^*}(\Omega, \mathbf{R}^d) \text{ and } \operatorname{div} w^n \rightharpoonup \operatorname{div} w \text{ in } L^{q^*}(\Omega) \right\}. \end{aligned}$$

Now assume that there exist $\varepsilon > 0$ and $w^0 \in \operatorname{dom} \nabla^*$ for which $\|w - w^0\|_{p^*} \geq \varepsilon$ or $\|\operatorname{div}(w - w^0)\|_{q^*} \geq \varepsilon$ for every $w \in \mathcal{D}(\Omega, \mathbf{R}^d)$. Then by the definition of the dual norm as a supremum, there exists $v \in L^p(\Omega, \mathbf{R}^d)$, $\|v\|_p \leq 1$, or $u \in L^q(\Omega)$, $\|u\|_q \leq 1$, with

$$\langle w - w^0, v \rangle \geq \frac{\varepsilon}{2} \quad \text{or} \quad \langle \operatorname{div}(w - w^0), u \rangle \geq \frac{\varepsilon}{2} \quad \text{for all } w \in \mathcal{D}(\Omega, \mathbf{R}^d).$$

This is a contradiction, and hence we can replace weak by strong convergence and get $\operatorname{dom} \nabla^* = \mathcal{D}_{\operatorname{div}}^1$, as desired. \square

Remark 6.89 The domain of definition of the adjoint ∇^* can be interpreted in a different way. To that end, we consider certain boundary values on $\partial\Omega$.

If for some $w \in \mathcal{C}^\infty(\overline{\Omega}, \mathbf{R}^d)$ and $v \in L_{\mathfrak{H}^{d-1}}^1(\partial\Omega)$ the identity

$$\int_{\partial\Omega} uv \, d\mathfrak{H}^{d-1} = \int_{\Omega} u \operatorname{div} w + \nabla u \cdot w \, dx$$

holds for every $u \in \mathcal{C}^\infty(\overline{\Omega})$, then $v = w \cdot v$ on $\partial\Omega$ \mathfrak{H}^{d-1} -almost everywhere, since by Gauss's Theorem (Theorem 2.81)

$$\int_{\partial\Omega} u(v - w \cdot v) \, d\mathfrak{H}^{d-1} = 0 \quad \text{for all } u \in \mathcal{C}^\infty(\overline{\Omega}),$$

and with Theorem 2.73 and the fundamental lemma of the calculus of variations (Lemma 2.75) we get that $v - w \cdot v = 0$ \mathfrak{H}^{d-1} -almost everywhere on $\partial\Omega$. This motivates a more general definition of the so-called normal trace on the boundary.

We say that $v \in L_{\mathfrak{H}^{d-1}}^1(\partial\Omega)$ is the *normal trace* of the vector field $w \in L^{p^*}(\Omega, \mathbf{R}^d)$ with $\operatorname{div} w \in L^{q^*}(\Omega)$, if there exists a sequence (w^n) in $\mathcal{C}^\infty(\overline{\Omega}, \mathbf{R}^d)$ with $w^n \rightarrow w$ in $L^{p^*}(\Omega, \mathbf{R}^d)$, $\operatorname{div} w^n \rightarrow \operatorname{div} w$ in $L^{q^*}(\Omega)$ and $w^n \cdot v \rightarrow v$ in $L_{\mathfrak{H}^{d-1}}^1(\partial\Omega)$, such that for all $u \in \mathcal{C}^\infty(\overline{\Omega})$,

$$\int_{\partial\Omega} uv \, d\mathfrak{H}^{d-1} = \int_{\Omega} u \operatorname{div} w + \nabla u \cdot w \, dx.$$

In this case we write $v = w \cdot v$ on $\partial\Omega$. Note that this definition corresponds to the closure of the normal trace for smooth vector fields w with respect to $\|w\|_{p^*} + \|\operatorname{div} w\|_{q^*}$ and $\|(w \cdot v)|_{\partial\Omega}\|_1$.

By the definition of the adjoint and Theorem 6.88 we see that $\operatorname{dom} \nabla^*$ is precisely the set on which the normal trace vanishes. Hence, we can say that

$$\nabla^* = -\operatorname{div} \text{ defined on } \{w \cdot v = 0 \text{ on } \partial\Omega\}.$$

Remark 6.90 The closed operator ∇^m from $L^q(\Omega)$ to $L^p(\Omega, \mathbf{R}^d)$ on the bounded Lipschitz domain Ω with $q \leq pd/(d-mp)$ if $mp < d$ has a closed range: If (u^n) is a sequence in $L^q(\Omega)$ such that $\lim_{n \rightarrow \infty} \nabla^m u^n = v$ for some $v \in L^p(\Omega, \mathbf{R}^{d^m})$, then $(P_m u^n)$ is a Cauchy sequence, since the Poincaré-Wirtinger inequality (6.35) and the embedding into $L^q(\Omega)$ (Theorem 6.76) lead to

$$\|P_m u^{n_1} - P_m u^{n_2}\|_q \leq C \|\nabla^m u^{n_1} - \nabla^m u^{n_2}\|_p$$

for $n_1, n_2 \geq n_0$ and arbitrary n_0 . Hence, there exists a limit $\lim_{n \rightarrow \infty} P_m u^n = u$ in $L^q(\Omega)$. Since $\nabla^m(P_m u^n) = \nabla^m u^n$ we get by closedness of the weak gradient that $\nabla^m u = v$. Hence, the range $\operatorname{rg}(\nabla^m)$ is also closed.

By the closed range theorem (Theorem 2.26) we get $\text{rg}((\nabla^m)^*) = \ker(\nabla^m)^\perp \subset L^{q^*}(\Omega)$. Since also $\ker \nabla^m = \Pi^m$ (see Lemma 6.79), this is equivalent to

$$\text{rg}((\nabla^m)^*) = (\Pi^m)^\perp = \left\{ w \in L^{q^*}(\Omega) \mid \int_{\Omega} w(x)x^\alpha dx = 0 \text{ for all } \alpha \in \mathbb{N} \text{ with } |\alpha| < m \right\}.$$

In the special case $m = 1$ and $p \geq q$ we can use the characterization of ∇^* to solve some divergence equations: the equation

$$w \in L^{q^*}(\Omega), \quad \int_{\Omega} w dx = 0 : \quad \begin{cases} -\operatorname{div} v = w & \text{in } \Omega, \\ v \cdot v = 0 & \text{on } \partial\Omega, \end{cases}$$

has a solution in $v \in L^{p^*}(\Omega, \mathbf{R}^d)$.

Finally, we can calculate the subgradient of the functional $u \mapsto \frac{1}{p} \|\nabla u\|_p^p$.

Lemma 6.91 *For a bounded Lipschitz domain $\Omega \subset \mathbf{R}^d$, $1 < p \leq q < \infty$, ∇ a closed mapping between $L^q(\Omega)$ and $L^p(\Omega, \mathbf{R}^d)$ with domain $H^{1,p}(\Omega)$ and*

$$\Psi(u) = \begin{cases} \frac{1}{p} \|\nabla u\|_p^p & \text{if } u \in H^{1,p}(\Omega), \\ \infty & \text{otherwise,} \end{cases}$$

one has for $u \in L^q(\Omega)$ that

$$\partial\Psi(u) = \begin{cases} \{-\operatorname{div}(|\nabla u|^{p-2}\nabla u)\} & \text{if } \begin{cases} \nabla u \in L^p(\Omega, \mathbf{R}^d), \\ \operatorname{div}(|\nabla u|^{p-2}\nabla u) \in L^{q^*}(\Omega), \\ \text{and } |\nabla u|^{p-2}\nabla u \cdot v = 0 \text{ on } \partial\Omega, \end{cases} \\ \emptyset & \text{otherwise.} \end{cases}$$

Proof The convex functional $F(v) = \frac{1}{p} \|v\|_p^p$ defined on $L^p(\Omega, \mathbf{R}^d)$ is, as a p th power of a norm, continuous everywhere, in particular at every point of $\text{rg}(\nabla)$. Since $\Psi = F \circ \nabla$, we can apply the identity $\partial\Psi = \nabla^* \circ \partial F \circ \nabla$ in the sense of Definition 6.41 (see Exercise 6.12). By the rule for subdifferentials for convex integrands (Example 6.50) as well as Gâteaux differentiability of $\xi \mapsto \frac{1}{p}|\xi|^p$ we get

$$F(v) = \int_{\Omega} \frac{1}{p} |v(x)|^p dx \quad \Rightarrow \quad \partial F(v) = \{|v|^{p-2}v\},$$

and it holds $\partial\Psi(u) \neq \emptyset$ if and only if $u \in \text{dom } \nabla = H^{1,p}(\Omega)$ and $|\nabla u|^{p-2}\nabla u \in \text{dom } \nabla^*$. By Theorem 6.88 and Remark 6.89, respectively, we can express the latter by $\operatorname{div}(|\nabla u|^{p-2}\nabla u) \in L^{q^*}(\Omega)$ with $|\nabla u|^{p-2}\nabla u \cdot v = 0$ on $\partial\Omega$. In this case we get $\partial\Psi(u) = \nabla^* \circ \partial F(\nabla u)$ which shows the desired identity. \square

Remark 6.92 It is easy to see that the case $p = 2$ leads to the negative Laplace operator for functions u that satisfy $\nabla u \cdot v = 0$ on the boundary; $\partial \frac{1}{2} \|\nabla \cdot\|_2^2 = -\Delta$.

Thus, the generalization for $p \in]1, \infty[$ is called *p-Laplace operator*; hence, one can say that the subgradient $\partial \frac{1}{p} \|\nabla \cdot\|_p^p$ is the *p-Laplace operator* for functions with the boundary conditions $|\nabla u|^{p-2} \nabla u \cdot v = 0$.

Example 6.93 (Solution of the p-Laplace Equation) An immediate application of the above result is the proof of existence and uniqueness of the *p-Laplace equation*. For a bounded Lipschitz domain Ω , $1 < p \leq q < \infty$, $q \leq d/(d-p)$ if $p < d$, and $f \in L^{q^*}(\Omega)$ we consider the minimization problem

$$\min_{u \in L^q(\Omega)} \frac{1}{p} \int_{\Omega} |\nabla u|^p dx - \int_{\Omega} f u dx + I_{\{v \in L^q(\Omega) \mid \int_{\Omega} v dx = 0\}}(u). \quad (6.38)$$

The restriction in the indicator function is exactly the condition $u \in (\Pi^1)^\perp$. We want to apply Theorem 6.84 to

$$\Phi(u) = - \int_{\Omega} f u dx + I_{(\Pi^1)^\perp}(u),$$

and $\varphi(t) = \frac{1}{p} t^p$. To that end we note that $\Phi(0) = 0$, and hence, Φ is proper on $H^{1,p}(\Omega)$. Convexity and lower semicontinuity are immediate, and coercivity follows from the fact that for $u \in L^q(\Omega)$ and $v \in \Pi^1$ (i.e. v is constant) with $v \neq 0$, one has $\Phi(u+v) = \infty$, since $\int_{\Omega} u + v dx \neq 0$. Thus, the assumptions in Theorem 6.84 are satisfied, and the minimization problem has a solution u^* , which is unique up to contributions from Π^1 . A solution u^{**} different from u^* would satisfy $u^{**} = u^* + v$ with $v \in \Pi^1$, $v \neq 0$, and this would imply $\Phi(u^{**}) = \infty$, a contradiction. Hence, the minimizer is unique.

Let us deduce the optimality conditions for u^* . Here we face a difficulty, since neither the Sobolev term Ψ nor Φ is continuous, i.e., the assumption for the sum rule in Theorem 6.51 are not satisfied. However, we see that both $u \mapsto \Psi(Q_1 u)$ and $u \mapsto \Phi(P_1 u)$ are continuous. Since for all $u \in L^q(\Omega)$ we have $u = P_1 u + Q_1 u$, we can apply the conclusion from Exercise 6.14 and get $0 \in \partial\Psi(u^*) + \partial\Phi(u^*)$. By Lemma 6.91 we know that $\partial\Psi$; let us compute $\partial\Phi(u^*)$. Since $u \mapsto \int_{\Omega} f u dx$ is continuous, Theorem 6.51 and Example 6.48 lead to

$$\partial\Phi(u) = \begin{cases} -f + \Pi^1 & \text{if } \int_{\Omega} u dx = 0, \\ \emptyset & \text{otherwise,} \end{cases}$$

since $((\Pi^1)^\perp)^\perp = \Pi^1$. Thus, u^* is optimal if and only if there exists some $\lambda^* \in \mathbf{R}$ such that

$$\begin{aligned} -\operatorname{div}(|\nabla u^*|^{p-2} \nabla u^*) &= f - \lambda^* \mathbf{1} && \text{in } \Omega, \\ |\nabla u^*|^{p-2} \nabla u^* \cdot v &= 0 && \text{on } \partial\Omega, \\ \int_{\Omega} u^* dx &= 0, \end{aligned}$$

where $\mathbf{1} \in L^{q^*}(\Omega)$ denotes the function that is constant 1. It is easy to calculate the value λ^* : we integrate the equation in Ω on both sides to get

$$\int_{\Omega} f \, dx - \lambda^* |\Omega| = \int_{\Omega} \nabla^* (|\nabla u^*|^{p-2} \nabla u^*) \, dx = 0,$$

and hence $\lambda^* = |\Omega|^{-1} \int_{\Omega} f \, dx$, which is the mean value of f . In conclusion, we have shown that for every $f \in L^{q^*}(\Omega)$ with $\int_{\Omega} f \, dx = 0$, there exists a unique solution $u \in L^q(\Omega)$ of the nonlinear partial differential equation

$$\begin{aligned} -\operatorname{div}(|\nabla u|^{p-2} \nabla u) &= f && \text{in } \Omega, \\ |\nabla u|^{p-2} \nabla u \cdot v &= 0 && \text{on } \partial\Omega, \\ \int_{\Omega} u \, dx &= 0. \end{aligned}$$

The solution u is exactly the minimizer of (6.38).

6.3.2 Practical Applications

The theory of convex minimization with Sobolev penalty that we have developed up to now gives a unified framework to treat the motivating examples from Sect. 6.1. In the following we revisit these problems and present some additional examples.

Application 6.94 (Denoising with L^q -Data and $H^{1,p}$ -Penalty) Consider the denoising problem on a bounded Lipschitz domain $\Omega \subset \mathbf{R}^d$. Further we assume that $1 < p \leq q < \infty$. Let $u^0 \in L^q(\Omega)$ be a noisy image and let $\lambda > 0$ be given. We aim to denoise u^0 by solving the minimization problem

$$\min_{u \in L^q(\Omega)} \frac{1}{q} \int_{\Omega} |u - u^0|^q \, dx + \frac{\lambda}{p} \int_{\Omega} |\nabla u|^p \, dx. \quad (6.39)$$

It is easy to see that this problem has a unique solution: the identity $A = \operatorname{id}$ is injective and has closed image, and since for $r = q$ the norm on $L^r(\Omega)$ is strictly convex, we obtain uniqueness and existence from Theorem 6.86.

Let us analyze the solutions u^* of (6.39) further. For example, it is simple to see that the mean values of u^* and u^0 are equal in the case $q = 2$, i.e., $Q_1 u^* = Q_1 u^0$ (see Exercise 6.27, which treats a more general case). It is a little more subtle to show that a maximum principle holds. We can derive this fact directly from the properties of the minimization problem:

Theorem 6.95 *Let $\Omega \subset \mathbf{R}^d$ be a bounded Lipschitz domain and let $1 < p \leq q < \infty$. Moreover, let $u^0 \in L^\infty(\Omega)$ with $L \leq u^0 \leq R$ almost everywhere and $\lambda > 0$.*

Then the solution u^ of (6.39) also satisfies $L \leq u^* \leq R$ almost everywhere.*

Proof Let F be the functional in (6.39). We plug in the function $u = \min(R, \max(L, u^*))$. If $u^*(x) \geq R$, then $|u(x) - u^0(x)| = R - u^0(x) \leq |u^*(x) - u^0(x)|$, and similarly we get $|u(x) - u^0(x)| \leq |u^*(x) - u^0(x)|$ if $u^*(x) \leq L$. This shows that

$$\frac{1}{q} \int_{\Omega} |u(x) - u^0(x)|^q dx \leq \frac{1}{q} \int_{\Omega} |u^*(x) - u^0(x)|^q dx.$$

Moreover, by Lemma 6.75 we get $u \in H^{1,p}(\Omega)$, and also that $\nabla u = \nabla u^*$ almost everywhere in $\{L \leq u^* \leq R\}$ and $\nabla u = 0$ almost everywhere else. Hence,

$$\frac{1}{p} \int_{\Omega} |\nabla u|^p dx \leq \frac{1}{p} \int_{\Omega} |\nabla u^*|^p dx,$$

and consequently $F(u) \leq F(u^*)$. Since u^* is the unique minimizer, we get $u^* = u$. The construction of u shows the claim. \square

The above shows that images $u^0 \in L^\infty(\Omega)$ are mapped to images in $L^\infty(\Omega)$, and in particular it cannot happen that u^* assumes values outside of the interval in which u^0 has its values. This can be seen in Figs. 6.11 and 6.12 which show and discuss some results for different parameters p, q .

Now consider the Euler-Lagrange equation for (6.39), which we derive using subgradients. The data term $\Phi = \frac{1}{q} \|\cdot - u^0\|_q^q$ is continuous, and hence we can use the sum rule (Theorem 6.51). The subgradient of Φ satisfies $\partial\Phi(u) = |u - u^0|^{q-2}(u - u^0)$, and the application of Lemma 6.91 together with the optimality condition for subgradients (Theorem 6.43) leads to

$$\begin{aligned} |u^* - u^0|^{q-2}(u^* - u^0) - \lambda \operatorname{div}(|\nabla u^*|^{p-2} \nabla u^*) &= 0 \quad \text{in } \Omega, \\ |\nabla u^*|^{p-2} \nabla u^* \cdot v &= 0 \quad \text{on } \partial\Omega. \end{aligned} \tag{6.40}$$

This shows, at least formally, that the solution u^* satisfies the nonlinear partial differential equation

$$-G(x, u^*(x), \nabla u^*(x), \nabla^2 u^*(x)) = 0 \quad \text{in } \Omega,$$

$$G(x, u, \xi, Q) = |u - u^0(x)|^{q-2}(u^0(x) - u) + \lambda |\xi|^{p-2} \operatorname{trace}\left(\left(\operatorname{id} + (p-2)\frac{\xi}{|\xi|} \otimes \frac{\xi}{|\xi|}\right)Q\right)$$

with respective boundary conditions. The function G is (degenerate) elliptic in the sense of Theorem 5.11 (see also Definition 5.10). For $p = 2$, the second-order term is the Laplace operator weighted by λ , which has favorable analytical properties. However, in the case $\xi = 0$ and $p < 2$ there is a singularity in G . On the other hand, the case $\xi = 0$ and $p > 2$ leads to a G that is independent of Q , i.e. the respective differential operator is degenerate there.

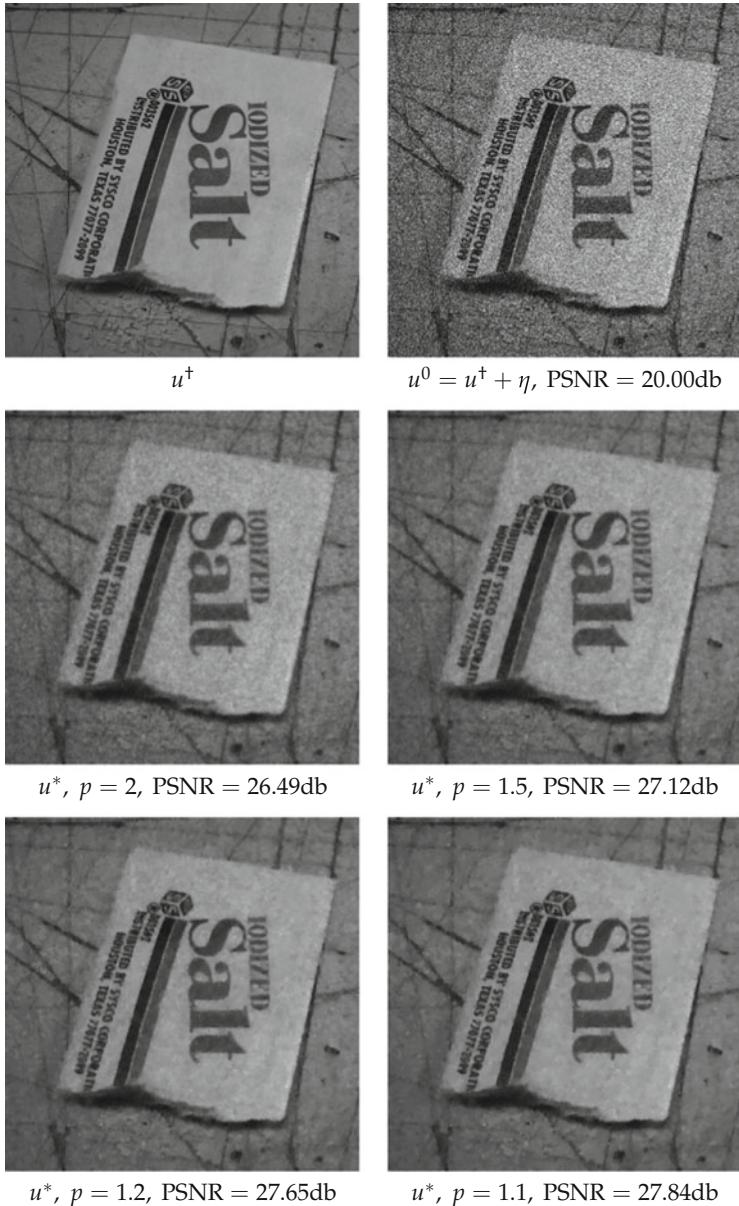


Fig. 6.11 Illustration of the denoising capabilities of variational denoising with Sobolev penalty. Top: Left the original, right its noisy version. Middle and bottom: The minimizer of (6.39) for $q = 2$ and different Sobolev exponents p . To allow for a comparison, the parameter λ has been chosen to maximize the PSNR with respect to the original image. Note that the remaining noise and the blurring of edges is less for $p = 1.1$ and $p = 1.2$ than it is for $p = 2$

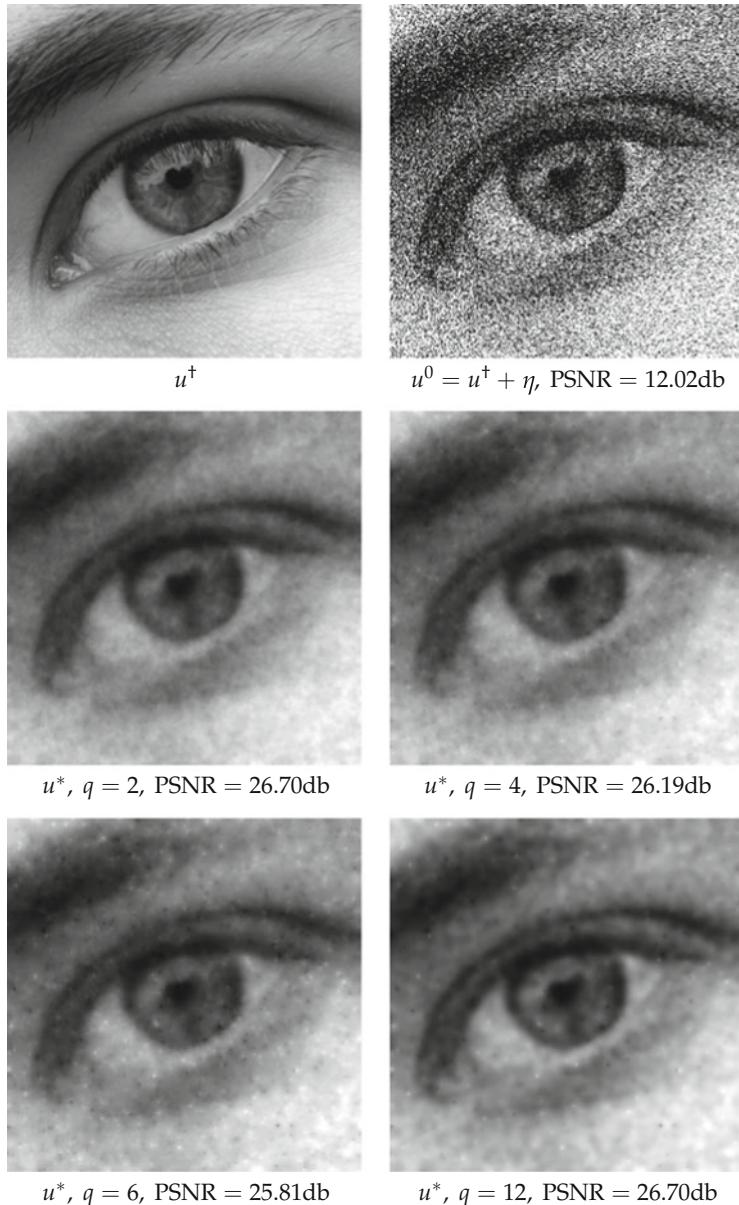


Fig. 6.12 Illustration of the influence of the exponent q in the data term of Application 6.94. Top: Left the original, right a version with strong noise. Middle and bottom: The minimizer of (6.39) for $p = 2$ and different exponents q , again with λ optimized with respect to the PSNR. For larger exponents q we see some “impulsive” noise artifacts, which again get less for $q = 6, q = 12$ due to the choice of λ . The image sharpness does not vary much

These facts complicate the analysis of solutions; however, in the case of $u^0 \in L^\infty(\Omega)$ one can show that u^* has to be more regular than $H^{1,p}(\Omega)$ in the interior of Ω : for $p \leq 2$ one has $u^* \in H^{2,p}(\Omega')$ for Ω' such that $\overline{\Omega'} \subset\subset \Omega$, and for $p > 2$ the solution is still in a suitable *Besov space* (see [131] for details).

Using Theorem 6.76 on Sobolev embeddings, this shows that the denoised image u^* for $d = 2$ is continuous in Ω (but probably does not have a continuous extension to the boundary): for $1 < p \leq 2$ this follows from the embedding $H^{2,p}(\Omega') \hookrightarrow C(\overline{\Omega'})$, and for $p > 2$ this is a consequence of $H^{1,p}(\Omega) \hookrightarrow C(\overline{\Omega})$. This shows that it is impossible to reconstruct images with discontinuities by solving (6.39) or (6.40), respectively. However, one notes that the solutions change qualitatively if p varies: for p close to 1, the solution appears to be less blurred; see again Fig. 6.11. This suggests having a closer look at the case $p = 1$, and we will do so in the next subsection. Figure 6.12 allows us to study the influence of the exponent q in the data term. It mostly influences the remaining noise, but does not change the overall smoothness properties of the solution.

Remark 6.96 Another variation of the denoising approach from Application 6.94 is to consider Sobolev penalties of higher order (see Exercise 6.27). Again, in the case $q = 2$ we see that u^* reproduces the respective polynomial parts of u^0 up to order $m - 1$.

Application 6.97 (Deconvolution with Sobolev Penalty) Let us analyze the reconstruction of images from a blurred and noisy image u^0 . We assume that we know u^0 on a bounded domain Ω' and that the blur results from a convolution with a kernel $k \in L^1(\Omega_0)$ with $\int_{\Omega_0} k \, dx = 1$. For the data in Ω' one needs information of u at most in the slightly larger set $\Omega' - \Omega_0$, and hence we assume that Ω is a bounded Lipschitz domain that satisfies $\Omega' - \Omega_0 \subset \Omega$. Hence, the forward operator A is defined as follows:

$$x \in \Omega' : \quad (Au)(x) = (u * k)(x) = \int_{\Omega} u(x - y)k(y) \, dy.$$

By Theorem 3.13, A maps $L^q(\Omega)$ to $L^q(\Omega')$ linearly and continuously for every $1 \leq q < \infty$. Moreover, A maps constant functions in Ω to constant functions in Ω' , and hence A is injective on Π^1 .

If we choose $1 < p \leq q < \infty$ and $q \leq pd/(d - p)$ if $p < d$, then Theorem 6.86 implies the existence of a unique minimizer of the Tikhonov functional

$$\min_{u \in L^q(\Omega)} \frac{1}{q} \int_{\Omega'} |u * k - u^0|^q \, dx + \frac{\lambda}{p} \int_{\Omega} |\nabla u|^p \, dx \quad (6.41)$$

for every $u^0 \in L^q(\Omega')$ and $\lambda > 0$.

Let us analyze the properties of the solution u^* of the deconvolution problem; see Fig. 6.13. Since the convolution maps constant functions to constant functions, we get in the case $q = 2$ the identity $\int_{\Omega} u^* \, dx = \int_{\Omega'} u^0 \, dx$ (see also Exercise 6.29).

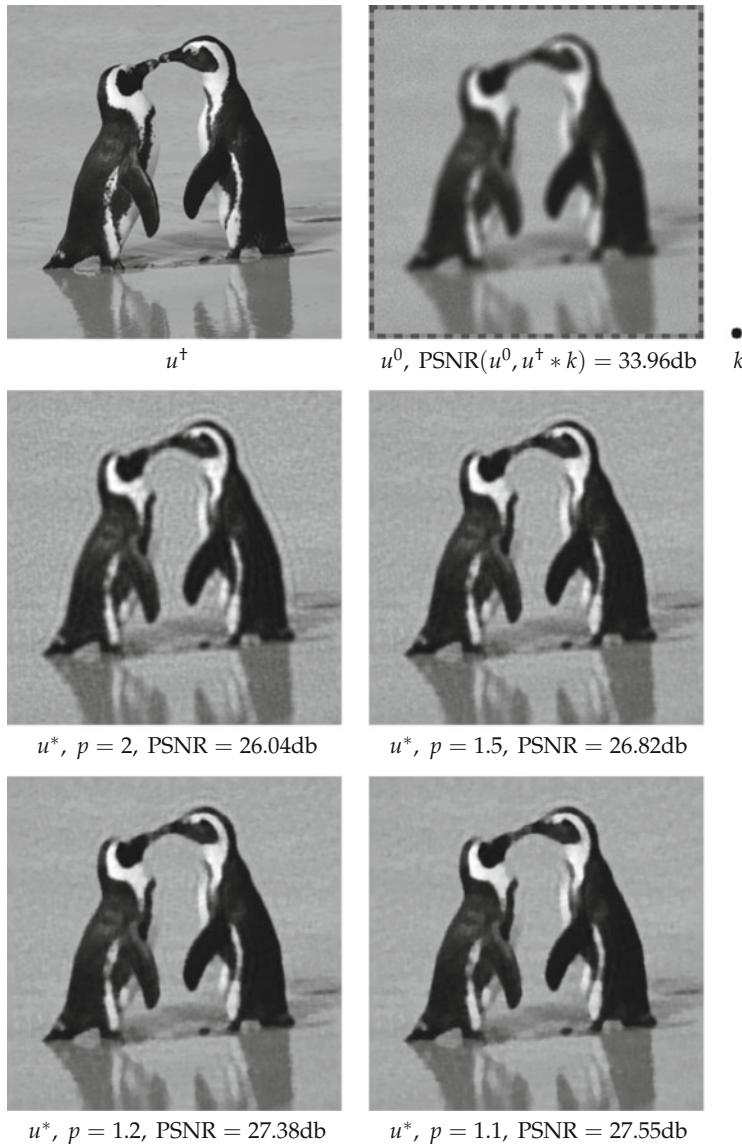


Fig. 6.13 Illustration of the method (6.41) for joint denoising and deblurring. Top: Left the original (320×320 pixels), right the measured data (310×310 pixels) obtained by convolution with an out-of-focus kernel (right, diameter of 11 pixels) and addition of noise. Middle and bottom: The minimizer of (6.41) for $q = 2$ and different exponents p with λ optimized for PSNR. For p close to 1 one sees, similarly to Fig. 6.11, a reduction of noise, fewer oscillating artifacts, and a sharper reconstruction of edges

However, a similar derivation of a maximum principle as in Theorem 6.95 is not possible.

The Euler-Lagrange equations for this problem are obtained similarly to Application 6.94. To find the subgradient of the data term $\Phi = \frac{1}{q} \|A \cdot -u^0\|_q^q$, we need the adjoint A^* (see Theorem 6.51): for some $w \in L^{q^*}(\Omega')$ we have

$$\begin{aligned} \int_{\Omega'} w(x)(u * k)(x) dx &= \int_{\Omega'} \int_{\Omega} u(y)k(x-y)w(x) dy dx \\ &= \int_{\Omega} \int_{\Omega'} w(x)k(x-y) dx u(y) dy \\ &= \int_{\Omega} (w * D_{-\text{id}}k)(y)u(y) dy, \end{aligned}$$

i.e., the operator A^* amounts to a zero-padding of w followed by convolution with the reflected kernel $\bar{k} = D_{-\text{id}}k$. If we introduce an additional variable for the elements of $\partial\Phi$, we can characterize the minimizer u^* by

$$\begin{aligned} v^* - \lambda \operatorname{div}(|\nabla u^*|^{p-2}\nabla u^*) &= 0, \\ (|u^* * k - u^0|^{q-2}(u^* * k - u^0)) * \bar{k} &= v^* \quad \text{in } \Omega, \\ |\nabla u^*|^{p-2}\nabla u^* \cdot v &= 0 \quad \text{on } \partial\Omega. \end{aligned} \tag{6.42}$$

In contrast to (6.40), we cannot see u^* as the solution of a partial differential equation: the system (6.42) contains a partial differential equation with a p -Laplace operator, and also an equation with an integral operator (to be more precise, a convolution). To analyze properties of u^* one can focus on the p -Laplace operator.

If we assume that $k \in L^q(\Omega_0)$, for example, then $v^* \in C(\overline{\Omega})$, since v^* is the convolution of $|u^* * k - u^0|^{q-2}(u^* * k - u^0) \in L^{q^*}(\Omega')$ and $\bar{k} \in L^q(-\Omega_0)$, and by Theorem 3.14 it is continuous. Similar to the analysis of the solution of (6.40) we can, for $p \leq 2$, deduce from the results of [131] that $u^* \in H^{2,p}(\Omega'')$ for all $\overline{\Omega''} \subset\subset \Omega$. Similarly, for $d = 2$, we know that every solution u^* is continuous, i.e. we expect qualitatively similar properties to those in the denoising case of (6.39).

This is confirmed by the numerical examples of this deconvolution method in Fig. 6.13. There, we fix $q = 2$ and vary p . For $p = 2$ one notes, besides the influence of the noise, also artifacts of the deconvolution: the solution oscillates in the neighborhood of large variations of contrast, especially along the contour of the penguins. Intuitively, this can be explained by the absence of a maximum principle: if such a principle held, these oscillations would lead to “overshooting” and hence would give a contradiction. For p close to 1 we again see a qualitative change of the solution similar to Fig. 6.11.

To enforce a maximum principle, we may employ, for u^0 with $L \leq u^0 \leq R$, the bounds $L \leq u \leq R$ simply by adding the respective indicator functional I_K . It is simple to prove the existence of minimizers in $L^q(\Omega)$ (Exercise 6.28), but the

derivation of optimality conditions leads to some difficulties: the subgradient of ∂I_K has to be seen as a subset of $H^{1,p}(\Omega)^*$ and not, as before, as a subset of $L^{q^*}(\Omega)$.

Finally, we point out the generalization to penalties with higher order, i.e., $m \geq 2$ (Exercise 6.29).

Application 6.98 (Inpainting with Sobolev Penalty) Now we turn to the reconstruction of missing image parts, i.e., to inpainting. Denote by Ω a bounded Lipschitz domain and by $\Omega' \subset \Omega$ a bounded Lipschitz subdomain with $\overline{\Omega'} \subset\subset \Omega$, on which we want to reconstruct an image that we know only on $\Omega \setminus \Omega'$. For simplicity we assume that we want to reconstruct only on a connected set Ω' ; the more general case of a union of finitely many connected Lipschitz subdomains is obvious.

Our model for the images is the Sobolev space $H^{1,p}(\Omega)$ with $p \in]1, \infty[$; we assume that the “true” image u^\dagger is in $H^{1,p}(\Omega)$ and also assume that there exists $u^0 \in H^{1,p}(\Omega)$ for which we know that $u^\dagger = u^0$ almost everywhere in $\Omega \setminus \Omega'$. The task of variational inpainting is to find an extension u^* to Ω with smallest Sobolev seminorm, i.e. the solution of

$$\min_{u \in L^q(\Omega)} \frac{1}{p} \int_{\Omega} |\nabla u|^p dx + I_{\{v \in L^q(\Omega) \mid v|_{\Omega \setminus \Omega'} = u^0|_{\Omega \setminus \Omega'}\}}(u) \quad (6.43)$$

for some $q \in]1, \infty[$ with $q \leq d/(d-p)$ if $p < d$. This is a minimization problem of a different type from the previously considered Tikhonov functionals, but the situation is still covered by Theorem 6.84. The set $K = \{v \in L^q(\Omega) \mid v = u^0 \text{ almost everywhere in } \Omega \setminus \Omega'\}$ is convex and bounded and has nonempty intersection with $H^{1,p}(\Omega)$, the corresponding indicator functional $\Phi = I_K$ has the needed properties, except coercivity. However, coercivity follows from the fact that for $u \in K$ and $v \in \Pi^1$ with $v \neq 0$ it is always the case that $u + v \notin K$. Hence, with $\varphi(t) = \frac{1}{p}t^p$, we have all assumptions of Theorem 6.84 satisfied and the existence of a minimizer u^* is guaranteed.

Since φ is strictly convex, solutions may differ only in Π^1 , but this is not allowed by the definition of K . Hence, the minimizer u^* is unique.

We study properties of the minimizer of (6.43). It is easy to see that u^* satisfies a maximum principle: if $L \leq u^0 \leq R$ almost everywhere in $\Omega \setminus \Omega'$, then $L \leq u^* \leq R$ has to hold almost everywhere in Ω , since one can use arguments similar those in the proof of Theorem 6.95 that the function $u = \min(R, \max(L, u^*))$ is also a minimizer. By uniqueness we obtain $u^* = u$, as desired.

If we derive the optimality condition with the help of subdifferentiation, we face a problem: the indicator functional for the constraint $v = u^0$ almost everywhere in $\Omega \setminus \Omega'$ is not continuous in $L^q(\Omega)$, as well as the p th power of the seminorm $\|\nabla u\|_p^p$. This poses difficulties in the applicability of Theorem 6.51. To prove additivity of the subdifferential nonetheless, we derive an equivalent version of the problem (6.43). The following lemma will be useful for that purpose.

Lemma 6.99 *Let Ω be a bounded Lipschitz domain, $m \geq 1$, and $p \in [1, \infty[$.*

1. *If, for $u \in H^{m,p}(\mathbf{R}^d)$ the identity $u|_{\mathbf{R}^d \setminus \Omega} = 0$ holds, then u is obtained by zero padding of some $u^0 \in H_0^{m,p}(\Omega)$.*
2. *If, for some $u \in H^{m,p}(\Omega)$, the traces of $u, \nabla u, \dots, \nabla^{m-1} u$ vanish on $\partial\Omega$, then $u \in H_0^{m,p}(\Omega)$.*

The *proof* uses the approximation arguments we already used in Theorems 6.74 and 6.88 and is a simple exercise (see Exercise 6.30).

Now consider $u \in H^{1,p}(\Omega)$ with $u = u^0$ almost everywhere in $\Omega \setminus \Omega'$. Then the extension of $v = u - u^0$ by zero is in $H^{1,p}(\mathbf{R}^d)$ with $v = 0$ almost everywhere outside of Ω' . By Lemma 6.99 we obtain that $v \in H_0^{1,p}(\Omega')$, and thus the traces of u and u^0 coincide on $\partial\Omega'$. If, conversely, for $u \in H^{1,p}(\Omega')$, the trace of u equals the trace of u^0 on $\partial\Omega'$, then Lemma 6.99 implies that $u - u^0 \in H_0^{1,p}(\Omega)$ has to hold. Thus, the extension of u by u^0 outside of Ω' is in $H^{1,p}(\Omega)$. We have shown that

$$\begin{aligned} \{u \in H^{1,p}(\Omega) \mid u|_{\Omega \setminus \Omega'} &= u^0|_{\Omega \setminus \Omega'}\} \\ &= \{u \in L^q(\Omega) \mid u|_{\Omega \setminus \Omega'} = u^0|_{\Omega \setminus \Omega'}, (u - u^0)|_{\Omega'} \in H_0^{1,p}(\Omega')\} \\ &= \{u \in L^q(\Omega) \mid u|_{\Omega \setminus \Omega'} = u^0|_{\Omega \setminus \Omega'}, u|_{\Omega'} \in H^{1,p}(\Omega'), u|_{\partial\Omega'} = u^0|_{\partial\Omega'}\}, \end{aligned}$$

where we have understood the restriction onto $\partial\Omega'$ as taking the trace with respect to Ω' . This motivates the definition of a linear map ∇_0 from $L^q(\Omega')$ to $L^p(\Omega')$:

$$\text{dom } \nabla_0 = H_0^{1,p}(\Omega') \subset L^q(\Omega'), \quad \nabla_0 u = \nabla u \text{ in } L^p(\Omega').$$

The space $L^q(\Omega')$ contains $H_0^{1,p}(\Omega')$ as a dense subspace, and thus ∇_0 is densely defined. The map is also closed: To see this, let $u^n \in H_0^{1,p}(\Omega')$ with $u^n \rightarrow u$ in $L^q(\Omega')$ and $\nabla_0 u^n \rightarrow v$ in $L^p(\Omega', \mathbf{R}^d)$. By Lemma 6.73 we see that $v = \nabla u$, and it remains to show that $u \in H_0^{1,p}(\Omega)$. To that end, note that $Q_1 u^n \rightarrow Q_1 u$ in $L^p(\Omega)$ by the equivalence of norms and by the Poincaré-Wirtinger inequality (see (6.35)) we get $\|P_1(u^n - u)\|_p \leq C \|\nabla(u^n - u)\|_p \rightarrow 0$ for $n \rightarrow \infty$. Hence, we get $u^n \rightarrow u$ in $H^{1,p}(\Omega')$, and since $H_0^{1,p}(\Omega')$ is a closed subspace, we get $u \in H_0^{1,p}(\Omega')$. This shows that ∇_0 is closed.

This allows us to reformulate problem (6.43) equivalently as

$$\min_{u \in L^q(\Omega)} \frac{1}{p} \int_{\Omega} |\nabla_0((u - u^0)|_{\Omega'}) + \nabla u^0|^p dx + I_{\{v \in L^q(\Omega) \mid v|_{\Omega \setminus \Omega'} = u^0|_{\Omega \setminus \Omega'}\}}(u), \quad (6.44)$$

where we implicitly extend the image of ∇_0 by 0 to all of Ω . The difference between this and the formulation in (6.43) is that the functional

$$F_1(u) = \frac{1}{p} \int_{\Omega} |\nabla_0((u - u^0)|_{\Omega'}) + \nabla u^0|^p dx$$

is continuous on the affine subspace $u^0 + X_1$, $X_1 = \{v \in L^q(\Omega) \mid v|_{\Omega'} = 0\}$. Also

$$F_2(u) = I_{\{v \in L^q(\Omega) \mid v|_{\Omega \setminus \Omega'} = u^0|_{\Omega \setminus \Omega'}\}}(u)$$

is continuous on the subspace $u^0 + X_2$, $X_2 = \{v \in L^q(\Omega) \mid v|_{\Omega \setminus \Omega'} = 0\}$. Since the restrictions $P_1 : u \mapsto u \chi_{\Omega \setminus \Omega'}$ and $P_2 : u \mapsto u \chi_{\Omega'}$, respectively, are continuous and they sum to the identity, we can apply the result of Exercise 6.14 and get $\partial(F_1 + F_2) = \partial F_1 + \partial F_2$.

If we denote by $A : L^q(\Omega) \rightarrow L^q(\Omega')$ the restriction to Ω' and by $E : L^p(\Omega', \mathbf{R}^d) \rightarrow L^p(\Omega, \mathbf{R}^d)$ the zero padding, we get

$$F_1 = \frac{1}{p} \|\cdot\|_p^p \circ T_{\nabla u^0} \circ E \circ \nabla_0 \circ A \circ T_{-u^0}.$$

The map A is surjective, and by the results of Exercises 6.11 and 6.12 for A and ∇_0 , respectively, as well as Theorem 6.51 for T_{-u^0} , E , and $T_{\nabla u^0}$, the subgradient satisfies

$$\partial F_1(u) = \begin{cases} \{A^* \nabla_0^* E^* J_p(\nabla u^0 + \nabla(u - u^0)|_{\Omega'})\} & \text{if } (u - u^0)|_{\Omega'} \in H_0^{1,p}(\Omega), \\ \emptyset & \text{otherwise,} \end{cases}$$

with $J_p(w) = w|w|^{p-2}$ for $w \in L^p(\Omega, \mathbf{R}^d)$. It is easy to see that E^* is a restriction onto Ω' (in the respective spaces), and similarly, A^* is a zero padding. Finally, we calculate ∇_0^* : If $w \in \text{dom } \nabla_0^* \subset L^{p^*}(\Omega')$, then for all $u \in \mathcal{D}(\Omega')$

$$\int_{\Omega'} w \cdot \nabla u \, dx = \int_{\Omega'} w \cdot \nabla_0 u \, dx = \int_{\Omega'} (\nabla_0^* w) u \, dx,$$

and this shows that $\nabla_0^* w = -\operatorname{div} w$ in the sense of the weak divergence. Conversely, let $w \in L^{p^*}(\Omega', \mathbf{R}^d)$ such that $-\operatorname{div} w \in L^{q^*}(\Omega')$. Then by the definition of $H_0^{1,p}(\Omega')$, we can choose for every $u \in H_0^{1,p}(\Omega')$ a sequence (u^n) in $\mathcal{D}(\Omega')$ such that $u^n \rightarrow u$ in $L^p(\Omega')$ as well as $\nabla u^n \rightarrow \nabla u$ in $L^p(\Omega', \mathbf{R}^d)$. Thus,

$$\int_{\Omega'} w \cdot \nabla u \, dx = \lim_{n \rightarrow \infty} \int_{\Omega'} w \cdot \nabla u^n \, dx = - \lim_{n \rightarrow \infty} \int_{\Omega'} (\operatorname{div} w) u^n \, dx = - \int_{\Omega'} (\operatorname{div} w) u \, dx,$$

and hence $w \in \text{dom } \nabla_0^*$ and $\nabla_0^* w = -\operatorname{div} w$. We have shown that

$$\nabla_0^* = -\operatorname{div}, \quad \text{dom } \nabla_0^* = \{w \in L^{p^*}(\Omega', \mathbf{R}^d) \mid \operatorname{div} w \in L^{q^*}(\Omega')\}$$

in other words, the adjoint of the gradient with zero boundary conditions is the weak divergence. In contrast to ∇^* , ∇_0^* operates on all vector fields for which the weak divergence exists and not only on those for which the normal trace vanishes at the boundary (cf. Theorem 6.88 and Remark 6.89).

For the subgradients of F_1 we get, using the convention that gradient and divergence are considered on Ω' and the divergence will be extended by zero, that

$$\partial F_1(u) = \begin{cases} \{-\operatorname{div}(\nabla(u|_{\Omega'})|\nabla(u|_{\Omega'})|^{p-2})\} & \text{if } (u - u^0)|_{\Omega'} \in H_0^{1,p}(\Omega'), \\ \emptyset & \text{otherwise.} \end{cases}$$

The calculation of the subgradient of the constraint $F_2 = I_K$ is simple: Using the subspace X_2 , defined above, we can write $K = u^0 + X_2$, and by Theorem 6.51 and Example 6.48, we get

$$\partial F_2(u) = \begin{cases} X_2^\perp & \text{if } u|_{\Omega \setminus \Omega'} = u^0|_{\Omega \setminus \Omega'}, \\ \emptyset & \text{otherwise,} \end{cases} \quad X_2^\perp = \{w \in L^{q^*}(\Omega) \mid w|_{\Omega'} = 0\}. \quad (6.45)$$

The optimality conditions for the minimizer u^* of (6.44) are

$$0 \in -\operatorname{div}(\nabla(u^*|_{\Omega})|\nabla(u^*|_{\Omega'})|^{p-2}) + X_2^\perp \quad \text{with} \quad \begin{cases} (u^* - u^0)|_{\Omega'} \in H_0^{1,p}(\Omega'), \\ u^*|_{\Omega \setminus \Omega'} = u^0|_{\Omega \setminus \Omega'}. \end{cases}$$

Since the divergence of $\Omega \setminus \Omega'$ is extended by zero and $(u^* - u^0)|_{\Omega'} \in H_0^{1,p}(\Omega')$ if and only if $u^*|_{\Omega'} \in H^{1,p}(\Omega')$ with $u^*|_{\partial\Omega'} = u^0|_{\partial\Omega'}$ in the sense of the trace, we conclude the characterization

$$\begin{aligned} u^* &= u^0 && \text{in } \Omega \setminus \Omega', \\ -\operatorname{div}(\nabla u^*|\nabla u^*|^{p-2}) &= 0 && \text{in } \Omega', \\ u^* &= u^0 && \text{on } \partial\Omega'. \end{aligned} \quad (6.46)$$

Note that the last equality has to be understood in the sense of the trace of u^0 on $\partial\Omega'$ with respect to Ω' . In principle, this could depend on the values of u^0 in the inpainting domain Ω' . However, it is simple to see that the traces of $\partial\Omega'$ with respect to Ω' and $\Omega \setminus \Omega'$ coincide for Sobolev functions $u^0 \in H^{1,p}(\Omega)$. Hence, the solution of the inpainting problem is independent of the auxiliary function u^0 .

Again, the optimality conditions (6.46) show that u^* has to be locally smooth in Ω' : by the same argument as in Applications 6.94 and 6.97, we get $u^* \in H^{2,p}(\Omega')$ for all $\overline{\Omega''} \subset\subset \Omega'$, if $p < 2$. For $p = 2$ we even get that the solution u^* in Ω' is harmonic (Example 6.4), and hence, $u^* \in C^\infty(\Omega')$. The case $p > 2$ is treated in some original papers (see, e.g., [60] and the references therein) and at least gives $u^* \in C^1(\Omega')$. Thus, the two-dimensional case ($d = 2$) always leads to continuous solutions, and this says that this method of inpainting is suited only for the reconstruction of homogeneous regions; see also Fig. 6.14.

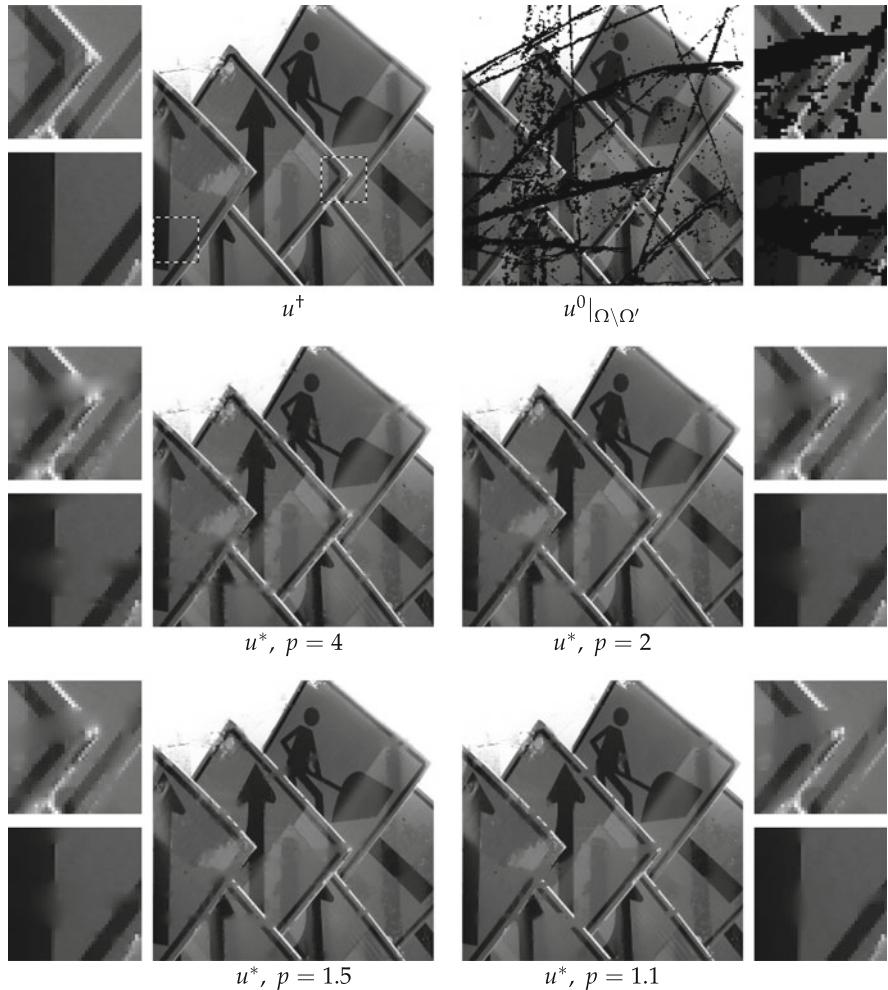


Fig. 6.14 Inpainting by solving (6.43) or (6.44), respectively. Top: Left the original u^\dagger together with two enlarged details (according to the marked regions in the image), right the given u^0 on $\Omega \setminus \Omega'$ (Ω' is given by the black region), again with details. Middle and bottom: The minimizer of the inpainting functional for different p together with enlarged details. While the reconstruction of homogeneous regions is good, edges get blurred in general. As the details show, this effect is more prominent for larger p ; on the other hand, for $p = 1.1$ some edges are extended in a sharp way (the edge of the arrow in the left, lower detail), but the geometry is not always reconstructed correctly (disconnected boundaries of the border of the sign in the upper right detail)

Application 6.100 (Variational Interpolation/Zooming) We again consider the problem to generate a continuous image $u^* : \Omega \rightarrow \mathbf{R}$, $\Omega =]0, N[\times]0, M[\subset \mathbf{R}^2$ from a discrete one $U^0 : \{1, \dots, N\} \times \{1, \dots, M\} \rightarrow \mathbf{R}$. In Sect. 3.1.1 we have already seen an efficient method to do so. While that method uses evaluation of the continuous image at specific points, we show a variational approach in the following.

To begin with, we assume that we are given a map $A : L^q(\Omega) \rightarrow \mathbf{R}^{N \times M}$, $q \in]1, \infty[$, that maps a continuous image to the discretely sampled points. In addition, we assume that A is linear, continuous, and surjective. Moreover, A should not map constant functions to zero, which would indeed not be appropriate, since A is a kind of “restriction operator.”

These assumptions on A imply immediately that $u \mapsto (Au)_{i,j}$ is an element in the dual space of $L^q(\Omega)$, and this means that there are functions $w^{i,j} \in L^{q^*}(\Omega)$ such that

$$(Au)_{i,j} = \langle w^{i,j}, u \rangle_{L^{q^*} \times L^q} = \int_{\Omega} w^{i,j} u \, dx$$

holds for all $u \in L^q(\Omega)$. The surjectivity of A is equivalent to the linear independence of the $w^{i,j}$; the assumption that constant functions are not in the kernel of A can be expressed with the vector $\bar{w} \in \mathbf{R}^{N \times M}$, defined by $\bar{w}_{i,j} = \int_{\Omega} w^{i,j} \, dx$, simply as $\bar{w} \neq 0$. In view of the above, the choice $w^{i,j}(x_1, x_2) = k(i - x_1, j - x_2)$ with suitable $k \in L^{q^*}(\mathbf{R}^2)$ seems natural. This amounts to a convolution with subsequent point sampling, and hence k should be a kind of low-pass filter, see Sect. 4.2.3. It is not hard to check that for example, $k = \chi_{]0, 1[\times]0, 1[}$ satisfies the assumptions for the map A . In this case the map A is nothing else than averaging u over the squares $]i - 1, i[\times]j - 1, j[$. Using $k(x_1, x_2) = \text{sinc}(x_1 - \frac{1}{2}) \text{sinc}(x_2 - \frac{1}{2})$ leads to the perfect low-pass filter for the sampling rate 1 with respect to the midpoints of the squares $(i - \frac{1}{2}, j - \frac{1}{2})$; see Theorem 4.35.

The assumption that the image u is an interpolation for the data U_0 is now expressed as $Au = U^0$. However, this is true for many images; indeed it is true for an infinite-dimensional affine subspace of $L^q(\Omega)$. We aim to find the image that is best suited for a given image model. Again we use the Sobolev space $H^{1,p}(\Omega)$ for $p \in]1, \infty[$ as a model, where we assume that $q \leq 2/(2-p)$ holds if $p < 2$. This leads to the minimization problem

$$\min_{u \in L^q(\Omega)} \frac{1}{p} \int_{\Omega} |\nabla u|^p \, dx + I_{\{v \in L^q(\Omega) \mid Av = U^0\}}(u). \quad (6.47)$$

Let us check for the existence of a solution using Theorem 6.84 and set, similar to Application 6.98, $\Phi = I_K$, this time with $K = \{u \in L^q(\Omega) \mid Au = U^0\}$. We check the assumption on Φ .

It is obvious that there is some $u^0 \in L^q(\Omega)$ such that $Au^0 = U^0$. However, we want to show the existence of some $u^1 \in H^{1,p}(\Omega)$ with this property. To that end, we note that $A : H^{1,p}(\Omega) \rightarrow \mathbf{R}^{N \times M}$ is well defined, linear and continuous

by the embedding $H^{1,p}(\Omega) \hookrightarrow L^q(\Omega)$ (see Theorem 6.76). Now assume that $A : H^{1,p}(\Omega) \rightarrow \mathbf{R}^{N \times M}$ is not surjective. Since $\mathbf{R}^{N \times M}$ is finite-dimensional, the image of A is a closed subspace, and hence it must be that $\|Au - U^0\| \geq \varepsilon$ for all $u \in H^{1,p}(\Omega)$ and some $\varepsilon > 0$. However, $H^{1,p}(\Omega)$ is dense in $L^q(\Omega)$, and hence there has to be some $\bar{u} \in H^{1,p}(\Omega)$ with $\|\bar{u} - u^0\|_q < \frac{\varepsilon}{2}\|A\|^{-1}$, and thus

$$\|A\bar{u} - U^0\| = \|A\bar{u} - Au^0\| \leq \|A\|\|\bar{u} - u^0\|_q < \varepsilon,$$

which is a contradiction. Hence, the operator A has to map $H^{1,p}(\Omega)$ onto $\mathbf{R}^{N \times M}$, and thus there is some $u^1 \in H^{1,p}(\Omega)$ with $Au^1 = U^0$. In particular Φ is proper on $H^{1,p}(\Omega)$.

Convexity of Φ is obvious, and the lower semicontinuity follows from the continuity of A and the representation $K = A^{-1}(\{U^0\})$. Finally, we assumed that A does not map constant functions to zero, i.e., for $u \in K$ and $v \in \Pi^1$ with $v \neq 0$, we have $u + v \notin K$. This shows the needed coercivity of Φ . We conclude with Theorem 6.84 that there exists a minimizer u^* of the functional in (6.47), and one can argue along the lines of Application 6.98 that it is unique.

If we try to apply convex analysis to study the minimizer u^* , we face similar problems to the one in Application 6.98: The restriction Φ is not continuous, and hence we cannot use the sum rule for subdifferentials without additional work. However, there is a remedy.

To see this, we note that by surjectivity of $A : H^{1,p}(\Omega) \rightarrow \mathbf{R}^{N \times M}$ there are NM linear independent vectors $u^{i,j} \in H^{1,p}(\Omega)$ ($1 \leq i \leq N$ and $1 \leq j \leq M$) such that the restriction of A to $V = \text{span}(u^{i,j}) \subset L^q(\Omega)$ is bijective. Hence, there exists A_V^{-1} and for $T_1 = A_V^{-1}A$, one has $T_1^2 = T_1$ and $\ker(T_1) = \ker(A)$. For $T_2 = \text{id} - T_1$ this implies that $\text{rg}(T_2) = \ker(A)$. With the above u^1 the function

$$u \in V : \quad u \mapsto \frac{1}{p} \int_{\Omega} |\nabla(u^1 + u)|^p \, dx$$

maps to \mathbf{R} and is continuous in the subspace topology (Theorem 6.25). Similarly we see the continuity of $u \mapsto I_K(u^1 + u) = 0$ for $u \in \text{rg}(T_2)$ in the subspace topology. Hence, the sum rule for subgradients holds in this case (again, see Exercise 6.14); some u^* is a solution of (6.47) if and only if $0 \in \partial\left(\frac{1}{p}\|\cdot\|_p^p \circ \nabla\right)(u^*) + \partial I_K(u^*)$. The first term is again the p -Laplace operator, while for the second we have

$$\partial I_K(u) = \begin{cases} \ker(A)^\perp & \text{if } Au = U^0, \\ \emptyset & \text{otherwise.} \end{cases}$$

Since A is surjective, the closed range theorem (Theorem 2.26) implies that $\ker A^\perp = \text{rg}(A^*) = \text{span}(w^{i,j})$, the latter since $w^{i,j} = A^*e^{i,j}$ for $1 \leq i \leq N$ and $1 \leq j \leq M$. Hence, the optimality conditions say that $u^* \in L^q(\Omega)$ is a solution

of (6.47) if there exists some $\lambda^* \in \mathbf{R}^{N \times M}$ such that

$$\begin{aligned} -\operatorname{div}(|\nabla u^*|^{p-2}\nabla u^*) &= \sum_{i=1}^N \sum_{j=1}^M \lambda_{i,j}^* w^{i,j} && \text{in } \Omega, \\ |\nabla u^*|^{p-2}\nabla u^* \cdot v &= 0 && \text{on } \partial\Omega, \\ \int_{\Omega} u^* w^{i,j} dx &= U_{i,j}^0 && \begin{aligned} 1 \leq i \leq N, \\ 1 \leq j \leq M. \end{aligned} \end{aligned} \tag{6.48}$$

The components of λ^* can be seen as Lagrange multipliers for the constraint $Au = U^0$ (cf. Example 6.48). In case of optimality of u^* , the λ^* obey another constraint: if we integrate the first equation in (6.48), similar to Example 6.93, on both sides we get

$$\sum_{i=1}^N \sum_{j=1}^M \lambda_{i,j}^* \int_{\Omega} w^{i,j} dx = \int_{\Omega} \nabla^* (|\nabla u^*|^{p-2}\nabla u^*) dx = 0,$$

i.e., $\lambda^* \cdot \bar{w} = 0$ with the above defined vector of integrals \bar{w} . Similar to the previous applications one can use some theory for the p -Laplace equation to show that the solution u^* has to be continuous if the functions $w^{i,j}$ are in $L^\infty(\Omega)$.

While (6.48) is a nonlinear partial differential equation for $p \neq 2$ that is coupled with linear equalities and the Lagrange multipliers, the case $p = 2$ leads to a linear system of equalities. This can be solved as follows. In the first step, we solve for every (i, j) the equation

$$\begin{cases} -\Delta z^{i,j} = w^{i,j} - \frac{1}{|\Omega|} \int_{\Omega} w^{i,j} dx & \text{in } \Omega, \\ \nabla z^{i,j} \cdot v = 0 & \text{on } \partial\Omega, \\ \int_{\Omega} z^{i,j} = 0. \end{cases} \tag{6.49}$$

Such $z^{i,j} \in H^1(\Omega)$ exist and are unique, see Example 6.93. Then we plug these into the optimality system and set $\lambda_0^* = |\Omega|^{-1} \int_{\Omega} u^* dx$, which then leads to

$$\begin{aligned} \int_{\Omega} u^* w^{i,j} dx &= \int_{\Omega} u^* w^{i,j} dx - \frac{1}{|\Omega|} \int_{\Omega} u^* dx \int_{\Omega} w^{i,j} dx + \frac{1}{|\Omega|} \int_{\Omega} u^* dx \int_{\Omega} w^{i,j} dx \\ &= \int_{\Omega} u^* (-\Delta z^{i,j}) dx + \lambda_0^* \bar{w}_{i,j} = \int_{\Omega} (-\Delta u^*) z^{i,j} dx + \lambda_0^* \bar{w}_{i,j} \\ &= \sum_{k=1}^N \sum_{l=1}^M \lambda_{k,l}^* \int_{\Omega} w^{k,l} z^{i,j} dx + \lambda_0^* \bar{w}_{i,j} = U_{i,j}^0. \end{aligned}$$

Here we used that $\nabla z^{i,j}$ and ∇u^* are contained in $\text{dom } \nabla^*$. We can simplify the scalar product further: using $\int_{\Omega} z^{i,j} dx = 0$, we get

$$\begin{aligned}\int_{\Omega} w^{k,l} z^{i,j} dx &= \int_{\Omega} \left(w^{k,l} - \frac{1}{|\Omega|} \int_{\Omega} w^{k,l} dy \right) z^{i,j} dx = \int_{\Omega} (-\Delta z^{k,l}) z^{i,j} dx \\ &= \int_{\Omega} \nabla z^{k,l} \cdot \nabla z^{i,j} dx.\end{aligned}$$

Setting $S_{(i,j),(k,l)} = \int_{\Omega} \nabla z^{k,l} \cdot \nabla z^{i,j} dx$ and using the constraint $\lambda^* \cdot \bar{w} = 0$, we obtain the finite-dimensional linear system of equations

$$\begin{cases} S\lambda^* + \bar{w}\lambda_0^* = U^0, \\ \bar{w}^T \lambda^* = 0, \end{cases} \quad (6.50)$$

for the Lagrange multipliers. This system has a unique solution (see Exercise 6.31), and hence λ^* and λ_0^* can be computed. The identity for $-\Delta u^*$ in (6.48) gives uniqueness of u^* up to constant functions, and the constant offset is determined by λ_0^* , namely

$$u^* = \sum_{i=1}^N \sum_{j=1}^M \lambda_{i,j}^* z^{i,j} + \lambda_0^* \mathbf{1}, \quad (6.51)$$

where $\mathbf{1}$ denotes the function that is equal to 1 on Ω . In conclusion, the method for H^1 interpolation reads as follows

1. For all (i, j) solve Eq. (6.49).
2. Calculate the matrix $S_{(i,j),(k,l)} = \int_{\Omega} \nabla z^{k,l} \cdot \nabla z^{i,j} dx$ and the vector $\bar{w}_{i,j} = \int_{\Omega} w^{i,j} dx$, and solve the linear system (6.50).
3. Calculate the solution u^* by plugging λ^* and λ_0^* into (6.51).

In practice, the solution of (6.47) and (6.48) is done numerically, i.e., the domain Ω is discretized accordingly. This allows one to obtain images of arbitrary resolution from the given image u^0 . Figure 6.15 compares the variational interpolation with the classical methods from Sect. 3.1.1. One notes that variational interpolation deals favorably with strong edges. Figure 6.16 illustrates the influence on the sampling operator A . The choice of this operator is especially important if the data U^0 does not exactly match the original image u^\dagger . In this case, straight edges are not reconstructed exactly, but change their shape in dependence of the sampling operator (most prominently seen for p close to 1).

One possibility to reduce these unwanted effects is to allow for $Au \neq U^0$. For example, one can replace problem (6.47) by the minimization of the Tikhonov functional

$$\min_{u \in L^q(\Omega)} \frac{\|Au - U^0\|_2^2}{2} + \frac{\lambda}{p} \int_{\Omega} |\nabla u|^p dx$$

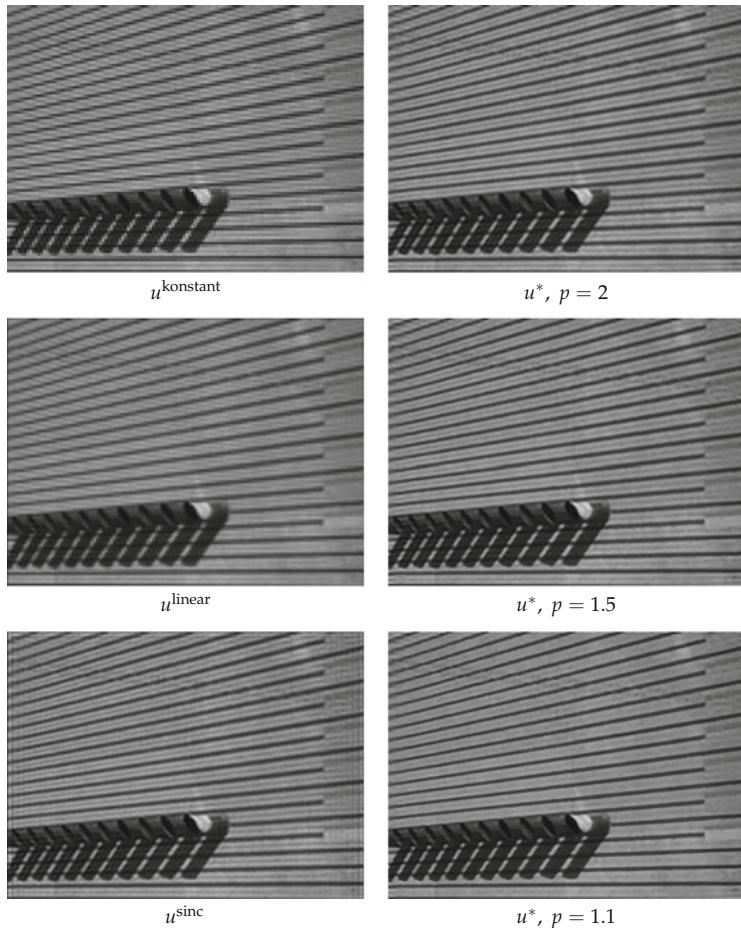


Fig. 6.15 Comparison of classical methods for interpolation and variational interpolation for eightfold zooming. Left column: the interpolation methods from Sect. 3.1.1 of an image with 128×96 pixels to 1024×768 pixels. Specifically, we show the result of constant interpolation (top), piecewise bilinear interpolation (middle), and tensor product interpolation with the sinc function (bottom). Right column: The minimizers of the variational interpolation method (6.47) for different exponents p . The used sampling operator A is the perfect low-pass filter. Constant and bilinear interpolation have problems with the line structures. These are handled satisfactorily, up to oscillation at the boundary, by u^{sinc} , and also u^* with $p = 2$ is comparable. Smaller p allows for sharp and almost perfect reconstruction of the dark lines, but at the expense of smaller details like the lighter lines between the dark lines

for some $\lambda > 0$ (with the norm $\|v\|_2^2 = \sum_{i=1}^N \sum_{j=1}^M |v_{i,j}|^2$ on the finite dimensional space $\mathbf{R}^{N \times M}$). As is generally the case with Tikhonov functionals, increasing λ lowers the influence of data errors with respect to the sampling operator A .

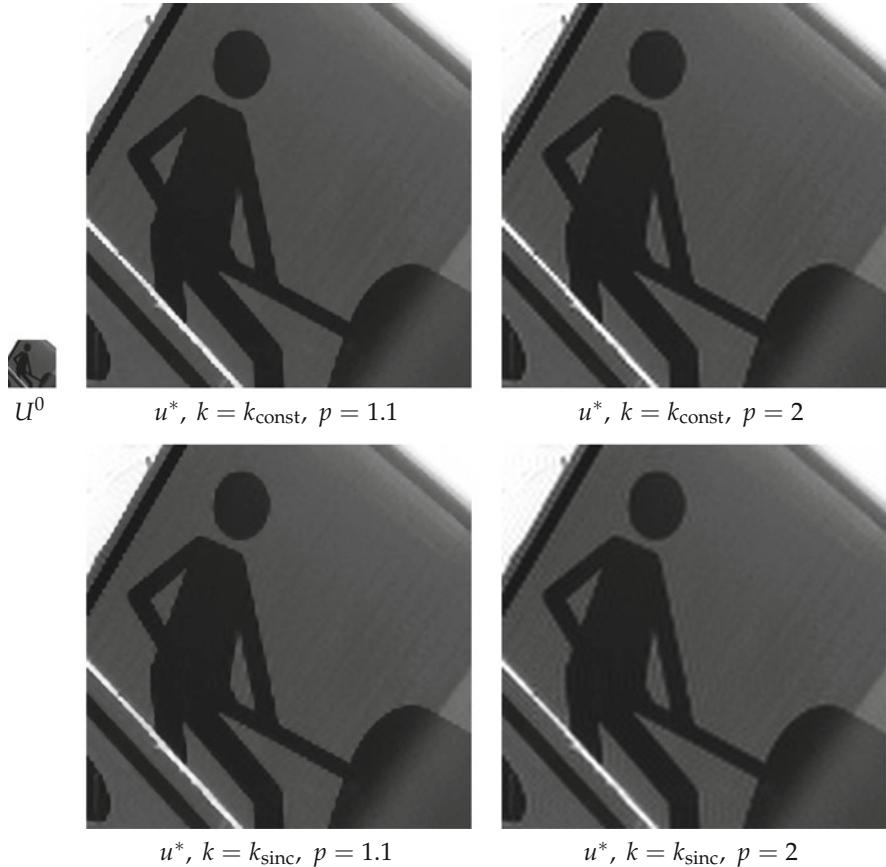


Fig. 6.16 Comparison of different sampling operators A for variational interpolation (6.47). Top left: Given data U^0 (96×96 pixel). Same row: Results of eightfold zooming by averaging over squares (i.e. $k_{\text{const}} = \chi_{[0,1] \times [0,1]}$) for different p . Bottom row: Respective results for sampling with the perfect low-pass filter $k_{\text{sinc}}(x_1, x_2) = \text{sinc}(x_1 - \frac{1}{2}) \text{sinc}(x_2 - \frac{1}{2})$. The operator that averages over squares favors solutions with “square structures”; the images look “blocky” at some places. This effect is not present for the perfect lowpass filter but there are some oscillation artifacts due to the non-locality of the sinc function

6.3.3 The Total Variation Penalty

The applications in the previous section always used $p > 1$. One reason for this constraint was that the image space of ∇^m , $L^p(\Omega, \mathbf{R}^{d^m})$, is a reflexive Banach space, and for F convex, lower semicontinuous and coercive, one also could deduce that $F \circ \nabla^m$ is lower semicontinuous (see Example 6.29). However, the illustrations indicated that $p \rightarrow 1$ leads to interesting effects. On the one hand, edges are more emphasized, and on the other hand, solutions appear to have more “linear” regions,

which is a favorable property for images with homogeneous regions. Hence, the question whether $p = 1$ can be used for a Sobolev penalty suggests itself, i.e. whether $H^{1,1}(\Omega)$ can be used as an image model. Unfortunately, this leads to problems in the direct method:

Theorem 6.101 (Failure of Lower Semicontinuity for the $H^{1,1}$ Semi-norm) *Let $\Omega \subset \mathbf{R}^d$ be a domain and $q \in [1, \infty[$. Then the functional $\Psi : L^q(\Omega) \rightarrow \mathbf{R}_\infty$ given by*

$$\Psi(u) = \begin{cases} \int_{\Omega} |\nabla u| \, dx & \text{if } u \in H^{1,1}(\Omega), \\ \infty & \text{otherwise,} \end{cases}$$

is not lower semicontinuous.

Proof We prove the claim for Ω with $\overline{B_1(0)} \subset\subset \Omega$, and the general case follows by translation and scaling. Let $\Omega' = B_1(0)$ and $u = \chi_{\Omega'}$. Clearly $u \in L^q(\Omega)$. We choose a mollifier $\varphi \in \mathcal{D}(B_1(0))$ and consider $u^n = u * \varphi_{n^{-1}}$ for $n \in \mathbf{N}$ such that $\overline{B_{1+n^{-1}}(0)} \subset\subset \Omega$. Then it holds that $u^n \rightarrow u$ in $L^q(\Omega)$. Moreover, $u^n \in H^{1,1}(\Omega)$ with

$$\nabla u^n(x) = \begin{cases} 0 & \text{if } |x| \leq \frac{n-1}{n} \text{ or } |x| \geq \frac{n+1}{n}, \\ u * \nabla \varphi_{n^{-1}} & \text{otherwise,} \end{cases}$$

where $\nabla \varphi_{n^{-1}}(x) = n^{d+1} \nabla \varphi(nx)$. Young's inequality for convolutions and the identities $|B_r(0)| = r^d |B_1(0)|$ and $(1 - n^{-1})^d = \sum_{k=0}^d \binom{d}{k} (-1)^k n^{-k}$ lead to

$$\begin{aligned} \int_{\Omega} |\nabla u^n| \, dx &\leq \left(\int_{\{\frac{n-1}{n} \leq |x| \leq \frac{n+1}{n}\}} u \, dx \right) \left(\int_{\mathbf{R}^d} |\nabla \varphi_{n^{-1}}| \, dx \right) = n \|\nabla \varphi\|_1 \int_{\{1-n^{-1} \leq |x| \leq 1\}} 1 \, dx \\ &\leq Cn \left(1 - (1 - n^{-1})^d \right) = Cn \left(\sum_{k=1}^d \binom{d}{k} (-1)^{k+1} n^{-k} \right) \leq C \sum_{k=0}^{d-1} \binom{d}{k+1} n^{-k}. \end{aligned}$$

The right-hand side is bounded for $n \rightarrow \infty$, and hence $\liminf_{n \rightarrow \infty} \Psi(u^n) < \infty$.

However, $u \in H^{1,1}(\Omega)$ cannot be true. If it were, we could test with $\phi \in \mathcal{D}(B_1(0))$ for $i = 1, \dots, d$ and get

$$\int_{\Omega} u \frac{\partial \phi}{\partial x_i} \, dx = \int_{\Omega} \frac{\partial \phi}{\partial x_i} \, dx = \int_{\partial \Omega} \varphi v_i \, dx = 0.$$

By the fundamental lemma of the calculus of variations we would get $\nabla u|_{B_1(0)} = 0$. Similarly one could conclude that $\nabla u|_{\Omega \setminus \overline{B_1(0)}} = 0$, i.e., $\nabla u = 0$ almost everywhere in Ω . This would imply that u is constant in Ω (see Lemma 6.79), a contradiction. By definition of Ψ this means $\Psi(u) = \infty$.

In other words, we have found a sequence $u^n \rightarrow u$, with $\Psi(u) > \liminf_{n \rightarrow \infty} \Psi(u^n)$, and hence Ψ is not lower semicontinuous. \square

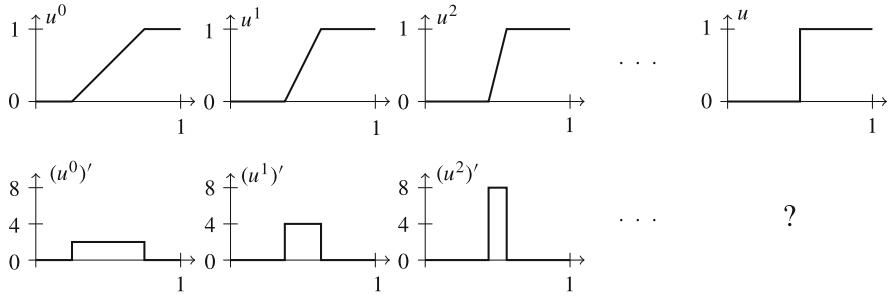


Fig. 6.17 Illustration of a sequence (u^n) in $H^{1,1}(]0, 1[)$, which violates the definition property of lower semicontinuity for $u \mapsto \int_0^1 |u'| dt$ in $L^q(]0, 1[)$. If the ramp part of the functions u^n gets arbitrarily steep while the total increase remains constant, the derivatives $((u^n)')$ are a bounded sequence in $L^1(]0, 1[)$, but the L^q limit u of (u^n) is only in $L^q(]0, 1[)$ and not in $H^{1,1}(]0, 1[)$. In particular, $((u^n)')$ does not converge in $L^1(]0, 1[)$, and one wonders in what sense there is still some limit

See also Fig. 6.17 for an illustration of a slightly different counterexample for the lower semicontinuity of the $H^{1,1}$ semi-norm in dimension one.

The above result prohibits direct generalizations of Theorems 6.84 and 6.86 to the case $p = 1$. If we have a closer look at the proof of lower semicontinuity of $F \circ A$ for $F : Y \rightarrow \mathbf{R}_\infty$ convex, lower semicontinuous, coercive and $A : X \supset \text{dom } A \rightarrow Y$ strongly-weakly closed in Example 6.29 we note that an essential ingredient is to deduce the existence of a weakly convergent subsequence of (Au_n) from the boundedness of $F(Au^n)$. However, this very property fails for ∇ as a strongly-weakly closed operator between $L^q(\Omega)$ and $L^1(\Omega, \mathbf{R}^d)$.

However, we can use this failure as a starting point to define a functional that is, is some sense, a generalization of the integral $\int_\Omega |\nabla u| dx$. More precisely, we replace $L^1(\Omega, \mathbf{R}^d)$ by the space of vector-valued Radon measures $\mathfrak{M}(\Omega, \mathbf{R}^d)$, which is, on the one hand, the dual space of a separable Banach space: by the Riesz-Markov theorem (Theorem 2.62) $\mathfrak{M}(\Omega, \mathbf{R}^d) = C_0(\Omega, \mathbf{R}^d)^*$. On the other hand, $L^1(\Omega, \mathbf{R}^d)$ is isometrically embedded in $\mathfrak{M}(\Omega, \mathbf{R}^d)$ by the map $u \mapsto u\mathcal{L}^d$, i.e., $\|u\mathcal{L}^d\|_{\mathfrak{M}} = \|u\|_1$ for all $u \in L^1(\Omega, \mathbf{R}^d)$ (see Example 2.60).

Since the norm on $\mathfrak{M}(\Omega, \mathbf{R}^d)$ is convex, weakly* lower-semicontinuous and coercive, it is natural to define the weak gradient ∇ on a subspace of $L^q(\Omega)$ with values in $\mathfrak{M}(\Omega, \mathbf{R}^d)$ and to consider the concatenation $\|\cdot\|_{\mathfrak{M}} \circ \nabla$. How can this weak gradient be defined? We simply use the most general notion of derivative that we have, namely the distributional gradient, and claim that this should have a representation as a finite vector-valued Radon measure.

Definition 6.102 (Weak Gradient in $\mathfrak{M}(\Omega, \mathbf{R}^d)$ /Total Variation) Let $\Omega \subset \mathbf{R}^d$ be a domain. Some $\mu \in \mathfrak{M}(\Omega, \mathbf{R}^d)$ is the weak gradient of some $u \in L^1_{\text{loc}}(\Omega)$ if for every $\varphi \in \mathcal{D}(\Omega, \mathbf{R}^d)$

$$\int_\Omega u \operatorname{div} \varphi dx = - \int_\Omega \varphi d\mu.$$

If it exists, we denote $\mu = \nabla u$ and call its norm, denoted by $\text{TV}(u) = \|\nabla u\|_{\mathfrak{M}}$, the *total variation* of u . If there does not exist a $\mu \in \mathfrak{M}(\Omega, \mathbf{R}^d)$ such that $\mu = \nabla u$, we define $\text{TV}(u) = \infty$.

It turns out that this definition is useful, and we can deduce several pleasant properties.

Lemma 6.103 *Let Ω be a domain and $q \in [1, \infty[$. Then for $u \in L^q(\Omega)$ one has the following.*

1. *If there exists a $\mu \in \mathfrak{M}(\Omega, \mathbf{R}^d)$ with $\mu = \nabla u$ as in Definition 6.102, it is unique.*
2. *It holds $\nabla u \in \mathfrak{M}(\Omega, \mathbf{R}^d)$ with $\|\nabla u\|_{\mathfrak{M}} \leq C$ if and only if for all $\varphi \in \mathcal{D}(\Omega, \mathbf{R}^d)$,*

$$\int_{\Omega} u \operatorname{div} \varphi \, dx \leq C \|\varphi\|_{\infty}.$$

In particular, we obtain

$$\text{TV}(u) = \sup \left\{ \int_{\Omega} u \operatorname{div} \varphi \, dx \mid \varphi \in \mathcal{D}(\Omega, \mathbf{R}^d), \|\varphi\|_{\infty} \leq 1 \right\}. \quad (6.52)$$

3. *If a sequence (u^n) converges to u in $L^q(\Omega)$ and $\nabla u^n \in \mathfrak{M}(\Omega, \mathbf{R}^d)$ for every n with $\nabla u^n \xrightarrow{*} \mu$ for some $\mu \in \mathfrak{M}(\Omega, \mathbf{R}^d)$, then $\mu = \nabla u$.*

The proof does not use any new techniques (see Exercise 6.33).

We remark that Eq. (6.52) is often used as a definition of the total variation. Since it can also take the value ∞ , we will use the total variation for general functions in $L^q(\Omega)$ in the following. If, however, it is clear from the context that $\nabla u \in \mathfrak{M}(\Omega, \mathbf{R}^d)$ exists, then we also write, equivalently, $\|\nabla u\|_{\mathfrak{M}}$.

Example 6.104 (Total Variation for Special Classes of Functions)

1. Sobolev functions

We already noted that every element $u \in H^{1,1}(\Omega)$ has a gradient in $\mathfrak{M}(\Omega, \mathbf{R}^d)$ with $(\nabla u)_{\mathfrak{M}} = (\nabla u)_{L^1} \mathfrak{L}^d$. Hence $\|\nabla u\|_{\mathfrak{M}} = \int_{\Omega} |\nabla u| \, dx$.

2. Characteristic functions

Let Ω' be a bounded Lipschitz subdomain of Ω and $u = \chi_{\Omega'}$. Then by the divergence theorem (Theorem 2.81), one has for $\varphi \in \mathcal{D}(\Omega, \mathbf{R}^d)$ with $\|\varphi\|_{\infty} \leq 1$ that

$$\int_{\Omega} u \operatorname{div} \varphi \, dx = \int_{\Omega'} \operatorname{div} \varphi \, dx = \int_{\partial \Omega'} \varphi \cdot v \, d\mathfrak{H}^{d-1}.$$

This means simply that $\nabla u = -v \mathfrak{H}^{d-1} \llcorner \partial \Omega'$, and this is a measure in $\mathfrak{M}(\Omega, \mathbf{R}^d)$, since for every $\varphi \in \mathcal{C}_0(\Omega, \mathbf{R}^d)$, $\varphi \cdot v$ is \mathfrak{H}^{d-1} integrable on $\partial \Omega'$, and

$$\left| - \int_{\Omega} \varphi \cdot v \, d\mathfrak{H}^{d-1} \right| \leq \mathfrak{H}^{d-1}(\partial \Omega') \|\varphi\|_{\infty},$$

i.e., $\nabla u \in \mathcal{C}_0(\Omega, \mathbf{R}^d)^*$, and the claim follows from the Riesz-Markov theorem (Theorem 2.62). We also see that $\|\nabla u\|_{\mathfrak{M}} \leq \mathfrak{H}^{d-1}(\partial\Omega')$ has to hold. In fact, we even have equality (see Exercise 6.36), i.e. $TV(u) = \mathfrak{H}^{d-1}(\partial\Omega')$.

In other words, the total variation of a characteristic function of Lipschitz sub-domains equals its perimeter. This motivates the following generalization of the *perimeter* for measurable sets $\Omega' \subset \Omega$:

$$\text{Per}(\Omega') = TV(\chi_{\Omega'}).$$

In this sense, bounded Lipschitz domains have finite perimeter. The study of sets Ω' with finite perimeter leads to the notion of *Caccioppoli sets*, which are a further slight generalization.

3. Piecewise smooth functions

Let $u \in L^q(\Omega)$ be piecewise smooth, i.e., we can write $\overline{\Omega}$ as a union of some $\overline{\Omega}_k$ with mutually disjoint bounded Lipschitz domains $\Omega_1, \dots, \Omega_K$, and for every $k = 1, \dots, K$ one has $u^k = u|_{\Omega_k} \in \mathcal{C}^1(\overline{\Omega}_k)$, i.e., u restricted to Ω_k can be extended to a differentiable function u^k on $\overline{\Omega}_k$. Let us see, whether the weak gradient is indeed a Radon measure.

To begin with, it is clear that $u \in H^{1,1}(\Omega_1 \cup \dots \cup \Omega_K)$ with $(\nabla u)_{L^1}|_{\Omega_k} = \nabla u^k$. For $\varphi \in \mathcal{D}(\Omega, \mathbf{R}^d)$ we note that

$$\int_{\Omega} u \operatorname{div} \varphi \, dx = \sum_{k=1}^K \int_{\partial\Omega_k \cap \Omega} \varphi \cdot u^k v^k \, d\mathfrak{H}^{d-1} - \int_{\Omega_k} \varphi \cdot \nabla u^k \, dx, \quad (6.53)$$

where v^k denotes the outer unit normal to Ω_k . We can rewrite this as follows. For pairs $1 \leq l < k \leq K$, let $\Gamma_{l,k} = \overline{\Omega_l} \cap \overline{\Omega_k} \cap \Omega$ and $v = v^k$ on $\Gamma_{l,k}$. Then

$$v^k = v \quad \text{on } \Gamma_{l,k} \cap \overline{\Omega_k}, \quad v^l = -v \quad \text{on } \Gamma_{l,k} \cap \overline{\Omega_l},$$

and for $1 \leq k \leq K$, one has

$$\partial\Omega_k \cap \Omega = \Gamma_{1,k} \cup \dots \cup \Gamma_{k-1,k} \cup \Gamma_{k,k+1} \cup \dots \cup \Gamma_{k,K},$$

and this implies

$$\int_{\partial\Omega_k \cap \Omega} \varphi \cdot u^k v^k \, d\mathfrak{H}^{d-1} = \sum_{l=1}^{k-1} \int_{\Gamma_{l,k}} \varphi \cdot u^k v \, d\mathfrak{H}^{d-1} - \sum_{l=k+1}^K \int_{\Gamma_{k,l}} \varphi \cdot u^k v \, d\mathfrak{H}^{d-1}.$$

Plugging this into (6.53), we get the representation

$$\int_{\Omega} u \operatorname{div} \varphi \, dx = - \sum_{k=1}^K \sum_{l=1}^{k-1} \int_{\Gamma_{l,k}} \varphi \cdot (u^l - u^k) v \, d\mathfrak{H}^{d-1} - \int_{\Omega} \varphi \cdot (\nabla u)_{L^1} \, dx,$$

which means nothing other than

$$\nabla u = (\nabla u)_{L^1} \mathfrak{L}^d + \sum_{l < k} (u^l - u^k) v \mathfrak{H}^{d-1} \llcorner \Gamma_{l,k}.$$

Note that this representation is independent of the order of the Ω_k : the product $(u^l - u^k)v$ stays the same if we swap Ω_k and Ω_l . It is easy to see (Exercise 6.37), that ∇u is a finite vector-valued Radon measure, and for the norm, one has

$$\text{TV}(u) = \|\nabla u\|_{\mathfrak{M}} = \|(\nabla u)_{L^1}\|_1 + \sum_{l < k} \int_{\Gamma_{l,k}} |u^l - u^k| d\mathfrak{H}^{d-1}.$$

Hence, the weak gradient is measured in the L^1 norm, and the “jumps” $u^l - u^k$ of the function u are integrated along the interfaces $\Gamma_{l,k}$; see Fig. 6.18 for a simple example.

These considerations show that the total variation, when used as a penalty in minimization problems, still allows functions with discontinuities. For images, this is an exceptionally favorable property, for then we may view images, somewhat oversimplified, as smooth within objects Ω_k and discontinuous along object boundaries (i.e., edges) $\partial\Omega_k$. We expect that solutions of suitable optimization problems have exactly these properties.

Prepared with the results of Lemma 6.103, we see that the total variation has properties relevant for the direct method, that the functional $\int_{\Omega} |\nabla \cdot| dx$ does not have.

Lemma 6.105 *Let $\Omega \in \mathbf{R}^d$ be bounded. The space of functions of bounded total variation*

$$\text{BV}(\Omega) = \{u \in L^1(\Omega) \mid \nabla u \in \mathfrak{M}(\Omega, \mathbf{R}^d)\}$$

equipped with the norm $\|u\|_{\text{BV}} = \|u\|_1 + \|\nabla u\|_{\mathfrak{M}}$ is a Banach space.

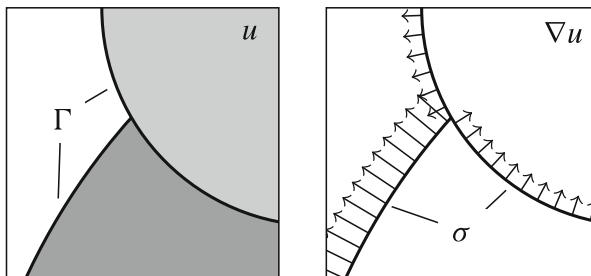


Fig. 6.18 A piecewise constant function u and its gradient as Radon measure. It holds that $\nabla u = \sigma \mathfrak{H}^1 \llcorner \Gamma$, where σ is vector field on the set of discontinuity Γ of the function in the left image (visualized in the right image). The point at the T-junction in Γ has \mathfrak{H}^1 measure zero and ∇u is not defined there

Moreover, if $\varphi : [0, \infty[\rightarrow \mathbf{R}_\infty$ is proper, convex, lower semicontinuous and increasing, then

$$\Psi(u) = \varphi(\text{TV}(u)) = \begin{cases} \varphi(\|\nabla u\|_{\mathfrak{M}}) & \text{for } u \in BV(\Omega), \\ \infty & \text{otherwise,} \end{cases}$$

is proper, convex, and (weakly) lower semicontinuous on every $L^q(\Omega)$, $1 \leq q < \infty$.

Proof First we show the properties of the functional Ψ . Since $\Psi(0) = \varphi(0)$, we see that Ψ is proper. Also, we see that $\|\cdot\|_{\mathfrak{M}}$ is convex, weakly* lower semicontinuous, and coercive. Since ∇ with $\text{dom } \nabla = BV(\Omega)$ is strongly-weakly* closed as a mapping from $L^q(\Omega)$ to $\mathfrak{M}(\Omega, \mathbf{R}^d)$ (we have $L^q(\Omega) \hookrightarrow L^1(\Omega)$), we obtain by the supplement in Example 6.29 that $\|\cdot\|_{\mathfrak{M}} \circ \nabla$ is convex and lower semicontinuous, and by Lemmas 6.14 and 6.21 also that of $\Psi = \varphi \circ \|\cdot\|_{\mathfrak{M}} \circ \nabla$.

It is clear that $BV(\Omega)$ is a normed space, so let us prove its completeness. Let (u^n) be a Cauchy sequence in $BV(\Omega)$. Hence, there exists a $u \in L^1(\Omega)$ such that $\lim_{n \rightarrow \infty} u^n = u$ in $L^1(\Omega)$. Moreover, for every $k \geq 1$ there exists a n_k such that for all $m \geq n_k$ one has $\|\nabla(u^{n_k} - u^m)\|_{\mathfrak{M}} \leq \frac{1}{k}$. By the lower semicontinuity of $\|\cdot\|_{\mathfrak{M}} \circ \nabla$ we get

$$\|\nabla(u^{n_k} - u)\|_{\mathfrak{M}} \leq \liminf_{m \rightarrow \infty} \|\nabla(u^{n_k} - u^m)\|_{\mathfrak{M}} \leq \frac{1}{k},$$

and in particular $u \in BV(\Omega)$ and $\|\nabla(u^{n_k} - u)\|_{\mathfrak{M}} \rightarrow 0$ converges for $k \rightarrow \infty$. The sequence $(\|\nabla(u^n - u)\|_{\mathfrak{M}})$ is a Cauchy sequence in \mathbf{R} , and hence, has at most one accumulation points, and hence $\lim_{n \rightarrow \infty} \|\nabla(u^n - u)\|_{\mathfrak{M}} = 0$ and consequently $\lim_{n \rightarrow \infty} u^n = u$ in $BV(\Omega)$. \square

To obtain a result on existence of minimizers of functionals with total variation penalty, we need some coercivity of Ψ . This will be deduced from some kind of Poincaré-Wirtinger inequality for the TV functional. To prove this, we need some preparations.

Lemma 6.106 *Let Ω be a bounded Lipschitz domain and $1 \leq q < \infty$. Then for every $u \in BV(\Omega) \cap L^q(\Omega)$, there exists a sequence (u^n) in $C^\infty(\overline{\Omega})$ such that*

$$\lim_{n \rightarrow \infty} u^n = u \quad \text{in } L^q(\Omega) \quad \text{and} \quad \lim_{n \rightarrow \infty} \int_{\Omega} |\nabla u^n| \, dx = \|\nabla u\|_{\mathfrak{M}}.$$

Proof We reuse the operators \mathcal{M}_n from Theorem 6.74, and show that the sequence $u^n = \mathcal{M}_n u$ for $u \in BV(\Omega) \cap L^q(\Omega)$ has the desired properties. Let us recall their definition:

$$\mathcal{M}_n u = \sum_{k=0}^K (T_{t_n \eta_k}(\varphi_k u)) * \psi^n$$

for a smooth partition of unity (φ_k) , translation vectors η_k , step sizes t_n , and scaled mollifiers ψ^n . By the arguments of Theorem 6.74 we get convergence $u^n \rightarrow u$ in $L^q(\Omega)$.

Now we choose $w \in \mathcal{D}(\Omega, \mathbf{R}^d)$ with $\|w\|_\infty \leq 1$ as a test function, and obtain

$$\begin{aligned}\int_{\Omega} u^n \operatorname{div} w \, dx &= \int_{\Omega} u (\mathcal{M}_n^* \operatorname{div} w) \, dx = \int_{\Omega} u \operatorname{div} (\mathcal{M}_n^* w) \, dx - \int_{\Omega} u (\mathcal{N}_n^* w) \, dx \\ &= \int_{\Omega} u \operatorname{div} (\mathcal{M}_n^* w) \, dx - \int_{\Omega} (\mathcal{N}_n u) \cdot w \, dx,\end{aligned}\quad (6.54)$$

where

$$\mathcal{M}_n^* w = \sum_{k=0}^K \varphi_k (T_{-t_n \eta_k} (w * \bar{\psi}^n)), \quad \mathcal{N}_n^* w = \sum_{k=0}^K \nabla \varphi_k \cdot (T_{-t_n \eta_k} (w * \bar{\psi}^n)),$$

and $\bar{\psi}^n = D_{-\operatorname{id}} \psi^n$ (the operators coincide with the adjoints of the operators \mathcal{M}_n and \mathcal{N}_n in the proof of Theorem 6.88). We have the following subgoals:

1. $\mathcal{M}_n^* w \in \mathcal{D}(\Omega, \mathbf{R}^d)$ with $\|\mathcal{M}_n^* w\|_\infty \leq 1$ for all $n \in \mathbf{N}$,
2. $\lim_{n \rightarrow \infty} \operatorname{div} (\mathcal{M}_n^* w) = \operatorname{div} w$ in $\mathcal{C}(\overline{\Omega})$ and
3. $\lim_{n \rightarrow \infty} \mathcal{N}_n u = 0$ in $L^1(\Omega)$.

Subgoal 1: From the properties of \mathcal{M}_n^* we already get $\mathcal{M}_n^* w \in \mathcal{D}(\Omega, \mathbf{R}^d)$ (see proof of Theorem 6.88), so we estimate for all $x \in \Omega$ (using the properties of the partition of unity and the mollifiers)

$$\begin{aligned}|(\mathcal{M}_n^* w)(x)| &\leq \sum_{k=0}^K \varphi_k \int_{\Omega} |w(y)| \psi^n(y - x + t_n \eta_k) \, dy \\ &\leq \|w\|_\infty \sum_{k=0}^K \varphi_k \int_{\mathbf{R}^d} \psi^n(y) \, dy \leq \|w\|_\infty \leq 1.\end{aligned}\quad (6.55)$$

This shows that $\|\mathcal{M}_n^* w\|_\infty \leq 1$.

Subgoal 2: From $\int_{\Omega} \psi^n \, dx = 1$ and $\sum_{k=0}^K \nabla \varphi_k \cdot w = 0$ we conclude that for all $x \in \Omega$,

$$\begin{aligned}(\operatorname{div} (\mathcal{M}_n^* w - w))(x) &= \sum_{k=0}^K \varphi_k(x) \int_{\Omega} (\operatorname{div} w(y) - \operatorname{div} w(x)) \psi^n(y - x + t_n \eta_k) \, dy \\ &\quad + \sum_{k=0}^K \nabla \varphi_k(x) \cdot \int_{\Omega} (w(y) - w(x)) \psi^n(y - x + t_n \eta_k) \, dy.\end{aligned}$$

Since both w and $\operatorname{div} w$ are uniformly continuous in Ω , we can find, for every $\varepsilon > 0$, a $\delta > 0$ such that $|w(x) - w(y)| \leq \varepsilon$ and $|\operatorname{div} w(x) - \operatorname{div} w(y)| \leq \varepsilon$ holds for all $x, y \in \Omega$ with $|x - y| \leq \delta$. Since (t_n) converges to zero and $\operatorname{supp} \psi^n$ becomes arbitrarily small, we can find some n_0 such that for all $n \geq n_0$, one has $|x - y + t_n \eta_k| \in \operatorname{supp} \psi^n \Rightarrow |x - y| \leq \delta$. For these n we obtain the estimate

$$|\operatorname{div}(\mathcal{M}_n^* w - w)(x)| \leq \varepsilon \sum_{k=0}^K \varphi_k(x) \int_{\mathbf{R}^d} \psi^n(y) dy + \varepsilon \sum_{k=0}^K |\nabla \varphi_k(x)| \int_{\mathbf{R}^d} \psi^n(y) dy \leq C\varepsilon.$$

The constant $C > 0$ can be chosen independently of n , and hence $\|\operatorname{div} \mathcal{M}_n^* w - \operatorname{div} w\|_\infty \rightarrow 0$ for $n \rightarrow \infty$.

Subgoal 3: We easily check that

$$\lim_{n \rightarrow \infty} \mathcal{N}_n u = \lim_{n \rightarrow \infty} \sum_{k=0}^K (T_{t_n \eta_k}(u \nabla \varphi_k)) * \psi^n = \sum_{k=0}^K \lim_{n \rightarrow \infty} T_{t_n \eta_k}(u \nabla \varphi_k) = u \sum_{k=0}^K \nabla \varphi_k = 0,$$

since smoothing with a mollifier converges in $L^1(\Omega)$ (see Lemma 3.16) and translation is continuous in $L^1(\Omega)$ (see Exercise 3.4).

Together with (6.54) our subgoals give the desired convergence $\int_{\Omega} |\nabla u^n| dx$: for every $\varepsilon > 0$ we can find some n_0 such that for $n \geq n_0$, $\|\mathcal{N}_n u\|_1 \leq \frac{\varepsilon}{3}$. For these n and arbitrary $w \in \mathcal{D}(\Omega, \mathbf{R}^d)$ with $\|w\|_\infty \leq 1$ we get from (6.54) and the total variation as defined in (6.52) that

$$\begin{aligned} \int_{\Omega} u^n \operatorname{div} w dx &= \int_{\Omega} u \operatorname{div}(\mathcal{M}_n^* w) dx - \int_{\Omega} (\mathcal{N}_n u) w dx \\ &\leq \|\nabla u\|_{\mathfrak{M}} + \|\mathcal{N}_n u\|_1 \|w\|_\infty \leq \|\nabla u\|_{\mathfrak{M}} + \frac{\varepsilon}{3}. \end{aligned}$$

Now we take, for every n , the supremum over all test functions w and get

$$\int_{\Omega} |\nabla u^n| dx \leq \|\nabla u\|_{\mathfrak{M}} + \varepsilon.$$

On the other hand, by (6.52), we can find some $w \in \mathcal{D}(\Omega, \mathbf{R}^d)$ with $\|w\|_\infty \leq 1$ such that $\int_{\Omega} u \operatorname{div} w dx \geq \|\nabla u\|_{\mathfrak{M}} - \frac{\varepsilon}{3}$. For this w we can ensure for some n_1 and all $n \geq n_1$ that

$$\left| \int_{\Omega} u \operatorname{div}(\mathcal{M}_n^* w - w) dx \right| \leq \|u\|_1 \|\operatorname{div}(\mathcal{M}_n^* w - w)\|_\infty \leq \frac{\varepsilon}{3}.$$

This implies

$$\begin{aligned} \int_{\Omega} u^n \operatorname{div} w \, dx &= \int_{\Omega} u \operatorname{div} w \, dx + \int_{\Omega} u \operatorname{div} (\mathcal{M}_n^* w - w) \, dx - \int_{\Omega} (\mathcal{N}_n u) w \, dx \\ &\geq \|\nabla u\|_{\mathfrak{M}} - \frac{\varepsilon}{3} - \frac{\varepsilon}{3} - \frac{\varepsilon}{3} \end{aligned}$$

and in particular, by the supremum definition (6.52),

$$\int_{\Omega} |\nabla u^n| \, dx \geq \|\nabla u\|_{\mathfrak{M}} - \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, we have shown the desired convergence $\lim_{n \rightarrow \infty} \int_{\Omega} |\nabla u^n| \, dx = \|\nabla u\|_{\mathfrak{M}}$. \square

Remark 6.107 The above theorem guarantees, for every $u \in \operatorname{BV}(\Omega)$, the existence of a sequence (u^n) in $C^\infty(\overline{\Omega})$ with $u^n \rightarrow u$ in $L^1(\Omega)$ and $\nabla u^n \xrightarrow{*} \nabla u$ in $\mathfrak{M}(\Omega, \mathbf{R}^d)$. By Lemma 6.103, the latter implies only the weaker property

$$\|\nabla u\|_{\mathfrak{M}} \leq \liminf_{n \rightarrow \infty} \int_{\Omega} |\nabla u^n| \, dx.$$

Hence, the convergence $u^n \rightarrow u$ in $L^1(\Omega)$ and $\|\nabla u^n\|_{\mathfrak{M}} \rightarrow \|\nabla u\|_{\mathfrak{M}}$ for some sequence (u^n) in $\operatorname{BV}(\Omega)$ and $u \in \operatorname{BV}(\Omega)$ has a special name; it is called *strict convergence*.

The approximation property from Lemma 6.106 allows us to transfer some properties of $H^{1,1}(\Omega)$ to $\operatorname{BV}(\Omega)$.

Lemma 6.108 *For a bounded Lipschitz domain $\Omega \subset \mathbf{R}^d$ and $1 \leq q \leq d/(d-1)$ with $d/(d-1) = \infty$ for $d = 1$, one has the following:*

1. *There is a continuous embedding $\operatorname{BV}(\Omega) \hookrightarrow L^q(\Omega)$, and in the case $q < d/(d-1)$ the embedding is also compact.*
2. *There exists $C > 0$ such that for every $u \in \operatorname{BV}(\Omega)$, the following Poincaré-Wirtinger inequality is satisfied:*

$$\|P_1 u\|_q = \left\| u - \frac{1}{|\Omega|} \int_{\Omega} u \, dx \right\|_q \leq C \|\nabla u\|_{\mathfrak{M}}.$$

Proof Assertion 1: Using Lemma 6.106 we choose, for every $u \in \operatorname{BV}(\Omega)$, a sequence (u^n) in $C^\infty(\overline{\Omega})$ that converges strictly to u . In particular, $u^n \in H^{1,1}(\Omega)$ for every n . Since $H^{1,1}(\Omega) \hookrightarrow L^q(\Omega)$ (see Theorem 6.76 and Remark 6.77), there is a $C > 0$ such that for all $v \in H^{1,1}(\Omega)$, $\|v\|_q \leq C \|v\|_{1,1} = C(\|v\|_1 + \|\nabla v\|_{\mathfrak{M}})$. The L^q -norm is lower semicontinuous in $L^1(\Omega)$ (to see this, we note that it can be expressed as $\|\cdot\|_q \circ A$, where A is the identity with domain $L^q(\Omega) \subset L^1(\Omega)$, see also Example 6.29 for a similar argument for $q > 1$). Thus, with the above C and

by strict convergence, we get

$$\|u\|_q \leq \liminf_{n \rightarrow \infty} \|u^n\|_q \leq C(\liminf_{n \rightarrow \infty} \|u^n\|_1 + \liminf_{n \rightarrow \infty} \|\nabla u^n\|_{\mathfrak{M}}) = C(\|u\|_1 + \|\nabla u\|_{\mathfrak{M}}),$$

which shows the desired inequality and embedding.

To show compactness of the embedding for $q < d/(d-1)$ we first consider the case $q = 1$. For a bounded sequence (u^n) in $BV(\Omega)$, choose $v^n \in \mathcal{C}^\infty(\overline{\Omega})$ such that $\|v^n - u^n\|_1 \leq \frac{1}{n}$ as well as $|\|\nabla v^n\|_{\mathfrak{M}} - \|\nabla u^n\|_{\mathfrak{M}}| \leq \frac{1}{n}$. Then, the sequence (v^n) is bounded in $H^{1,1}(\Omega)$, and there exists, since $H^{1,1}(\Omega) \hookrightarrow L^1(\Omega)$ is compact, a subsequence (v^{n_k}) with $\lim_{k \rightarrow \infty} v^{n_k} = v$ for some $v \in L^1(\Omega)$. For the respective subsequence (u^{n_k}) , one has by construction that $\lim_{k \rightarrow \infty} u^{n_k} = v$, which shows that $BV(\Omega) \hookrightarrow L^1(\Omega)$ compactly.

For the case $1 < q < d/(d-1)$ we recall Young's inequality for numbers, which is

$$ab \leq \frac{a^p}{p} + \frac{b^{p^*}}{p^*}, \quad a, b \geq 0, \quad p \in]1, \infty[.$$

We choose $r \in]q, d/(d-1)[$ and $p = (r-1)/(q-1)$, and get that for all $a \geq 0$ and $\delta > 0$,

$$a^q = (\delta^{\frac{1}{p}} a^{\frac{r(q-1)}{r-1}})(\delta^{-\frac{1}{p}} a^{\frac{r-q}{r-1}}) \leq \frac{q-1}{r-1} \delta a^r + \frac{r-q}{r-1} \delta^{-\frac{p^*}{p}} a,$$

since $p^* = (r-1)/(r-q)$ and $\frac{r}{p} + \frac{1}{p^*} = q$. Consequently, there exists for every $\delta > 0$ a $C_\delta > 0$, such that for all $a \geq 0$,

$$a^q \leq \delta a^r + C_\delta a. \quad (6.56)$$

Now let again (u^n) be bounded in $BV(\Omega)$, hence also bounded in $L^r(\Omega)$ with bound $L > 0$ on the norm. By the above, we can assume without loss of generality that $u^n \rightarrow u$ in $L^1(\Omega)$. For every $\varepsilon > 0$ we choose $0 < \delta < \frac{\varepsilon^q}{2L^r}$ and $C_\delta > 0$ such that (6.56) holds. Since (u^n) is a Cauchy sequence in $L^1(\Omega)$, there exists an n_0 such that

$$\int_{\Omega} |u^{n_1} - u^{n_2}| \, dx \leq \frac{\varepsilon^q}{2C_\delta}$$

for all $n_1, n_2 \geq n_0$. Consequently, we see that for these n_1 and n_2 , using (6.56),

$$\begin{aligned} \int_{\Omega} |u^{n_1} - u^{n_2}|^q \, dx &\leq \delta \int_{\Omega} |u^{n_1} - u^{n_2}|^r \, dx + C_\delta \int_{\Omega} |u^{n_1} - u^{n_2}| \, dx \\ &\leq \frac{\varepsilon^q}{2L^r} L^r + C_\delta \frac{\varepsilon^q}{2C_\delta} = \varepsilon^q, \end{aligned}$$

and hence (u^n) is a Cauchy sequence in $L^q(\Omega)$ and hence convergent. By the continuous embedding $L^q(\Omega) \hookrightarrow L^1(\Omega)$ the limit has to be u .

Assertion 2: We begin the proof for $q = 1$ with a preliminary remark. If $\nabla u = 0$ for some $u \in \text{BV}(\Omega)$ then obviously $u \in H^{1,1}(\Omega)$, and by Lemma 6.79 we see that u is constant. The rest of the argument can be done similarly to Lemma 6.81. If the inequality did not hold, there would exist a sequence (u^n) in $\text{BV}(\Omega)$ with $\int_{\Omega} u^n dx = 0$, $\|u^n\|_1 = 1$, and $\|\nabla u^n\|_{\mathfrak{M}} \leq \frac{1}{n}$. In particular, we would have $\lim_{n \rightarrow \infty} \nabla u^n = 0$. By the compact embedding $\text{BV}(\Omega) \hookrightarrow L^1(\Omega)$ we get the existence of a $u \in L^1(\Omega)$ and a subsequence (u^{n_k}) with $\lim_{k \rightarrow \infty} u^{n_k} = u$. The gradient is a closed mapping, and hence, $\nabla u = 0$, and by the preliminary remark, u is constant in Ω . Since also $\int_{\Omega} u dx = \lim_{n \rightarrow \infty} \int_{\Omega} u^n dx = 0$ holds, we get $u = 0$. This shows that $\lim_{k \rightarrow \infty} u^{n_k} = 0$, which contradicts $\|u^n\|_1 = 1$ and, hence the inequality is proved for $q = 1$.

For $1 < q \leq d/(d-1)$ the claim follows using assertion 1:

$$\|P_1 u\|_q \leq C(\|P_1 u\|_1 + \|\nabla u\|_{\mathfrak{M}}) \leq C\|\nabla u\|_{\mathfrak{M}}.$$

□

Corollary 6.109 *In the situation of Lemma 6.108, let $\varphi : [0, \infty[\rightarrow \mathbf{R}_{\infty}$ be coercive and $1 \leq q \leq d/(d-1)$. The function $\Psi(u) = \varphi(\text{TV}(u))$ is coercive in the sense $\|P_1 u\|_q \rightarrow \infty \Rightarrow \Psi(u) \rightarrow \infty$ (since for all $u \in L^q(\Omega)$ one has $\|P_1 u\|_q \leq C \text{TV}(u)$ for some $C > 0$, the claim follows from the coercivity of φ).*

If φ is strongly coercive, one gets “strong” coercivity, i.e., $\|P_1 u\|_q \rightarrow \infty \Rightarrow \Psi(u)/\|P_1 u\|_q \rightarrow \infty$, in a similar way.

The theory of functions of bounded total variation is a large field, which is used not only in image processing, but also for so-called free discontinuity problems (see, e.g., [5]). It is closely related to geometric measure theory (the standard treatment of which is [62]). For the sake of completeness we cite three results of this theory that do not need any further notions.

Theorem 6.110 (Traces of $\text{BV}(\Omega)$ Functions) *For a bounded Lipschitz domain $\Omega \subset \mathbf{R}^d$ there exists a unique linear and continuous mapping $T : \text{BV}(\Omega) \rightarrow L^1_{\mathfrak{H}^{d-1}}(\partial\Omega)$ such that for all $u \in \text{BV}(\Omega) \cap \mathcal{C}(\overline{\Omega})$, $Tu = u|_{\partial\Omega}$.*

Moreover, this map is continuous with respect to strict convergence, i.e.,

$$\left. \begin{aligned} u^n &\rightarrow u \quad \text{in } L^1(\Omega) \\ \|\nabla u^n\|_{\mathfrak{M}} &\rightarrow \|\nabla u\|_{\mathfrak{M}} \end{aligned} \right\} \quad \Rightarrow \quad Tu^n \rightarrow Tu \quad \text{in } L^1_{\mathfrak{H}^{d-1}}(\partial\Omega).$$

The notion of trace allows one to formulate and prove a property that distinguishes $\text{BV}(\Omega)$ from the spaces $H^{1,p}(\Omega)$.

Theorem 6.111 (Zero Extension of BV Functions) *For a bounded Lipschitz domain $\Omega \subset \mathbf{R}^d$, the zero extension $E : \text{BV}(\Omega) \rightarrow \text{BV}(\mathbf{R}^d)$ is continuous, and*

$$\nabla(Eu) = \nabla u - vTu\mathfrak{H}^{d-1} \llcorner \partial\Omega,$$

where v is the outer unit normal and Tu is the trace on $\partial\Omega$ defined in Theorem 6.110.

Corollary 6.112 For a bounded Lipschitz domain $\Omega \subset \mathbf{R}^d$ and a Lipschitz subdomain $\Omega' \subset \Omega$ with $\overline{\Omega'} \subset\subset \Omega$, for $u^1 \in BV(\Omega')$, $u^2 \in BV(\Omega \setminus \overline{\Omega'})$, and $u = u^1 + u^2$ (with implicit zero extension) one has that $u \in BV(\Omega)$ and

$$\nabla u = \nabla(u^1)|_{\Omega'} + \nabla(u^2)|_{\Omega \setminus \Omega'} + (u^2|_{\partial(\Omega \setminus \Omega')} - u^1|_{\partial\Omega'})v\mathfrak{H}^{d-1} \llcorner \partial\Omega'.$$

Here we take the trace of u^1 on $\partial\Omega'$ with respect to Ω' , and the trace of u^2 on $\partial(\Omega \setminus \Omega')$ with respect to $\Omega \setminus \overline{\Omega'}$.

The following result relates functions of bounded total variation and the perimeter of their sub-level sets

Theorem 6.113 (Co-area Formula) For a bounded Lipschitz domain $\Omega \subset \mathbf{R}^d$ and $u \in BV(\Omega)$ it holds that

$$\|\nabla u\|_{\mathfrak{M}} = \int_{\Omega} 1 \, d|\nabla u| = \int_{\mathbf{R}} \text{Per}(\{x \in \Omega \mid u(x) \leq t\}) \, dt = \int_{\mathbf{R}} \text{TV}(\chi_{\{u \leq t\}}) \, dt.$$

In other words, the total variation is the integral over all perimeters of the sublevel sets.

The proofs of the previous three theorems can be found, e.g., in [5] and [61].

Using the co-area formula, one sees that for $u \in BV(\Omega)$ and $h : \mathbf{R} \rightarrow \mathbf{R}$ strongly increasing and continuously differentiable with $\|h'\|_{\infty} < \infty$, the scaled versions $h \circ u$ are also contained in $BV(\Omega)$. The value $\text{TV}(h \circ u)$ depends only on the sublevel sets of u ; see Exercise 6.34.

Now we can start to use TV as an image model in variational problems.

Theorem 6.114 (Existence of Solutions with Total Variation Penalty) Let $\Omega \subset \mathbf{R}^d$ be a bounded Lipschitz domain, $q \in]1, \infty[$, with $q \leq d/(d-1)$ and $\Phi : L^q(\Omega) \rightarrow \mathbf{R}_{\infty}$ proper on $BV(\Omega)$, convex, lower semicontinuous on $L^q(\Omega)$, and coercive in Π^1 , i.e.,

$$\left\| u - \frac{1}{|\Omega|} \int_{\Omega} u \, dx \right\|_q \text{ bounded} \quad \text{and} \quad \left| \int_{\Omega} u \, dx \right| \rightarrow \infty \quad \Rightarrow \quad \Phi(u) \rightarrow \infty.$$

Moreover, let $\varphi : [0, \infty[\rightarrow \mathbf{R}_{\infty}$ be proper, convex, lower semicontinuous, increasing, and strongly coercive. Then for every $\lambda > 0$ there exists a solution u^* of the problem

$$\min_{u \in L^q(\Omega)} \Phi(u) + \lambda \varphi(\text{TV}(u)).$$

The assertion is also true if Φ is bounded from below and φ is only coercive.

Proof One argues similarly to the proof of Theorem 6.84, the respective properties for the TV functional follow from Lemmas 6.105 and 6.108 as well as Corollary 6.109. \square

Analogously to Theorem 6.86, we obtain the following (see also [1, 31]):

Theorem 6.115 (Tikhonov Functionals with Total Variation Penalty) *For a bounded Lipschitz domain Ω , $q \in]1, \infty[$, a Banach space Y , and $A \in \mathcal{L}(L^q(\Omega), Y)$ one has the following implication: If*

1. $q \leq d/(d-1)$ and A does not vanish for constant functions or
2. A is injective and $\text{rg}(A)$ closed,

then there exists for every $u^0 \in Y$, $r \in [1, \infty[$, and $\lambda > 0$ a solution u^ of the problem*

$$\min_{u \in L^q(\Omega)} \frac{\|Au - u^0\|_Y^r}{r} + \lambda \text{TV}(u). \quad (6.57)$$

If A is injective, $r > 1$, and the norm in Y is strictly convex, then u^ is unique.*

The *proof* can be done with the help of Theorem 6.114, and one needs to show only the needed properties of $\Phi = \frac{1}{r} \|Au - u^0\|_Y^r$, which goes along the lines of the proof of Theorem 6.86. The uniqueness follows from general consideration about Tikhonov functionals (see Example 6.32). \square

Now we aim to derive optimality conditions for minimization problems involving the total variation. Hence, we are interested in the subgradient of TV. To develop some intuition on what it looks like, we write $\text{TV} = \|\cdot\|_{\mathfrak{M}} \circ \nabla$, where ∇ is considered as a closed mapping from $L^q(\Omega)$ to $\mathfrak{M}(\Omega, \mathbf{R}^d)$. By the result of Exercise 6.12 we get

$$\partial \text{TV} = \nabla^* \circ \partial \|\cdot\|_{\mathfrak{M}} \circ \nabla.$$

Let us analyze $\partial \|\cdot\|_{\mathfrak{M}} \subset \mathfrak{M}(\Omega, \mathbf{R}^d) \times \mathfrak{M}(\Omega, \mathbf{R}^d)^*$. According to Example 6.49, this subgradient is, for some $\mu \in \mathfrak{M}(\Omega, \mathbf{R}^d)$, the set

$$\partial \|\cdot\|_{\mathfrak{M}}(\mu) = \begin{cases} \{\|\sigma\|_{\mathfrak{M}^*} \leq 1\} & \text{if } \mu = 0, \\ \{\|\sigma\|_{\mathfrak{M}^*} = 1, \langle \sigma, \mu \rangle_{\mathfrak{M}^* \times \mathfrak{M}} = \|\mu\|_{\mathfrak{M}}\} & \text{otherwise,} \end{cases} \quad (6.58)$$

as a subset of $\mathfrak{M}(\Omega, \mathbf{R}^d)^*$. However, there is no simple characterization of the space $\mathfrak{M}(\Omega, \mathbf{R}^d)^*$ as a space of functions or measures, and hence it is difficult to describe these sets. We note, though, that for some $\sigma \in \mathfrak{M}(\Omega, \mathbf{R}^d)^*$ with $\|\sigma\|_{\mathfrak{M}^*} \leq 1$ we can use the following construction: with the total variation measure $|\mu|$ (see Definition 2.57) and an arbitrary $v \in L^1_{|\mu|}(\Omega, \mathbf{R}^d)$, the finite vector-valued measure $v|\mu|$ is a Radon measure with $\|v|\mu|\|_{\mathfrak{M}} = \|v\|_1$, where the latter norm is the norm

in $L_{|\mu|}^1(\Omega, \mathbf{R}^d)$. It follows that

$$\langle \sigma, v|\mu| \rangle_{\mathfrak{M}^* \times \mathfrak{M}} \leq \|v|\mu|\|_{\mathfrak{M}} = \|v\|_1.$$

Since $v \in L_{|\mu|}^1(\Omega, \mathbf{R}^d)$ was arbitrary and $|\mu|$ is finite, we can identify σ by duality with an element in $(\sigma)_{|\mu|} \in L_{|\mu|}^\infty(\Omega, \mathbf{R}^d)$ with $\|(\sigma)_{|\mu|}\|_\infty \leq 1$. By the polar decomposition $\mu = \sigma_\mu |\mu|$, $\sigma_\mu \in L_{|\mu|}^\infty(\Omega, \mathbf{R}^d)$ (see Theorem 2.58) we obtain

$$\langle \sigma, \mu \rangle_{\mathfrak{M}^* \times \mathfrak{M}} = \|\mu\|_{\mathfrak{M}} \iff \int_{\Omega} (\sigma)_{|\mu|} \cdot \sigma_\mu \, d|\mu| = \int_{\Omega} 1 \, d|\mu|.$$

Since $\|(\sigma)_{|\mu|}\|_\infty \leq 1$, it follows that $0 \leq 1 - (\sigma)_{|\mu|} \cdot \sigma_\mu \, d|\mu|$ almost everywhere, and hence the latter is equivalent to $(\sigma)_{|\mu|} \cdot \sigma_\mu = 1$ $|\mu|$ -almost everywhere and hence, to $(\sigma)_{|\mu|} = \sigma_\mu \, d|\mu|$ -almost everywhere (by the Cauchy-Schwarz inequality).

We rewrite the result more compactly: since the meaning is clear from context, we write σ instead of $(\sigma)_{|\mu|}$, and also we set $\sigma_\mu = \frac{\mu}{|\mu|}$, since σ_μ is the $|\mu|$ almost everywhere uniquely defined density of μ with respect to $|\mu|$. This gives the characterization

$$\partial \|\cdot\|_{\mathfrak{M}}(\mu) = \left\{ \sigma \in \mathfrak{M}(\Omega, \mathbf{R}^d)^* \mid \|\sigma\|_{\mathfrak{M}^*} \leq 1, \sigma = \frac{\mu}{|\mu|} \, d|\mu| \text{ almost everywhere} \right\}. \quad (6.59)$$

Now we discuss the adjoint operator ∇^* as a closed mapping between $\mathfrak{M}(\Omega, \mathbf{R}^d)^*$ and $L^{q^*}(\Omega)$. We remark that it is not clear whether ∇^* is densely defined (since $\mathfrak{M}(\Omega, \mathbf{R}^d)$ is not reflexive). Testing with $u \in \mathcal{D}(\Omega)$, we get for $\sigma \in \text{dom } \nabla^*$ that

$$\int_{\Omega} (\nabla^* \sigma) u \, dx = \langle \sigma, \nabla u \rangle_{\mathfrak{M}^* \times \mathfrak{M}},$$

hence in the sense of distributions we have $\nabla^* \sigma = -\text{div } \sigma$. Moreover, for $\sigma \in \mathcal{C}^\infty(\overline{\Omega}, \mathbf{R}^d)$ and all $u \in \mathcal{C}^\infty(\overline{\Omega})$ one has that

$$\int_{\Omega} \sigma \cdot \nabla u \, dx = \int_{\partial\Omega} u \sigma \cdot v \, d\mathfrak{H}^{d-1} - \int_{\Omega} u \text{div } \sigma \, dx.$$

The right-hand side can be estimated by the L^q norm only if $\sigma \cdot v = 0$ on $\partial\Omega$. If this is the case, then $\mu \mapsto \int_{\Omega} \sigma \, d\mu$ induces a continuous linear map $\bar{\sigma}$ on $\mathfrak{M}(\Omega, \mathbf{R}^d)$, which satisfies for every $u \in \mathcal{C}^\infty(\overline{\Omega})$ that

$$|\langle \bar{\sigma}, \nabla u \rangle_{\mathfrak{M}^* \times \mathfrak{M}}| \leq \|\text{div } \sigma\|_{q^*} \|u\|_q.$$

Since test functions u are dense in $L^q(\Omega)$ (see Theorem 6.74), $\bar{\sigma} \in \text{dom } \nabla^*$ has to hold. In some sense, $\text{dom } \nabla^*$ contains only the elements for which the normal

trace vanishes on the boundary. Since we will use another approach later, we will not try to formulate a normal trace for certain elements in $\mathfrak{M}(\Omega, \mathbf{R}^d)^*$, but for the time being, use the slightly sloppy characterization

$$\nabla^* = -\operatorname{div}, \quad \operatorname{dom} \nabla^* = \{\sigma \in \mathfrak{M}(\Omega, \mathbf{R}^d)^* \mid \operatorname{div} \sigma \in L^{q^*}(\Omega), \sigma \cdot v = 0 \text{ on } \partial\Omega\}.$$

Collecting the previous results, we can describe the subgradient of the total variation as

$$\partial \operatorname{TV}(u) = \left\{ -\operatorname{div} \sigma \mid \|\sigma\|_{\mathfrak{M}^*} \leq 1, \sigma \cdot v = 0 \text{ on } \partial\Omega, \sigma = \frac{\nabla u}{|\nabla u|} |\nabla u| \text{ almost everywhere} \right\}. \quad (6.60)$$

Unfortunately, some objects in the representation are not simple to deal with. As already noted, the space $\mathfrak{M}(\Omega, \mathbf{R}^d)^*$ does not have a characterization as a function space (there are attempts to describe the biduals of $\mathcal{C}(K)$ for compact Hausdorff spaces K [82], though). Also, we would like to have a better understanding of the divergence operator on $\mathfrak{M}(\Omega, \mathbf{R}^d)^*$, especially under what circumstances one can speak of a vanishing normal trace on the boundary. Therefore, we present a different approach to characterizing $\partial \operatorname{TV}$, which does not use the dual space of the space of vector-valued Radon measures but only regular functions.

At the core of the approach lies the following normed space, which can be seen as a generalization of the sets defined in (6.37) for $p = 1$:

$$\begin{aligned} \mathcal{D}_{\operatorname{div}, \infty} &= \left\{ \sigma \in L^\infty(\Omega, \mathbf{R}^d) \mid \exists \operatorname{div} \sigma \in L^{q^*}(\mathbf{R}^d) \text{ and a sequence } (\sigma^n) \text{ in } \mathcal{D}(\Omega, \mathbf{R}^d) \right. \\ &\quad \left. \text{with } \lim_{n \rightarrow \infty} \|\sigma^n - \sigma\|_{q^*} + \|\operatorname{div}(\sigma^n - \sigma)\|_{q^*} = 0 \right\}, \\ \|\sigma\|_{\operatorname{div}, \infty} &= \|\sigma\|_\infty + \|\operatorname{div} \sigma\|_{q^*}. \end{aligned} \quad (6.61)$$

We omit the dependence on q also for these spaces. In some sense, the elements of $\mathcal{D}_{\operatorname{div}, \infty}$ satisfy $\sigma \cdot v = 0$ on $\partial\Omega$, cf. Remark 6.89. Let us analyze this space further; we first prove its completeness.

Lemma 6.116 *Let $\Omega \subset \mathbf{R}^d$ be a bounded Lipschitz domain and $q \in]1, \infty[$. Then $\mathcal{D}_{\operatorname{div}, \infty}$ according to the definition in (6.61) is a Banach space.*

Proof For a Cauchy sequence (σ^n) in $\mathcal{D}_{\operatorname{div}, \infty}$ we have the convergence $\sigma^n \rightarrow \sigma$ in $L^\infty(\Omega, \mathbf{R}^d)$ as well as $\operatorname{div} \sigma^n \rightarrow w$ in $L^{q^*}(\Omega)$. By the closedness of the weak divergence we have $w = \operatorname{div} \sigma$, hence it remains to show the needed approximation property from the definition (6.61). To that end, choose for every n a sequence $(\sigma^{n,k})$ of test functions (i.e., in $\mathcal{D}(\Omega, \mathbf{R}^d)$), that approximate σ^n in the sense of (6.61). Now, for every n there is a k_n such that

$$\|\sigma^{n,k_n} - \sigma^n\|_{q^*} + \|\operatorname{div}(\sigma^{n,k_n} - \sigma^n)\|_{q^*} \leq \frac{1}{n}.$$

Since Ω is bounded, we have $\|\sigma^n - \sigma\|_{q^*} \leq |\Omega|^{(q-1)/q} \|\sigma^n - \sigma\|_\infty$. For $\varepsilon > 0$ we choose n_0 such that for all $n \geq n_0$ we have the inequalities

$$\frac{1}{n} \leq \frac{\varepsilon}{3}, \quad \|\sigma^n - \sigma\|_\infty \leq \frac{|\Omega|^{\frac{q}{q-1}} \varepsilon}{3}, \quad \|\operatorname{div}(\sigma^n - \sigma)\|_{q^*} \leq \frac{\varepsilon}{3},$$

and for these n we have

$$\begin{aligned} \|\sigma^{n,k_n} - \sigma\|_{q^*} + \|\operatorname{div}(\sigma^{n,k_n} - \sigma)\|_{q^*} &\leq \|\sigma^{n,k_n} - \sigma^n\|_{q^*} + \|\operatorname{div}(\sigma^{n,k_n} - \sigma^n)\|_{q^*} \\ &\quad + |\Omega|^{\frac{q-1}{q}} \|\sigma^n - \sigma\|_\infty + \|\operatorname{div}(\sigma^n - \sigma)\|_{q^*} \\ &\leq \frac{\varepsilon}{3} + |\Omega|^{\frac{q-1}{q}} \frac{|\Omega|^{\frac{q}{q-1}} \varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon. \end{aligned}$$

Hence, we have found a sequence in $\mathcal{D}(\Omega, \mathbf{R}^d)$ that approximates σ in the desired way, which shows that $\sigma \in \mathcal{D}_{\operatorname{div},\infty}$. \square

Remark 6.117 According to (6.37) with $p = q$, the construction (6.61) differs from the set $\mathcal{D}_{\operatorname{div}}^1$ only in the fact that its elements additionally lie in $L^\infty(\Omega, \mathbf{R}^d)$. Together with Proposition 6.88 and Remark 6.89, we can therefore claim that

$$\begin{aligned} \mathcal{D}_{\operatorname{div},\infty} &= \operatorname{dom} \nabla^* \cap L^\infty(\Omega, \mathbf{R}^d) \\ &= \{\sigma \in L^\infty(\Omega, \mathbf{R}^d) \mid \operatorname{div} \sigma \in L^{q^*}(\Omega), \sigma \cdot v = 0 \text{ on } \partial\Omega\}, \end{aligned} \quad (6.62)$$

where the gradient is regarded as a closed mapping from $L^q(\Omega)$ to $L^q(\Omega, \mathbf{R}^d)$ with domain $H^{1,q}(\Omega)$. In particular, the proof of Proposition 6.88 demonstrates that for $\sigma \in \mathcal{D}_{\operatorname{div},\infty}$, the sequence (σ^n) in $\mathcal{D}(\Omega, \mathbf{R}^d)$ defined by $\sigma^n = \mathcal{M}_n^* \sigma$ is approximating in the sense of (6.61). For these σ^n , one can now estimate the L^∞ -norm: $|\sigma^n(x)| = |(\mathcal{M}_n^* \sigma)(x)| \leq \|\sigma\|_\infty$ for $x \in \Omega$; cf. (6.55) in the proof of Lemma 6.106. We can therefore also write

$$\mathcal{D}_{\operatorname{div},\infty} = \left\{ \sigma \in L^\infty(\Omega, \mathbf{R}^d) \mid \exists \operatorname{div} \sigma \in L^{q^*}(\mathbf{R}^d), (\sigma^n) \text{ in } \mathcal{D}(\Omega, \mathbf{R}^d) \text{ with} \right.$$

$$\left. \|\sigma^n\|_\infty \leq \|\sigma\|_\infty, \lim_{n \rightarrow \infty} \|\sigma^n - \sigma\|_{q^*} + \|\operatorname{div}(\sigma^n - \sigma)\|_{q^*} = 0 \right\}.$$

The Banach space $\mathcal{D}_{\operatorname{div},\infty}$ is well suited for the description of the subgradient of TV, as the following lemma will show.

Lemma 6.118 ($\mathcal{D}_{\operatorname{div},\infty}$ -Vector Fields and ∂ TV) *For $\Omega \subset \mathbf{R}^d$ a bounded Lipschitz domain, $q \in]1, \infty[$ and $u \in BV(\Omega) \cap L^q(\Omega)$, we have that $w \in L^{q^*}(\Omega)$ lies in $\partial TV(u)$ if and only if there exists $\sigma \in \mathcal{D}_{\operatorname{div},\infty}$ such that*

$$\|\sigma\|_\infty \leq 1, \quad -\operatorname{div} \sigma = w \text{ and } - \int_{\Omega} u \operatorname{div} \sigma \, dx = TV(u).$$

Proof Let us first prove that $w \in \partial \text{TV}(u)$ if and only if

$$\int_{\Omega} vw \, dx \leq \text{TV}(v) \quad \text{for all } v \in \text{BV}(\Omega) \cap L^q(\Omega) \quad \text{and} \quad \int_{\Omega} uw \, dx = \text{TV}(u). \quad (6.63)$$

We use similar arguments to those in Example 6.49. Let $w \in \partial \text{TV}(u)$. According to the subgradient inequality, we have for all $v \in \text{BV}(\Omega) \cap L^q(\Omega)$ and $\lambda > 0$ that after inserting λv , we have

$$\lambda \int_{\Omega} vw \, dx \leq \lambda \text{TV}(v) + \left(\int_{\Omega} uw \, dx - \text{TV}(u) \right).$$

Since $\int_{\Omega} uw - \text{TV}(u)$ does not depend on λ , on dividing by λ and considering the limit as $\lambda \rightarrow \infty$, we obtain the inequality $\int_{\Omega} vw \, dx \leq \text{TV}(v)$. This shows the first property asserted, which in particular implies $\int_{\Omega} uw \, dx \leq \text{TV}(u)$. On the other hand, inserting $v = 0$ into the subgradient inequality yields $\text{TV}(u) \leq \int_{\Omega} uw \, dx$. Therefore, equality has to hold.

For the converse, assume that $w \in L^{q^*}(\Omega)$ satisfies $\int_{\Omega} vw \, dx \leq \text{TV}(v)$ for all $v \in \text{BV}(\Omega) \cap L^q(\Omega)$ and $\int_{\Omega} uw \, dx = \text{TV}(u)$. Then for arbitrary $v \in \text{BV}(\Omega) \cap L^q(\Omega)$, one has

$$\text{TV}(u) + \int_{\Omega} w(v - u) \, dx = \int_{\Omega} vw \, dx \leq \text{TV}(v).$$

Hence, the subgradient inequality is satisfied, and we have $w \in \partial \text{TV}(u)$.

Next, we show, for $w \in L^{q^*}(\Omega)$, the equivalence

$$\int_{\Omega} vw \, dx \leq \text{TV}(v) \quad \text{for all } v \in \text{BV}(\Omega) \cap L^q(\Omega) \iff \begin{cases} \text{there exists } \sigma \in \mathcal{D}_{\text{div},\infty} \text{ with} \\ w = -\text{div } \sigma \text{ and } \|\sigma\|_{\infty} \leq 1. \end{cases}$$

This assertion is true if and only if equality holds for the sets K_1 and K_2 defined by

$$K_1 = \left\{ w \in L^{q^*}(\Omega) \mid \int_{\Omega} vw \, dx \leq \text{TV}(v) \text{ for all } v \in \text{BV}(\Omega) \cap L^q(\Omega) \right\},$$

$$K_2 = \{ -\text{div } \sigma \mid \sigma \in \mathcal{D}_{\text{div},\infty} \text{ with } \|\sigma\|_{\infty} \leq 1 \}.$$

Being an intersection of convex closed halfspaces, K_1 is convex, closed, and, of course, nonempty. Analogously, it is easy to conclude that K_2 is nonempty as well as convex, and we will now show that it is also closed. For this purpose, we choose a sequence (w^n) in K_2 with $\lim_{n \rightarrow \infty} w^n = w$ for some $w \in L^{q^*}(\Omega)$. Every w^n can be represented as $w^n = -\text{div } \sigma^n$ with $\sigma^n \in \mathcal{D}_{\text{div},\infty}$ and $\|\sigma^n\|_{\infty} \leq 1$. Note that the sequence (σ^n) is also bounded in $L^{q^*}(\Omega, \mathbf{R}^d)$, and hence there exists a weakly convergent subsequence, which we again index by n . Since the div operator

is weakly-strongly closed, we obtain $w = -\operatorname{div} \sigma$, and furthermore, the weak lower semicontinuity of the L^∞ -norm yields

$$\|\sigma\|_\infty \leq \liminf_{n \rightarrow \infty} \|\sigma^n\|_\infty \leq 1.$$

It remains to show that σ can be approximated by $\mathcal{D}(\Omega, \mathbf{R}^d)$. Since every σ^n is contained in $\mathcal{D}_{\operatorname{div}, \infty}$, we obtain, analogously to the argumentation in Lemma 6.116, a sequence $(\bar{\sigma}^n)$ in $\mathcal{D}(\Omega, \mathbf{R}^d)$ with

$$\bar{\sigma}^n \rightharpoonup \sigma \text{ in } L^{q^*}(\Omega, \mathbf{R}^d) \text{ and } \operatorname{div} \bar{\sigma}^n \rightarrow \operatorname{div} \sigma \text{ in } L^{q^*}(\Omega).$$

Using the same argument as in the end of the proof of Proposition 6.88, weak convergence can be replaced by strong convergence, possibly yielding a different sequence. Finally, this implies $\sigma \in \mathcal{D}_{\operatorname{div}, \infty}$, and thus K_2 is closed.

Let us now show that $K_2 \subset K_1$: For $w = -\operatorname{div} \sigma \in K_2$, according to Remark 6.117, there exists a sequence (σ^n) in $\mathcal{D}(\Omega, \mathbf{R}^d)$ with $\|\sigma^n\|_\infty \leq \|\sigma\|_\infty \leq 1$ and $\lim_{n \rightarrow \infty} -\operatorname{div} \sigma^n = -\operatorname{div} \sigma = w$ in $L^{q^*}(\Omega)$. Using the supremum definition of TV (6.52), this implies for arbitrary $v \in BV(\Omega) \cap L^q(\Omega)$ that

$$\int_{\Omega} v w \, dx = \lim_{n \rightarrow \infty} - \int_{\Omega} v \operatorname{div} \sigma^n \, dx \leq TV(v),$$

i.e., $w \in K_1$. Conversely, assume there existed $w^0 \in K_1 \setminus K_2$. Then according to the Hahn-Banach theorem (Proposition 2.29), there would exist an element $u^0 \in L^q(\Omega)$, that separates the compact convex set $\{w^0\}$ and K_2 , i.e.,

$$\sup_{w \in K_2} \int_{\Omega} u^0 w \, dx < \int_{\Omega} u^0 w^0 \, dx.$$

However, since $-\operatorname{div} \sigma \in K_2$ for every $\sigma \in \mathcal{D}(\Omega, \mathbf{R}^d)$ with $\|\sigma\|_\infty \leq 1$, we infer $u^0 \in BV(\Omega)$ and $TV(u^0) < \int_{\Omega} u^0 w^0 \, dx$, a contradiction to $w^0 \in K_1$. Hence, we have $K_1 = K_2$.

From this together with (6.63), the assertion follows. \square

This result already shows that we can use the function space $\mathcal{D}_{\operatorname{div}, \infty}$ and the weak divergence instead of ∇^* on $\mathfrak{M}(\Omega, \mathbf{R}^d)^*$. In order to obtain a representation as in (6.60) with $\mathcal{D}_{\operatorname{div}, \infty}$, we still need a characterization analogous to (6.59). As we have seen before, an essential ingredient for that is taking the trace $\sigma \mapsto (\sigma)_{|\mu|}$ from $\mathfrak{M}(\Omega, \mathbf{R}^d)^*$ to $L_{|\mu|}^\infty(\Omega, \mathbf{R}^d)$. An analogous statement will not be possible for general $\sigma \in \mathcal{D}_{\operatorname{div}, \infty}$ and $\mu \in \mathfrak{M}(\Omega, \mathbf{R}^d)$, but according to (6.63), it suffices to consider $\mu = \nabla u$ with $u \in BV(\Omega)$. This is the purpose of the following lemma.

Lemma 6.119 (Traces of $\mathcal{D}_{\text{div},\infty}$ -Vector Fields) For $\Omega \subset \mathbf{R}^d$ a bounded Lipschitz domain, $q \in]1, \infty[$ and $u \in \text{BV}(\Omega) \cap L^q(\Omega)$, there is a linear and continuous mapping

$$T_u^\nu : \mathcal{D}_{\text{div},\infty} \rightarrow L_{|\nabla u|}^\infty(\Omega) \text{ with } \|T_u^\nu \sigma\|_\infty \leq \|\sigma\|_\infty,$$

such that for every $\sigma \in \mathcal{D}(\Omega, \mathbf{R}^d)$, one has

$$T_u^\nu \sigma = \sigma \cdot \frac{\nabla u}{|\nabla u|} \text{ in } L_{|\nabla u|}^\infty(\Omega)$$

where $\frac{\nabla u}{|\nabla u|}$ is the sign of the polar decomposition of ∇u . Furthermore, T_u^ν is weakly continuous in the sense that

$$\left. \begin{array}{ll} \sigma^n \rightharpoonup \sigma & \text{in } L^{q^*}(\Omega, \mathbf{R}^d) \\ \text{div } \sigma^n \rightharpoonup \text{div } \sigma & \text{in } L^{q^*}(\Omega) \end{array} \right\} \implies T_u^\nu \sigma^n \xrightarrow{*} T_u^\nu \sigma \text{ in } L_{|\nabla u|}^\infty(\Omega).$$

Proof Let $u \in \text{BV}(\Omega) \cap L^q(\Omega)$. For $\sigma \in \mathcal{D}_{\text{div},\infty}$, we choose, according to Remark 6.117, a sequence (σ^n) in $\mathcal{D}(\Omega, \mathbf{R}^d)$ with $\|\sigma^n\|_\infty \leq \|\sigma\|_\infty$, $\lim_{n \rightarrow \infty} \sigma^n = \sigma$, and $\lim_{n \rightarrow \infty} \text{div } \sigma^n = \text{div } \sigma$ in the respective spaces. Furthermore, we construct an associated linear form as follows: for every $\varphi \in \mathcal{C}^\infty(\overline{\Omega})$, let

$$L(\varphi) = - \int_{\Omega} u(\varphi \text{div } \sigma + \nabla \varphi \cdot \sigma) \, dx = \lim_{n \rightarrow \infty} - \int_{\Omega} u(\varphi \text{div } \sigma^n + \nabla \varphi \cdot \sigma^n) \, dx$$

where the last equality is due to the convergence in $L^{q^*}(\Omega)$ and $u \in L^q(\Omega)$. Additionally, $\varphi \sigma^n \in \mathcal{D}(\Omega, \mathbf{R}^d)$ with $\text{div}(\varphi \sigma^n) = \varphi \text{div } \sigma^n + \nabla \varphi \cdot \sigma^n$. Due to $u \in \text{BV}(\Omega)$ and the characterization of TV in (6.52), this implies

$$\begin{aligned} |L(\varphi)| &= \lim_{n \rightarrow \infty} \left| \int_{\Omega} u \text{div}(\varphi \sigma^n) \, dx \right| = \lim_{n \rightarrow \infty} \left| \int_{\Omega} \varphi \sigma^n \, d\nabla u \right| \\ &\leq \liminf_{n \rightarrow \infty} \|\sigma^n\|_\infty \int_{\Omega} |\varphi| \, d|\nabla u| \leq \|\sigma\|_\infty \|\varphi\|_1, \end{aligned}$$

where the latter norm is taken in $L_{|\nabla u|}^1(\Omega)$. Note that the set $\mathcal{D}(\Omega) \subset \mathcal{C}^\infty(\overline{\Omega})$ is densely contained in this space (cf. Exercise 6.35). Therefore, L can be uniquely extended to an element in $T_u^\nu \sigma \in L_{|\nabla u|}^1(\Omega)^* = L_{|\nabla u|}^\infty(\Omega)$ with $\|T_u^\nu \sigma\|_\infty \leq \|\sigma\|_\infty$, and hence the linear mapping $T_u^\nu : \mathcal{D}_{\text{div},\infty} \rightarrow L_{|\nabla u|}^\infty(\Omega)$ is continuous.

For $\sigma \in \mathcal{D}(\Omega, \mathbf{R}^d)$, we can choose $\sigma^n = \sigma$, and the construction yields for all $\varphi \in \mathcal{C}^\infty(\overline{\Omega})$

$$\int_{\Omega} (T_u^\nu \sigma) \varphi \, d|\nabla u| = L(\varphi) = - \int_{\Omega} u \operatorname{div}(\varphi \sigma) \, dx = \int_{\Omega} \varphi \left(\sigma \cdot \frac{\nabla u}{|\nabla u|} \right) \, d|\nabla u|.$$

Since the test functions are dense in $L_{|\nabla u|}^1(\Omega)$, this implies the identity $T_\sigma^\nu u = \sigma \cdot \frac{\nabla u}{|\nabla u|}$ in $L_{|\nabla u|}^\infty(\Omega)$.

In order to establish the weak continuity, let (σ^n) and σ in $\mathcal{D}_{\operatorname{div}, \infty}$ be given as in the assertion. For $\varphi \in \mathcal{C}^\infty(\overline{\Omega})$, we infer, due to the construction as well as the weak convergence of (σ^n) and $(\operatorname{div} \sigma^n)$, that

$$\begin{aligned} \lim_{n \rightarrow \infty} \int_{\Omega} (T_u^\nu \sigma^n) \varphi \, d|\nabla u| &= \lim_{n \rightarrow \infty} - \int_{\Omega} u (\varphi \operatorname{div} \sigma^n + \nabla \varphi \cdot \sigma^n) \, dx \\ &= - \int_{\Omega} u (\varphi \operatorname{div} \sigma + \nabla \varphi \cdot \sigma) \, dx = \int_{\Omega} (T_u^\nu \sigma) \varphi \, d|\nabla u|. \end{aligned}$$

Resultingly, $T_u^\nu \sigma^n \xrightarrow{*} T_u^\nu \sigma$ converges in $L_{|\nabla u|}^\infty(\Omega)$; again due to the density of test functions. \square

Remark 6.120 (T_u^ν as Normal Trace Operator) The mapping T_u^ν is the unique element in $\mathcal{L}(\mathcal{D}_{\operatorname{div}, \infty}, L_{|\nabla u|}^\infty(\Omega))$ that satisfies $T_u^\nu \sigma = \sigma \cdot \frac{\nabla u}{|\nabla u|}$ for all $\sigma \in \mathcal{D}(\Omega, \mathbf{R}^d)$ and that exhibits the weak continuity described in Lemma 6.119. (This fact is implied by the approximation property specified in the definition of $\mathcal{D}_{\operatorname{div}, \infty}$.) If we view $\frac{\nabla u}{|\nabla u|}$ as a vector field of outer normals with respect to the level-sets of u , we can also interpret T_u^ν as the *normal trace operator*.

Thus, we write $\sigma \cdot \frac{\nabla u}{|\nabla u|} = T_u^\nu \sigma$ for $\sigma \in \mathcal{D}_{\operatorname{div}, \infty}$. In particular, due to the weak continuity, the following generalization of the *divergence theorem* holds:

$$u \in \operatorname{BV}(\Omega), \sigma \in \mathcal{D}_{\operatorname{div}, \infty} : \quad - \int_{\Omega} u \operatorname{div} \sigma \, dx = \int_{\Omega} \left(\sigma \cdot \frac{\nabla u}{|\nabla u|} \right) \, d|\nabla u|. \quad (6.64)$$

The assertions of Lemmas 6.118 and 6.119 are the crucial ingredients for the desired characterization of the subdifferential of the total variation.

Theorem 6.121 (Characterization of $\partial \operatorname{TV}$ with Normal Trace) *Let $\Omega \subset \mathbf{R}^d$ be a bounded Lipschitz domain, $q \in]1, \infty[$, $q \leq d/(d-1)$, and $u \in \operatorname{BV}(\Omega)$. Then for $w \in L^{q^*}(\Omega)$, one has the equivalence*

$$w \in \partial \operatorname{TV}(u) \iff \text{there exists } \sigma \in \mathcal{D}_{\operatorname{div}, \infty} \text{ with } \begin{cases} \|\sigma\|_\infty \leq 1, \\ -\operatorname{div} \sigma = w, \\ \sigma \cdot \frac{\nabla u}{|\nabla u|} = 1, \end{cases}$$

where $\sigma \cdot \frac{\nabla u}{|\nabla u|} = 1$ represents the $|\nabla u|$ -almost everywhere identity for the normal trace of σ .

In particular, using the alternative notation (6.62), we have that

$$\partial \text{TV}(u) = \left\{ -\operatorname{div} \sigma \mid \|\sigma\|_\infty \leq 1, \sigma \cdot v = 0 \text{ on } \partial \Omega \text{ and } \sigma \cdot \frac{\nabla u}{|\nabla u|} = 1 \text{ } |\nabla u| \text{-almost everywhere} \right\},$$

with $\sigma \in L^\infty(\Omega, \mathbf{R}^d)$ and $\operatorname{div} \sigma \in L^{q^*}(\Omega)$.

Proof In view of the assertion in Lemma 6.118, we first consider an arbitrary $\sigma \in \mathcal{D}_{\operatorname{div},\infty}$ with $\|\sigma\|_\infty \leq 1$. For this σ , there exists, according to Lemma 6.119, the normal trace $T_u^\nu \sigma = \sigma \cdot \frac{\nabla u}{|\nabla u|} \in L_{|\nabla u|}^\infty(\Omega)$ with $|\sigma \cdot \frac{\nabla u}{|\nabla u|}(x)| \leq 1$ for $|\nabla u|$ -almost all $x \in \Omega$. Together with (6.64), this implies

$$-\int_\Omega u \operatorname{div} \sigma \, dx = \int_\Omega 1 \, d|\nabla u| \iff \int_\Omega 1 - \left(\sigma \cdot \frac{\nabla u}{|\nabla u|} \right) \, d|\nabla u| = 0,$$

and since the integrand on the right-hand side is $|\nabla u|$ -almost everywhere nonpositive, we infer the equivalence to $\sigma \cdot \frac{\nabla u}{|\nabla u|} = 1$ $|\nabla u|$ -almost everywhere.

According to Lemma 6.118, $w \in \partial \text{TV}(u)$ if and only if there exists $\sigma \in \mathcal{D}_{\operatorname{div},\infty}$ with $\|\sigma\|_\infty \leq 1$ and $-\operatorname{div} \sigma = w$ such that

$$-\int_\Omega u \operatorname{div} \sigma \, dx = \int_\Omega 1 \, d|\nabla u|.$$

Due to the observation above, this is equivalent to the existence of a $\sigma \in \mathcal{D}_{\operatorname{div},\infty}$ with $\|\sigma\|_\infty \leq 1$, $-\operatorname{div} \sigma = w$ and $\sigma \cdot \frac{\nabla u}{|\nabla u|} = 1$ $|\nabla u|$ -almost everywhere.

The last assertion simply expresses the equivalence just proved in set notation, using the interpretation of $\mathcal{D}_{\operatorname{div},\infty}$ in (6.62). \square

Remark 6.122 (Characterization of ∂TV with Full Trace) According to an idea in [77], we can define a “full” trace as follows: for $u \in \text{BV}(\Omega) \cap L^q(\Omega)$, $\sigma \in \mathcal{D}_{\operatorname{div},\infty}$ with $\|\sigma\|_\infty \leq 1$ and $\sigma \cdot \frac{\nabla u}{|\nabla u|} = 1$ $|\nabla u|$ -almost everywhere, it is also the case in a certain sense that $\sigma = \frac{\nabla u}{|\nabla u|}$ $|\nabla u|$ -almost everywhere. For an approximating sequence (σ^n) in $\mathcal{D}(\Omega, \mathbf{R}^d)$ with $\sigma^n \rightarrow \sigma$ in $L^{q^*}(\Omega, \mathbf{R}^d)$, $\operatorname{div} \sigma^n \rightarrow \operatorname{div} \sigma$ in $L^{q^*}(\Omega)$, and $\|\sigma^n\|_\infty \leq \|\sigma\|_\infty$, we infer for all n that the norm of $\sigma^n \in L_{|\nabla u|}^\infty(\Omega, \mathbf{R}^d)$ is not greater than 1. Furthermore, due to $1 - |\sigma^n|^2 \geq 0$ $|\nabla u|$ -almost everywhere, one has that

$$\begin{aligned} \frac{1}{2} \left| \sigma^n - \frac{\nabla u}{|\nabla u|} \right|^2 &= \frac{1}{2} |\sigma^n|^2 - \sigma^n \cdot \frac{\nabla u}{|\nabla u|} + \frac{1}{2} \\ &\leq \frac{1}{2} + \frac{1}{2} |\sigma^n|^2 - \sigma^n \cdot \frac{\nabla u}{|\nabla u|} + \frac{1}{2} - \frac{1}{2} |\sigma^n|^2 = 1 - \sigma^n \cdot \frac{\nabla u}{|\nabla u|}. \end{aligned}$$

The weak continuity of the normal trace now implies the convergence

$$\lim_{n \rightarrow \infty} \int_{\Omega} \left| \sigma^n - \frac{\nabla u}{|\nabla u|} \right|^2 d|\nabla u| \leq \lim_{n \rightarrow \infty} 2 \int_{\Omega} 1 - \sigma^n \cdot \frac{\nabla u}{|\nabla u|} d|\nabla u| = 0,$$

i.e., we have $\lim_{n \rightarrow \infty} \sigma^n = \frac{\nabla u}{|\nabla u|}$ in $L^2_{|\nabla u|}(\Omega, \mathbf{R}^d)$ and, due to the finiteness of the measure $|\nabla u|$, also in $L^1_{|\nabla u|}(\Omega, \mathbf{R}^d)$.

We now say that $\sigma \in \mathcal{D}_{\text{div}, \infty}$ has a (full) *trace* $T_u^d \sigma \in L^\infty_{|\nabla u|}(\Omega, \mathbf{R}^d)$ if and only if for every approximating sequence (σ^n) as above, one has $\sigma^n \rightarrow T_u^d \sigma \in L^1_{|\nabla u|}(\Omega, \mathbf{R}^d)$. This yields, by a simple calculation, a densely defined, closed operator T_u^d between $\mathcal{D}_{\text{div}, \infty}$ and $L^\infty_{|\nabla u|}(\Omega, \mathbf{R}^d)$ with $T_u^d \sigma = \sigma$ for all $\sigma \in \mathcal{D}(\Omega, \mathbf{R}^d)$. In contrast to the normal trace operator T_u^v , T_u^d does not have to be continuous.

Using this notion of a trace and writing, slightly abusing notation, $\sigma = T_u^d \sigma$, we can express $\partial \text{TV}(u)$ by

$$\partial \text{TV}(u) = \left\{ -\operatorname{div} \sigma \mid \|\sigma\|_\infty \leq 1, \sigma \cdot v = 0 \text{ on } \partial\Omega, \sigma = \frac{\nabla u}{|\nabla u|} |\nabla u| \text{-almost everywhere} \right\},$$

where throughout, we assume $\sigma \in L^\infty(\Omega, \mathbf{R}^d)$, $\operatorname{div} \sigma \in L^{q^*}(\Omega)$, and the existence of the full trace of σ in $L^\infty_{|\nabla u|}(\Omega, \mathbf{R}^d)$.

We now have three equivalent characterizations of the subgradient of the total variation at hand. In order to distinguish them, let us summarize them here again. Recall that for $u \in \text{BV}(\Omega) \cap L^q(\Omega)$, $\frac{\nabla u}{|\nabla u|} \in L^\infty_{|\nabla u|}(\Omega, \mathbf{R}^d)$ denotes the unique element of the polar decomposition of the Radon measure ∇u , i.e., $\nabla u = \frac{\nabla u}{|\nabla u|} |\nabla u|$.

1. Dual space representation: (Equation (6.60))

$$\partial \text{TV}(u) = \left\{ -\operatorname{div} \sigma \mid \|\sigma\|_{\mathfrak{M}^*} \leq 1, \sigma \cdot v = 0 \text{ on } \partial\Omega, \sigma = \frac{\nabla u}{|\nabla u|} |\nabla u| \text{-a. e.} \right\}.$$

In this case, $-\operatorname{div} \sigma \in L^q(\Omega)$ and $\sigma \cdot v = 0$ on $\partial\Omega$ hold in the sense of $\sigma \in \operatorname{dom} \nabla^* \subset \mathfrak{M}(\Omega, \mathbf{R}^d)^*$ with ∇ as a closed mapping between $L^q(\Omega)$ and $\mathfrak{M}(\Omega, \mathbf{R}^d)$. The equality $\sigma = \frac{\nabla u}{|\nabla u|}$ holds in the sense of $\sigma = (\sigma)_{|\nabla u|} \in L^\infty_{|\nabla u|}(\Omega, \mathbf{R}^d)$ as a restriction of σ on $L^1_{|\nabla u|}(\Omega, \mathbf{R}^d) \subset \mathfrak{M}(\Omega, \mathbf{R}^d)$.

2. $\mathcal{D}_{\text{div}, \infty}$ normal trace representation: (Proposition 6.121)

$$\partial \text{TV}(u) = \left\{ -\operatorname{div} \sigma \mid \|\sigma\|_\infty \leq 1, \sigma \cdot v = 0 \text{ on } \partial\Omega, \sigma \cdot \frac{\nabla u}{|\nabla u|} = 1 |\nabla u| \text{-a. e.} \right\}.$$

The existence of $-\operatorname{div} \sigma$ and $\sigma \cdot v = 0$ is considered to mean that $\sigma \in \mathcal{D}_{\text{div}, \infty}$ according to (6.61), whereas we interpret $\sigma \cdot \frac{\nabla u}{|\nabla u|} = 1$ as an equation for

$\sigma \cdot \frac{\nabla u}{|\nabla u|} = T_u^\nu \sigma$ with the continuous normal trace operator $T_u^\nu : \mathcal{D}_{\text{div},\infty} \rightarrow L_{|\nabla u|}^\infty(\Omega)$ according to Lemma 6.119.

3. $\mathcal{D}_{\text{div},\infty}$ **trace representation:** (Remark 6.122)

$$\partial \text{TV}(u) = \left\{ -\operatorname{div} \sigma \mid \|\sigma\|_\infty \leq 1, \sigma \cdot v = 0 \text{ on } \partial\Omega, \sigma = \frac{\nabla u}{|\nabla u|} |\nabla u| \text{-a.e.} \right\}.$$

Again, the existence of $-\operatorname{div} \sigma$ and $\sigma \cdot v = 0$ on $\partial\Omega$ is an alternative expression for $\sigma \in \mathcal{D}_{\text{div},\infty}$. The identity $\sigma = \frac{\nabla u}{|\nabla u|}$ implicitly expresses that $\sigma \in \operatorname{dom} T_u^d \subset \mathcal{D}_{\text{div},\infty}$ with the not necessarily continuous trace operator T_u^d of Remark 6.119 and that the equation $T_u^d \sigma = \frac{\nabla u}{|\nabla u|}$ in $L_{|\nabla u|}^\infty(\Omega, \mathbf{R}^d)$ is satisfied.

Remark 6.123 (∂TV and the Mean Curvature) For $u \in H^{1,1}(\Omega)$, one has $(\nabla u)_\mathfrak{M} = (\nabla u)_{L^1} \mathcal{L}^d$, which implies that $|(\nabla u)_\mathfrak{M}|$ -almost everywhere is equivalent to (Lebesgue)-almost everywhere in $\{(\nabla u)_{L^1} \neq 0\}$. Furthermore, we have the agreement of the sign

$$\left(\frac{(\nabla u)_\mathfrak{M}}{|(\nabla u)_\mathfrak{M}|} \right)(x) = \frac{(\nabla u)_{L^1}(x)}{|(\nabla u)_{L^1}(x)|} \quad \text{for almost all } x \in \{(\nabla u)_{L^1} \neq 0\}.$$

Let us further note that the trace $T_u^d \sigma$ (cf. Remark 6.122) exists for every $\sigma \in \mathcal{D}_{\text{div},\infty}$. This implies, writing $\nabla u = (\nabla u)_{L^1}$, that

$$\partial \text{TV}(u) = \left\{ -\operatorname{div} \sigma \mid \|\sigma\|_\infty \leq 1, \sigma \cdot v = 0 \text{ on } \partial\Omega, \sigma = \frac{\nabla u}{|\nabla u|} \text{ a.e. in } \{|\nabla u| \neq 0\} \right\}.$$

For sufficiently smooth u , we therefore have that $w \in \partial \text{TV}(u)$ on $\{|\nabla u| \neq 0\}$ satisfies the identity $w = -\operatorname{div} \left(\frac{\nabla u}{|\nabla u|} \right) = -\kappa$, where for $x \in \Omega$ with $\nabla u(x) \neq 0$, $\kappa(x)$ represents the mean curvature of the level-set $\{y \in \Omega \mid u(y) = u(x)\}$ in x (cf. Exercise 5.9).

Therefore, the subgradient of the total variation provides a generalization of the *mean curvature* of the level-sets of u :

$$\kappa = -\partial \text{TV}(u).$$

We will get back to this interpretation in the examples below.

Finally, let us consider the application of the total variation image model to the image-processing problems introduced in Sect. 6.3.2.

Example 6.124 (Denoising with TV Penalty Term) As in Application 6.94, we desire to denoise an image $u^0 \in L^q(\Omega)$, $q \in]1, \infty[$ on a bounded Lipschitz domain $\Omega \subset \mathbf{R}^d$, this time using the total variation as an image model. For $\lambda > 0$, we look

for a solution to

$$\min_{u \in L^q(\Omega)} \frac{1}{q} \int_{\Omega} |u - u^0|^q dx + \lambda \operatorname{TV}(u). \quad (6.65)$$

According to Proposition 6.115, there exists a unique solution to this problem.

The Euler-Lagrange equations can be derived analogously to Application 6.94; we have only to substitute the subgradient of $\frac{1}{p} \|\nabla \cdot\|_p^p$ accordingly. Therefore, u^* is a solution of (6.65) if and only if there exists $\sigma^* \in \mathcal{D}_{\operatorname{div}, \infty}$ such that the equations

$$\left\{ \begin{array}{ll} |u^* - u^0|^{q-2}(u^* - u^0) - \lambda \operatorname{div} \sigma^* = 0 & \text{in } \Omega, \\ \sigma^* \cdot \nu = 0 & \text{on } \partial\Omega, \\ \|\sigma^*\|_{\infty} \leq 1 \quad \text{and} \quad \sigma^* = \frac{\nabla u^*}{|\nabla u^*|} & |\nabla u^*| \text{-almost everywhere} \end{array} \right. \quad (6.66)$$

are satisfied. Writing $\kappa^* = \operatorname{div} \sigma^*$ and interpreting this as the mean curvature of the level-sets of u^* , we can consider u^* as the solution of the equation

$$|u^* - u^0|^{q-2}(u^* - u^0) - \lambda \kappa^* = 0.$$

Remark 6.125 The total variation penalty term was presented in the form of the denoising problem with quadratic data term for the first time in [121]. For that reason, the cost functional in (6.65) is also named after the authors as the *Rudin-Osher-Fatemi functional*. Since then, TV has become one of the standard models in image processing.

By means of Eq. (6.66) or rather its interpretation as an equation for the mean curvature, we can gain a qualitative understanding of the solutions of problem (6.65). For this purpose, we first derive a maximum principle again.

Lemma 6.126 (Maximum Principle for L^q -TV Denoising) *If in the situation of Example 6.124 $L \leq u^0 \leq R$ almost everywhere in Ω for some $L, R \in \mathbf{R}$, then the solution u^* of (6.65) also satisfies $L \leq u^* \leq R$ almost everywhere in Ω .*

Proof Let (u^n) be a sequence in $C^\infty(\overline{\Omega})$ such that $u^n \rightarrow u^*$ in $L^q(\Omega)$ as well as $\operatorname{TV}(u^n) \rightarrow \operatorname{TV}(u^*)$. According to Lemma 6.106, such a sequence exists; without loss of generality, we can assume that $u^n \rightarrow u^*$ even holds pointwise almost everywhere in Ω (by applying the theorem of Fischer-Riesz, Proposition 2.48). As seen in the proof of the maximum principle in Proposition 6.95, we set $v^n = \min(R, \max(L, u^n))$ as well as $v^* = \min(R, \max(L, u^*))$. Analogously to that situation, we see that one always has $|v^n - u^0| \leq |u^n - u^0|$ almost everywhere in Ω . Together with the pointwise almost everywhere convergence and

the boundedness of (v^n) , we infer that $v^n \rightarrow v^*$ in $L^q(\Omega)$ and

$$\int_{\Omega} |v^* - u^0|^q dx = \lim_{n \rightarrow \infty} \int_{\Omega} |v^n - u^0|^q dx \leq \lim_{n \rightarrow \infty} \int_{\Omega} |u^n - u^0|^q dx = \int_{\Omega} |u^* - u^0|^q dx.$$

Furthermore, due to the chain rule for Sobolev functions (Lemma 6.75), we have $v^n \in H^{1,1}(\Omega)$ with $\nabla v^n = \nabla u^n$ on $\{\nabla u^n \neq 0\}$ and $\nabla v^n = 0$ otherwise, i.e.,

$$\text{TV}(v^n) = \int_{\Omega} |\nabla v^n| dx \leq \int_{\Omega} |\nabla u^n| dx = \text{TV}(u^n).$$

If F denotes the cost functional in (6.65), then the choice of (u^n) , the properties of (v^n) , and the lower semicontinuity of F in $L^q(\Omega)$ imply

$$F(v^*) \leq \liminf_{n \rightarrow \infty} F(v^n) \leq \frac{1}{q} \int_{\Omega} |u^* - u^0|^q dx + \liminf_{n \rightarrow \infty} \lambda \text{TV}(u^n) = F(u^*).$$

Therefore, v^* is a minimizer and due to uniqueness, we have $u^* = \min(R, \max(L, u^*))$, which proves the assertion. \square

Like the method in Application 6.94, the variational denoising presented in Example 6.124 yields in particular a solution $u^* \in L^\infty(\Omega)$ if $u^0 \in L^\infty(\Omega)$. In this case, we also obtain $|u^0 - u^*|^{q-2}(u^0 - u^*) \in L^\infty(\Omega)$ and since the Euler-Lagrange equation (6.66) holds, we have $\kappa^* = \text{div } \sigma^* \in L^\infty(\Omega)$. Interpreting κ^* as the mean curvature, we conclude that the mean curvatures of the level-sets of u^* therefore have to be essentially bounded. This condition still allows u^* to exhibit discontinuities (in contrast to the solutions associated with the Sobolev penalty term in (6.39)), but corners and objects with high curvature cannot be reproduced.

Figure 6.19 shows some numerical examples of this method. We can see that it yields very good results for piecewise constant functions, especially in terms of the reconstruction of the object boundaries while simultaneously removing the noise. If this condition is violated, which usually happens with natural images, artifacts arise in many cases, which let the result appear “blocky” or “staircased.” In this context, this is again referred to as “staircasing artifacts” or the “staircasing effect.”

The effect of the regularization parameter λ can exemplarily be observed in Fig. 6.20 for $q = 2$: for higher values, also the size of the details, that are not reconstructed in the smoothed images anymore, increases. Relevant edges, however, are preserved. An undesired effect, which appears for large λ , is a reduction of contrast. One can show that this is a consequence of the quadratic error term and that it is possible to circumvent it by a transition to the L^1 -norm (L^1 -TV, cf. [36, 55]). Solutions of this problem even satisfy a variant of the gray value scaling invariance [GSI] of Chap. 5: for suitable strictly increasing $h : \mathbf{R} \rightarrow \mathbf{R}$ and for a minimizer u^* of the L^1 -TV problem for the data u^0 , the scaled version $h \circ u^*$ is a minimizer for the scaled data $h \circ u^0$; see Exercise 6.38 for more details.

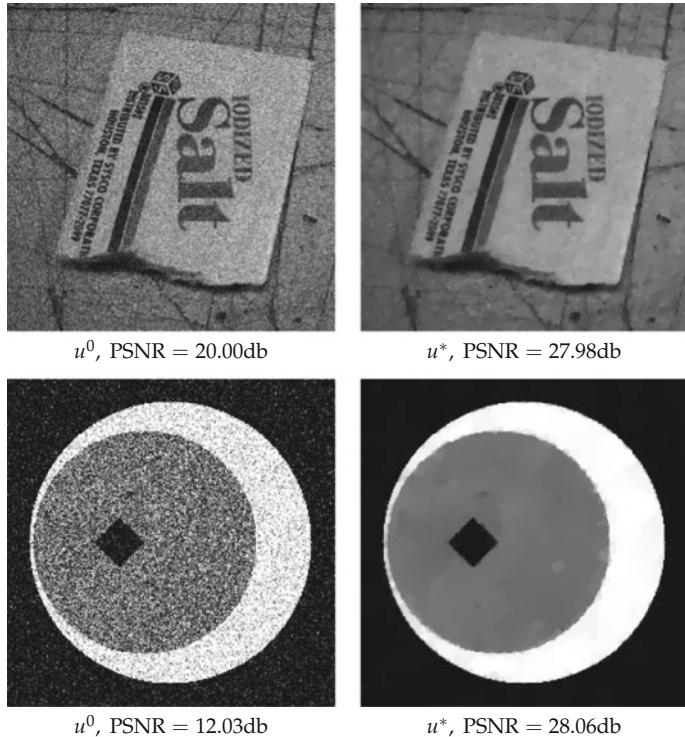


Fig. 6.19 Denoising with total variation penalty. Top: Left a noisy natural image (see Fig. 6.11 for the original), right the solution of (6.65). Bottom: The noisy version of a piecewise constant artificial image (original in Fig. 5.7), right the result of TV denoising. In both cases $q = 2$ as well as λ chosen to optimize PSNR. For the natural image we see effects similar to Fig. 6.11 for $p = 1.1$. The reconstruction for the artificial image is particularly good: The original has small total variation and hence, fits exactly to the modeling assumptions for (6.65)

Example 6.127 (Deconvolution with Total Variation Penalty) Of course, we can use the TV image model also for deconvolution. As in Application 6.97, let $\Omega' \subset \mathbf{R}^d$ be a domain, $k \in L^1(\Omega_0)$ with $\int_{\Omega_0} k \, dx = 1$, and Ω a bounded Lipschitz domain such that $\Omega' - \Omega_0 \subset \Omega$. Moreover, assume that $q \in]1, \infty[$ satisfies $q \leq d/(d-1)$. Then by Theorem 6.115 we get that for every $u^0 \in L^q(\Omega')$ and $\lambda > 0$ there is a solution of the minimization problem

$$\min_{u \in L^q(\Omega)} \frac{1}{q} \int_{\Omega} |u * k - u^0|^q \, dx + \lambda \operatorname{TV}(u). \quad (6.67)$$

Similarly to Application 6.97, we can derive the Euler-Lagrange equations for u^* , and the only difference is the use of the subdifferential of the TV semi-norm: u^* is

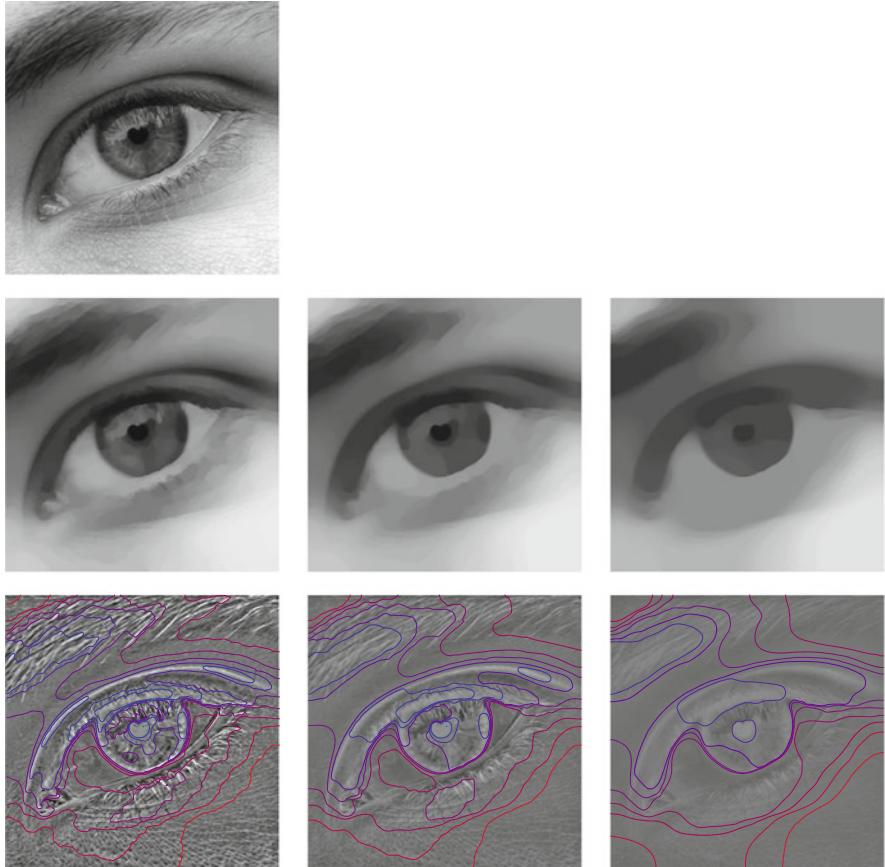


Fig. 6.20 Effect of L^2 -TV denoising with varying regularization parameter. Top left: Original. Second row: Solutions u^* for different λ . Third row: The difference images $(u^* - u^0)/\lambda$ with contours of the level sets of u^* . In both rows $\lambda = 0.1$, $\lambda = 0.3$, and $\lambda = 0.9$ have been used. The Euler-Lagrange equation (6.66) states that the difference images coincide with the curvature of the level sets, and indeed, this can be seen for the depicted contours

an optimal solution of (6.67) if and only if there exists $\sigma^* \in \mathcal{D}_{\text{div},\infty}$ such that

$$\left\{ \begin{array}{ll} \left(|u^* * k - u^0|^{q-2} (u^* * k - u^0) \right) * \bar{k} = \lambda \operatorname{div} \sigma^* & \text{in } \Omega, \\ \sigma^* \cdot v = 0 & \text{on } \partial\Omega, \\ \|\sigma^*\|_\infty \leq 1 \quad \text{and} \quad \sigma^* = \frac{\nabla u^*}{|\nabla u^*|} & |\nabla u^*| \text{ almost everywhere,} \end{array} \right. \quad (6.68)$$

where $\bar{k} = D_{-\text{id}}k$.

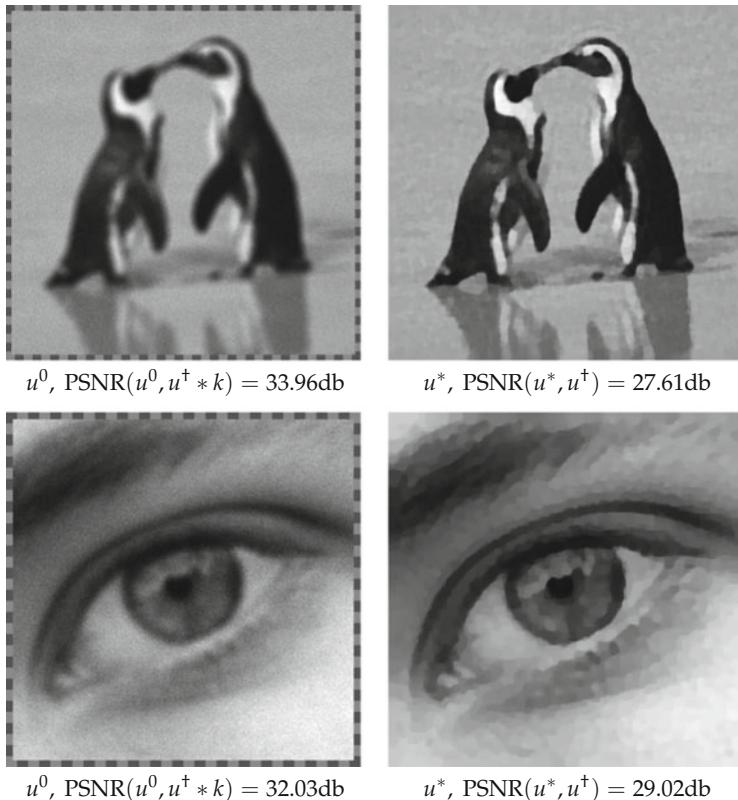


Fig. 6.21 Deconvolution with total variation penalty. Top: Left the convolved and noisy data u^0 from Fig. 6.13, right a minimizer u^* of the functional (6.67) for $q = 2$. Bottom: Left convolved and noisy data (comparable to u^0 in Fig. 6.3), right the respective deconvolved version. The parameter λ has been optimized for maximal PSNR

Qualitatively, the results are comparable with those of the denoising problem with total variation; in particular one can see that the mean curvature is essentially bounded if $k \in L^q(\Omega_0)$ (cf. the arguments in Application 6.97). Numerical results for this method are reported in Fig. 6.21. Despite the noise, a considerably sharper version of the noisy image could be reconstructed. However, details smaller than a certain size have been lost. Moreover, the staircasing effect is a little bit stronger than in Example 6.124 and leads to a blocky appearance which is typical for total variation methods.

Example 6.128 (Total Variation Inpainting) Let $\Omega \subset \mathbf{R}^d$ be a bounded Lipschitz domain and Ω' with $\overline{\Omega'} \subset\subset \Omega$ a Lipschitz subdomain on which the “true image” $u^\dagger : \Omega \rightarrow \mathbf{R}$ is to be reconstructed. Moreover, we assume that $u^\dagger|_{\Omega \setminus \Omega'} \in BV(\Omega \setminus \overline{\Omega'})$, and so the zero extension u^0 of $u^\dagger|_{\Omega \setminus \Omega'}$ is, by Theorem 6.111, in $BV(\Omega)$. As image

model we choose the TV functional, i.e., we want to find u^* that solves

$$\min_{u \in L^q(\Omega)} \text{TV}(u) + I_{\{v \in L^q(\Omega) \mid v|_{\Omega \setminus \Omega'} = u^0|_{\Omega \setminus \Omega'}\}}(u) \quad (6.69)$$

for some given $q \in]1, \infty[$ with $q \leq d/(d-1)$. Using Theorem 6.114 and the same techniques as in Application 6.98, we can prove the existence of such a u^* ; however, we do not know whether it is unique, since TV is not strictly convex.

For the optimality conditions we analyze the total variation functional on the set

$$K = \{v \in L^q(\Omega) \mid v = u^0 \text{ almost everywhere on } \Omega \setminus \Omega'\}.$$

A $v \in K$ with $\text{TV}(v) < \infty$ has to be of the form $v = u^0 + u$, where u is the zero extension of some element in $\text{BV}(\Omega')$. Conversely, Theorem 6.111 implies that the zero extension of every $u \in \text{BV}(\Omega')$ is in $\text{BV}(\Omega)$. Hence, we have that

$$\{u \in \text{BV}(\Omega) \mid u|_{\Omega \setminus \Omega'} = u^0|_{\Omega \setminus \Omega'}\} = \{u \in L^q(\Omega) \mid u|_{\Omega \setminus \Omega'} = u^0|_{\Omega \setminus \Omega'}, u|_{\Omega'} \in \text{BV}(\Omega')\}.$$

This allows us to rewrite problem (6.69) as

$$\min_{u \in L^q(\Omega)} \text{TV}(u\chi_{\Omega'} + u^0) + I_{\{v \in L^q(\Omega) \mid v|_{\Omega \setminus \Omega'} = u^0|_{\Omega \setminus \Omega'}\}}(u). \quad (6.70)$$

The first summand in (6.70), let us denote it by F_1 , is constant on the affine subspace $u^0 + X_1$, $X_1 = \{u \in L^q(\Omega) \mid u|_{\Omega'} = 0\}$ and hence continuous, while the second one, the indicator functional $F_2 = I_K$, is continuous on $u^0 + X_2$ with $X_2 = \{u \in L^q(\Omega) \mid u|_{\Omega \setminus \Omega'} = 0\}$ for the same reason. This ensures that the sum rule for subgradients can be applied in this situation (see Exercise 6.14). If we denote by A the mapping $u \mapsto u\chi_{\Omega'}$, we have that $\text{rg}(A) = X_2$, and according to Exercise 6.15, we can apply the chain rule for subgradients to

$$F_1 = \text{TV} \circ T_{u^0} \circ A$$

and obtain with Theorem 6.51 that

$$\partial F_1(u) = A^* \partial \text{TV}(u\chi_{\Omega'} + u^0),$$

where A^* is the zero extension $L^{q^*}(\Omega') \rightarrow L^{q^*}(\Omega)$. Applying the characterization of ∂TV leads to $w \in \partial F_1(u)$, and that $u|_{\Omega'} \in \text{BV}(\Omega')$ is equivalent to the existence of some $\sigma \in \mathcal{D}_{\text{div}, \infty}$ such that

$$\begin{cases} w = 0 & \text{in } \Omega \setminus \Omega', \\ \|\sigma\|_\infty \leq 1, & \\ -\operatorname{div} \sigma = w & \text{in } \Omega', \end{cases} \quad \begin{aligned} \sigma \cdot v = 0 & \quad \text{on } \partial\Omega, \\ \sigma = \frac{\nabla \bar{u}}{|\nabla \bar{u}|}, & \quad |\nabla \bar{u}| \text{ almost everywhere,} \end{aligned}$$

with $\bar{u} = u\chi_{\Omega'} + u^0$. Corollary 6.112 gives

$$\nabla \bar{u} = \nabla(u\chi_{\Omega'} + u^0) = \nabla(u|_{\Omega'}) + \nabla(u^0|_{\Omega \setminus \Omega'}) + (u^0|_{\partial(\Omega \setminus \Omega')} - u|_{\partial\Omega'})\nu \mathfrak{H}^{d-1} \llcorner \partial\Omega',$$

where $u^0|_{\partial(\Omega \setminus \Omega')}$ is the trace of u^0 on $\partial\Omega'$ with respect to $\Omega \setminus \Omega'$ and $u|_{\partial\Omega'}$ is the trace on $\partial\Omega'$ with respect to Ω' . Since $\operatorname{div} \sigma$ is arbitrary on $\Omega \setminus \Omega'$, σ plays no role there, and we can modify the condition for the trace of σ accordingly to

$$\begin{cases} \sigma = \frac{\nabla(u|_{\Omega'})}{|\nabla(u|_{\Omega'})|} & |\nabla(u|_{\Omega'})| \text{ almost everywhere,} \\ \sigma = \nu & \text{on } \{u|_{\partial\Omega'} < u^0|_{\partial(\Omega \setminus \Omega')}\}, \\ \sigma = -\nu & \text{on } \{u|_{\partial\Omega'} > u^0|_{\partial(\Omega \setminus \Omega')}\}, \end{cases}$$

and the latter equalities hold \mathfrak{H}^{d-1} -almost everywhere on $\partial\Omega'$.

The subdifferential of F_2 is similar to (6.45), and with the characterization of $\partial \operatorname{TV}$ we obtain, similarly to Application 6.98, that u^* is a minimizer of (6.69) if and only if there exists a $\sigma^* \in \mathcal{D}_{\operatorname{div}, \infty}$ such that

$$\begin{cases} u^* = u^0 & \text{in } \Omega \setminus \Omega', \\ -\operatorname{div} \sigma^* = 0, \quad \|\sigma^*\|_\infty \leq 1 & \text{in } \Omega', \\ \sigma^* \cdot \nu = 0 & \text{on } \partial\Omega, \\ \sigma^* = \frac{\nabla(u^*|_{\Omega'})}{|\nabla(u^*|_{\Omega'})|} & |\nabla(u^*|_{\Omega'})| \text{ almost everywhere,} \\ \sigma^* = \nu & \text{on } \{u^*|_{\partial\Omega'} < u^0|_{\partial(\Omega \setminus \Omega')}\}, \\ \sigma^* = -\nu & \text{on } \{u^*|_{\partial\Omega'} > u^0|_{\partial(\Omega \setminus \Omega')}\}. \end{cases} \quad (6.71)$$

In Ω' we can see $\operatorname{div} \sigma^*$ as the mean curvature of the level sets of u^* , and we see that the optimality conditions tell us that it has to vanish there. Hence, by definition, every level set is a so-called *minimal surface*, and this notion underlies a whole theory for BV functions; see, e.g., [65]. These minimal surfaces are connected to the level sets of u^0 on the boundary $\partial\Omega'$ where the traces of $u^*|_{\partial\Omega'}$ and $u^0|_{\partial(\Omega \setminus \Omega')}$ coincide. In this case we get the impression that u^* indeed connects the boundaries of objects. However, it can happen that u^* jumps at some parts of $\partial\Omega'$. According to (6.71) this can happen only under special circumstances, but still, in this case some level sets “end” at that point, and this gives results in which some objects appear to be “cut off.”

In the case of images, i.e., for $d = 2$, we even have that vanishing mean curvature means that TV inpainting connects object boundaries by straight lines (see Exercise 6.40). In fact, this can be seen in Figs. 6.22 and 6.23.

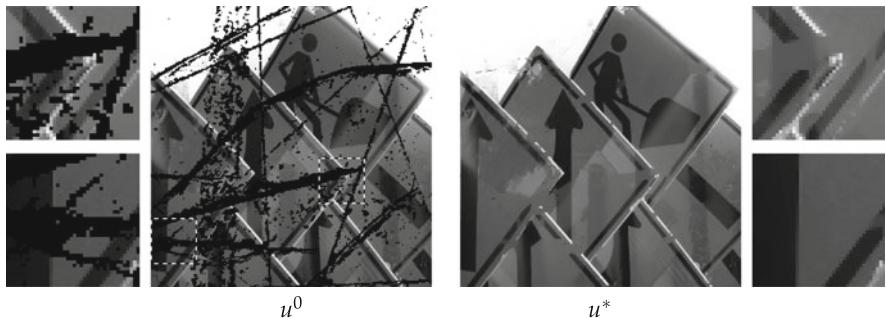


Fig. 6.22 Illustration of total variation inpainting. Left: Given data; the black region is the region which is to be restored (see also Fig. 6.14). Right: The solution of the minimization problem (6.69). Edges of larger objects are reconstructed well, but finer structures are disconnected more often

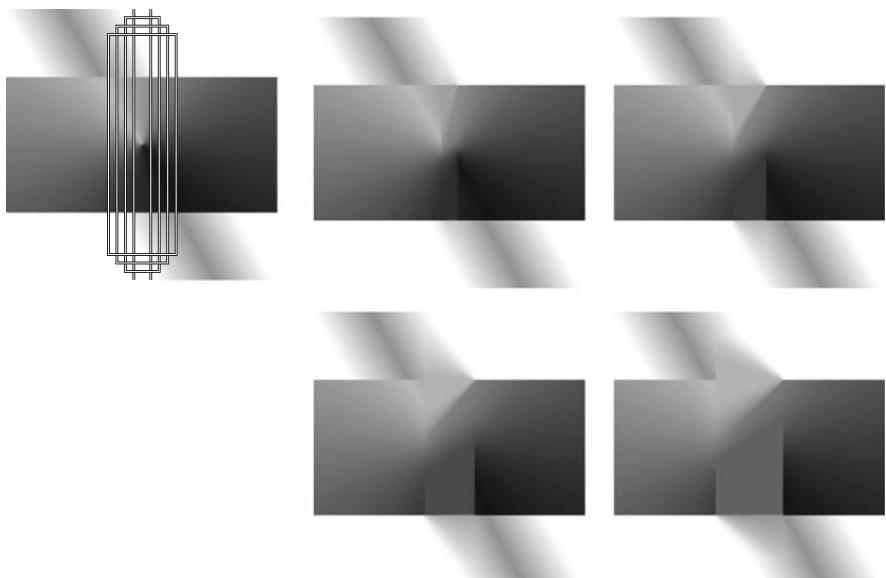


Fig. 6.23 TV inpainting connects level-sets with straight lines. Top row, left: An artificial image with marked inpainting domains Ω' of increasing width. Right and below: Solutions u^* of the inpainting problem (6.69) for these regions. One clearly sees that some level sets are connected by straight lines. At the points on the boundary $\partial\Omega'$ where this does not happen, which are the points where the solutions jump, such a connection would increase the total variation

In conclusion, we note that the TV model has favorable properties for the reconstruction of image data, but the solutions of TV inpainting problems like (6.69) may jump at the boundary of the inpainting domain, and hence the inpainting domain may still be visible after inpainting. Moreover, object boundaries can be connected only by straight lines which is not necessarily a good fit for the rest of the object's shape.

We also note that the solutions of (6.69) obey a maximum principle similar to inpainting with Sobolev semi-norm (Application 6.98). Moreover, a variant of gray value scaling invariance [GSI] from Chap. 5 is satisfied, similar to L^1 -TV denoising (Exercise 6.39).

Example 6.129 (Interpolation with Minimal Total Variation) We can also use the TV functional in the context of Application 6.100. We recall the task: for a discrete image $U^0 \in \mathbf{R}^{N \times M}$ we want to find a continuous $u^* : \Omega \rightarrow \mathbf{R}$ with $\Omega =]0, N[\times]0, M[$ such that u^* is an interpolation of the data U^0 . For a linear, continuous and surjective sampling operator $A : L^q(\Omega) \rightarrow \mathbf{R}^{N \times M}$ with $q \in]1, 2]$ we require that $Au^* = U^0$ holds. Moreover, u^* should correspond to some image model, in this case the total variation model. This leads to the minimization problem

$$\min_{u \in L^q(\Omega)} \text{TV}(u) + I_{\{v \in L^q(\Omega) \mid Av = U^0\}}(u). \quad (6.72)$$

If constant functions are not in the kernel of A , we can argue similarly to Application 6.100 using Theorem 6.114 and obtain the existence of a minimizer u^* . As in Example 6.128, minimizers are not necessarily unique.

We have all techniques to obtain optimality conditions for u^* from Application 6.100, we have only to modify (6.48) for ∂TV . With the $w^{i,j} \in L^{q^*}(\Omega)$ associated to the linear maps $u \mapsto (Au)_{i,j}$, i.e., $(Au)_{i,j} = \int_{\Omega} u w^{i,j} dx$, we can say that $u^* \in L^q(\Omega)$ is optimal in the sense of (6.72) if and only if there exist $\sigma^* \in \mathcal{D}_{\text{div}, \infty}$ and some $\lambda^* \in \mathbf{R}^{N \times M}$ such that

$$\left\{ \begin{array}{ll} \|\sigma^*\|_{\infty} \leq 1, \\ -\operatorname{div} \sigma^* = \sum_{i=1}^N \sum_{j=1}^M \lambda_{i,j}^* w^{i,j} & \text{in } \Omega, \\ \sigma^* \cdot \nu = 0 & \text{on } \partial \Omega, \\ \sigma^* = \frac{\nabla u^*}{|\nabla u^*|} & |\nabla u^*| \text{ almost everywhere,} \\ \int_{\Omega} u^* w^{i,j} dx = U_{i,j}^0, & 1 \leq i \leq N, \\ & 1 \leq j \leq M. \end{array} \right. \quad (6.73)$$

Hence, optimal solutions u^* satisfy an equation for the mean curvature that also has to lie in the subspace spanned by the $\{w^{i,j}\}$. If $w^{i,j} \in L^{\infty}(\Omega)$ for all i, j , then the mean curvature of the level sets of u^* has to be essentially bounded.

The actual form of solutions depends on the choice of the functions $w^{i,j}$, i.e., on the sampling operator A ; see Fig. 6.24 for some numerical examples. If one chooses the mean value over squares, i.e. $w^{i,j} = \chi_{[i-1, i[\times]j-1, j[}$, the level sets necessarily have constant curvature on these squares. The curvature is determined by $\lambda_{i,j}^*$. The level sets of u^* on $]i-1, i[\times]j-1, j[$ are, in the case of $\lambda_{i,j}^* = 0$, line segments (similar to Example 6.128), and are segments of circles in other cases

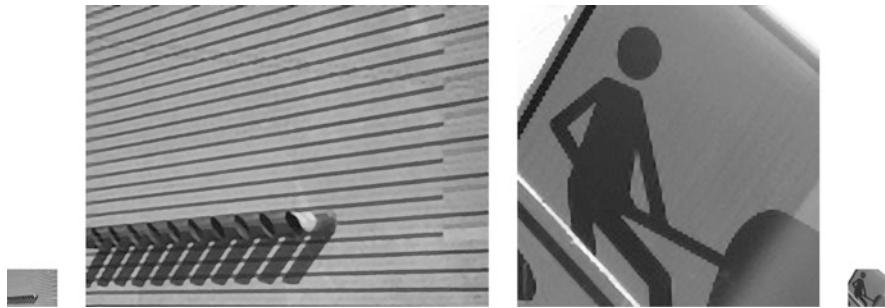


Fig. 6.24 Examples of TV interpolation with perfect low-pass filter. Outer left and right: Original images U^0 . Middle: Solutions u^* of the TV interpolation problem with eightfold magnification. All images are shown with the same resolution

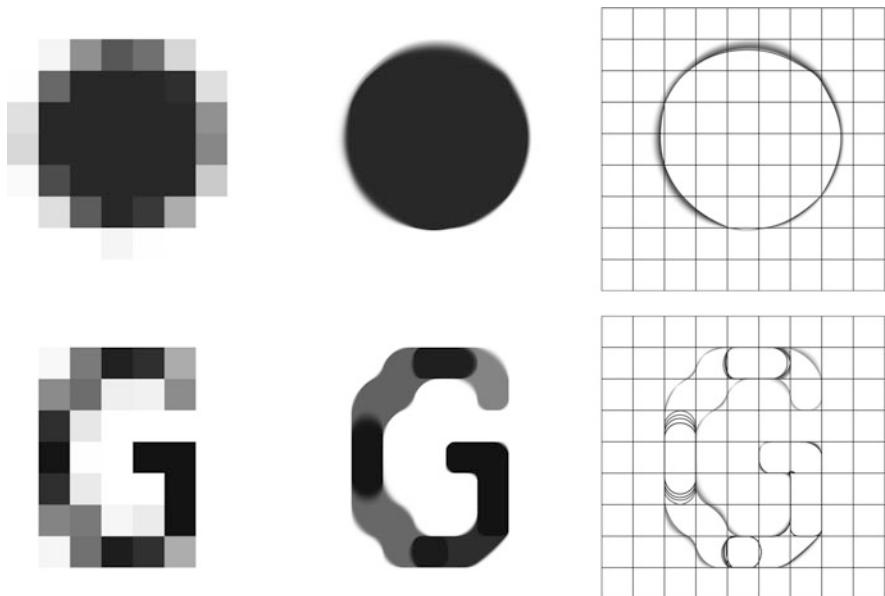


Fig. 6.25 TV interpolation with mean values over image valued leads to solutions with piecewise constant curvature of the level sets. Left: Original images U^0 (9×9 pixels). Middle: Solutions u^* of the respective TV interpolation problem with 60-fold magnification. Right: Level sets of u^* together with a grid of the original image

(see Exercise 6.40). Hence, TV interpolation is well suited for images that fulfill these characteristics. For images with a more complex geometry, however, we may still hope that the geometry is well approximated by these segments. On the other hand, it may also happen that straight lines are interpolated by segments of varying curvature. This effect is most strongly if the given Data U^0 only fits loosely to the unknown “true” image; see Fig. 6.25.

6.3.4 Generalization to Color Images

In the following we will briefly show how one can extend the introduced variational methods to color images. We recall: depending on the choice of the color space, a color image has N components, and hence can be modeled as a function $u : \Omega \rightarrow \mathbf{R}^N$. We discussed some effects of the choice of the color space, may it be RGB or HSV, already in Chap. 5. We could also note that methods that couple these color components in some way, usually lead to better results than methods without such coupling. Hence, we focus on the development of methods that couple the color components; moreover, we restrict ourselves to the RBG color space.

As an example, let us discuss the denoising problem with Sobolev semi-norm or total variation from Application 6.94 and Example 6.124, respectively. Let $u^0 \in L^q(\Omega, \mathbf{R}^N)$, $N \geq 1$, be a given noisy color image. The result of the variational problem (6.39) applied to all color channels separately amounts to the solution of

$$\min_{u \in L^q(\Omega, \mathbf{R}^N)} \frac{1}{q} \int_{\Omega} \left(\sum_{i=1}^N |u_i - u_i^0|^q \right) dx + \begin{cases} \frac{\lambda}{p} \int_{\Omega} \left(\sum_{i=1}^N |\nabla u_i|^p \right) dx & \text{if } p > 1, \\ \lambda \sum_{i=1}^N \text{TV}(u_i) & \text{if } p = 1, \end{cases} \quad (6.74)$$

respectively. To couple the color channels, we can choose different vector norms in \mathbf{R}^N for the data term and different matrix norms in $\mathbf{R}^{N \times d}$ for the penalty term. We want to do this in a way such that the channels do not separate, i.e., such that the terms are not both sums over the contributions of the channels $i = 1, \dots, N$. We focus on the pointwise matrix norm for ∇u , since one can see the influence of the norms more easily in this case, and choose the usual pointwise Euclidean vector norm on $L^q(\Omega, \mathbf{R}^N)$:

$$1 \leq q < \infty : \quad \|u\|_q = \left(\int_{\Omega} \left(\sum_{i=1}^N |u_i(x)|^2 \right)^{\frac{q}{2}} dx \right)^{\frac{1}{q}}, \quad \|u\|_{\infty} = \text{ess sup}_{x \in \Omega} \left(\sum_{i=1}^N |u_i(x)|^2 \right)^{\frac{1}{2}}.$$

The analogous choice for the matrix norm, i.e., the sum of the squares of entries, seems suitable; this amounts to the so-called *Frobenius norm* $|\nabla u(x)|_F^2 = |\nabla u(x)|^2 = \sum_{i=1}^N \sum_{j=1}^d |\partial_{x_j} u_i(x)|^2$, i.e., for $1 \leq p < \infty$

$$\|\nabla u\|_p = \left(\int_{\Omega} \left(\sum_{i=1}^N \sum_{j=1}^d \left| \frac{\partial u_i}{\partial x_j}(x) \right|^2 \right)^{\frac{p}{2}} dx \right)^{\frac{1}{p}}, \quad \|\nabla u\|_{\infty} = \text{ess sup}_{x \in \Omega} \left(\sum_{i=1}^N \sum_{j=1}^d \left| \frac{\partial u_i}{\partial x_j}(x) \right|^2 \right)^{\frac{1}{2}}.$$

If we define the divergence of a matrix-valued function componentwise, i.e., for $v : \Omega \rightarrow \mathbf{R}^{N \times d}$ we set $(\text{div } v)_i = \sum_{j=1}^d \partial_{x_j} v_{i,j}$, then we obtain the following

generalization of the total variation to the vectorial total variation:

$$\text{TV}(u) = \sup \left\{ \int_{\Omega} u \cdot \operatorname{div} v \, dx \mid v \in \mathcal{D}(\Omega, \mathbf{R}^{N \times d}), \|v\|_{\infty} \leq 1 \right\}. \quad (6.75)$$

For $q \neq 2$ or $p \neq 2$ the Sobolev and total variation denoising

$$\min_{u \in L^q(\Omega, \mathbf{R}^N)} \frac{1}{q} \int_{\Omega} |u - u^0|^q \, dx + \begin{cases} \frac{\lambda}{p} \|\nabla u\|_p^p & \text{if } p > 1, \\ \lambda \text{TV}(u) & \text{if } p = 1, \end{cases} \quad (6.76)$$

couples the color channels in the intended way.

We derive another possibility to couple the color channels and follow an idea from [124]. For a fixed point $x \in \Omega$ we consider a *singular value decomposition* of $\nabla u(x)$, i.e.

$$\nabla u(x) = \sum_{k=1}^K \sigma_k(x) (\eta_k(x) \otimes \xi_k(x)), \quad \begin{cases} \xi_1(x), \dots, \xi_K(x) \in \mathbf{R}^d & \text{orthonormal,} \\ \eta_1(x), \dots, \eta_K(x) \in \mathbf{R}^N & \text{orthonormal,} \\ \sigma_1(x), \dots, \sigma_K(x) \geq 0 & \text{singular values} \end{cases}$$

with $K = \min(d, N)$ and $(\eta \otimes \xi)_{i,j} = \eta_i \xi_j$ (see, e.g., [102]). The values $\sigma_k(x)$ are uniquely determined up to reordering. In the case $N = 1$ one such decomposition is $\sigma_1(x) = |\nabla u(x)|$, $\xi_1(x) = \frac{\nabla u}{|\nabla u|}(x)$ and $\eta_1(x) = 1$. For $N > 1$ we can interpret $\xi_k(x)$ as a “generalized” normal direction for which the color $u(x)$ changes in the direction $\eta_k(x)$ at the rate $\sigma_k(x)$. If $\sigma_k(x)$ is large (or small, respectively), then the color changes in the direction $\xi_k(x)$ a lot (or only slightly, respectively). In particular, $\max_{k=1, \dots, K} \sigma_k(x)$ quantifies the intensity of the largest change in color. The way to define a suitable matrix norm for ∇u is, to use this intensity as a norm:

$$|\nabla u(x)|_{\text{spec}} = \max_{k=1, \dots, K} \sigma_k(x), \quad \sigma_1(x), \dots, \sigma_K(x) \text{ singular values of } \nabla u(x).$$

This is indeed a matrix norm on $\mathbf{R}^{N \times d}$, the so-called *spectral norm*. It coincides with the operator norm of $\nabla u(x)$ as a linear mapping from \mathbf{R}^d to \mathbf{R}^N . To see this, note that for $z \in \mathbf{R}^d$ with $|z| \leq 1$ we get by orthonormality of $\eta_k(x)$ and $\xi_k(x)$, the Pythagorean theorem, and Parseval’s identity that

$$|\nabla u(x)z|^2 = \sum_{k=1}^K \sigma_k(x)^2 |\xi_k(x) \cdot z|^2 \leq \|\nabla u(x)\|_{\text{spec}}^2 \sum_{k=1}^K |\xi_k(x) \cdot z|^2 \leq |\nabla u(x)|_{\text{spec}}^2.$$

The supremum over all $|z| \leq 1$ is assumed at $z = \xi_k(x)$ with $\sigma_k(x) = \max_{l=1, \dots, K} \sigma_l(x)$, and hence $|\nabla u(x)|_{\text{spec}}$ equals the operator norm.

This allows us to define respective Sobolev semi-norms and total variation: For $1 \leq p < \infty$ let

$$\|\nabla u\|_{p,\text{spec}} = \left(\int_{\Omega} |\nabla u(x)|_{\text{spec}}^p dx \right)^{\frac{1}{p}}, \quad \|u\|_{\infty,\text{spec}} = \text{ess sup}_{x \in \Omega} |\nabla u(x)|_{\text{spec}},$$

as well as

$$\text{TV}_{\text{spec}}(u) = \sup \left\{ \int_{\Omega} u \cdot \text{div } v \, dx \mid v \in \mathcal{D}(\Omega, \mathbf{R}^{N \times d}), \|v\|_{\infty,\text{spec}} \leq 1 \right\}. \quad (6.77)$$

The denoising problem with this penalty term reads

$$\min_{u \in L^q(\Omega, \mathbf{R}^N)} \frac{1}{q} \int_{\Omega} |u - u^0|^q \, dx + \begin{cases} \frac{\lambda}{p} \int_{\Omega} |\nabla u(x)|_{\text{spec}}^p dx & \text{if } p > 1, \\ \lambda \text{TV}_{\text{spec}}(u) & \text{if } p = 1. \end{cases} \quad (6.78)$$

The minimization problems (6.74)–(6.78) are convex optimization problems in a Banach space. They can be treated with the methods developed in this chapter. In contrast to the case of gray value images, we have to use results for Sobolev spaces of vector valued functions and the space of vector valued functions of bounded total variation, respectively. Basically, the theory for existence and uniqueness of solutions carries over without problems, and hence we can speak of (unique) minimizers. Also, convex analysis, especially subgradient calculus, carries over, and we can derive Euler-Lagrange equations for optimality and discuss these equations. These are, depending on the coupling of the minimization problems, a coupled system of partial differential equations with couplings similar to the coupling of the color components for the Perona-Malik equation (see Sect. 5.3.1).

Figure 6.26 shows numerical results for the three variational methods for total variation denoising with quadratic data term. All methods give good results, but there are some differences in the details: there are color artifacts at sharp transitions, and the image seems have “too many colors” in these regions. This effect is most prominent for the separable functional (6.74); the coupling by the Frobenius norm (6.76) reduces the effect significantly. A further visual improvement is seen for the reconstruction with the spectral norm (6.78), where areas with constant color are reconstructed without any disturbances in the colors.

Other variational problems for color images can be formulated and solved, e.g., deblurring as in Example 6.127. This approach leads to similar results; see Fig. 6.27. It is instructive to observe the different effects for the different inpainting problems in Figs. 6.28 and 6.29 and see how different models for color images lead to different reconstructions for color transitions. Figure 6.28 clearly shows that different total variation models with different pointwise matrix norms make a large difference. The Frobenius norm (6.75) leads to many smooth color transitions in the reconstruction, while the spectral norm (6.77) leads to the formation of edges,

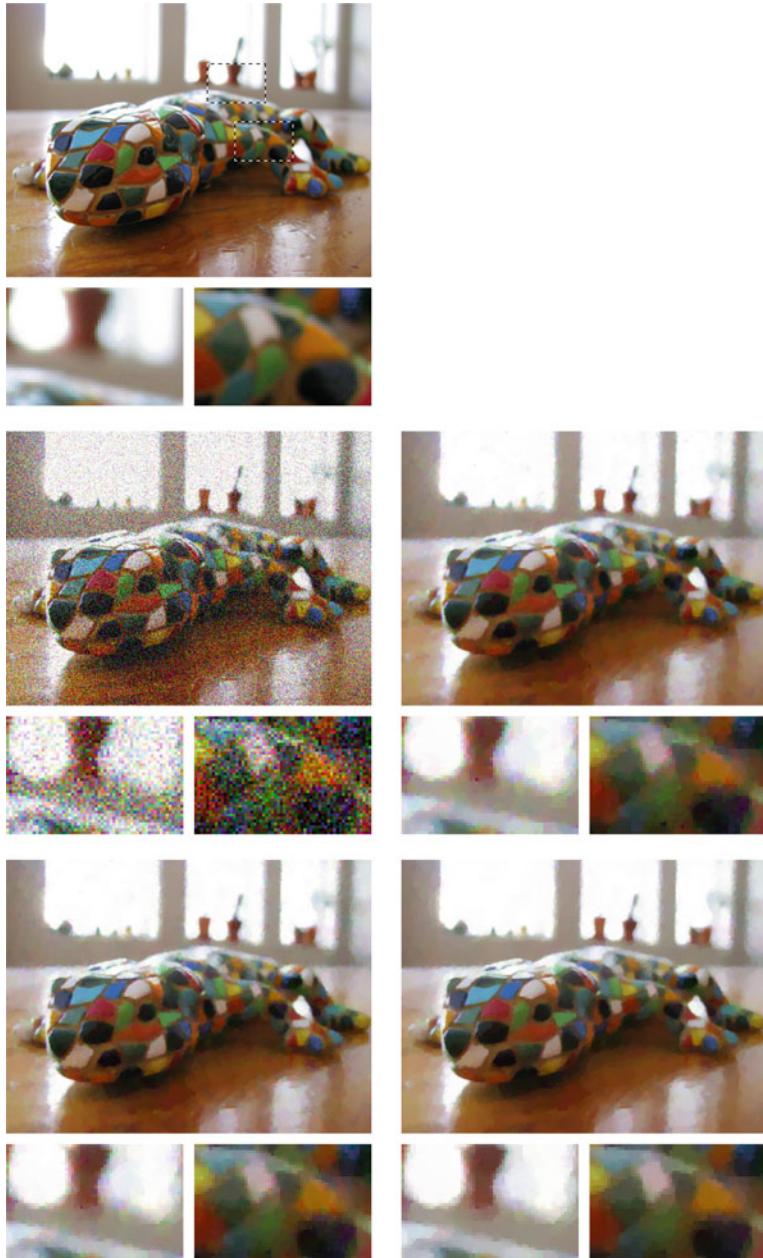


Fig. 6.26 Illustration of variational L^2 -TV denoising of color images. Top: The original u^\dagger with some marked details. Middle: Left the images with additive chromatic noise u^0 ($\text{PSNR}(u^0, u^\dagger) = 16.48 \text{ dB}$), right the solution u_{sep}^* with separable penalty term (6.74) ($\text{PSNR}(u_{\text{sep}}^*, u^\dagger) = 27.84 \text{ dB}$). Bottom: Left the solution u^* for the pointwise Frobenius matrix norm (6.76) ($\text{PSNR}(u^*, u^\dagger) = 28.46 \text{ dB}$), right the solution u_{spec}^* for the pointwise spectral norm (6.78) ($\text{PSNR}(u_{\text{spec}}^*, u^\dagger) = 28.53 \text{ dB}$)



Fig. 6.27 Solution of the variational deconvolution problem with monochromatic noise and blurred color data. Top: Original image u^\dagger . Bottom, left to right: the convolution kernel k , the given data u^0 , and the reconstruction u^* obtained by L^2 -TV deconvolution (with Frobenius matrix norm)



Fig. 6.28 Inpainting of color images with different models. From left to right: Given data with inpainting region, solutions for H^1 , TV, and TV_{spec} penalty, respectively

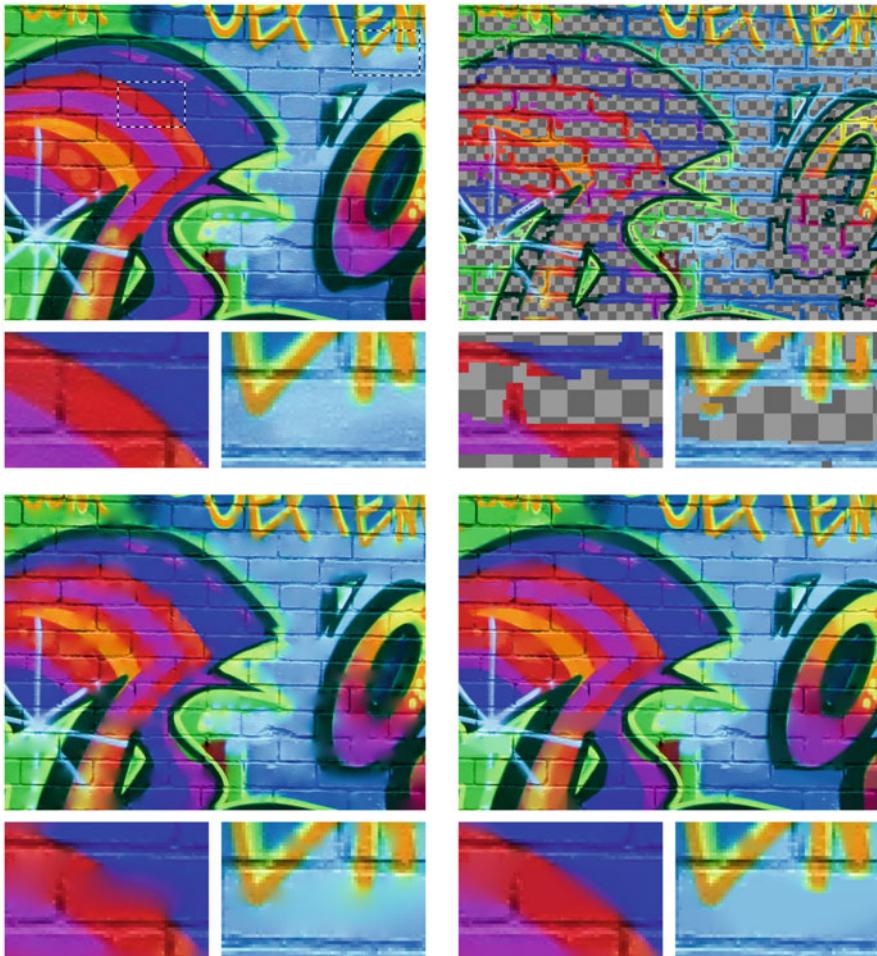


Fig. 6.29 Reconstruction of a color image from information along edges. Top: Left original u^\dagger with two highlighted regions, right the given data u^0 along edges and the inpainting region. Bottom: Reconstruction with H^1 inpainting (left) and TV inpainting (with Frobenius matrix norm, right)

which is characteristic for total variation method and which we have seen for gray images; cf. Fig. 6.23.

The use of the total variation with Frobenius matrix norm for the inpainting of mostly homogeneous regions has some advantages over H^1 inpainting (which is separable). In general, the reconstruction of edges is better and the “cropping effect” is not as strong as in the case of scalar gray-valued TV inpainting; see Fig. 6.29. For the sake of completeness, we mention that similar effects can be observed for variational interpolation. Figure 6.30 shows a numerical example in which the TV_{spec} penalty leads to sharper color transitions than the TV penalty with Frobenius matrix norm.



Fig. 6.30 Interpolation for color images. Top: Left the color image U^0 to be interpolated, right the sinc interpolation (zoom factor 4). Bottom: Solution of the TV interpolation problem (left) and the TV_{spec} interpolation problem (right) for fourfold zooming

6.4 Numerical Methods

Our ultimate goal is, as it was in Chap. 5, where we developed methods based on partial differential equations, to apply the method to concrete images. Since our variational methods are genuinely minimization problems for functions on continuous domains, we are faced with the question of appropriate discretization, but we also need numerical methods to solve the respective optimization problems. There exists a vast body of work on this topic, some of it developing methods for special problems in variational imaging while other research develops fairly abstract optimization concepts. In this section we will mainly introduce tools that allow us to solve the variational problems we developed in this chapter. Our focus is more on broad applicability of the tools than on best performance, high speed or efficiency. However, these latter aspects should not be neglected, but we refer to the original literature on these topics.

Let us start with the problem to find a solution for the convex minimization problem

$$\min_{u \in X} F(u)$$

over a Banach space X . We assume that F is proper, convex, lower semicontinuous, and that there exists a solution. By Theorem 6.43 it is necessary and sufficient

to solve the Euler-Lagrange equation $0 \in \partial F(u)$. If we assume that we can calculate the inverse graph $(\partial F)^{-1}$ by elementary operations that our computer can perform, we can just choose $u^* \in (\partial F)^{-1}(0)$ as a solution. However, in all our cases this is not the case; ∂F may be a nonlinear differential operator (as in the optimality conditions in (6.40), (6.46)), or contain further linear operators (see (6.42) and Application 6.97), and, on top of that, may even be multivalued (as in Application 6.100 and in all examples in Sect. 6.3.3). Moreover, neither continuity nor differentiability can be assumed, and hence classical numerical methods for the solution of nonlinear equations, such as Newton's methods, are of limited use.

6.4.1 Solving a Partial Differential Equation

A first very simple idea for solving Euler-Lagrange equations for some F is motivated by the fact that these equations are often partial differential equations of the form

$$-G(x, u(x), \nabla u(x), \nabla^2 u(x)) = 0 \quad \text{in } \Omega \quad (6.79)$$

with $G : \Omega \times \mathbf{R} \times \mathbf{R}^d \times S^{d \times d} \rightarrow \mathbf{R}$ and respective boundary conditions on $\partial\Omega$. Since the right-hand side is zero, we can interpret solutions of the equation as stationary points of the scale space associated to the differential operator G . Hence, we introduce a scale parameter $t \geq 0$ and consider

$$\frac{\partial u(t, x)}{\partial t} = G(x, u(t, x), \nabla u(t, x), \nabla^2 u(t, x)) \quad \text{in }]0, \infty[\times \Omega, \quad u(0, x) = f(x) \quad \text{in } \Omega \quad (6.80)$$

with arbitrary initial value $f : \Omega \rightarrow \mathbf{R}$ and boundary conditions on $[0, \infty[\times \partial\Omega$. If we assume that Eq. (6.80) is solvable and $\frac{\partial u}{\partial t}(t, \cdot) \rightarrow 0$ for $t \rightarrow \infty$ (both in some suitable sense), then the solution $u(T, \cdot)$, for some $T > 0$ large enough, is an approximation to the solution of (6.79). Equation (6.80) can be discretized and solved numerically with some of the methods from Sect. 5.4, e.g., with finite-difference approximations and a semi-implicit time stepping method. Then, one iterates until the difference of two consecutive steps is small enough (which makes sense, since this condition corresponds to a “small” discrete time derivative).

Example 6.130 (Variational Denoising) The Euler-Lagrange equation (6.40) of the minimization problem

$$\min_{u \in L^q(\Omega)} \frac{1}{q} \int_{\Omega} |u - u^0|^q \, dx + \frac{\lambda}{p} \int_{\Omega} |\nabla u|^p \, dx$$

from Application 6.94 is a nonlinear elliptic equation. The instationary version with initial value $f = u^0$ reads

$$\begin{cases} \frac{\partial u}{\partial t} - \operatorname{div}(|\nabla u|^{p-2}\nabla u) = |u^0 - u|^{q-2}(u^0 - u) & \text{in }]0, \infty[\times \Omega, \\ |\nabla u|^{p-2}\nabla u \cdot v = 0 & \text{on }]0, \infty[\times \partial\Omega, \\ u(0, \cdot) = u^0 & \text{in } \Omega, \end{cases}$$

and is a nonlinear diffusion equation. The methods from Sect. 5.4.1 allow a numerical solution as follows. We choose a spatial stepsize $h > 0$, a time stepsize $\tau > 0$, denote by U^n the discrete solution at time $n\tau$ (and with U^0 the discrete and noisy data), and we discretize the diffusion coefficient $|\nabla u|^{p-2}$ by

$$A(U)_{i \pm \frac{1}{2}, j} = \left(\frac{|\nabla U|_{i \pm 1, j} + |\nabla U|_{i, j}}{2} \right)^{p-2}, \quad |\nabla U|_{i, j}^2 = \frac{(U_{i+1, j} - U_{i, j})^2}{h^2} + \frac{(U_{i, j+1} - U_{i, j})^2}{h^2},$$

and $A(U)_{i, j \pm \frac{1}{2}}$ similarly and obtain, with the matrix $\mathbf{A}(U)$ from (5.23), the following semi-implicit method:

$$U^{n+1} = \left(\operatorname{id} - \frac{\tau}{h^2} \mathbf{A}(U^n) \right)^{-1} (U^n + \tau |U^0 - U^n|^{q-2} (U^0 - U^n)).$$

In every time step we have to solve a linear system of equations, which can be done efficiently.

In the case $p < 2$, the entries in A can become arbitrarily large or even become infinite. This leads to numerical problems. A simple workaround is the following trick: we choose a “small” $\varepsilon > 0$ and replace $|\nabla U|$ by

$$|\nabla U|_{i, j}^2 = \frac{(U_{i+1, j} - U_{i, j})^2}{h^2} + \frac{(U_{i, j+1} - U_{i, j})^2}{h^2} + \varepsilon^2.$$

This eliminates the singularities in A , and with $p = 1$ we can even approximate the denoising with total variation penalty from Example 6.124. However, this approach leads to reduced accuracy, and a “wrong” choice of ε may lead to results of bad quality.

Example 6.131 (Variational Inpainting) In Application 6.98 we faced an Euler-Lagrange equation that was a partial differential equation in Ω' , and its scale space version is

$$\begin{cases} \frac{\partial u}{\partial t} = \operatorname{div}(|\nabla u|^{p-2}\nabla u) & \text{in }]0, \infty[\times \Omega', \\ u = u^0 & \text{on }]0, \infty[\times \partial\Omega', \\ u(0, \cdot) = u^0|_{\Omega'} & \text{in } \Omega', \end{cases}$$

where $u^0 \in H^{1,p}(\Omega)$ on $\Omega \setminus \Omega'$ is the image to be extended.

We obtain a numerical method as follows. We assume that the image data U^0 is given on a rectangular grid. By Ω'_h we denote the set of discrete grid points in Ω' . For simplicity we assume that Ω'_h does not contain any boundary points of the rectangular grid. We define the *discrete boundary* as

$$(\partial\Omega')_h = \{(i, j) \notin \Omega_h \mid \{(i-1, j), (i+1, j), (i, j-1), (i, j+1)\} \cap \Omega_h \neq \emptyset\}.$$

The discrete images U will be defined on $\Omega'_h \cup (\partial\Omega')_h$. For a given U we denote by $\mathbf{A}(U)|_{\Omega'_h}$ the restriction of the matrix $A(U)$ from Example 6.130 on Ω'_h , i.e., the matrix in which we eliminated the rows and columns that belong to the indices that are *not* in Ω'_h . A semi-implicit method is then given by the successive solution of the linear system of equations

$$\begin{aligned} \left(\text{id}_{\Omega'_h} - \frac{\tau}{h^2} \mathbf{A}(U^n)|_{\Omega'_h} \right) U^{n+1}|_{\Omega'_h} &= U^n|_{\Omega'_h}, \\ \text{id}_{(\partial\Omega')_h} U^{n+1}|_{(\partial\Omega')_h} &= U^0|_{(\partial\Omega')_h}. \end{aligned}$$

As in Example 6.130, we may need to replace the diffusion coefficient $A(U)$ by a smoothed version in the case $p < 2$.

The above approach can also be applied for Euler-Lagrange equations that are not only partial differential equations.

Example 6.132 (Variational Deblurring) If we transform the optimality condition (6.42) of the deconvolution problem from Application 6.97 into an instationary equation and use, for simplicity, $q = 2$, we obtain the problem

$$\begin{cases} \frac{\partial u}{\partial t} - \text{div}(|\nabla u|^{p-2} \nabla u) = (u^0 - u * k) * \bar{k} & \text{in }]0, \infty[\times \Omega, \\ |\nabla u|^{p-2} \nabla u \cdot v = 0 & \text{on }]0, \infty[\times \partial\Omega, \\ u(0, \cdot) = f & \text{in } \Omega. \end{cases}$$

The convolution introduced implicit integral terms that appear in addition to the differential terms. Hence, we also need to discretize the convolution in addition to the partial differential equation (which can be done similarly to Example 6.130). From Sect. 3.3.3 we know how this can be done. We denote the matrix that implements the convolution with k on discrete images by \mathbf{B} , and its adjoint, i.e., the matrix for the convolution with \bar{k} , by \mathbf{B}^* . The right-hand side of the equation is affine linear in u , and we discretize it implicitly in the context of the semi-implicit method, i.e.,

$$\frac{U^{n+1} - U^n}{\tau} - \frac{1}{h^2} \mathbf{A}(U^n) U^{n+1} = \mathbf{B}^* U^0 - \mathbf{B}^* \mathbf{B} U^{n+1}.$$

We rearrange this into the iteration

$$U^{n+1} = \left(\text{id} + \tau \mathbf{B}^* \mathbf{B} - \frac{\tau}{h^2} \mathbf{A}(U^n) \right)^{-1} (U^n + \tau \mathbf{B}^* U^0).$$

It is simple to derive that the linear system has a unique solution, i.e., the iteration is well defined.

Remark 6.133 More abstractly, we note that the approach in this section realizes a so-called *gradient flow*. Let X be a Hilbert space and $F : X \rightarrow \mathbf{R}$ a continuously differentiable functional. Then using the Riesz mapping, we define a continuous “vector field” by $u \mapsto J_X^{-1}DF(u)$. Now we pose the problem to find some $u : [0, \infty[\rightarrow X$ that satisfies

$$\frac{\partial u}{\partial t} = -J_X^{-1}DF(u) \quad \text{for } t > 0, \quad u(0) = u^0,$$

with some $u^0 \in X$. We see that for every solution, one has

$$\begin{aligned} \left(\frac{\partial F \circ u}{\partial t} \right)(t) &= \left\langle DF(u(t)), \frac{\partial u}{\partial t}(t) \right\rangle = -(DF(u(t)), DF(u(t)))_{X^*} \\ &= -\|DF(u(t))\|_{X^*}^2 \leq 0. \end{aligned}$$

This shows that $u(t)$ reduces the functional values with increasing t . This justifies the use of gradient flows for minimization processes.

For real Hilbert spaces X we can generalize the notion of gradient flow to subgradients. If $F : X \rightarrow \overline{\mathbf{R}_\infty}$ is proper, convex, and lower semicontinuous, one can show that for every $u^0 \in \overline{\text{dom } \partial F}$ there exists a function $u : [0, \infty[\rightarrow X$ that solves, in some sense, the differential inclusion

$$-\frac{\partial u}{\partial t}(t) \in \partial F(u(t)) \quad \text{for } t > 0, \quad u(0) = u^0.$$

The study of such problems is the subject of the theory of nonlinear semigroups and monotone operators in Hilbert spaces, see, e.g., [21, 130].

The above examples show how one can use the well-developed theory of numerics of partial differential equations to easily obtain numerical methods to solve the optimality conditions of the variational problems. As we have seen in Examples 6.130–6.132, sometimes some modifications are necessary in order to avoid undesired numerical effects (cf. the case $p < 2$ in the diffusion equations). The reason for this is the discontinuity of the differential operators or, more abstractly, the discontinuity of the subdifferential. This problem occurs, in contrast to linear maps, even in finite dimensions and leads to problems with numerical methods.

Hence, we are going to develop another approach that does not depend on the evaluation of general subdifferentials.

6.4.2 Primal-Dual Methods

Again we consider the problem to minimize a proper, convex, and lower semicontinuous functional F , now on a real Hilbert space X . In the following we identify, via the Riesz map, $X = X^*$. In particular, we will consider the subdifferential as a graph in $X \times X$, i.e., $\partial F(u)$ is a subset of X .

Suppose that F is of the form $F = F_1 + F_2 \circ A$, where $F_1 : X \rightarrow \mathbf{R}_\infty$ is proper, convex, and lower semicontinuous, $A \in \mathcal{L}(X, Y)$ for some Banach space Y , and $F_2 : Y \rightarrow \mathbf{R}$ is convex and continuously differentiable. Moreover, we assume that F_1 has a “simple” structure, which we explain in more detail later. Theorem 6.51 states that the Euler-Lagrange equation for the optimality of u^* for the problem

$$\min_{u \in X} F_1(u) + F_2(Au) \quad (6.81)$$

is given by

$$0 \in \partial F_1(u^*) + A^*DF_2(Au^*).$$

Now we reformulate this equivalently: for an arbitrary $\sigma > 0$ we have

$$\begin{aligned} 0 &\in \partial F_1(u^*) + A^*DF_2(Au^*) \\ \Leftrightarrow -\sigma A^*DF_2(Au^*) &\in \sigma \partial F_1(u^*) \\ \Leftrightarrow u^* - \sigma A^*DF_2(Au^*) &\in (\text{id} + \sigma \partial F_1)(u^*) \\ \Leftrightarrow u^* &\in ((\text{id} + \sigma \partial F_1)^{-1} \circ (\text{id} - \sigma A^* \circ DF_2 \circ A))(u^*). \end{aligned} \quad (6.82)$$

Somewhat surprisingly, it turns out that the operation on the right-hand side of the last formulation is single-valued.

Lemma 6.134 *Let X be a real Hilbert space and $F : X \rightarrow \mathbf{R}_\infty$ proper, convex, and lower semicontinuous. For every $\sigma > 0$, one has that $(\text{id} + \sigma \partial F)^{-1}$ is characterized by the mapping that maps u to the unique minimizer of*

$$\min_{v \in X} \frac{\|v - u\|_X^2}{2} + \sigma F(v). \quad (6.83)$$

Moreover, the map $(\text{id} + \sigma \partial F)^{-1}$ is nonexpansive, i.e., for $u^1, u^2 \in X$,

$$\|(\text{id} + \sigma \partial F)^{-1}(u^1) - (\text{id} + \lambda \partial F)^{-1}(u^2)\|_X \leq \|u^1 - u^2\|_X.$$

Proof For $u \in H$ we consider the minimization problem (6.83). The objective functional is proper, convex, lower semicontinuous, and coercive, and by Theorem 6.31 there exists a minimizer $v^* \in X$. By strict convexity of the norm, this minimizer is X . By Theorems 6.43 and 6.51 (the norm is continuous in X), $v \in X$ is a minimizer if and only if

$$0 \in \partial\left(\frac{1}{2}\|\cdot\|_X^2 \circ T_{-u}\right)(v) + \sigma\partial F(v) \iff 0 \in v - u + \sigma\partial F(v).$$

The latter is equivalent to $v \in (\text{id} + \sigma\partial F)^{-1}(u)$, and by uniqueness of the minimizer we obtain that $v \in (\text{id} + \sigma\partial F)^{-1}(u)$ if and only if $v = v^*$. In particular, $(\text{id} + \sigma\partial F)^{-1}(u)$ is single-valued.

To prove the inequality we first show *monotonicity* of ∂F (cf. Theorem 6.33 and its proof). Let $v^1, v^2 \in X$ and $w^1 \in \partial F(v^1)$ as well as $w^2 \in \partial F(v^2)$. By the respective subgradient inequalities we get

$$(w^1, v^2 - v^1) \leq F(v^2) - F(v^1), \quad (w^2, v^1 - v^2) \leq F(v^1) - F(v^2).$$

Adding both inequalities leads to $(w^1 - w^2, v^2 - v^1) \leq 0$, and hence $(w^1 - w^2, v^1 - v^2) \geq 0$.

Now let $v^i = (\text{id} + \sigma\partial F)^{-1}(u^i)$ for $i = 1, 2$. This means that $u^i = v^i + \sigma w^i$ with $w^i \in \partial F(v^i)$ and the monotonicity of the subdifferential leads to

$$\|u^1 - u^2\|_X^2 = \|v^1 - v^2\|_X^2 + 2\sigma \underbrace{(v^1 - v^2, w^1 - w^2)_X}_{\geq 0} + \sigma^2 \underbrace{\|w^1 - w^2\|_X^2}_{\geq 0} \geq \|v^1 - v^2\|_X^2$$

as desired. \square

Remark 6.135 The mapping $(\text{id} + \sigma\partial F)^{-1}$ is also called the *resolvent* of ∂F for $\sigma > 0$.

Another name is the *proximal mapping* of σF , and denoted by $\text{prox}_{\sigma F}$ but we will stick to the resolvent notation in this book.

We revisit the result in (6.82) and observe that we have obtained an equivalent formulation of the optimality condition for u^* as a fixed point equation

$$u^* = ((\text{id} + \sigma\partial F_1)^{-1} \circ (\text{id} - \sigma A^* \circ DF_2 \circ A))(u^*).$$

This leads to a numerical method immediately, namely to the fixed point iteration

$$u^{n+1} = T(u^n) = ((\text{id} + \sigma\partial F_1)^{-1} \circ (\text{id} - \sigma A^* \circ DF_2 \circ A))(u^n). \quad (6.84)$$

This method is known as “forward backward splitting”, see [46, 93]. It is a special case of *splitting methods*, and depending on how the sum $\partial F_1 + A^*DF_2A$ is split up, one obtains different methods, e.g., the *Douglas-Rachford splitting method* or the *alternating direction method of multipliers*; see [56].

To justify the fixed point iteration, we assume that the iterates (u^n) converge to some u . By assumptions on A and F_2 we see that T is continuous, and hence we also have the convergence $T(u^n) \rightarrow T(u)$. Since $T(u^n) = u^{n+1}$, we see that $T(u) = u$, and this gives the optimality of u . Hence, the fixed point iteration is, in case of convergence, a continuous numerical method for the minimization of sums of the form $F_1 + F_2 \circ A$ with convex functionals under the additional assumption that F_2 is continuously differentiable.

Let us analyze the question whether and when $(\text{id} + \sigma \partial F_1)^{-1}$ can be computed by elementary operations in more detail. We cannot expect this to be possible for general functionals, but for several functionals that are interesting in our context, there are indeed formulae. We begin with some elementary rules of calculus for resolvents and look at some concrete examples.

Lemma 6.136 (Calculus for Resolvents) *Let $F_1 : X \rightarrow \mathbf{R}_\infty$ be a proper, convex, and lower semicontinuous functional on the real Hilbert space X , with Y another real Hilbert space and $\sigma > 0$.*

1. For $\alpha \in \mathbf{R}$

$$F_2 = F_1 + \alpha \quad \Rightarrow \quad (\text{id} + \sigma \partial F_2)^{-1} = (\text{id} + \sigma \partial F_1)^{-1},$$

i.e., if $F_2(u) = F_1(u) + \alpha$, then $(\text{id} + \sigma \partial F_2)^{-1}(u) = (\text{id} + \sigma \partial F_1)^{-1}(u)$,

2. for $\tau, \lambda > 0$

$$F_2 = \tau F_1 \circ \lambda \text{id} \quad \Rightarrow \quad (\text{id} + \sigma \partial F_2)^{-1} = \lambda^{-1} \text{id} \circ (\text{id} + \sigma \tau \lambda^2 \partial F_1)^{-1} \circ \lambda \text{id},$$

i.e., if $F_2(u) = \tau F_1(\lambda u)$, then $(\text{id} + \sigma \partial F_2)^{-1}(u) = \lambda^{-1} (\text{id} + \sigma \tau \lambda^2 \partial F_1)^{-1}(\lambda u)$,

3. for $u^0 \in X$, $w^0 \in X$

$$F_2 = F_1 \circ T_{u^0} + (w^0, \cdot) \quad \Rightarrow \quad (\text{id} + \sigma \partial F_2)^{-1} = T_{-u^0} \circ (\text{id} + \sigma \partial F_1)^{-1} \circ T_{u^0 - \sigma w^0},$$

i.e., if $F_2(u) = F_1(u + u^0) + (w^0, u)$, then $(\text{id} + \sigma \partial F_2)^{-1}(u) = (\text{id} + \sigma \partial F_1)^{-1}(u + u^0 - \sigma w^0) - u^0$,

4. for an isometric isomorphism $A \in \mathcal{L}(Y, X)$

$$F_2 = F_1 \circ A \quad \Rightarrow \quad (\text{id} + \sigma \partial F_2)^{-1} = A^* \circ (\text{id} + \sigma \partial F_1)^{-1} \circ A,$$

i.e., if $F_2(u) = F_1(Au)$, then $(\text{id} + \sigma \partial F_2)^{-1}(u) = A^* (\text{id} + \sigma \partial F_1)^{-1}(Au)$,

5. if $F_2 : Y \rightarrow \mathbf{R}_\infty$ is proper, convex, and lower semicontinuous, then

$$F_3(u, w) = F_1(u) + F_2(w) \quad \Rightarrow \quad (\text{id} + \sigma \partial F_3)^{-1}(u, w) = \begin{pmatrix} (\text{id} + \sigma \partial F_1)^{-1}(u) \\ (\text{id} + \sigma \partial F_2)^{-1}(w) \end{pmatrix}.$$

Proof Assertions 1–3: The proof of the identities consists of obvious and elementary steps, which we omit here.

Assertion 4: We reformulate the problem (6.83) for the calculation of the resolvent of ∂F_2 equivalently as

$$\begin{aligned} v^* \quad & \text{solves} \quad \min_{v \in X} \frac{\|v - u\|_X^2}{2} + \sigma F_1(Av) \\ \Leftrightarrow \quad v^* \quad & \text{solves} \quad \min_{v \in \text{rg}(A^*)} \frac{\|A(v - u)\|_Y^2}{2} + \sigma F_1(Av) \\ \Leftrightarrow \quad Av^* \quad & \text{solves} \quad \min_{w \in Y} \frac{\|w - Au\|_Y^2}{2} + \sigma F_1(w). \end{aligned}$$

Here we used both the bijectivity of A and $A^*A = \text{id}$. The last formulation is equivalent to $v^* = A^*(\text{id} + \sigma \partial F_1)^{-1}(Au)$, and thus

$$(\text{id} + \sigma \partial F_2)^{-1} = A^* \circ (\text{id} + \sigma \partial F_1)^{-1} \circ A.$$

Assertion 5: Similar to Assertion 4 we write

$$\begin{aligned} (v^*, \omega^*) \quad & \text{solves} \quad \min_{\substack{v \in X \\ \omega \in Y}} \frac{\|v - u\|_X^2}{2} + \sigma F_1(v) + \frac{\|\omega - w\|_Y^2}{2} + \sigma F_2(\omega) \\ \Leftrightarrow \quad & \begin{cases} v^* \quad \text{solves} \quad \min_{v \in X} \frac{\|v - u\|_X^2}{2} + \sigma F_1(v), \\ \omega^* \quad \text{solves} \quad \min_{\omega \in Y} \frac{\|\omega - w\|_Y^2}{2} + \sigma F_2(\omega). \end{cases} \end{aligned}$$

The first minimization problem is solved only by $v^* = (\text{id} + \sigma \partial F_1)^{-1}(u)$, and the second only by $\omega^* = (\text{id} + \sigma \partial F_2)^{-1}(w)$, i.e.,

$$(\text{id} + \sigma \partial F_3)^{-1}(u, w) = \left(\begin{array}{l} (\text{id} + \sigma \partial F_1)^{-1}(u) \\ (\text{id} + \sigma \partial F_2)^{-1}(w) \end{array} \right)$$

as desired. \square

Example 6.137 (Resolvent Maps)

1. Functionals in \mathbf{R}

For $F : \mathbf{R} \rightarrow \mathbf{R}_\infty$ proper, convex, and lower semicontinuous, $\text{dom } \partial F$ has to be an interval (open, half open, or closed), and every $\partial F(t)$ is a closed interval which we denote by $[G^-(t), G^+(t)]$ (where the values $\pm\infty$ are explicitly allowed, but then are excluded from the interval). The functions G^- , G^+ are monotonically increasing in the sense that $G^+(s) \leq G^-(t)$ for $s < t$.

Often it is possible to approximate $(\text{id} + \sigma \partial F)^{-1}(t)$ numerically. Suppose that we know an initial interval in which the solution lies, i.e., we know $s^- < s^+$ with

$$s^+ + \sigma G^+(s^-) < t < s^+ + \sigma G^-(s^+). \quad (6.85)$$

Then we can perform bisection as follows:

- Choose $s = \frac{1}{2}(s^- + s^+)$ and check whether $t \in s + \sigma[G^-(s), G^+(s)]$: in this case set $s = (\text{id} + \sigma \partial F)^{-1}(t)$.
- In the case $t < s + \sigma G^-(s)$ replace s^+ by s .
- In the case $t > s + \sigma G^+(s)$ replace s^- by s .

One immediately sees that, in the case the solution has not been found, the new bounds s^-, s^+ again satisfy (6.85). Hence, one can simply iterate until a desired precision has been reached.

Moreover, we remark that for continuously differentiable F the calculation of the resolvent reduces to solving

$$s + \sigma F'(s) = t.$$

If we have an approximate solution s_0 , in which F' is differentiable, we can use Newton's method to obtain a better approximation:

$$s = s_0 + \frac{t - s_0 - \sigma F'(s_0)}{1 + \sigma F''(s_0)}.$$

If F is well behaved enough and s_0 is close enough to the solution, then the iteration converges faster than bisections and this method should be preferred. However, the prerequisites may not always be satisfied.

2. Norm functionals

Let F be of the form $F(u) = \varphi(\|u\|_X)$ with $\varphi : [0, \infty[\rightarrow \mathbf{R}_\infty$ proper, convex, lower semicontinuous, and increasing. Then we can express $(\text{id} + \sigma \partial F)^{-1}$ in terms of the resolvent of $\partial \varphi$: we observe that $v = (\text{id} + \sigma \partial F)^{-1}(u)$ holds if and only if

$$\begin{aligned} u \in v + \sigma \partial F(v) &= \partial\left(\frac{1}{2}\|\cdot\|_X^2 + \sigma F\right)(v) = \partial\left(\left(\frac{1}{2}\|\cdot\|^2 + \sigma \varphi\right) \circ \|\cdot\|_X\right)(v) \\ &= \{w \in X \mid (w, v)_X = \|w\|_X \|v\|_X, \|w\|_X \in (\text{id} + \sigma \partial \varphi)(\|v\|_X)\} \end{aligned}$$

(cf. Example 6.49). Moreover, $(u, v)_X = \|u\|_X \|v\|_X$ if and only if $v = \lambda u$ for some $\lambda \geq 0$, and hence the properties that define the set in the last equation are equivalent to $\|v\|_X = (\text{id} + \sigma \partial \varphi)^{-1}(\|u\|_X)$ and $v = \lambda u$ for a $\lambda \geq 0$, which in turn is equivalent to

$$v = \begin{cases} (\text{id} + \sigma \partial \varphi)^{-1}(\|u\|_X) \frac{u}{\|u\|_X} & \text{if } u \neq 0, \\ 0 & \text{if } u = 0. \end{cases}$$

By the monotonicity of φ , we have $(\text{id} + \sigma \partial \varphi)^{-1}(0) = 0$, and we can write the resolvent, using the notation $\frac{0}{0} = 1$, as

$$(\text{id} + \sigma \partial(\varphi \circ \|\cdot\|_X))^{-1}(u) = (\text{id} + \sigma \partial \varphi)^{-1}(\|u\|_X) \frac{u}{\|u\|_X}.$$

3. Linear quadratic functionals

Let $Q : X \rightarrow X$ be a linear, continuous, self-adjoint, and positive semidefinite map, i.e., $Q \in \mathcal{L}(X, X)$ with $(Qu, v) = (u, Qv)$ for all $u, v \in X$ and $(Qu, u) \geq 0$ for all u . Moreover, let $w \in X$. We consider the functional

$$F(u) = \frac{(Qu, u)_X}{2} + (w, u)_X,$$

which is differentiable with derivative $DF(u) = Qu + w$. By the assumption on Q ,

$$\begin{aligned} F(u) + (DF(u), v - u)_X &= \frac{(Qu, u)_X}{2} + (w, u)_X + (Qu + w, v - u)_X \\ &= -\frac{(Qu, u)_X}{2} + (Qu + w, v)_X - \frac{(Qv, v)_X}{2} + \frac{(Qv, v)_X}{2} \\ &= -\frac{(Q(u - v), (u - v))_X}{2} + \frac{(Qv, v)_X}{2} + (w, v)_X \\ &\leq F(v) \end{aligned}$$

for all $u, v \in X$, and this implies, using Theorem 6.33, the convexity of F . For given σ and $u \in X$ we would like to calculate the resolvent $(\text{id} + \sigma \partial F)^{-1}(u)$. One has $u = v + \sigma Qv + \sigma w$ if and only if $v = (\text{id} + \sigma Q)^{-1}(u - \sigma w)$, and hence

$$(\text{id} + \sigma \partial F)^{-1}(u) = (\text{id} + \sigma Q)^{-1}(u - \sigma w).$$

Thus in finite-dimensional spaces, the resolvent is a translation followed by the solution of a linear system of equations. It is easy to see that $\text{id} + \sigma Q$ corresponds to a positive definite matrix, and hence one can use, for example, the method of conjugate gradients to compute $(\text{id} + \sigma Q)^{-1}(u - \sigma w)$, and one may even use a preconditioner; cf. [66].

4. Indicator functionals

Let $K \subset X$ be nonempty, convex, and closed, and let $F = I_K$ be the respective indicator functional. The resolvent in $u \in X$ amounts to the solution of problem (6.83), which is equivalent to the projection

$$\min_{v \in K} \frac{\|v - u\|_X^2}{2}.$$

Hence,

$$(\text{id} + \sigma \partial I_K)^{-1}(u) = P_K(u),$$

and in particular, the resolvent does not depend on σ . Without further information on K we cannot say how the projection can be calculated, but in several special cases, simple ways to do so exist.

- (a) Let K be a nonempty closed interval in \mathbf{R} . Depending on the boundedness, we get

$$\begin{aligned} K = [a, \infty[&\Rightarrow (\text{id} + \sigma \partial I_K)^{-1}(t) = \max(a, t), \\ K =]-\infty, b] &\Rightarrow (\text{id} + \sigma \partial I_K)^{-1}(t) = \min(b, t), \\ K = [a, b] &\Rightarrow (\text{id} + \sigma \partial I_K)^{-1}(t) = \max(a, \min(b, t)). \end{aligned}$$

- (b) Let K be a closed subspace. With a complete orthonormal system V of K we can express the projection as $P_K(u) = \sum_{v \in V} (v, u)v$. The formula becomes simpler if K is separable, since then V is at most countable and

$$P_K(u) = \sum_{n=1}^{\dim K} (v^n, u)_X v^n.$$

If one has only some basis $\{v^1, \dots, v^N\}$ for a subspace K with $\dim K < \infty$ and Gram-Schmidt orthonormalization would be too expensive, one could still calculate the projection $P_K(u)$ by solving the linear system of equations

$$x \in \mathbf{R}^N : \quad Mx = b, \quad M_{i,j} = (v^i, v^j)_X, \quad b_j = (u, v^j)_X$$

and setting $P_K(u) = \sum_{i=1}^N x_i v^i$. By definition, M is positive definite and hence invertible. However, depending on the basis $\{v^i\}$, it may be badly conditioned.

5. Convex integrands and summands

Let $X = L^2(\Omega, \mathbf{R}^N)$ be the Lebesgue-Hilbert space with respect to a measure space $(\Omega, \mathfrak{F}, \mu)$ and let $\varphi : \mathbf{R}^N \rightarrow \mathbf{R}_\infty$ be proper, convex, and lower semicontinuous with known resolvent $(\text{id} + \sigma \partial \varphi)^{-1}$. Moreover, assume $\varphi \geq 0$ and $\varphi(0) = 0$ if Ω has infinite measure, and φ bounded from below otherwise (cf. Examples 6.23 and 6.29).

The minimization problem (6.83) corresponding to the resolvent of the subgradient of

$$F(u) = \int_{\Omega} \varphi(u(x)) \, dx$$

reads

$$\min_{v \in L^2(\Omega, \mathbf{R}^N)} \int_{\Omega} \frac{|v(x) - u(x)|^2}{2} + \sigma \varphi(v(x)) \, dx,$$

and by Example 6.50 we can express the subdifferential pointwise almost everywhere. Hence, we have $v^* = (\text{id} + \sigma \partial F)^{-1}(u)$ if and only if

$$0 \in v^*(x) - u(x) + \sigma \partial \varphi(v^*(x)) \quad \text{for almost all } x \in \Omega,$$

which in turn is equivalent to

$$v^*(x) = (\text{id} + \sigma \partial \varphi)^{-1}(u(x)) \quad \text{for almost all } x \in \Omega.$$

Thus, the resolvent satisfies the identity

$$(\text{id} + \sigma \partial F)^{-1}(u) = (\text{id} + \sigma \partial \varphi)^{-1} \circ u.$$

In the special case of finite sums, we have an even more general result: for a given family $\varphi_1, \dots, \varphi_M : \mathbf{R}^N \rightarrow \mathbf{R}_{\infty}$ of proper, convex, and lower semicontinuous functionals one gets for $u : \{1, \dots, M\} \rightarrow \mathbf{R}^N$,

$$F(u) = \sum_{j=1}^M \varphi_j(u_j) \quad \Rightarrow \quad (I + \sigma \partial F)^{-1}(u)_j = (I + \sigma \partial \varphi_j)^{-1}(u_j).$$

Applying the rules of the calculus of resolvents and using the previous examples, we obtain a fairly large class of functionals for which we can evaluate the resolvent by elementary means and which can be used for practical numerical methods. One important case is the p th power of the L^p norms.

Example 6.138 (Resolvent of $\partial \frac{1}{p} \|\cdot\|_p^p$) For some measure space $(\Omega, \mathfrak{F}, \mu)$ we consider $X = L^2(\Omega, \mathbf{R}^N)$ and the functionals

$$F(u) = \frac{1}{p} \int_{\Omega} |u(x)|^p \, dx$$

for $p \in [1, \infty[$ and

$$F(u) = \int_{\Omega} I_{\{\cdot \leq 1\}}(u(x)) \, dx$$

for $p = \infty$, respectively. We aim to calculate the resolvent of the subgradient $\sigma \partial F$ for all $\sigma > 0$. By item 5 in Example 6.137 we restrict our attention to the resolvent of the subgradient of $x \mapsto \frac{1}{p}|x|^p$ in \mathbf{R}^N . This functional depends only on the Euclidean

norm $|\cdot|$, and by item 2 in Example 6.137 we have only to determine the resolvent of $\partial\varphi_p$ with

$$\varphi_p(t) = \begin{cases} \frac{1}{p}t^p & \text{if } t \geq 0, \\ 0 & \text{otherwise,} \end{cases} \quad \text{for } p < \infty, \quad \varphi_\infty(t) = I_{]-\infty, 1]}(t) \quad \text{for } p = \infty,$$

as a function on \mathbf{R} , and moreover, this only for positive arguments. Let us discuss this for fixed $t \geq 0$ and varying p : In the case $p = 1$ we can reformulate $s = (\text{id} + \sigma \partial\varphi)^{-1}(t)$ by definition as

$$t \in \begin{cases} [0, \sigma] & \text{if } s = 0 \\ \{s + \sigma\} & \text{otherwise.} \end{cases}$$

It is easy to see that this is equivalent to $s = \max(0, t - \sigma)$. In the case $p \in]1, \infty[$ we have that $s = (\text{id} + \sigma \partial\varphi)^{-1}(t)$ is equivalent to solving

$$t = s + \sigma s^{p-1},$$

for s . For $p = 2$ this is the same as $s = t/(1 + \sigma)$ and in all other cases, this is a nonlinear equation, but for $t = 0$ we have that $s = 0$ is always a solution. If $p > 2$ is an integer, the problem is equivalent to finding a root of a polynomial of degree $p - 1$. As is known, this problem can be solved in closed form using roots in the cases $p = 3, 4, 5$ (with the quadratic formula, Cardano's method, and Ferrari's method, respectively). By the substitution $s^{p-1} = \xi$ we can extend this method to the range $p \in]1, 2[$; in this way, we can also treat the cases $p = \frac{3}{2}, \frac{4}{3}, \frac{5}{4}$ exactly. For all other $p \in]1, \infty[$ we resort to a numerical method. For $t > 0$ we can employ Newton's method, for example, and get the iteration

$$s_{n+1} = s_n + \frac{t - s_n - \sigma s_n^{p-1}}{1 + \sigma(p-1)s_n^{p-2}}.$$

The iterates are decreasing and converge, giving high precision after a few iterations, if of initializes the iteration with

$$s_0 \geq t \quad \text{and} \quad s_0 < \left(\frac{t}{\sigma(2-p)}\right)^{\frac{1}{p-1}} \quad \text{if } p < 2;$$

see Exercise 6.41. The only case remaining is $p = \infty$, but this is treated already in item 4 of Example 6.137 and gives $s = \min(1, t)$.

In conclusion, we can calculate the resolvent of ∂F at a point $u \in L^2(\Omega, \mathbf{R}^N)$ as follows (using Newton's method for all $p \in]1, \infty[$ not explicitly mentioned otherwise):

- **The case $p = 1$**

$$F(u) = \int_{\Omega} |u(x)| \, dx \quad \Rightarrow \quad (\text{id} + \sigma \partial F)^{-1}(u) = \max(0, |u| - \sigma) \frac{u}{|u|}.$$

This is exactly the so-called soft thresholding.

- **The case $p = 2$**

$$F(u) = \frac{1}{2} \int_{\Omega} |u(x)|^2 \, dx \quad \Rightarrow \quad (\text{id} + \sigma \partial F)^{-1}(u) = \frac{u}{1 + \sigma}.$$

- **The case $p = \infty$**

$$F(u) = I_{\{\|v\|_{\infty} \leq 1\}}(u) \quad \Rightarrow \quad (\text{id} + \sigma \partial F)^{-1}(u) = \min(1, |u|) \frac{u}{|u|} = \frac{u}{\max(1, |u|)}.$$

Hence, the resolvent is the pointwise projection onto the closed unit ball in \mathbf{R}^N .

- **The case $1 < p < \infty$**

Newton's method amounts to the following procedure:

1. Set $v = |u|$ and choose some $v^0 \in L^2(\Omega)$ with

$$\begin{cases} v^0(x) \geq v(x) & \text{almost everywhere in } \Omega, \\ v^0(x) < \left(\frac{v(x)}{\sigma(2-p)}\right)^{\frac{1}{p-1}} & \text{almost everywhere in } \{v(x) \neq 0\} \quad \text{if } p < 2. \end{cases}$$

2. Iterate

$$v^{n+1} = v^n + \frac{v - v^n - \sigma |v^n|^{p-1}}{1 + \sigma(p-1)|v^n|^{p-2}}.$$

The sequence converges monotonically decreasing and pointwise almost everywhere to some $v^* \in L^2(\Omega)$, and by Lebesgue's dominated convergence theorem (Theorem 2.47) also in $L^2(\Omega)$.

3. With this scalar-valued limit function $v^* = (\text{id} + \sigma |\cdot|^{p-1})^{-1} \circ u$, one has

$$F = \frac{1}{p} \|\cdot\|_p^p \quad \Rightarrow \quad (\text{id} + \sigma \partial F)^{-1}(u) = v^* \frac{u}{|u|}.$$

We continue our discussion of suitable optimization methods. If the assumption of continuous differentiability that were needed in the derivation of the fixed point iteration (6.84) are fulfilled, then this fixed point iteration is a suitable method.

Unfortunately, the assumption of continuous differentiability of one term is often not satisfied for problems in imaging. We consider the simple example of Tikhonov functionals (cf. Example 6.32)

$$\min_{u \in X} F_1(u) + F_2(Au), \quad F_1(u) = \frac{\lambda \|u\|_X^p}{p}, \quad F_2(v) = \frac{\|v - u^0\|_Y^q}{q}.$$

Then F_2 may be continuous but not continuously differentiable. Hence, we aim for a method that can treat more general functionals F_2 .

At this point, Fenchel duality comes in handy. We assume that F_2 is the Fenchel conjugate with respect to a Hilbert space Y , i.e., $F_2 = F_2^{**}$ in Y . In the following we identify the spaces $Y = Y^*$ also for conjugation, i.e.,

$$w \in Y : \quad F_2^*(w) = \sup_{v \in Y} (w, v)_Y - F_2(v), \quad v \in Y : \quad F_2(v) = \sup_{w \in Y} (v, w)_Y - F_2^*(w),$$

and similarly for conjugation in X . We postulate that also F_2^* is “simple enough,” but do not make assumptions on continuity or differentiability (note that lower semicontinuity is already implied, though).

In the following we also need that the conclusion of Fenchel-Rockafellar duality (see Theorem 6.68 for sufficient conditions) holds:

$$\max_{w \in Y} -F_1^*(-A^*w) - F_2^*(w) = \min_{u \in X} F_1(u) + F_2(Au). \quad (6.86)$$

By Remark 6.72 we know that a simultaneous solution of the primal-dual problem is equivalent to finding a saddle point of the Lagrange functional $L : \text{dom } F_1 \times \text{dom } F_2^* \rightarrow \mathbf{R}$, which is defined by

$$L(u, w) = (w, Au)_Y + F_1(u) - F_2^*(w). \quad (6.87)$$

Recall that $(u^*, w^*) \in \text{dom } F_1 \times \text{dom } F_2^*$ is a saddle point of L if and only if for all $(u, w) \in \text{dom } F_1 \times \text{dom } F_2^*$,

$$L(u^*, w) \leq L(u^*, w^*) \leq L(u, w^*).$$

For the Lagrange functional we define for every pair $(u^0, w^0) \in \text{dom } F_1 \times \text{dom } F_2^*$ the restrictions $L_{w^0} : X \rightarrow \mathbf{R}_\infty$, $L_{u^0} : Y \rightarrow \mathbf{R} \cup \{-\infty\}$ by

$$L_{w^0}(u) = \begin{cases} L(u, w^0) & \text{if } u \in \text{dom } F_1, \\ \infty & \text{otherwise,} \end{cases} \quad L_{u^0}(w) = \begin{cases} L(u^0, w) & \text{if } w \in \text{dom } F_2^*, \\ -\infty & \text{otherwise.} \end{cases}$$

It is simple to see, using the notions of Definition 6.60 and the result of Lemma 6.57, that $L_{w^0} \in \Gamma_0(X)$ and $-L_{u^0} \in \Gamma_0(Y)$. Hence by Lemma 6.134 the resolvents $(\text{id} + \sigma \partial L_{w^0})^{-1}$ and $(\text{id} + \tau \partial (-L_{u^0}))^{-1}$ exist, and with the help of Lemma 6.136

one easily checks that

$$\begin{aligned} (\text{id} + \sigma \partial L_{w^0})^{-1}(u) &= (\text{id} + \sigma \partial F_1)^{-1}(u - \sigma A^* w^0), \\ (\text{id} + \tau \partial(-L_{u^0}))^{-1}(w) &= (\text{id} + \tau \partial F_2^*)^{-1}(w + \tau A u^0). \end{aligned}$$

Thus, the property of $(u^*, w^*) \in \text{dom } F_1 \times \text{dom } F_2^*$ being a saddle point of L is equivalent to

$$u^* \text{ solves } \min_{u \in X} L_{w^*} \quad \text{and} \quad w^* \text{ solves } \min_{w \in Y} (-L_{u^*}).$$

The optimality conditions from Theorem 6.43 allow, for arbitrary $\sigma, \tau > 0$, the following equivalent formulations:

$$\begin{aligned} (u^*, w^*) \text{ saddle point} &\Leftrightarrow \left\{ 0 \in \partial L_{w^*}(u^*), \quad 0 \in \partial(-L_{u^*})(w^*) \right. \\ &\Leftrightarrow \left\{ \begin{array}{l} u^* \in u^* + \sigma \partial L_{w^*}(u^*), \\ w^* \in w^* + \tau \partial(-L_{u^*})(w^*) \end{array} \right. \\ &\Leftrightarrow \left\{ \begin{array}{l} u^* = (\text{id} + \sigma \partial L_{\bar{w}^*})^{-1}(u^*), \\ w^* = (\text{id} + \tau \partial(-L_{\bar{u}^*}))^{-1}(w^*), \\ \bar{u}^* = u^*, \quad \bar{w}^* = w^*. \end{array} \right. \end{aligned}$$

The pair $(\bar{u}^*, \bar{w}^*) \in X \times Y$ has been introduced artificially and will denote the points at which we take the resolvents of $\partial L_{\bar{w}^*}$ and $\partial(-L_{\bar{u}^*})$. Hence, we have again a formulation of optimality in terms of a fixed point equation: the saddle points $(u^*, w^*) \in \text{dom } F_1 \times \text{dom } F_2^*$ are exactly the elements of $X \times Y$ that satisfy the coupled fixed point equations

$$\begin{cases} u^* = (\text{id} + \sigma \partial F_1)^{-1}(u^* - \sigma A^* \bar{w}^*), \\ w^* = (\text{id} + \tau \partial F_2^*)^{-1}(w^* + \tau A \bar{u}^*), \\ \bar{u}^* = u^*, \quad \bar{w}^* = w^*. \end{cases} \quad (6.88)$$

We obtain a numerical method by fixed point iteration: to realize the right hand side, we need only the resolvents of ∂F_1 and ∂F_2^* and the application of A and its adjoint A^* . Since the resolvents are nonexpansive (cf. Lemma 6.134) and A and A^* are continuous, the iteration is even Lipschitz continuous. We recall that no assumptions on differentiability of F_2 have been made.

From Eq. (6.88) we can derive a number of numerical methods, and we give a few (see also [7]).

- **Explicit Arrow-Hurwicz method**

Here we set $\bar{u}^n = u^n$, $\bar{w}^n = w^n$ and iterate the joint resolvent in (6.88), i.e.,

$$\begin{cases} u^{n+1} = (\text{id} + \sigma \partial F_1)^{-1}(u^n - \sigma A^* w^n), \\ w^{n+1} = (\text{id} + \tau \partial F_2^*)^{-1}(w^n + \tau A u^n). \end{cases}$$

- **Semi-implicit Arrow-Hurwicz method**

This method differs from the explicit method in that we use the “new” primal vector u^{n+1} to calculate w^{n+1} :

$$\begin{cases} u^{n+1} = (\text{id} + \sigma \partial F_1)^{-1}(u^n - \sigma A^* w^n), \\ w^{n+1} = (\text{id} + \tau \partial F_2^*)^{-1}(w^n + \tau A u^{n+1}). \end{cases}$$

In practice one can save some memory, since we can overwrite u^n directly with u^{n+1} . Moreover, we note that one can interchange the roles of u and w , of course.

- **Modified Arrow-Hurwicz method/Extra gradient method**

The idea of the *modified Arrow-Hurwicz method* is not to use $\bar{u}^n = u^n$ and $\bar{w}^n = w^n$ but to carry these on and update them with explicit Arrow-Hurwicz steps [113]:

$$\begin{cases} u^{n+1} = (\text{id} + \sigma \partial F_1)^{-1}(u^n - \sigma A^* \bar{w}^n), \\ w^{n+1} = (\text{id} + \tau \partial F_2^*)^{-1}(w^n + \tau A \bar{u}^n), \\ \bar{u}^{n+1} = (\text{id} + \sigma \partial F_1)^{-1}(u^{n+1} - \sigma A^* \bar{w}^n), \\ \bar{w}^{n+1} = (\text{id} + \tau \partial F_2^*)^{-1}(w^{n+1} + \tau A \bar{u}^n). \end{cases}$$

The earlier *extra gradient method*, proposed in [89], uses a similar idea, but differs in that in the calculations of \bar{u}^{n+1} and \bar{w}^{n+1} we evaluate the operators A^* and A at w^{n+1} and u^{n+1} , respectively.

- **Arrow-Hurwicz method with linear primal extra gradient/Chambolle-Pock method**

A method that seems to work well for imaging problems, proposed in [34], is based on the semi-implicit Arrow-Hurwicz method with a primal “extra gradient sequence” (\bar{u}^n). Instead of an Arrow-Hurwicz step, one uses a well-chosen linear combination (based on $\bar{u}^* = 2u^* - u^*$).

$$\begin{cases} w^{n+1} = (\text{id} + \tau \partial F_2^*)^{-1}(w^n + \tau A \bar{u}^n), \\ u^{n+1} = (\text{id} + \sigma \partial F_1)^{-1}(u^n - \sigma A^* w^{n+1}), \\ \bar{u}^{n+1} = 2u^{n+1} - u^n. \end{cases} \quad (6.89)$$

Note that only the primal variable uses an “extra gradient,” and that the order in which we update the primal and dual variable is important, i.e., we have to update the dual variable first. Of course we can swap the roles of the primal and dual variables, and in this case we could speak of a dual extra gradient. The memory requirements are comparable to those in the explicit Arrow-Hurwicz method.

For all the above methods there are conditions on the steps sized σ and τ that guarantee convergence in some sense. In comparison to the classical Arrow-Hurwicz method, the modified Arrow-Hurwicz method and the extra gradient method need weaker conditions to ensure convergence, and hence they are of practical interest, despite the slightly higher memory requirements. For details we refer to the original papers and treat only the case of convergence of the Arrow-Hurwicz method with linear primal extra gradient. We follow the presentation of [34] and first derive an estimate for one Arrow-Hurwicz step with fixed (\bar{u}, \bar{w}) .

Lemma 6.139 *Let X, Y be real Hilbert spaces, $A \in \mathcal{L}(X, Y)$, $F_1 \in \Gamma_0(X)$, $F_2^* \in \Gamma_0(Y)$, and $\sigma, \tau > 0$. Moreover, let $Z = X \times Y$, and for elements $z = (u, w)$, $\bar{z} = (\bar{u}, \bar{w})$ in $X \times Y$, let*

$$(z, \bar{z})_Z = \frac{(u, \bar{u})_X}{\sigma} + \frac{(w, \bar{w})_Y}{\tau}, \quad \|z\|_Z^2 = \frac{\|u\|_X^2}{\sigma} + \frac{\|w\|_Y^2}{\tau}.$$

If $\bar{z}, z^n, z^{n+1} \in Z$ with $\bar{z} = (\bar{u}, \bar{w})$, $z^n = (u^n, w^n)$, and $z^{n+1} = (u^{n+1}, w^{n+1})$, and the equations

$$\begin{cases} u^{n+1} = (\text{id} + \sigma \partial F_1)^{-1}(u^n - \sigma A^* \bar{w}) \\ w^{n+1} = (\text{id} + \tau \partial F_2^*)^{-1}(w^n + \tau A \bar{u}) \end{cases} \quad (6.90)$$

are satisfied, then for every $z = (u, w)$, $u \in \text{dom } F_1$, $w \in \text{dom } F_2^*$, one has the estimate

$$\begin{aligned} \frac{\|z^n - z^{n+1}\|_Z^2}{2} + (w^{n+1} - w, A(u^{n+1} - \bar{u}))_Y - (w^{n+1} - \bar{w}, A(u^{n+1} - u))_Y \\ + L(u^{n+1}, w) - L(u, w^{n+1}) \leq \frac{\|z - z^n\|_Z^2}{2} - \frac{\|z - z^{n+1}\|_Z^2}{2}, \end{aligned} \quad (6.91)$$

where L denotes the Lagrange functional (6.87).

Proof First we note that the equation for u^{n+1} means nothing other than $u^n - \sigma A^* \bar{w} - u^{n+1} \in \sigma \partial F_1(u^{n+1})$. The respective subgradient inequality in u leads to

$$\sigma F_1(u^{n+1}) + (u^n - \sigma A^* \bar{w} - u^{n+1}, u - u^{n+1})_X \leq \sigma F_1(u),$$

which we rearrange to

$$\frac{(u^n - u^{n+1}, u - u^{n+1})_X}{\sigma} \leq F_1(u) - F_1(u^{n+1}) + (\bar{w}, A(u - u^{n+1}))_Y.$$

Proceeding similarly for w^{n+1} , we obtain the inequality

$$\frac{(w^n - w^{n+1}, w - w^{n+1})_Y}{\tau} \leq F_2^*(w) - F_2^*(w^{n+1}) - (w - w^{n+1}, A\bar{u})_Y.$$

Adding the right-hand sides, using the definition of L , and inserting $\pm(w^{n+1}, Au^{n+1})_Y$, we obtain

$$\begin{aligned} & F_1(u) - F_2^*(w^{n+1}) + F_2^*(w) - F_1(u^{n+1}) + (\bar{w}, A(u - u^{n+1}))_Y - (w - w^{n+1}, A\bar{u})_Y \\ &= L(u, w^{n+1}) - L(u^{n+1}, w) + (\bar{w}, A(u - u^{n+1}))_Y - (w^{n+1}, Au)_Y \\ &\quad - (w - w^{n+1}, A\bar{u})_Y + (w, Au^{n+1})_Y \\ &= L(u, w^{n+1}) - L(u^{n+1}, w) + (\bar{w}, A(u - u^{n+1}))_Y - (w^{n+1}, Au)_Y + (w^{n+1}, Au^{n+1})_Y \\ &\quad - (w - w^{n+1}, A\bar{u})_Y + (w, Au^{n+1})_Y - (w^{n+1}, Au^{n+1})_Y \\ &= L(u, w^{n+1}) - L(u^{n+1}, w) + (\bar{w} - w^{n+1}, A(u - u^{n+1}))_Y \\ &\quad - (w - w^{n+1}, A(\bar{u} - u^{n+1}))_Y. \end{aligned} \tag{6.92}$$

Adding the respective left-hand sides, we obtain the scalar product in Z , which we reformulate as

$$\begin{aligned} (z^n - z^{n+1}, z - z^{n+1})_Z &= -\frac{\|z^n - z^{n+1}\|_Z^2}{2} + (z^n - z^{n+1}, z - z^{n+1})_Z - \frac{\|z - z^{n+1}\|_Z^2}{2} \\ &\quad + \frac{\|z^n - z^{n+1}\|_Z^2}{2} + \frac{\|z - z^{n+1}\|_Z^2}{2} \\ &= \frac{\|z^n - z^{n+1}\|_Z^2}{2} + \frac{\|z - z^{n+1}\|_Z^2}{2} - \frac{\|z - z^n\|_Z^2}{2}. \end{aligned}$$

Moving $\frac{1}{2}\|z - z^{n+1}\|_Z^2 - \frac{1}{2}\|z - z^n\|_Z^2$ to the right-hand side and all terms in (6.92) to the left-hand side, we obtain the desired inequality. \square

Next we remark that the iteration (6.89) can be represented by (6.90): the choice $\bar{u} = 2u^n - u^{n-1}$ and $\bar{w} = w^{n+1}$ with $u^{-1} = u^0$ is obviously the correct one, given that we initialize with $\bar{u}^0 = u^0$, which we assume in the following. Hence, we can use the estimate (6.91) to analyze the convergence. To that end, we analyze the scalar products in the estimate and aim to estimate them from below. Ultimately we would like to estimate in a way such that we can combine them with the norms in the expression.

Lemma 6.140 *In the situation of Lemma 6.139 let (z^n) be a sequence in $Z = X \times Y$ with components $z^n = (u^n, w^n)$ and $u^{-1} = u^0$. If $\sigma\tau\|A\|^2 \leq 1$, then for all $w \in Y$ and $M, N \in \mathbb{N}$ with $M \leq N$, one has*

$$\begin{aligned} - \sum_{n=M}^{N-1} (w^{n+1} - w, A(u^{n+1} - 2u^n + u^{n-1}))_Y &\leq \delta_M(w) + \sigma\tau\|A\|^2 \frac{\|w^N - w\|_Y^2}{2\tau} \\ &+ \frac{\|u^M - u^{M-1}\|_X^2}{2\sigma} + \sqrt{\sigma\tau}\|A\| \left(\sum_{n=M}^{N-2} \frac{\|z^{n+1} - z^n\|_Z^2}{2} \right) + \frac{\|z^N - z^{N-1}\|_Z^2}{2}, \end{aligned} \quad (6.93)$$

where $\delta_n(w) = (w^n - w, A(u^n - u^{n-1}))_Y$.

Proof First we fix $n \in \mathbb{N}$. Then by definition of $\delta_n(w)$ and the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned} -(w^{n+1} - w, A(u^{n+1} - 2u^n + u^{n-1}))_Y \\ &= (w^{n+1} - w, A(u^n - u^{n-1}))_Y - (w^{n+1} - w, A(u^{n+1} - u^n))_Y \\ &= \delta_n(w) - \delta_{n+1}(w) + (w^{n+1} - w^n, A(u^n - u^{n-1}))_Y \\ &\leq \delta_n(w) - \delta_{n+1}(w) + \|A\| \|w^{n+1} - w^n\|_Y \|u^n - u^{n-1}\|_X. \end{aligned}$$

Young's inequality for numbers $ab \leq \lambda a^2/2 + b^2/(2\lambda)$ applied with $\lambda = \sqrt{\sigma/\tau}$ leads to

$$\|A\| \|w^{n+1} - w^n\|_Y \|u^n - u^{n-1}\|_X \leq \sqrt{\sigma\tau}\|A\| \left(\frac{\|u^n - u^{n-1}\|_X^2}{2\sigma} + \frac{\|w^{n+1} - w^n\|_Y^2}{2\tau} \right).$$

Summing over n , we get, using the condition that $\sigma\tau\|A\|^2 \leq 1$,

$$\begin{aligned} - \sum_{n=M}^{N-1} (w^{n+1} - w, A(u^{n+1} - 2u^n + u^{n-1}))_Y \\ &\leq \delta_M(w) - \delta_N(w) + \sqrt{\sigma\tau}\|A\| \sum_{n=M}^{N-1} \left(\frac{\|u^n - u^{n-1}\|_X^2}{2\sigma} + \frac{\|w^{n+1} - w^n\|_Y^2}{2\tau} \right) \\ &\leq \delta_M(w) - \delta_N(w) + \frac{\|u^M - u^{M-1}\|_X^2}{2\sigma} \\ &+ \sqrt{\sigma\tau}\|A\| \left(\sum_{n=M}^{N-2} \frac{\|z^{n+1} - z^n\|_Z^2}{2} \right) + \frac{\|w^N - w^{N-1}\|_Y^2}{2\tau}. \end{aligned}$$

We estimate the value $-\delta_N(w)$ also with the Cauchy-Schwarz inequality, the operator norm, and Young's inequality, this time with $\lambda = \sigma$ to get

$$-\delta_N(w) \leq (\|A\| \|w^N - w\|_Y) \|u^N - u^{N-1}\|_X \leq \sigma \tau \|A\|^2 \frac{\|w^N - w\|_Y^2}{2\tau} + \frac{\|u^N - u^{N-1}\|_X^2}{2\sigma}.$$

This proves the claim. \square

Together, the inequalities (6.91) and (6.93) are the essential ingredients to prove convergence in finite-dimensional spaces.

Theorem 6.141 (Convergence of the Arrow-Hurwicz Method with Primal Extra Gradient) *Let X, Y be finite-dimensional real Hilbert spaces, $A \in \mathcal{L}(X, Y)$, $F_1 \in \Gamma_0(X)$, $F_2^* \in \Gamma_0(Y)$ and let $\sigma, \tau > 0$ be step sizes that satisfy $\sigma \tau \|A\|^2 < 1$. Moreover, assume that the Lagrange functional L from (6.87) has a saddle point.*

Then the iteration

$$\begin{cases} w^{n+1} = (\text{id} + \tau \partial F_2^*)^{-1}(w^n + \tau A \bar{u}^n), \\ u^{n+1} = (\text{id} + \sigma \partial F_1)^{-1}(u^n - \sigma A^* w^{n+1}), \\ \bar{u}^{n+1} = 2u^{n+1} - u^n, \end{cases}$$

converges for arbitrary initial values $(u^0, w^0) \in X \times Y$, $\bar{u}^0 = u^0$ to a saddle point $(u^, w^*) \in X \times Y$ of L .*

Proof Let $z \in Z = X \times Y$ with components $z = (u, w)$. For fixed n let $\bar{u} = 2u^n - u^{n-1}$ (with $u^{-1} = u^0$), $\bar{w} = w^{n+1}$, and apply Lemma 6.139. The estimate (6.91) becomes

$$\begin{aligned} & \frac{\|z^{n+1} - z^n\|_Z^2}{2} + (w^{n+1} - w, A(u^{n+1} - 2u^n + u^{n-1}))_Y \\ & + L(u^{n+1}, w) - L(u, w^{n+1}) \leq \frac{\|z - z^n\|_Z^2}{2} - \frac{\|z - z^{n+1}\|_Z^2}{2}. \end{aligned} \quad (6.94)$$

For some $N \in \mathbf{N}$ we sum from 0 to $N - 1$ and use (6.93) from Lemma 6.140 to estimate from below (noting that $\delta_0(w) = 0$ and $\|u^0 - u^{-1}\|_X = 0$) to get

$$\begin{aligned} & \left(\sum_{n=0}^{N-1} \frac{\|z^{n+1} - z^n\|_Z^2}{2} \right) - \sigma \tau \|A\|^2 \frac{\|w^N - w\|_Y^2}{2\tau} - \sqrt{\sigma \tau} \|A\| \left(\sum_{n=0}^{N-2} \frac{\|z^{n+1} - z^n\|_Z^2}{2} \right) \\ & - \frac{\|z^N - z^{N-1}\|_Z^2}{2} + \sum_{n=0}^{N-1} (L(u^{n+1}, w) - L(u, w^{n+1})) \leq \frac{\|z - z^0\|_Z^2}{2} - \frac{\|z - z^N\|_Z^2}{2}. \end{aligned}$$

Rearranging, using the definition of the norm in Z and $-\sigma\tau\|A\|^2 \leq 0$, leads to

$$(1 - \sqrt{\sigma\tau}\|A\|) \left(\sum_{n=0}^{N-2} \frac{\|z^{n+1} - z^n\|_Z^2}{2} \right) + (1 - \sigma\tau\|A\|^2) \frac{\|z - z^N\|_Z^2}{2} \\ + \sum_{n=0}^{N-1} (L(u^{n+1}, w) - L(u, w^{n+1})) \leq \frac{\|z - z^0\|_Z^2}{2}.$$

If we plug in an arbitrary saddle point $\tilde{z} = (\tilde{u}, \tilde{w})$ of L (which exists by assumption), we obtain that $L(u^{n+1}, \tilde{w}) - L(\tilde{u}, w^{n+1}) \geq 0$ for all n . This means that all three terms on the left-hand side are nonnegative, and the right-hand side is independent of N , which implies convergence and boundedness, respectively, i.e.,

$$\sum_{n=0}^{\infty} \frac{\|z^n - z^{n+1}\|_Z^2}{2} \leq \frac{\|\tilde{z} - z^0\|_Z^2}{2(1 - \sqrt{\sigma\tau}\|A\|)}, \quad \frac{\|\tilde{z} - z^N\|_Z^2}{2} \leq \frac{\|\tilde{z} - z^0\|_Z^2}{2(1 - \sigma\tau\|A\|^2)}.$$

The first estimate implies $\lim_{n \rightarrow \infty} \|z^{n+1} - z^n\|_Z^2 = 0$. Consequently, the sequence (z^n) is bounded, and by the finite dimensionality of Z there exists a convergent subsequence (z^{n_k}) with $\lim_{k \rightarrow \infty} z^{n_k} = z^*$ for some $z^* = (u^*, w^*) \in Z$. Moreover, we have convergence of neighboring subsequences $\lim_{k \rightarrow \infty} z^{n_k+1} = \lim_{k \rightarrow \infty} z^{n_k-1} = z^*$, and also, by continuity of A , A^* , $(\text{id} + \sigma\partial F_1)^{-1}$ and $(\text{id} + \tau\partial F_2^*)^{-1}$ (see Lemma 6.134), we conclude that

$$\begin{aligned} \bar{u}^* &= \lim_{k \rightarrow \infty} \bar{u}^{n_k} = \lim_{k \rightarrow \infty} 2u^{n_k} - u^{n_k-1} = u^*, \\ w^* &= \lim_{k \rightarrow \infty} (\text{id} + \tau\partial F_2^*)^{-1}(w^{n_k} + \tau A\bar{u}^{n_k}) = (\text{id} + \tau\partial F_2^*)^{-1}(w^* + \tau Au^*), \\ u^* &= \lim_{k \rightarrow \infty} (\text{id} + \sigma\partial F_1)^{-1}(u^{n_k} - \sigma A^*w^{n_k+1}) = (\text{id} + \sigma\partial F_1)^{-1}(u^* - \sigma A^*w^*). \end{aligned}$$

Thus, the pair (u^*, w^*) satisfies Eq. (6.88), and this shows that it is indeed a saddle point of L .

It remains to show that the whole sequence (z^n) converges to z^* . To that end fix $k \in \mathbb{N}$ and $N \geq n_k + 1$. Summing (6.94) from $M = n_k$ to $N - 1$ and repeating the above steps, we get

$$(1 - \sqrt{\sigma\tau}\|A\|) \left(\sum_{n=n_k}^{N-2} \frac{\|z^{n+1} - z^n\|_Z^2}{2} \right) + (1 - \sigma\tau\|A\|^2) \frac{\|z - z^N\|_Z^2}{2} \\ + \sum_{n=n_k}^{N-1} (L(u^{n+1}, w) - L(u, w^{n+1})) \leq \delta_{n_k}(w) + \frac{\|u^{n_k} - u^{n_k-1}\|_X^2}{2\sigma} + \frac{\|z - z^{n_k}\|_Z^2}{2}.$$

Plugging in z^* , using that $\sigma\tau\|A\|^2 < 1$ and that (u^*, w^*) is a saddle point, we arrive at

$$\|z^* - z^N\|_Z^2 \leq \frac{2\sigma\delta_{n_k}(w^*) + \|u^{n_k} - u^{n_k-1}\|_X^2 + \sigma\|z^* - z^{n_k}\|_Z^2}{\sigma(1 - \sigma\tau\|A\|^2)}.$$

Obviously, $\lim_{k \rightarrow \infty} \delta_{n_k}(w^*) = (w^* - w^*, A(u^* - u^*))_Y = 0$, and hence the right-hand side converges to 0 for $k \rightarrow \infty$. In particular, for every $\varepsilon > 0$ there exists k such that the right-hand side is smaller than ε^2 . This means, that for all $N \geq n_k$,

$$\|z^* - z^N\|_Z^2 \leq \varepsilon^2,$$

and we can conclude that $\lim_{N \rightarrow \infty} z^N = z^*$, which was to be shown. \square

Unfortunately, the proof of convergence does not reveal anything about the speed of convergence of (u^n, w^n) , and hence, it is difficult to decide when to stop the iteration. There are several approaches to solving this problem, one of which is based on the *duality gap*. If (\tilde{u}, \tilde{w}) is a saddle point of L , then

$$\inf_{u \in X} L(u, w^n) \leq L(\tilde{u}, w^n) \leq L(\tilde{u}, \tilde{w}) \leq L(u^n, \tilde{w}) \leq \sup_{w \in Y} L(u^n, w).$$

The infimum on the left-hand side is exactly the dual objective value $-F_1^*(-A^*w^n) - F_2^*(w^n)$, while the supremum on the right-hand side is the primal objective value $F_1(u^n) + F_2(Au^n)$. The difference is always nonnegative and vanishes exactly at the saddle points of L . Hence, one defines the duality gap $\mathcal{G} : X \times Y \rightarrow \mathbf{R}_\infty$ as follows:

$$\mathcal{G}(u, w) = F_1(u) + F_2(Au) + F_1^*(-A^*w) + F_2^*(w). \quad (6.95)$$

This also shows that \mathcal{G} is proper, convex, and lower semicontinuous. In particular,

$$\begin{aligned} \mathcal{G}(u^n, w^n) &\geq (F_1(u^n) + F_2(Au^n)) - (\min_{u \in X} F_1(u) + F_2(Au)), \\ \mathcal{G}(u^n, w^n) &\geq (\max_{w \in Y} -F_1^*(-A^*w) - F_2^*(w)) - (-F_1^*(-A^*w^n) - F_2^*(w^n)). \end{aligned} \quad (6.96)$$

Hence, a small duality gap implies that the differences between the functional values of u^n and w^n and the respective optima of the primal and dual problems, respectively, are also small. Hence, the condition $\mathcal{G}(u^n, w^n) < \varepsilon$ for some given tolerance $\varepsilon > 0$ is a suitable criterion to terminate the iteration.

If (u^n, w^n) converges to a saddle point (u^*, w^*) , then the lower semicontinuity \mathcal{G} gives only $0 \leq \liminf_{n \rightarrow \infty} \mathcal{G}(u^n, w^n)$, i.e., the duality gap does not necessarily converge to 0. However, this may be the case, and a necessary condition is the continuity of F_1 , F_2 and their Fenchel conjugates. In these cases, \mathcal{G} gives a stopping

criterion that guarantees for the primal-dual method (6.89) (given its convergence) the optimality of the primal and dual objective values up to a given tolerance.

6.4.3 Application of the Primal-Dual Methods

Now we want to apply a discrete version of the method we derived in the previous subsection to solve the variational problems numerically. To that end, we have to discretize the respective minimization problems and check whether Fenchel-Rockafellar duality holds, in order to guarantee the existence of a saddle point of the Lagrange functional. If we succeed with this, we have to identify how to implement the steps of the primal-dual method (6.89) and then, by Theorem 6.141, we have a convergent numerical method to solve our problems. Let us start with the discretization of the functionals

As in Sect. 5.4 we assume that we have rectangular discrete images, i.e., $N \times M$ matrices ($N, M \geq 1$). The discrete indices (i, j) always satisfy $1 \leq i \leq N$ and $1 \leq j \leq M$. For a fixed “pixel size” $h > 0$, the (i, j) th entry corresponds to the function value at (ih, jh) . The associated space is denoted by $\mathbf{R}^{N \times M}$ and equipped with the scalar product

$$(u, v) = h^2 \sum_{i=1}^N \sum_{j=1}^M u_{i,j} v_{i,j}.$$

To form the discrete gradient we also need images with multidimensional values, and hence we denote by $\mathbf{R}^{N \times M \times K}$ the space of images with K -dimensional values. For $u, v \in \mathbf{R}^{N \times M \times K}$ we define a pointwise and a global scalar product by

$$u_{i,j} \cdot v_{i,j} = \sum_{k=1}^K u_{i,j,k} v_{i,j,k}, \quad (u, v) = h^2 \sum_{i=1}^N \sum_{j=1}^M u_{i,j} \cdot v_{i,j},$$

respectively. These definitions give a pointwise absolute value $|u_{i,j}| = \sqrt{u_{i,j} \cdot u_{i,j}}$ and a norm $\|u\| = \sqrt{(u, u)}$.

With this notation, we can write the L^p norms simply as “pointwise summation”: For $u \in \mathbf{R}^{N \times M \times K}$ and $p \in [1, \infty[$ we have

$$\|u\|_p = \left(h^2 \sum_{i=1}^N \sum_{j=1}^M |u_{i,j}|^p \right)^{\frac{1}{p}}, \quad \|u\|_\infty = \max_{i=1, \dots, N} \max_{j=1, \dots, M} |u_{i,j}|.$$

To discretize the Sobolev semi-norm, we need a discretization of ∇ . We assume, as we have done before, a finite difference approximation with constant boundary treatment:

$$(\partial_1 u)_{i,j} = \begin{cases} \frac{u_{i+1,j} - u_{i,j}}{h} & \text{if } i < N, \\ 0 & \text{otherwise} \end{cases} \quad (\partial_2 u)_{i,j} = \begin{cases} \frac{u_{i,j+1} - u_{i,j}}{h} & \text{if } j < M, \\ 0 & \text{otherwise.} \end{cases}$$

The discrete gradient $\nabla_h u$ is an element in $\mathbf{R}^{N \times M \times 2}$ given by

$$(\nabla_h u)_{i,j,k} = (\partial_k u)_{i,j}. \quad (6.97)$$

As a linear operator, it maps $\nabla_h : \mathbf{R}^{N \times M} \rightarrow \mathbf{R}^{N \times M \times 2}$. Since we aim to use it as $A = \nabla_h$ in method (6.89), we need its adjoint as well as its norm. The former is related to the discrete divergence operator $\operatorname{div}_h : \mathbf{R}^{N \times M \times 2} \rightarrow \mathbf{R}^{N \times M}$ with zero boundary values:

$$(\operatorname{div}_h v)_{i,j} = \begin{cases} \frac{v_{1,j,1}}{h} & \text{if } i = 1, \\ \frac{v_{i,j,1} - v_{i-1,j,1}}{h} & \text{if } 1 < i < N, \\ \frac{-v_{N-1,j,1}}{h} & \text{if } i = N, \end{cases} + \begin{cases} \frac{v_{i,1,2}}{h} & \text{if } j = 1, \\ \frac{v_{i,j,2} - v_{i,j-1,2}}{h} & \text{if } 1 < j < M, \\ \frac{-v_{i,M-1,2}}{h} & \text{if } j = M. \end{cases} \quad (6.98)$$

Lemma 6.142 (Nullspace, Adjoint, and Norm Estimate for ∇_h) *The linear map $\nabla_h : \mathbf{R}^{N \times M} \rightarrow \mathbf{R}^{N \times M \times 2}$ has the following properties:*

1. *The nullspace is $\ker(\nabla_h) = \operatorname{span}(\mathbf{1})$ with the constant vector $\mathbf{1} \in \mathbf{R}^{N \times M}$,*
2. *the adjoint is $\nabla_h^* = -\operatorname{div}_h$, and*
3. *the norm satisfies $\|\nabla_h\|^2 < \frac{8}{h^2}$.*

Proof Assertion 1: Let $u \in \mathbf{R}^{N \times M}$ with $\nabla_h u = 0$. By induction we get for every $1 \leq i \leq N-1$ and $1 \leq j \leq M$ by definition of ∇_h that $u_{i,j} = u_{i+1,j} = \dots = u_{N,j}$. Hence, u is constant along the first component. For $1 \leq j \leq M-1$ we argue similarly to see that $u_{N,j} = u_{N,j+1} = \dots = u_{N,M}$, and consequently u is a scalar multiple of the constant image $\mathbf{1}$.

Assertion 2: For $u \in \mathbf{R}^{N \times M}$ and $v \in \mathbf{R}^{N \times M \times 2}$ let $\partial_1^- v^1$ and $\partial_2^- v^2$ be the first and second summand in (6.98), respectively. We form the scalar product $(\nabla_h u, v)$ and evaluate

$$\begin{aligned} (\nabla_h u, v) &= h \left(\sum_{i=1}^{N-1} \sum_{j=1}^M (u_{i+1,j} - u_{i,j}) v_{i,j,1} \right) + h \left(\sum_{i=1}^N \sum_{j=1}^{M-1} (u_{i,j+1} - u_{i,j}) v_{i,j,2} \right) \\ &= h \sum_{j=1}^M \left(\left(\sum_{i=2}^N u_{i,j} v_{i-1,j,1} \right) - \left(\sum_{i=1}^{N-1} u_{i,j} v_{i,j,1} \right) \right) \end{aligned}$$

$$\begin{aligned}
& + h \sum_{i=1}^N \left(\left(\sum_{j=2}^M u_{i,j} v_{i,j-1,2} \right) - \left(\sum_{j=1}^{M-1} u_{i,j} v_{i,j,2} \right) \right) \\
& = h \sum_{j=1}^M \left(-v_{1,j,1} u_{1,j} + \left(\sum_{i=2}^{N-1} (v_{i-1,j,1} - v_{i,j,1}) u_{i,j} \right) + v_{N-1,j,1} u_{N,j} \right) \\
& + h \sum_{i=1}^N \left(-v_{i,1,2} u_{i,1} + \left(\sum_{j=2}^{M-1} (v_{i,j-1,2} - v_{i,j,2}) u_{i,j} \right) + v_{i,M-1,2} u_{i,M} \right) \\
& = h^2 \sum_{i=1}^N \sum_{j=1}^M \left(-(\partial_1^- v^1)_{i,j} - (\partial_2^- v^2)_{i,j} \right) u_{i,j} = (u, -\operatorname{div}_h v).
\end{aligned}$$

Assertion 3: To begin with, we show that for $u \in \mathbf{R}^{N \times M}$ with $\|u\| = 1$,

$$-h^2 \sum_{i=1}^{N-1} \sum_{j=1}^M u_{i+1,j} u_{i,j} < 1, \quad -h^2 \sum_{i=1}^N \sum_{j=1}^{M-1} u_{i,j+1} u_{i,j} < 1.$$

Assume that the first inequality is not satisfied. Let $v_{i,j} = -u_{i+1,j}$ if $i < N$ and $v_{N,j} = 0$. By the Cauchy-Schwarz inequality we obtain $(v, u) \leq \|v\| \|u\| \leq 1$, i.e. the scalar product has to be equal to 1. This means that the Cauchy-Schwarz inequality is tight and hence $u = v$. In particular, we have $u_{N,j} = 0$, and recursively we get for $i < N$ that

$$u_{i,j} = -u_{i+1,j} = (-1)^{N-i} u_{N,j} = 0,$$

i.e., $u = 0$, which is a contradiction. For the second inequality we argue similarly.

Now we estimate

$$\begin{aligned}
\|\nabla_h u\|^2 & = h^2 \left(\sum_{i=1}^{N-1} \sum_{j=1}^M \left(\frac{u_{i+1,j} - u_{i,j}}{h} \right)^2 \right) + h^2 \left(\sum_{i=1}^N \sum_{j=1}^{M-1} \left(\frac{u_{i,j+1} - u_{i,j}}{h} \right)^2 \right) \\
& = \left(\sum_{i=1}^{N-1} \sum_{j=1}^M u_{i+1,j}^2 + u_{i,j}^2 - 2u_{i+1,j} u_{i,j} \right) + \left(\sum_{i=1}^N \sum_{j=1}^{M-1} u_{i,j+1}^2 + u_{i,j}^2 - 2u_{i,j+1} u_{i,j} \right) \\
& \leq 4 \left(\sum_{i=1}^N \sum_{j=1}^M u_{i,j}^2 \right) - 2 \left(\sum_{i=1}^{N-1} \sum_{j=1}^M u_{i+1,j} u_{i,j} \right) - 2 \left(\sum_{i=1}^N \sum_{j=1}^{M-1} u_{i,j+1} u_{i,j} \right) \\
& < \frac{4}{h^2} + \frac{2}{h^2} + \frac{2}{h^2} = \frac{8}{h^2}.
\end{aligned}$$

Since the set $\{\|u\| = 1\}$ is compact, there exists some u^* with $\|u^*\| = 1$ where the value of the operator norm is attained, showing that $\|\nabla_h\|^2 = \|\nabla_h u^*\|^2 < \frac{8}{h^2}$. \square

The Sobolev semi-norm of u is discretized as $\|\nabla_h u\|_p$, and the total variation corresponds to the case $p = 1$, i.e., $\text{TV}_h(u) = \|\nabla_h u\|_1$. This allows one to discretize the applications with Sobolev penalty from Sect. 6.3.2 and their counterparts with total variation from Sect. 6.3.3. If Fenchel-Rockafellar duality (6.86) holds, method (6.89) yields a saddle point of the Lagrange functional and the primal component is a solution of the original problem. With a little intuition and some knowledge about the technical feasibility of resolvent maps it is possible to derive practical algorithms for a large number of convex minimization problems in imaging. To get an impression, how this is done in concrete cases, we discuss the applications from Sects. 6.3.2 and 6.3.3 in detail.

Example 6.143 (Primal-Dual Method for Variational Denoising) For $1 \leq p < \infty$, $1 < q < \infty$, $X = \mathbf{R}^{N \times M}$, and a discrete, noisy image $U^0 \in \mathbf{R}^{N \times M}$ and $\lambda > 0$ the discrete denoising problem reads

$$\min_{u \in X} \frac{\|u - U^0\|_q^q}{q} + \frac{\lambda \|\nabla_h u\|_p^p}{p}.$$

Obviously, the objective is continuous at X , and also coercive, since in finite dimensions all norms are equivalent. By the direct method (Theorem 6.17) we get the existence of minimizers also for this special case. We introduce $Y = \mathbf{R}^{N \times M \times 2}$ and define

$$u \in X : \quad F_1(u) = \frac{1}{q} \|u - U^0\|_q^q, \quad v \in Y : \quad F_2(v) = \frac{\lambda}{p} \|v\|_p^p, \quad A = \nabla_h,$$

and note that the assumptions for Fenchel-Rockafellar duality from Theorem 6.68 are satisfied, and hence the associated Lagrange functional has a saddle point. To apply method (6.89) we need the resolvents of ∂F_1 and ∂F_2^* . Note that Lemma 6.65 and Example 6.64 applied to F_2^* lead to

$$F_2 = \frac{\lambda}{p} \|\cdot\|_p^p \quad \Rightarrow \quad F_2^* = \begin{cases} \frac{\lambda}{p^*} \|\cdot\|_{p^*}^{p^*} \circ \lambda^{-1} \text{id} = \frac{\lambda^{-p^*/p}}{p^*} \|\cdot\|_{p^*}^{p^*} & \text{if } p > 1, \\ \lambda I_{\{\|v\|_\infty \leq 1\}} \circ \lambda^{-1} \text{id} = I_{\{\|v\|_\infty \leq \lambda\}} & \text{if } p = 1. \end{cases}$$

Now let $\sigma, \tau > 0$. For the resolvent $(\text{id} + \sigma \partial F_1)^{-1}$ we have by Lemma 6.136 and Example 6.138,

$$((\text{id} + \sigma \partial F_1)^{-1}(u))_{i,j} = U_{i,j}^0 + \text{sgn}(u_{i,j} - U_{i,j}^0) (\text{id} + \sigma |\cdot|^{q-1})^{-1}(|u_{i,j} - U_{i,j}^0|).$$

The same tools lead to

$$((\text{id} + \tau \partial F_2^*)^{-1}(w))_{i,j} = \begin{cases} \left(\text{id} + \tau \lambda^{-\frac{p^*}{p}} |\cdot|^{p^*-1} \right)^{-1}(|w_{i,j}|) \frac{w_{i,j}}{|w_{i,j}|} & \text{if } p > 1, \\ \min(\lambda, |w_{i,j}|) \frac{w_{i,j}}{|w_{i,j}|} = \frac{w_{i,j}}{\max(1, |w_{i,j}|/\lambda)} & \text{if } p = 1. \end{cases}$$

Table 6.1 Primal-dual method for the numerical solution of the discrete variational denoising problem

Primal-dual method for the solution of the variational denoising problem

$$\min_{u \in \mathbb{R}^{N \times M}} \frac{\|u - U^0\|_q^q}{q} + \frac{\lambda \|\nabla_h u\|_p^p}{p}.$$

1. *Initialize*

Let $n = 0$, $\bar{u}^0 = u^0 = U^0$, $w^0 = 0$. Choose $\sigma, \tau > 0$ with $\sigma\tau \leq \frac{8}{h^2}$.

2. *Dual step*

$$\bar{w}^{n+1} = w^n + \tau \nabla_h \bar{u}^n,$$

$$w_{i,j}^{n+1} = \begin{cases} \left(\text{id} + \tau \lambda^{-\frac{p^*}{p}} |\cdot|^{p^*-1} \right)^{-1}(|\bar{w}_{i,j}^{n+1}|) \frac{\bar{w}_{i,j}^{n+1}}{|\bar{w}_{i,j}^{n+1}|} & \text{if } p > 1, \\ \frac{\bar{w}_{i,j}^{n+1}}{\max(1, |\bar{w}_{i,j}^{n+1}|/\lambda)} & \text{if } p = 1, \end{cases} \quad \begin{matrix} 1 \leq i \leq N, \\ 1 \leq j \leq M. \end{matrix}$$

3. *Primal step and extra gradient*

$$v^{n+1} = u^n + \sigma \operatorname{div}_h w^{n+1} - U^0,$$

$$u_{i,j}^{n+1} = U_{i,j}^0 + \operatorname{sgn}(v_{i,j}^{n+1}) (\text{id} + \sigma |\cdot|^{q-1})^{-1}(|v_{i,j}^{n+1}|), \quad \begin{matrix} 1 \leq i \leq N, \\ 1 \leq j \leq M. \end{matrix}$$

$$\bar{u}^{n+1} = 2u^{n+1} - u^n.$$

4. *Iterate*

Update $n \leftarrow n + 1$ and continue with Step 2.

Finally, we note that $\nabla_h^* = -\operatorname{div}_h$ holds and that the step size restriction $\sigma\tau \|\nabla_h\|^2 < 1$ is satisfied for $\sigma\tau \leq \frac{8}{h^2}$; see Lemma 6.142. Now we have all ingredients to fully describe the numerical method for variational denoising; see Table 6.1. By Theorem 6.141 this method yields a convergent sequence $((u^n, w^n))$.

Let us briefly discuss a stopping criterion based on the duality gap (6.95). By Lemma 6.65 and Example 6.64, we have

$$F_1 = \frac{1}{q} \|\cdot\|_q^q \circ T_{-U^0} \quad \Rightarrow \quad F_1^* = \frac{1}{q^*} \|\cdot\|_{q^*}^{q^*} + (U^0, \cdot).$$

For $p > 1$ we get

$$G(u^n, w^n) = \frac{\|u^n - U^0\|_q^q}{q} + \frac{\|\operatorname{div}_h w^n\|_{q^*}^{q^*}}{q^*} + (U^0, \operatorname{div}_h w^n) + \frac{\lambda}{p} \|\nabla_h u^n\|_p^p + \frac{\lambda^{-\frac{p^*}{p}}}{p^*} \|w^n\|_{p^*}^{p^*}$$

which is a continuous functional, and we conclude in particular that $\mathcal{G}(u^n, w^n) \rightarrow 0$ for $n \rightarrow \infty$.

In the case $p = 1$ every w^n satisfies the condition $\|w^n\|_\infty \leq \lambda$, and thus

$$\mathcal{G}(u^n, w^n) = \frac{\|u^n - U^0\|_q^q}{q} + \frac{\|\operatorname{div}_h w^n\|_{q^*}^{q^*}}{q^*} + (U^0, \operatorname{div}_h w^n) + \lambda \operatorname{TV}_h(u^n),$$

and the convergence $\mathcal{G}(u^n, w^n) \rightarrow 0$ is satisfied as well. Thus, stopping the iteration as soon as $\mathcal{G}(u^n, w^n) < \varepsilon$, and this has to happen at some point, one has arrived at some u^n such that the optimal primal objective value is approximated up to a tolerance of ε .

Example 6.144 (Primal-Dual Method for Tikhonov Functionals/Deconvolution)

Now we consider the situation with a discretized linear operator A_h . Let again $1 \leq p < \infty$, $1 \leq q < \infty$ (this time we also allow $q = 1$) and $X = \mathbf{R}^{N_1 \times M_1}$, moreover, let $A_h \in \mathcal{L}(X, Y)$ with $Y = \mathbf{R}^{N_2 \times M_2}$ a forward operator that does not map constant images to zero, and let $U^0 \in Z$ be noisy measurements. The problem is the minimization of the Tikhonov functional

$$\min_{u \in X} \frac{\|A_h u - U^0\|_q^q}{q} + \frac{\lambda \|\nabla_h u\|_p^p}{p}.$$

Noting that the nullspace of ∇_h consists of the constant image only (see Lemma 6.142), we can argue similarly to Theorems 6.86 and 6.115 and Example 6.143 that minimizers exist. Application 6.97 and Example 6.127 motivate the choice of A_h as a discrete convolution operator as in Sect. 3.3.3, i.e., $A_h u = u * k_h$ with a discretized convolution kernel k_h with nonnegative entries that sum to one. The norm is, in this case, estimated by $\|A_h\| \leq 1$, and the adjoint A_h^* amounts to a convolution with $\bar{k}_h = D_{-\operatorname{id}k}$. In the following the concrete operator will not be of importance, and we will discuss only the general situation.

Since the minimization problem features both A_h and ∇_h , we dualize with $Z = \mathbf{R}^{N \times M \times 2}$, but in the following way:

$$u \in X : F_1(u) = 0, \quad (v, w) \in Y \times Z : F_2(v, w) = \frac{1}{q} \|v - U^0\|_q^q + \frac{\lambda}{p} \|w\|_p^p, \quad A = \begin{bmatrix} A_h \\ \nabla_h \end{bmatrix}.$$

Again, the assumption in Theorem 6.68 are satisfied and hence we only need to find saddle points of the Lagrange functional. We dualize F_2 , which, similar to Example 6.143, leads to

$$F_2^*(\bar{v}, \bar{w}) = \begin{cases} \frac{1}{q^*} \|\bar{v}\|_{q^*}^{q^*} + (U^0, \bar{v}) & \text{if } q > 1, \\ I_{\{\|\bar{v}\|_\infty \leq 1\}}(\bar{v}) + (U^0, \bar{v}) & \text{if } q = 1, \end{cases} + \begin{cases} \frac{\lambda^{-p^*/p}}{p^*} \|\bar{w}\|_{p^*}^{p^*} & \text{if } p > 1, \\ I_{\{\|\bar{w}\|_\infty \leq \lambda\}}(\bar{w}) & \text{if } p = 1. \end{cases}$$

By Lemma 6.136 item 5 we can apply the resolvent of ∂F_2^* for σ by applying the resolvents for \bar{v} and \bar{w} componentwise. We calculated both already in Example 6.143, and we only need to translate the variable \bar{v} by $-\sigma U^0$. The resolvent for ∂F_1 is trivial: $(\text{id} + \tau \partial F_1)^{-1} = \text{id}$.

Let us note how $\sigma \tau$ can be chosen: we get an estimate of $\|A\|^2$ by

$$\|Au\|^2 = \|A_h u\|^2 + \|\nabla_h u\|^2 \leq (\|A_h\|^2 + \|\nabla_h\|^2) \|u\|^2 < (\|A_h\|^2 + \frac{8}{h^2}) \|u\|^2,$$

and hence $\sigma \tau \leq (\|A_h\|^2 + \frac{8}{h^2})^{-1}$ is sufficient for the estimate $\sigma \tau \|A\|^2 < 1$.

The whole method is described in Table 6.2. The duality gap is, in the case of $p > 1$ and $q > 1$, as follows (note that $F_1^* = I_{\{0\}}$):

$$\begin{aligned} \mathcal{G}(u^n, v^n, w^n) &= I_{\{0\}}(\operatorname{div}_h w^n - A_h^* v^n) + \frac{1}{q} \|A_h u^n - U^0\|_q^q + \frac{\lambda}{p} \|\nabla_h u^n\|_p^p \\ &\quad + \frac{1}{q^*} \|v^n\|_{q^*}^{q^*} + (U^0, v^n) + \frac{\lambda^{-\frac{-p^*}{p}}}{p^*} \|w^n\|_{p^*}^{p^*}. \end{aligned}$$

Due to the presence of the indicator functional it will be ∞ in general, since we cannot guarantee that $\operatorname{div}_h w^n = A_h^* v^n$ ever holds. However, if we knew that the iterates (u^n) lay in some norm ball, i.e., we had an a-priori estimate $\|u^n\| \leq C$ for all n , we could use the following trick: we change F_1 to $F_1 = I_{\{\|u\| \leq C\}}$, and the resolvent $(\text{id} + \tau \partial F_1)^{-1}$ becomes the projection onto $\{\|u\| \leq C\}$ (see Example 6.137), and especially, the method is unchanged. The Fenchel dual becomes $F_1^* = C \|\cdot\|$ (see Example 6.64). Hence, the duality gap of the iterates is now given by the continuous function

$$\begin{aligned} \tilde{\mathcal{G}}(u^n, v^n, w^n) &= C \|\operatorname{div}_h w^n - A_h^* v^n\| + \frac{1}{q} \|A_h u^n - U^0\|_q^q + \frac{\lambda}{p} \|\nabla_h u^n\|_p^p \\ &\quad + \frac{1}{q^*} \|v^n\|_{q^*}^{q^*} + (U^0, v^n) + \frac{\lambda^{-\frac{-p^*}{p}}}{p^*} \|w^n\|_{p^*}^{p^*}, \end{aligned}$$

and can be used as a stopping criterion. It is possible, but tedious, to obtain an a priori estimate $\|u^n\| \leq C$ for this problem, e.g., using the coercivity of F_2 and F_2^* and the estimate of the iterates in the proof of Theorem 6.141. We omit this estimate and just note that in practice it is possible to use a crude estimate $C > 0$ without dramatic consequences: if C is too small, $\tilde{\mathcal{G}}(u^n, v^n, w^n)$ still goes to zero, but it may become negative and the estimates (6.96) may be invalid.

Table 6.2 Primal-dual method for the minimization of the Tikhonov functional with Sobolev- or total variation penalty

Primal-dual method for the minimization of Tikhonov functionals

$$\min_{u \in \mathbf{R}^{N_1 \times M_1}} \frac{\|A_h u - U^0\|_q^q}{q} + \frac{\lambda \|\nabla_h u\|_p^p}{p}.$$

1. *Initialization*

Let $n = 0$, $\bar{u}^0 = u^0 = 0$, $v^0 = 0$ and $w^0 = 0$.

Choose $\sigma, \tau > 0$ such that $\sigma \tau \leq (\|A_h\|^2 + \frac{8}{h^2})^{-1}$.

2. *Dual step*

$$\bar{v}^{n+1} = v^n + \tau(A_h \bar{u}^n - U^0),$$

$$v_{i,j}^{n+1} = \begin{cases} \operatorname{sgn}(\bar{v}_{i,j}^{n+1})(\operatorname{id} + \tau |\cdot|^{q^*-1})^{-1}(|\bar{v}_{i,j}^{n+1}|) & \text{if } q > 1, \\ \min(1, \max(-1, \bar{v}_{i,j}^{n+1})) & \text{if } q = 1, \end{cases} \quad 1 \leq i \leq N_2,$$

$$\bar{w}^{n+1} = w^n + \tau \nabla_h \bar{u}^n,$$

$$w_{i,j}^{n+1} = \begin{cases} (\operatorname{id} + \tau \lambda^{-\frac{p^*}{p}} |\cdot|^{p^*-1})^{-1}(|\bar{w}_{i,j}^{n+1}|) \frac{\bar{w}_{i,j}^{n+1}}{|\bar{w}_{i,j}^{n+1}|} & \text{if } p > 1, \\ \frac{\bar{w}_{i,j}^{n+1}}{\max(1, |\bar{w}_{i,j}^{n+1}|/\lambda)} & \text{if } p = 1, \end{cases} \quad 1 \leq i \leq N_1, \quad 1 \leq j \leq M_1.$$

3. *Primal step and extra gradient*

$$u^{n+1} = u^n + \sigma (\operatorname{div}_h w^{n+1} - A_h^* v^{n+1}),$$

$$\bar{u}^{n+1} = 2u^{n+1} - u^n.$$

4. *Iteration*

Set $n \leftarrow n + 1$ and continue with step 2.

Example 6.145 (Primal-Dual Method for Variational Inpainting) Let $\Omega_h = \{1, \dots, N\} \times \{1, \dots, M\}$, $\Omega'_h \subsetneq \Omega_h$ be an inpainting region, $U^0 : \Omega_h \setminus \Omega'_h \rightarrow \mathbf{R}$ an image to be inpainted, and consider for $1 \leq p < \infty$ the discrete inpainting problem in $X = \mathbf{R}^{N \times M}$:

$$\min_{u \in X} \frac{\|\nabla_h u\|_p^p}{p} + I_{\{v|_{\Omega_h \setminus \Omega'_h} = U^0\}}(u).$$

We want to apply the primal-dual method to this problem. First note that a minimizer exists. This can be shown, as in the previous examples, using arguments similar to those for the existence of solutions in Application 6.98 and Example 6.128. To

dualize we choose $Y = \mathbf{R}^{N \times M \times 2}$ and

$$u \in X : \quad F_1(u) = I_{\{v|_{\Omega_h \setminus \Omega'_h} = U^0\}}(u), \quad v \in Y : \quad F_2(v) = \frac{\lambda}{p} \|v\|_p^p, \quad A = \nabla_h.$$

Again similarly to the continuous case, we use the sum and the chain rule for subgradients: F_2 is continuous everywhere and F_1 and $F_2 \circ \nabla_h$ satisfy the assumption in Exercise 6.14. We obtain

$$F = F_1 + F_2 \circ \nabla_h \quad \Rightarrow \quad \partial F = \partial F_1 + \nabla_h^* \circ \partial F_2 \circ \nabla_h,$$

and by Remark 6.72 there exists a saddlepoint of the respective Lagrange functionals.

To apply the primal-dual method we need $(\text{id} + \sigma \partial F_1)^{-1}$, and this amounts to the projection onto $K = \{v \in X \mid v|_{\Omega_h \setminus \Omega'_h} = U^0\}$ (see Example 6.137). To evaluate this we do

$$(\text{id} + \sigma \partial F_1)^{-1}(u) = \chi_{\Omega_h \setminus \Omega'_h} U^0 + \chi_{\Omega'_h} u$$

after extending U^0 to Ω_h arbitrarily. The resulting method (6.89) is shown in Table 6.3.

Example 6.146 (Methods for Linear Equality Constraints/Interpolation) Our final example will deal with general linear equality constraints combined with minimizing a Sobolev semi-norm or the total variation. Let $1 \leq p < \infty$, $X = \mathbf{R}^{N_1 \times M_1}$, and $Y = \mathbf{R}^{N_2 \times M_2}$. For $U^0 \in Y$ we encode the linear equality constraint with a map $A_h : \mathbf{R}^{N_1 \times M_1} \rightarrow \mathbf{R}^{N_2 \times M_2}$ as $A_h u = U^0$. We assume that A_h is surjective and that $A_h \mathbf{1} \neq 0$ holds; see also Application 6.100 and Example 6.128. Our aim is to solve the minimization problem

$$\min_{u \in X} \frac{\|\nabla_h u\|_p^p}{p} + I_{\{A_h v = U^0\}}(u)$$

numerically. Having the interpolation problem in mind we consider the equality constraint in the equivalent form $(A_h u)_{i,j} = (v^{i,j}, u) = U^0_{i,j}$, where the $v^{i,j}$ are the images of the normalized standard basis vectors $v^{i,j} = \frac{1}{h} A_h^* e^{i,j}$ for $1 \leq i \leq N_2$, $1 \leq j \leq M_2$.

Similarly to Example 6.145 one sees that this problem has a solution. With $Z = \mathbf{R}^{N \times M \times 2}$ we write

$$u \in X : \quad F_1(u) = I_{\{A_h v = U^0\}}(u), \quad v \in Z : \quad F_2(v) = \frac{\lambda}{p} \|v\|_p^p, \quad A = \nabla_h,$$

such that Fenchel-Rockafellar duality holds for the sum $F = F_1 + F_2 \circ A$ (cf. Example 6.145). The projection $(\text{id} + \sigma \partial F_1)^{-1}(u)$ can be realized by the solution of

Table 6.3 Primal-dual method for the numerical solution of the variational inpainting problem**Primal-dual method for variational inpainting**

$$\min_{u \in \mathbf{R}^{N \times M}} \frac{\|\nabla_h u\|_p^p}{p} + I_{\{v|_{\Omega_h \setminus \Omega'_h} = U^0\}}(u).$$

1. Initialization

Let $n = 0$, $\bar{u}^0 = u^0 = U^0$, $w^0 = 0$. Choose $\sigma, \tau > 0$ with $\sigma\tau \leq \frac{8}{h^2}$.

2. Dual step

$$\bar{w}^{n+1} = w^n + \tau \nabla_h \bar{u}^n,$$

$$w_{i,j}^{n+1} = \begin{cases} (\text{id} + \tau |\cdot|^{p^*-1})^{-1}(|\bar{w}_{i,j}^{n+1}|) \frac{\bar{w}_{i,j}^{n+1}}{|\bar{w}_{i,j}^{n+1}|} & \text{if } p > 1, \\ \frac{\bar{w}_{i,j}^{n+1}}{\max(1, |\bar{w}_{i,j}^{n+1}|)} & \text{if } p = 1, \end{cases} \quad \begin{matrix} 1 \leq i \leq N, \\ 1 \leq j \leq M. \end{matrix}$$

3. Primal step and extra gradient

$$u^{n+1} = \chi_{\Omega_h \setminus \Omega'_h} U^0 + \chi_{\Omega'_h} (u^n + \sigma \operatorname{div}_h w^{n+1}),$$

$$\bar{u}^{n+1} = 2u^{n+1} - u^n.$$

4. Iteration

Set $n \leftarrow n + 1$ and continue with step 2.

the linear system

$$\lambda \in Y : \quad (A_h A_h^*) \lambda = U^0 - A_h u \quad \Rightarrow \quad (\text{id} + \sigma \partial F_1)^{-1}(u) = u + A_h^* \lambda;$$

see Example 6.137. Table 6.4 shows the resulting numerical method.

Now let us consider the concrete case of the interpolation problem from Application 6.100, i.e., the k -fold zooming of an image $U^0 \in \mathbf{R}^{N \times M}$. Here it is natural to restrict oneself to positive integers k . Thus, the map A_h models a k -fold downsizing from $N_1 \times M_1$ to $N_2 \times M_2$ with $N_1 = kN$, $M_1 = kM$, $N_2 = N$, and $M_2 = M$. If we choose, for example, the averaging over $k \times k$ squares, then the $v^{i,j}$ are given by

$$v^{i,j} = \frac{1}{hk} \chi_{[(i-1)k+1, ik] \times [(j-1)k+1, jk]}, \quad 1 \leq i \leq N, \quad 1 \leq j \leq M.$$

The factor $\frac{1}{hk}$ makes these vectors orthonormal, and hence the matrix $A_h A_h^*$ is the identity. Thus, there is no need to solve a linear system and we can set $\lambda = \mu$ in step 3 of the method in Table 6.4. For the perfect low-pass filter one can leverage the orthogonality of the discrete Fourier transform in a similar way.

Table 6.4 Primal-dual method for the numerical solution of minimization problems with Sobolev semi-norm or total variation and linear equality constraints

Primal-dual method for linear equality constraints

$$\min_{u \in \mathbf{R}^{N_1 \times M_1}} \frac{\|\nabla_h u\|_p^p}{p} + I_{\{A_h v = U^0\}}(u).$$

1. *Initialization*

Let $n = 0$, $\bar{u}^0 = u^0 = U^0$, $w^0 = 0$. Choose $\sigma, \tau > 0$ with $\sigma\tau \leq \frac{8}{h^2}$.

2. *Dual step*

$$\bar{w}^{n+1} = w^n + \tau \nabla_h \bar{u}^n,$$

$$w_{i,j}^{n+1} = \begin{cases} (\text{id} + \tau |\cdot|^{p^*-1})^{-1}(|\bar{w}_{i,j}^{n+1}|) \frac{\bar{w}_{i,j}^{n+1}}{|\bar{w}_{i,j}^{n+1}|} & \text{if } p > 1, \\ \frac{\bar{w}_{i,j}^{n+1}}{\max(1, |\bar{w}_{i,j}^{n+1}|)} & \text{if } p = 1, \end{cases} \quad \begin{matrix} 1 \leq i \leq N_1, \\ 1 \leq j \leq M_1. \end{matrix}$$

3. *Primal step and extra gradient*

$$\mu^{n+1} = U^0 - A_h(u^n + \sigma \operatorname{div}_h w^{n+1}),$$

$$\lambda^{n+1} = (A_h A_h^*)^{-1} \mu^{n+1},$$

$$u^{n+1} = u^n + A_h^* \lambda^{n+1},$$

$$\bar{u}^{n+1} = 2u^{n+1} - u^n.$$

4. *Iteration*

Set $n \leftarrow n + 1$ and continue with step 2.

6.5 Further Developments

Variational methods continue to develop rapidly, and it seems impossible to come close to a complete overview of further development. We only try to sketch some selected topics, which we describe briefly in the following. We also like to point the reader to the monographs [8, 37, 126].

One early proposal of a variational method is the so-called Mumford-Shah functional for *image segmentation* [101]. This functional is based on the model that an image consists of smooth (twice differentiable) parts that are separated by jump discontinuities. Mumford and Shah proposed to segment an image $u^0 : \Omega \rightarrow \mathbf{R}$ with

an edge set $\Gamma \subset \Omega$ and a piecewise smooth function u that minimize the following functional:

$$E(u, \Gamma) = \int_{\Omega} (u - u^0)^2 dx + \lambda \int_{\Omega \setminus \Gamma} |\nabla u|^2 dx + \mu \mathfrak{H}^{d-1}(\Gamma),$$

where $\lambda, \mu > 0$ are again some regularization parameters. The third term penalizes the length of the edge set and prohibits the trivial and unnatural minimizer $\Gamma = \Omega$ and $u^0 = u$. The Mumford-Shah functional is a mathematically challenging object, and both its analysis and numerical minimization have been the subject of numerous studies [5, 6, 9, 18, 33, 49, 76, 112]. On big challenge in the analysis of the minimization of E results from an appropriate description of the objects Γ and the functions $u \in H^1(\Omega \setminus \Gamma)$ (with changing Γ) in an appropriate functional-analytic context. It turned out that the space of special functions of bounded total variation $\text{SBV}(\Omega)$ is well suited. It consists of these functions in $\text{BV}(\Omega)$, where the derivative can be written as

$$\nabla u = (\nabla u)_{L^1} \mathfrak{L}^d + (u^+ - u^-) v \mathfrak{H}^{d-1} \llcorner \Gamma,$$

where $(u^+ - u^-)$ denotes the *jump*, v the *measure-theoretic normal*, and Γ the *jump set*. The vector space $\text{SBV}(\Omega)$ is a proper subspace of $\text{BV}(\Omega)$ (for BV functions the gradient also contains a so-called *Cantor part*), and the Mumford-Shah functional is well defined on this space but is not convex. Moreover, it is possible to prove the existence of minimizers of the Mumford-Shah functional in $\text{SBV}(\Omega)$, but due to the nonconvexity, the proof is more involved than in the cases we treated in this chapter.

A simplified variational model for segmentation with piecewise constant images has been proposed by Chan and Vese in [38]. Formally it is a limiting case of the Mumford-Shah problem with the parameters $(\alpha\lambda, \alpha\mu)$, where $\alpha \rightarrow \infty$, and hence it mainly contains the geometric parameter Γ . In the simplest case of two gray values one has to minimize the functional

$$F(\Gamma, c_1, c_2) = \int_{\Omega} (\chi_{\Omega'} c_1 + \chi_{\Omega \setminus \Omega'} c_2 - u^0)^2 dx + \mu \mathfrak{H}^{d-1}(\Gamma),$$

where $\Gamma = \partial\Omega' \cap \Omega$. For a fixed Γ , the optimal constants are readily calculated as $c_1 = |\Omega'|^{-1} \int_{\Omega'} u^0 dx$, and $c_2 = |\Omega \setminus \Omega'|^{-1} \int_{\Omega \setminus \Omega'} u^0 dx$ and the difficult part is to find the edge set Γ . For this problem so-called *level-set methods* are popular [105, 128]. Roughly speaking, the main idea is to represent Γ as the zero level-set of a function $\phi : \Omega \rightarrow \mathbf{R}$ that is positive in Ω' and negative in $\Omega \setminus \Omega'$. With the Heaviside function $H = \chi_{[0, \infty]}$ we can write F as

$$F(\phi, c_1, c_2) = \int_{\Omega} (c_1 - u^0)^2 (H \circ \phi) + (c_2 - u^0)^2 ((1 - H) \circ \phi) dx + \mu \text{TV}(H \circ \phi),$$

which is to be minimized. In practice one replaces H by a smoothed version H_ε , formally derives the Euler-Lagrange equations, and performs a gradient descent as described in Sect. 6.4.1:

$$\frac{\partial \phi}{\partial t} = (H'_\varepsilon \circ \phi) \left(- (c_1 - u^0)^2 + (c_2 - u^0)^2 + \operatorname{div} \left(\frac{\nabla \phi}{|\nabla \phi|} \right) \right).$$

The numerical solution of this equation implicitly defines the evolution of the edge set Γ . If one updates the mean values c_1 and c_2 during the iteration, this approach results, after a few more numerical tricks, in a solution method [38, 73].

To *decompose images* into different parts there are more elaborate approaches such as the denoising method we have seen in this chapter, especially with respect to texture. The model behind these methods assumes that the dominant features of an image u^0 are given by a piecewise smooth “cartoon” part and a texture part. The texture part should contain mainly fine and repetitive structures, but not contain noise. If one assumes that u^0 contains noise, one postulates

$$u^0 = u^{\text{cartoon}} + u^{\text{texture}} + \eta.$$

We mention specifically the so-called G -norm model by Meyer [70, 87, 99]. In this model the texture part u^{texture} is described by the semi-norm that is dual to the total variation:

$$\|u\|_* = \inf \left\{ \|\sigma\|_\infty \mid \sigma \in \mathcal{D}_{\operatorname{div}, \infty}, \operatorname{div} \sigma = u \right\}.$$

The G -norm has several interesting properties and is well suited to describe oscillating repetitive patterns. In this context u^{cartoon} is usually modeled with Sobolev semi-norms or the total variation, which we discussed extensively in this chapter. An associated minimization problem can be, for example,

$$\min_{u^{\text{cartoon}}, u^{\text{texture}}} \frac{1}{q} \int_{\Omega} |u^0 - u^{\text{cartoon}} - u^{\text{texture}}|^q \, dx + \frac{\lambda}{p} \int_{\Omega} |\nabla u^{\text{cartoon}}|^p \, dx + \mu \|u^{\text{texture}}\|_*$$

with two parameters $\lambda, \mu > 0$ and exponents $1 < p, q < \infty$ or $\lambda \operatorname{TV}$ instead of $\|\nabla u\|_p^p$. Approaches to modeling textures that are different from the G -norm also use dual semi-norms, e.g., the negative Sobolev semi-norm

$$1 \leq p < \infty : \quad \|u\|_{H^{-1,p}} = \sup \left\{ \int_{\Omega} uv \, dx \mid v \in H^{1,p^*}(\Omega), \|\nabla v\|_{p^*} \leq 1 \right\}.$$

There are also theoretical results and several numerical methods for this approach available [10, 92, 107].

The problem to determine the *optical flow* can be cast as a variational problem in different ways; see, for example, [17, 24, 75]. We show the classical approach of [78]. For a given image sequence $u : [0, 1] \times \Omega \rightarrow \mathbf{R}$ on a domain Ω one aims to

find a velocity field $v : [0, 1] \times \Omega \rightarrow \mathbf{R}^d$ for which $v(t, \cdot)$ gives the directions and velocities of the objects in each image $u(t, \cdot)$. The respective variational problem is then derived from the following considerations. If we traced the movement of every point $x \in \Omega$ in time, the path would follow the trajectory $\varphi_x : [0, 1] \rightarrow \Omega$ with $\varphi_x(0) = x$. Now we assume that the overall brightness of the sequence u^0 does not change, and even that all the points do not change their brightness over time. If we plug this assumption into the trajectory, we see that $u(t, \varphi_x(t)) = u(0, x)$ has to hold. Differentiating this equation with respect to t leads to

$$\frac{\partial u}{\partial t}(t, \varphi_x(t)) + \nabla u(t, \varphi_x(t)) \cdot \varphi'_x(t) = 0.$$

The derivative $\varphi'_x(t)$ is exactly the velocity $v(t, \varphi_x(t))$. If we claim that the above equation is satisfied throughout $[0, 1] \times \Omega$ we obtain the *optical flow constraint*

$$\frac{\partial u}{\partial t} + \nabla u \cdot v = 0. \quad (6.99)$$

On the other hand, the velocity field v should have a certain smoothness in space, i.e., we expect that the objects mainly follow rigid motions and deform only slightly, e.g., by a change of the viewpoint. The idea from [78] is to enforce the smoothness by an H^1 penalty. Since the brightness will not be exactly constant, one does not enforce the optical flow constraint exactly. This leads to the following minimization problem for the optical flow at time t :

$$F(v) = \frac{1}{2} \int_{\Omega} \left(\frac{\partial u}{\partial t}(t) + \nabla u(t) \cdot v \right)^2 dx + \frac{\lambda}{2} \int_{\Omega} |\nabla v|^2 dx.$$

There are many variants of this theme. On the one hand, one could use the whole time interval in the optimization, and hence the minimization problem would contain an integral over the time interval $[0, 1]$. On the other hand, one can consider numerous other data and regularization terms [24].

In practice there are only discrete images. There are several approaches in the literature that determine the optical flow in the case that only $u^0 = u(0)$ and $u^1 = u(1)$ are known. Here one assumes again that $u(t)$ satisfies the optical flow constraint. If v is known, one can set the initial condition $u(0) = u^0$ and solve the transport equation (6.99) (e.g., by the method of characteristics; see Chap. 5) to obtain $u(1) = u^1$. If this is not satisfied, one can use the discrepancy $u(1) - u^1$ to determine how well the unknown v fits the data. This motivates the optimization problem

$$\min_v \frac{1}{2} \int_{\Omega} |u(1) - u^0|^2 dx + \lambda \int_0^1 \int_{\Omega} \varphi \left(\frac{\partial v}{\partial t}, \nabla v \right) dx dt \quad \text{with} \quad \begin{cases} \frac{\partial u}{\partial t} + \nabla u \cdot v = 0, \\ u(0) = u^0; \end{cases}$$

see [17]. The penalty for v contains a regularization in space, but also the time derivative $\frac{\partial v}{\partial t}$. Another possibility is to fix both endpoints $u(0)$ and $u(1)$ and to

look for an interpolation $u : [0, 1] \rightarrow \Omega$ and measure the deviation of the optical flow constraint. In this case one optimizes over both u and v ; for this approach, see [42, 43, 84]:

$$\min_{u,v} \frac{1}{2} \int_0^1 \int_{\Omega} \left(\frac{\partial u}{\partial t} + v \cdot \nabla u \right)^2 + \lambda \varphi \left(\frac{\partial v}{\partial t}, \nabla v \right) dx dt \quad \text{with} \quad \begin{cases} u(0) = u^0, \\ u(1) = u^1. \end{cases}$$

The *registration* of an image $u^0 : \Omega \rightarrow \mathbf{R}$ with another $u^1 : \Omega \rightarrow \mathbf{R}$ uses similar functionals. The main difference lies in the fact that the vector field $v : \Omega \rightarrow \mathbf{R}^d$ now is a deformation field, i.e., it is time independent and does not need to satisfy the optical flow constraint, but is considered to be a stationary coordinate transformation. The forward model is the equation $u^0 \circ (\text{id} + v) = u^1$. Variational methods use a minimization problem that measures both the data fit $u^0 \circ (\text{id} + v) - u^1$ and the smoothness of v . For more details we refer to the survey article [63] and the book [100].

Finally, we give an outlook on recent developments for *regularization terms* for images. As developed in Sects. 6.3.1 and 6.3.3, spaces of weakly differentiable functions are well suited for the regularization of variational imaging methods. The total variation is especially interesting as a penalty due to its ability to preserve edges. It is used in virtually all imaging problems. Unfortunately, there are some problems with the TV semi-norm. The most prominent drawback is the “staircasing effect,” i.e. the problem that new edges arise in places where no new edges should be (see, e.g., Figs. 6.19 and 6.21). These are two sides of the same coin. On the one hand, we like to have new edges to appear at places where the original image has jumps, but on the other hand they also appear where the original image was smooth. The TV semi-norm cannot distinguish between these two cases.

One possibility to overcome the staircasing effect is to incorporate terms with higher-order derivatives. This should be done in a way that still allows one to reconstruct jump discontinuities. Unfortunately, these are competing demands: If a function in $\text{BV}(\Omega)$ has a jump, then its derivative ∇u is singular with respect to the Lebesgue measure, and it is differentiable only in the sense of distributions. If, on the other hand, $\nabla^2 u$ is, for example, a Radon measure or an L^p function, one can show that ∇u is an L^p function, i.e., u can no longer have edges. There are many different approaches to enabling higher-order methods to respect edges, both with convex and non-convex functionals [19, 31, 40, 41, 74, 95, 125, 129]. We present only some convex approaches, since there is a well-developed theory in this case.

A simple generalization of TV is the *total variation of second order* [125]:

$$\text{TV}^2(u) = \sup \left\{ \int_{\Omega} u \operatorname{div}^2 v dx \mid v \in \mathcal{D}(\Omega, \mathbf{R}^{d \times d}), \|v\|_{\infty} \leq 1 \right\}.$$

Of course, one can use other differential operators for a similar definition, e.g., the Laplace operator

$$\|\Delta u\|_{\mathfrak{M}} = \sup \left\{ \int_{\Omega} u \Delta v dx \mid v \in \mathcal{D}(\Omega), \|v\|_{\infty} \leq 1 \right\}$$

or the variant [95]

$$\|\text{diag } \nabla^2 u\|_{\mathfrak{M}} = \sup \left\{ \int_{\Omega} u \left(\sum_{i=1}^d \frac{\partial^2 v_i}{\partial x_i^2} \right) dx \mid v \in \mathcal{D}(\Omega, \mathbf{R}^d), \|v\|_{\infty} \leq 1 \right\}.$$

As such, these regularization terms in a denoising problem give smooth solutions, but in comparison with TV they lead to notably less sharp edges (see Fig. 6.31).

To employ the advantages of the total variation, one can use the *infimal convolution*. There are the following variants proposed in the literature [31, 41]:

$$(\text{TV} \triangle \alpha \text{TV}^2)(u) = \inf_{u^1 + u^2 = u} \text{TV}(u^1) + \alpha \text{TV}^2(u^2),$$

$$(\text{TV} \triangle \alpha \|\cdot\|_{\mathfrak{M}} \circ \Delta)(u) = \inf_{u^1 + u^2 = u} \text{TV}(u^1) + \alpha \|\Delta u^2\|_{\mathfrak{M}}.$$

Both approaches are well suited for preserving edges with reduced staircasing, and the different second-order differential operators lead to slightly different

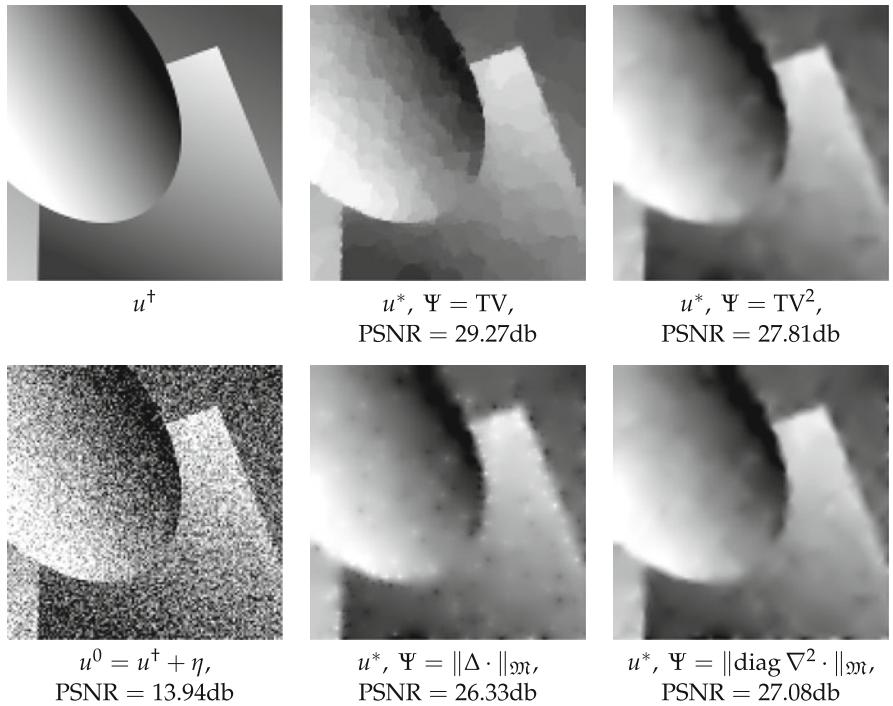


Fig. 6.31 Illustration of variational denoising with second-order penalties. Left: The original on top, below the noisy version. Middle and right: The minimizers of the L^2 - Ψ denoising problem with PSNR-optimal parameter λ

characteristics. The reduction of staircasing artifacts is even more prominent for functionals that constrain the zeroth and first order derivatives in the predual form, i.e., in the functional [129]

$$\Psi_{\text{diag}}(u) = \sup \left\{ \int_{\Omega} u \left(\sum_{i=1}^d \frac{\partial^2 v_i}{\partial x_i^2} \right) dx \mid v \in \mathcal{D}(\Omega, \mathbf{R}^d), \|v\|_{\infty} \leq \alpha, \|\text{diag } \nabla v\|_{\infty} \leq 1 \right\}$$

or in the “total generalized variation” of second order

$$\text{TGV}_{\alpha}^2(u) = \sup \left\{ \int_{\Omega} u \operatorname{div}^2 v dx \mid v \in \mathcal{D}(\Omega, S^{d \times d}), \|v\|_{\infty} \leq \alpha, \|\operatorname{div} v\|_{\infty} \leq 1 \right\},$$

which easily generalized to higher orders [19]. Figure 6.32 shows results for these functionals in a denoising problem.

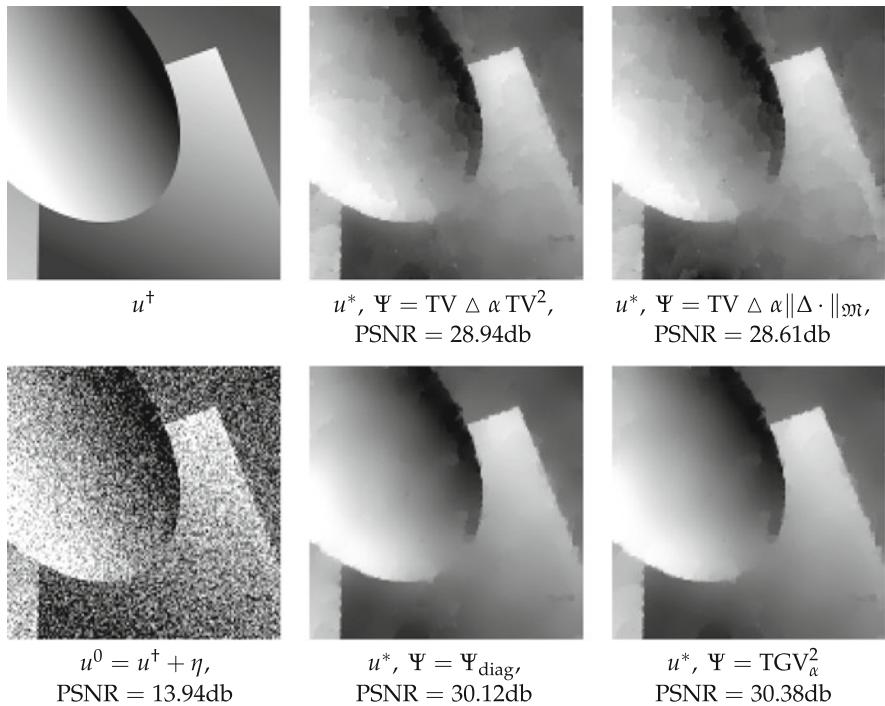


Fig. 6.32 Illustration of variational denoising with penalties that combine first and second order. Left: The original on top, below the noisy version. Middle and right: The minimizer of the L^2 - Ψ denoising problems with PSNR-optimal parameters λ

6.6 Exercises

Exercise 6.1 Let $\widehat{P}_\lambda(\xi) = (2\pi)^{-d/2}(1 + \lambda|\xi|^2)^{-1}$ with $\lambda > 0$.

1. Show that the inverse Fourier transform P_λ in the sense of distributions is given by

$$P_\lambda(x) = \frac{|x|^{1-d/2}}{(2\pi)^{d-1}\lambda^{(d+2)/4}} K_{d/2-1}\left(\frac{2\pi|x|}{\sqrt{\lambda}}\right).$$

The following result from [127] may come in handy:

$$\mathcal{F}\left((1 + |\cdot|^2)^{-\frac{m}{2}}\right) = \frac{(2\pi)^{\frac{m-d}{2}}}{\Gamma(\frac{m}{2})} |\cdot|^{\frac{m-d}{2}} K_{\frac{d-m}{2}}(2\pi|\cdot|).$$

2. For odd d , use

$$K_{-\nu} = K_\nu, \quad \nu \in \mathbf{C}, \quad K_{n+1/2}(z) = \sqrt{\frac{\pi}{2z}} e^{-z} \sum_{k=0}^n \frac{(n+k)!}{k!(n-k)!(2z)^k}, \quad n \in \mathbf{N},$$

from [138] to derive a formula P_λ that does not use $K_{d/2-1}$.

3. Finally, show that $P_\lambda \in L^1(\mathbf{R}^d)$ for every dimension $d \geq 1$. You may use the estimates

$$|K_\nu(z)| = \begin{cases} \mathcal{O}(|z|^{-\nu}) & \text{if } \operatorname{Re} \nu > 0, \\ \mathcal{O}(\log|z|) & \text{if } \nu = 0, \end{cases} \quad \text{for } |z| \rightarrow 0,$$

$$|K_\nu(z)| = \mathcal{O}\left(\frac{e^{-|z|}}{|z|^{1/2}}\right) \quad \text{for } |z| \rightarrow \infty,$$

from [138].

Exercise 6.2 Let $\Omega \subset \mathbf{R}^d$ be a domain and $\Omega' \subset \Omega$ a bounded Lipschitz subdomain such that $\overline{\Omega'} \subset\subset \Omega$.

1. Show the identities

$$\{u \in H^1(\Omega) \mid u = 0 \text{ on } \Omega \setminus \Omega'\} = \{u \in H^1(\Omega) \mid u|_{\Omega'} \in H_0^1(\Omega')\} = \overline{\mathcal{D}(\Omega')},$$

where the closure is to be understood with respect to the H^1 norm.

2. Prove that the traces of the functions $u^1 = u|_{\Omega \setminus \overline{\Omega'}}$ and $u^2 = u|_{\Omega'}$ in $L^2_{\mathfrak{H}^{d-1}}(\partial\Omega')$ are equal for all functions in $u \in H^1(\Omega)$.

Exercise 6.3 Show that on every nontrivial Banach space X there is a noncoercive functional $F : X \rightarrow \mathbf{R}_\infty$ that has a unique minimizer.

Exercise 6.4 Let X^* be the dual space of a separable normed space. Show that a functional $F : X^* \rightarrow \mathbf{R}_\infty$, that is bounded from below, coercive, and sequentially weak* lower semicontinuous has a minimizer in X^* .

Exercise 6.5 Let $\Phi : X \rightarrow Y$ be an affine linear map between vector spaces X, Y , i.e.,

$$\Phi(\lambda x + (1 - \lambda)y) = \lambda\Phi(x) + (1 - \lambda)\Phi(y) \quad \text{for all } x, y \in X \quad \text{and } \lambda \in \mathbf{K}.$$

Show that there exist a $y^0 \in Y$ and a linear $F : X \rightarrow Y$ such that $\Phi(x) = y^0 + Fx$ holds for all $x \in X$.

Exercise 6.6 Let $A \in \mathcal{L}(X, Y)$ be a linear and continuous map between Banach spaces X and Y . Show that the following assertions are equivalent:

1. For some $u^0 \in Y$ and $p \geq 1$ the functional $F : X \rightarrow \mathbf{R}$ given by

$$F(u) = \frac{\|Au - u^0\|_Y^p}{p}$$

is coercive on X .

2. The operator A is injective, $\text{rg}(A)$ is closed, and in particular $A^{-1} : \text{rg}(A) \rightarrow X$ is continuous.

Exercise 6.7 Let X and Y be Banach spaces and let X be reflexive. Moreover let $|\cdot|_X : X \rightarrow \mathbf{R}$ be an admissible semi-norm on X , i.e., there exist a linear and continuous $P : X \rightarrow X$ and constants $0 < c \leq C < \infty$ such that

$$c\|Pu\|_X \leq |u|_X \leq C\|Pu\|_X \quad \text{for all } u \in X.$$

Show that if $A \in \mathcal{L}(X, Y)$ is continuously invertible on $\{u \in X \mid |u|_X = 0\} = \ker(P)$ then there exists a minimizer of the Tikhonov functional F with

$$F(u) = \frac{\|Au - u^0\|_Y^p}{p} + \lambda \frac{|u|_X^q}{q}$$

for every $u^0 \in Y$, $p, q \geq 1$, and $\lambda > 0$.

Addition: The assertion remains true if X is the dual space of a separable normed space and A is weakly*-to-weakly continuous.

Exercise 6.8 Let X be a real Hilbert space, and $u \in X$ with $\|u\|_X = 1$.

Show $w \in X$ with $(w, v) \geq 0$ for all $v \in X$ with $(u, v) < 0$ then there exists some $\mu \geq 0$ such that $w + \mu u = 0$.

Exercise 6.9 Let $K \subset X$ be a nonempty convex subset of a normed space and assume that $B_r(x^1) \subset K$ for some $x^1 \in X$ and $r > 0$. Show that for every $x^0 \in K$

and $\lambda \in]0, 1]$ also $B_{\lambda r}(x^\lambda) \subset K$, where $x^\lambda = \lambda x^1 + (1 - \lambda)x^0$. Deduce that if $\text{int}(K) \neq \emptyset$, then $\overline{\text{int}(K)} = K$.

Exercise 6.10 Let $F : X \rightarrow \mathbf{R}_\infty$ be convex on a real normed space X . Show that:

1. If $w^1 \in \partial F(u)$ and $w^2 \in \partial F(u)$ for some $u \in X$, then $\lambda w^1 + (1 - \lambda)w^2 \in \partial F(u)$ for all $\lambda \in [0, 1]$.
2. If F is lower semicontinuous, then for a sequence $((u^n, w^n))$ in $X \times X^*$ with $w^n \in \partial F(u^n)$ and $u^n \rightarrow u$ and $w^n \xrightarrow{*} w$ one has $w \in \partial F(u)$.
3. The assertion in item 2 remains true if $u^n \rightharpoonup u$ and $w^n \rightarrow w$.
4. The set $\partial F(u)$, $u \in X$ is sequentially weakly* closed.

Exercise 6.11 Let X, Y be real Banach spaces. Prove that $\partial(F \circ A) = A^* \circ \partial F \circ A$ holds for $F : X \rightarrow \mathbf{R}_\infty$ convex and $A : Y \rightarrow X$ linear, continuous, and surjective.

[Hint:] Use the open mapping theorem (Theorem 2.16) for $B : Y \times \mathbf{R} \rightarrow X \times \mathbf{R}$ with $B : (v, t) \mapsto (Av, \langle w, v \rangle_{Y^* \times Y} + t)$ and $w \in Y^*$ to show that

$$K_2 = \{(Av, s) \in X \times \mathbf{R} \mid s \leq F(Au) + \langle w, v - u \rangle_{Y^* \times Y}\}$$

has nonempty interior for every $u \in \text{dom } F$.

Exercise 6.12 Let X, Y be real normed spaces, $\text{dom } A \subset Y$ a dense subspace of Y , $A : \text{dom } A \rightarrow X$ linear, and $F : X \rightarrow \mathbf{R}_\infty$ convex. Show that the following criteria are sufficient for $\partial(F \circ A) = A^* \circ \partial F \circ A$ to hold (with $A^* : X^* \supset \text{dom } A^* \rightarrow Y^*$ from Definition 2.25):

1. F is continuous at some point $u^0 \in \text{dom } F \cap \text{rg}(A)$,
2. X, Y are complete and A is surjective.

[Hint:] Show that A is an open mapping in this situation.

Exercise 6.13 Let $K_1, K_2 \subset X$ be nonempty convex subsets of a normed space X and $X = X_1 + X_2$ with subspaces X_1, X_2 and continuous $P_i \in \mathcal{L}(X, X_i)$ such that $\text{id} = P_1 + P_2$. Show that:

1. If the K_i are disjoint and both relatively open with respect to X_i , i.e., $(K_i - x) \cap X_i$ is relatively open in X_i for every $x \in X$, then there exist $x^* \in X^*$, $x^* \neq 0$, and $\lambda \in \mathbf{R}$ such that

$$\operatorname{Re} \langle x^*, x \rangle \leq \lambda \quad \text{for all } x \in K_1 \quad \text{and} \quad \operatorname{Re} \langle x^*, x \rangle \geq \lambda \quad \text{for all } x \in K_2.$$

[Hint:] Separate $\{0\}$ from the open set $K_1 - K_2$.

2. The assertion remains true if $\text{int}_{X_i}(K_i)$ are nonempty and disjoint. Here $\text{int}_{X_i}(K_i)$ is the set of all $x \in K_i$ such that 0 is a relative interior point of $(K_i - x) \cap X_i$ in X_i .

[Hint:] First prove that $\overline{K_i} = \overline{\text{int}_{X_i}(K_i)}$ (see also Exercise 6.9).

Exercise 6.14 Let $F_1, F_2 : X \rightarrow \mathbf{R}_\infty$ be convex functionals on a real normed space X with the property that $X = X_1 + X_2$ for subspaces X_i and continuous $P_i \in \mathcal{L}(X, X_i)$ with $\text{id} = P_1 + P_2$.

Furthermore, assume that there exists a point $x^0 \in \text{dom } F_1 \cap \text{dom } F_2$ such that $x^i \mapsto F_i(x^0 + x^i)$ with $x^i \in X_i$ continuous at 0. Prove that the identity

$$\partial(F_1 + F_2) = \partial F_1 + \partial F_2$$

holds.

Exercise 6.15 Show that for real Banach spaces X, Y and $A \in \mathcal{L}(Y, X)$ such that $\text{rg}(A)$ is closed and there exists a subspace $X_1 \subset X$ with $X = X_1 + \text{rg}(A)$, as well as mappings $P_1 \in \mathcal{L}(X, X_1)$ and $P_2 \in \mathcal{L}(X, \text{rg}(A))$ with $\text{id} = P_1 + P_2$, one has for convex $F : X \rightarrow \mathbf{R}_\infty$ for which there exists a point $x^0 \in \text{dom } F \cap \text{rg}(A)$ such that $x^1 \mapsto F(x^0 + x^1)$, $x^1 \in X_1$ is continuous at the origin that

$$\partial(F \circ A) = A^* \circ \partial F \circ A.$$

Exercise 6.16 Use the results of Exercise 6.15 to find an alternative proof for the third point in Theorem 6.51.

Exercise 6.17 Let X, Y be real Banach spaces and $A \in \mathcal{L}(X, Y)$ injective and $\text{rg}(A)$ dense in Y . Show:

1. A^{-1} with $\text{dom } A^{-1} = \text{rg}(A)$ is a closed and densely defined linear map,
2. $(A^*)^{-1}$ with $\text{dom } (A^*)^{-1} = \text{rg}(A^*)$ is a closed and densely defined linear map,
3. it holds that $(A^{-1})^* = (A^*)^{-1}$.

Exercise 6.18 Let $\Omega \subset \mathbf{R}^d$ be a domain and $F : L^2(\Omega) \rightarrow \mathbf{R}$ convex and Gâteaux-differentiable. Consider the problems of minimizing the functionals

$$\min_{u \in L^2(\Omega)} F_1(u), \quad F_1 = F + I_{\{\|v\|_2 \leq 1\}}, \quad \min_{u \in L^2(\Omega)} F_2(u), \quad F_2 = F + I_{\{\|v\|_\infty \leq 1\}}.$$

Use subdifferential calculus to calculate the subgradients of F_1 and F_2 . Derive the optimality conditions and verify that they are equivalent to those in Example 6.37.

Exercise 6.19 Let X, Y be nonempty sets and $F : X \times Y \rightarrow \mathbf{R}_\infty$. Show that

$$\sup_{y \in Y} \inf_{x \in X} F(x, y) \leq \inf_{x \in X} \sup_{y \in Y} F(x, y),$$

where the supremum of a set that is unbounded from above is ∞ and the infimum of a set that is unbounded from below is $-\infty$.

Exercise 6.20 Prove the claims in Lemma 6.59.

Exercise 6.21 For $p \in [1, \infty[$ let $F : X \rightarrow \mathbf{R}_\infty$ be a positively p -homogeneous and proper functional on the real Banach space X , i.e.,

$$F(\alpha u) = \alpha^p F(u) \quad \text{for all } u \in X \text{ and } \alpha \geq 0.$$

Show:

1. For $p > 1$, one has that F^* is positively p^* -homogeneous with $\frac{1}{p} + \frac{1}{p^*} = 1$.
2. For $p = 1$, one has that $F^* = I_K$ with a convex and closed set $K \subset X^*$.
3. For “ $p = \infty$,” i.e., $F = I_K$ with $K \neq \emptyset$ positively absorbing, i.e., $\alpha K \subset K$ for $\alpha \in [0, 1]$, one has that F^* is positively 1-homogeneous.

Exercise 6.22 Let X be a real Banach space and $F : X \rightarrow \mathbf{R}_\infty$ strongly coercive.

Show that the Fenchel conjugate $F^* : X \rightarrow \mathbf{R}$ is continuous.

Exercise 6.23 Let X be a real Banach space, $I \neq \emptyset$ an index set, and $F_i : X \rightarrow \mathbf{R}_\infty$ a family of proper functionals with $\bigcap_{i \in I} \text{dom } F_i \neq \emptyset$. Prove the identity

$$\left(\sup_{i \in I} F_i \right)^* = \left(\inf_{i \in I} F_i^* \right)^{**}.$$

Exercise 6.24 Let X be a real reflexive Banach space and let $F_1, F_2 : X \rightarrow \mathbf{R}_\infty$ be proper functionals. Moreover, assume that there exists a $u^0 \in \text{dom } F_1 \cap \text{dom } F_2$ such that F_1 is continuous in u^0 .

Show:

1. For every $L \in \mathbf{R}$ and $R > 0$, the set

$$M = \{(w^1, w^2) \in X^* \times X^* \mid F_1^*(w^1) + F_2^*(w^2) \leq L, \|w^1 + w^2\|_{X^*} \leq R\}$$

is bounded.

[Hint:] Use the uniform boundedness principle. To that end, use the fact that for every $(v^1, v^2) \in X \times X$ one can find a representation $v^1 - \alpha u^0 = v^2 - \alpha u^1$ with $\alpha > 0$ and $u^1 \in \text{dom } F_1$ and apply the Fenchel inequality to $\langle w^1, u^1 \rangle$ and $\langle w^2, u^0 \rangle$.

2. The infimal convolution $F_1^* \triangle F_2^*$ is proper, convex, and lower semicontinuous.

[Hint:] For the lower semicontinuity use the result of the first point to conclude that for sequences $w^n \rightarrow w$ with $(F_1^* \triangle F_2^*)(w^n) \rightarrow t$, sequences $((w^1)^n), ((w^2)^n)$ with $(w^1)^n + (w^2)^n = w^n$ are bounded and $F_1^*((w^1)^n) + F_2^*((w^2)^n) \leq (F_1^* \triangle F_2^*)(w^n) + \frac{1}{n}$. Also use the lower semicontinuity of F_1^* and F_2^* .

3. Moreover, the infimal convolution is *exact*, i.e., for every $w \in X^*$ the minimum in

$$\min_{w^1 + w^2 = w} F_1^*(w^1) + F_2^*(w^2)$$

is attained.

[Hint:] Use the first point and the direct method to show that for a given $w \in X^*$ a respective minimizing sequence $((w^1)^n, (w^2)^n)$ with $(w^1)^n + (w^2)^n = w$ is bounded.

4. If F_1 and F_2 are convex and lower semicontinuous, then $(F_1 + F_2)^* = F_1^* \triangle F_2^*$.
5. The exactness of $F_1^* \triangle F_2^*$ and $(F_1 + F_2)^* = F_1^* \triangle F_2^*$ implies that for convex and lower semicontinuous F_1, F_2 , one has that $\partial(F_1 + F_2) = \partial F_1 + \partial F_2$.

Exercise 6.25 Consider the semi-norm

$$\|\nabla^m u\|_p = \left(\int_{\mathbf{R}^d} \left(\sum_{|\alpha|=m} \binom{m}{\alpha} \left| \frac{\partial^m u}{\partial x^\alpha}(x) \right|^2 \right)^{p/2} dx \right)^{1/p}$$

on the space $H^{m,p}(\mathbf{R}^d)$, $m \geq 0$, $p \in [1, \infty]$.

Show that for every $u \in H^{m,p}(\mathbf{R}^d)$, $x^0 \in \mathbf{R}^d$, and all isometries $O \in \mathbf{R}^{d \times d}$ one has

$$\|\nabla^m \cdot\|_p = \|\nabla^m \cdot\|_p \circ T_{x^0} \circ O.$$

In other words, $\|\nabla^m u\|_p = \|\nabla^m(u(O \cdot + x^0))\|_p$, i.e., $\|\nabla^m \cdot\|_p$ is invariant under translation and the application of isometries.

[Hint:] Consider $\nabla^m u(x)$ as m -linear map $\mathbf{R}^d \times \cdots \times \mathbf{R}^d \rightarrow \mathbf{R}$ and use the identities

$$\begin{aligned} \nabla^m(T_{x^0} D_O u)(x)(h_1, \dots, h_m) &= \nabla^m u(Ox + x^0)(Oh_1, \dots, Oh_m), \\ \sum_{|\alpha|=m} \binom{m}{\alpha} \left| \frac{\partial^m u}{\partial x^\alpha}(x) \right|^2 &= \sum_{i_m=1}^d \cdots \sum_{i_1=1}^d \left| \frac{\partial^m u}{\partial x_{i_1} \cdots \partial x_{i_m}}(x) \right|^2, \end{aligned}$$

which hold for smooth functions.

Exercise 6.26 Let the assumptions in Theorem 6.86 be satisfied. Moreover, assume that Y is a Hilbert space. Denote by $T : Y \rightarrow A\Pi^m$ the orthogonal projection onto the image of the polynomials of degree up to $m - 1$ under A and further set $S = A^{-1}TA$, where A is inverted on $A\Pi^m$.

Show that every solution u^* of the problem (6.36) satisfies $ASu^* = Tu^*$.

Exercise 6.27 Let Ω be a bounded Lipschitz domain, $m \geq 1$, and $p, q \in]1, \infty[$.

1. Using appropriate conditions for q , show that the denoising problem

$$\min_{u \in L^q(\Omega)} \frac{1}{q} \int_{\Omega} |u - u^0|^q dx + \frac{\lambda}{p} \int_{\Omega} |\nabla^m u|^p dx$$

has solutions for every $u^0 \in L^q(\Omega)$ and $\lambda > 0$.

2. Show that for $q = 2$, every minimizer u^* satisfies $Q_m u^* = Q_m u^0$.
3. Does a maximum principle as in Theorem 6.95 hold for $m \geq 2$?

Exercise 6.28 Let Ω' be a bounded domain, $k \in L^1(\Omega_0)$, Ω a bounded Lipschitz domain such that $\Omega' - \Omega_0 \subset \Omega$ holds, and let $m \geq 1$ and $p, q \in]1, \infty[$.

Furthermore, let $\lambda > 0$ and $u^0 \in L^q(\Omega)$ satisfy the estimates $L \leq u^0 \leq R$ almost everywhere in Ω for some $L, R \in \mathbf{R}$.

Prove that the problem

$$\min_{u \in L^q(\Omega)} \frac{1}{q} \int_{\Omega'} |u * k - u^0|^q dx + \frac{\lambda}{p} \int_{\Omega} |\nabla^m u|^p dx + I_{\{v \in L^q(\Omega) \mid L \leq v \leq R \text{ almost everywhere}\}}(u)$$

has a unique solution.

Exercise 6.29 Let Ω' be a bounded domain, $k \in L^1(\Omega_0)$, Ω a bounded Lipschitz domain such that $\Omega' - \Omega_0 \subset \Omega$ holds, and let $m \geq 1$ and $p, q \in]1, \infty[$.

1. Show that $A : L^q(\Omega) \rightarrow L^q(\Omega')$ defined by $Au = (u * k)|_{\Omega'}$ maps polynomials in Ω to polynomials in Ω' .
2. Assume that k satisfies, for all multi-indices α with $|\alpha| < m$, that

$$\int_{\Omega_0} k(x) x^\alpha dx = \begin{cases} 1 & \text{if } \alpha = 0, \\ 0 & \text{otherwise.} \end{cases} \quad (6.100)$$

Prove that $Au = u|_{\Omega'}$ for all polynomials $u \in \Pi^m$.

[Hint:] You may use the multinomial theorem $(x + y)^\alpha = \sum_{\beta \leq \alpha} \binom{\alpha}{\beta} x^\beta y^{\alpha-\beta}$.

3. Prove: If k satisfies the assumption (6.100) and q satisfies appropriate assumptions, then the minimization problem

$$\min_{u \in L^q(\Omega)} \frac{1}{q} \int_{\Omega'} |u * k - u^0|^q dx + \frac{\lambda}{p} \int_{\Omega} |\nabla^m u|^p dx$$

has a unique solution for every $u^0 \in L^q(\Omega')$ and $\lambda > 0$.

Exercise 6.30 Prove the claims in Lemma 6.99.

1. For the first claim you may proceed as follows: consider the smoothing operators \mathcal{M}_n^* from Theorem 6.88, i.e.,

$$\mathcal{M}_n^* u = \sum_{k=0}^K \varphi_k(T_{-t_n \eta_k}(u * \bar{\psi}_{k,n})),$$

and show that $\mathcal{M}_n^* u \in \mathcal{D}(\Omega)$ as well as $\mathcal{M}_n^* u|_{\Omega} \rightarrow u^0 = u|_{\Omega}$ in $H^{m,p}(\Omega)$.

2. For the second claim you may use Gauss's theorem (Theorem 2.81) to reduce it to the first claim.

Exercise 6.31 Let $M, N \in \mathbf{N}$ with $N, M \geq 1$ and let $S \in \mathbf{R}^{N \times N}$ and $W \in \mathbf{R}^{M \times N}$ be matrices.

Show that if S is positive definite on $\ker(W)$, i.e., for $Wx = 0$ and $x \neq 0$, one has that $x^T S x > 0$, then the block matrix

$$A = \begin{pmatrix} S & W^T \\ W & 0 \end{pmatrix}$$

is invertible.

Exercise 6.32 Let $\Omega \subset \mathbf{R}^d$ be a domain, $N \in \mathbf{N}$, and $L : \mathcal{D}(\Omega, \mathbf{R}^N) \rightarrow \mathbf{R}$ such that there exists a constant $C \geq 0$ such that $|L(\varphi)| \leq C\|\varphi\|_\infty$ for every $\varphi \in \mathcal{D}(\Omega, \mathbf{R}^N)$. Show that L has a unique continuous extension to $\mathcal{C}_0(\Omega, \mathbf{R}^N)$ and hence that there exists a vector-valued finite Radon measure $\mu \in \mathfrak{M}(\Omega, \mathbf{R}^N)$ such that

$$L(\varphi) = \int_{\Omega} \varphi \, d\mu \quad \text{for all } \varphi \in \mathcal{D}(\Omega, \mathbf{R}^N).$$

[Hint:] Use the definition of $\mathcal{C}_0(\Omega, \mathbf{R}^N)$ as well as Theorems 3.13 and 2.62.

Exercise 6.33 Prove the assertions in Lemma 6.103.

[Hint:]

1. Use the results of Exercise 6.32 and the characterization of $\mathfrak{M}(\Omega, \mathbf{R}^d)$ as a dual space.
2. Combine the definition of the weak gradient with the identification $\mathfrak{M}(\Omega, \mathbf{R}^d) = \mathcal{C}_0(\Omega, \mathbf{R}^d)^*$.
3. Follow the proof of Lemma 6.73.

Exercise 6.34 Let $\Omega \subset \mathbf{R}^d$ be a bounded Lipschitz domain, and let $h : \mathbf{R} \rightarrow \mathbf{R}$ be strictly increasing, continuous, differentiable, and satisfy $\|h'\|_\infty < \infty$.

Prove that for every $u \in \text{BV}(\Omega)$ one has also $h \circ u$ in $\text{BV}(\Omega)$ and

$$\text{TV}(h \circ u) = \int_{\mathbf{R}} h'(t) \text{Per}(\{u \leq t\}) \, dt \leq \|h'\|_\infty \text{TV}(u).$$

Exercise 6.35 Let $\Omega \subset \mathbf{R}^d$ be a domain, $N \in \mathbf{N}$, and μ a σ -finite positive Radon measure on Ω . Prove that $\mathcal{D}(\Omega, \mathbf{R}^N)$ is dense in every $L_\mu^p(\Omega, \mathbf{R}^N)$ with $1 \leq p < \infty$.

[Hint:] For $u \in L_\mu^p(\Omega, \mathbf{R}^N)$ use an argument based on truncation, Lusin's theorem, and the inner regularity of Borel measures (see, e.g., [5]) to find an approximating sequence (u^n) in $\mathcal{C}_c(\Omega, \mathbf{R}^N)$. Then use a mollifier to uniformly approximate $u \in \mathcal{C}_c(\Omega, \mathbf{R}^d)$ by a sequence in $\mathcal{D}(\Omega, \mathbf{R}^N)$.

Exercise 6.36 Let $\Omega \subset \mathbf{R}^d$ be a domain and Ω' a bounded Lipschitz domain. Show that the total variation of $u = \chi_{\Omega'}$ is given by $\text{TV}(u) = \mathfrak{H}^{d-1}(\partial\Omega')$.

[Hint:] Use the definition of a Lipschitz domain to extend the field of inner normals $-\nu$ to a neighborhood Ω_0 of $\partial\Omega'$ such that $\|\nu\|_\infty \leq 1$ holds. Then use mollifiers and truncation to construct a sequence $-\nu_\varepsilon \in \mathcal{D}(\Omega, \mathbf{R}^d)$ with $\|\nu_\varepsilon\|_\infty \leq 1$, that converges on $\partial\Omega'$ pointwise \mathfrak{H}^{d-1} -almost everywhere to $-\nu$. Finally, show that this sequence is a maximizing sequence in the definition of TV .

Exercise 6.37 Let $\Omega, \Omega_1, \dots, \Omega_K \subset \mathbf{R}^d$ be bounded Lipschitz domains with $\overline{\Omega} = \bigcup_{k=1}^K \overline{\Omega_k}$ and Ω_k mutually disjoint. Moreover, let $u : \Omega \rightarrow \mathbf{R}$ be such that every $u^k = u|_{\Omega_k}$ can be extended to an element in $C^1(\overline{\Omega_k})$. Show that with $\Gamma_{l,k} = \overline{\Omega_l} \cap \overline{\Omega_k} \cap \Omega$, one has

$$\text{TV}(u) = \sum_{k=1}^K \int_{\Omega_k} |\nabla u_k| \, dx + \sum_{l < k} \int_{\Gamma_{l,k}} |u^l - u^k| \, d\mathfrak{H}^{d-1}.$$

[Hint:] For every $\varepsilon > 0$ choose a neighborhood $\Omega'_{l,k}$ of the part $\Gamma_{l,k}$ of the boundary with $|\Omega'_{l,k}| < \varepsilon$. Approximate $\text{sgn}(u^k - u^l)v$ there by smooth functions (similar to Exercise 6.36) and also approximate on $\Omega_k \setminus (\bigcup_{1 \leq l < k \leq K} \Omega'_{l,k})$ almost everywhere the negative sign $-\frac{\nabla u_k}{|\nabla u_k|}$. Patch these piecewise functions by smooth cutoff functions to construct a sequence (φ^n) in $\mathcal{D}(\Omega, \mathbf{R}^d)$ with $\|\varphi^n\|_\infty \leq 1$ that converges almost everywhere on Ω_k to $-\frac{\nabla u^k}{|\nabla u^k|}$ and \mathfrak{H}^{d-1} -almost everywhere on $\Gamma_{l,k}$ to $(u^k - u^l)v$.

Exercise 6.38 Let $\Omega \subset \mathbf{R}^d$ be a bounded Lipschitz domain, $q \in]1, \infty[$ with $q \leq d/(d-1)$, and $u^0 \in L^1(\Omega)$ as well as $\lambda > 0$.

1. Prove that there exists a minimizer of the L^1 -TV denoising problem

$$\min_{u \in L^q(\Omega)} \int_{\Omega} |u - u^0| \, dx + \lambda \text{TV}(u). \quad (6.101)$$

2. Show that if $h : \mathbf{R} \rightarrow \mathbf{R}$ is strictly increasing, continuously differentiable with $\|h'\|_\infty < \infty$, and u^* is a solution of (6.101) with data u^0 , then $h \circ u^*$ is a solution of (6.101) with data $h \circ u^0$.

[Hint:] Use Fubini's theorem and the result of Exercise 6.34.

Exercise 6.39 Let $\Omega \subset \mathbf{R}^d$ be a bounded Lipschitz domain, Ω' with $\overline{\Omega'} \subset\subset \overline{\Omega}$ a bounded Lipschitz subdomain, $q \in]1, \infty[$ with $q \leq d/(d-1)$ and $u^0 \in \text{BV}(\Omega \setminus \overline{\Omega'})$ with $L \leq u^0 \leq R$ almost everywhere in $\Omega \setminus \overline{\Omega'}$. Moreover, let $u^* \in \text{BV}(\Omega)$ be a solution of the TV-inpainting problem (6.69).

1. Prove that $L \leq u^* \leq R$ holds almost everywhere in Ω .
2. Moreover, show that for every strictly increasing and continuously differentiable $h : \mathbf{R} \rightarrow \mathbf{R}$ with $\|h'\|_\infty < \infty$, one has that $h \circ u^*$ is a solution of the TV-inpainting problem with $h \circ u^0 \in \text{BV}(\Omega \setminus \overline{\Omega'})$.

Exercise 6.40 Let $\Omega \subset \mathbf{R}^2$ be a domain and $u : \Omega \rightarrow \mathbf{R}$ twice continuously differentiable in a neighborhood of $(x_0, y_0) \in \Omega$ with $u(x_0, y_0) = 0$. Moreover, let the zero level-set $\{u = 0\}$ be locally parameterized by arclength by functions $x, y :]-\varepsilon, \varepsilon[\rightarrow \mathbf{R}$, i.e., $x(0) = x_0$, $y(0) = y_0$, $|x'|^2 + |y'|^2 = 1$, and $u(x(s), y(s)) = 0$.

Show that if

$$(\nabla u)(x(s), y(s)) \neq 0 \quad \text{as well as} \quad \operatorname{div}\left(\frac{(\nabla u)(x(s), y(s))}{|(\nabla u)(x(s), y(s))|}\right) = \kappa$$

for some $\kappa \in \mathbf{R}$ and all $s \in]-\varepsilon, \varepsilon[$, then there exists a $\varphi_0 \in \mathbf{R}$ such that (x, y) can be written as

$$\begin{cases} x(s) = x_0 + \sin(\kappa s + \varphi_0), \\ y(s) = y_0 + \cos(\kappa s + \varphi_0), \end{cases}$$

for all $s \in]-\varepsilon, \varepsilon[$. In particular, (x, y) parameterizes a piece of a line or circle with curvature κ .

Exercise 6.41 Let $t > 0$, $p \in]1, \infty[$, $\sigma > 0$, and s_0 be such that

$$s_0 \geq t \quad \text{and} \quad s_0 < \left(\frac{t}{\sigma(2-p)}\right)^{\frac{1}{p-1}} \quad \text{if } p < 2.$$

Show that the sequence (s_n) defined by the iteration

$$s_{n+1} = s_n + \frac{t - s_n - \sigma s_n^{p-1}}{1 + \sigma(p-1)s_n^{p-2}},$$

is well defined, fulfills $s_n > 0$ for all n and is decreasing. Moreover, it converges to the unique s , which fulfills the equation $s + \sigma s^{p-1} = t$.

Exercise 6.42 Implement the primal-dual method for variational denoising (Table 6.1).

Exercise 6.43 For $K \geq 1$ let the matrix $\kappa \in \mathbf{R}^{(2K+1) \times (2K+1)}$ represent a convolution kernel that is indexed by $-K \leq i, j \leq K$ and satisfies $\sum_{i=-K}^K \sum_{j=-K}^K \kappa_{i,j} = 1$. Moreover, for $N, M \geq 1$ let $A_h : \mathbf{R}^{(N+2K) \times (M+2K)} \rightarrow \mathbf{R}^{N \times M}$ be a discrete convolution operator,

$$(A_h u)_{i,j} = \sum_{k=-K}^K \sum_{l=-K}^K u_{(i+K-k), (j+K-k)Kk, l}.$$

1. Implement the primal-dual method from Table 6.2 for the solution of the variational deconvolution problem

$$\min_{u \in \mathbf{R}^{(N+2K) \times (M+2K)}} \frac{\|A_h u - U^0\|_q^q}{q} + \frac{\lambda \|\nabla_h u\|_p^p}{p}$$

for given data $U^0 \in \mathbf{R}^{N \times M}$ and parameter $\lambda > 0$.

2. Derive a method that takes additional constraints $\underline{U}^0 \leq u_{i,j} \leq \overline{U}^0$ for $1 \leq i \leq N + 2K, 1 \leq j \leq M + 2K$ with $\underline{U}^0 = \min_{i,j} U_{i,j}^0$ and $\overline{U}^0 = \max_{i,j} U_{i,j}^0$ into account.
3. Implement and test the method with additional constraints. Do you notice differences in the results of the two methods?

Exercise 6.44 Let $(u^*, w^*) \in \mathbf{R}^{N \times M} \times \mathbf{R}^{N \times M \times 2}$ be a saddle point of the Lagrange functional for the discrete inpainting problems

$$L(u, w) = (\nabla_h u, w) + I_{\{v|_{\Omega_h \setminus \Omega'_h} = U^0\}}(u) - \begin{cases} \frac{1}{p^*} \|w\|_{p^*}^{p^*} & \text{if } p > 1, \\ I_{\{\|w\|_\infty \leq 1\}}(w) & \text{if } p = 1, \end{cases}$$

from Example 6.145.

1. Prove the discrete maximum principle $\underline{U}^0 \leq u_{i,j}^* \leq \overline{U}^0$ for all $1 \leq i \leq N$ and $1 \leq j \leq M$, where \underline{U}^0 and \overline{U}^0 are the minimum and maximum values of $U_{i,j}^0$, respectively. Derive an a priori estimate $\|u^*\|_\infty \leq \|U^0\|_\infty$.
2. For $p > 1$ use the dual problem and Young's inequality for numbers to derive the following a priori estimate for w^* :

$$\|w^*\|_{p^*} \leq \begin{cases} \frac{2^{p-1}}{p-1} \|\nabla_h U^0\|_p^{p-1} & \text{if } p \geq 2, \\ \left(\frac{p-1}{2-p}\right)^p \|\nabla_h U^0\|_p^{p-1} & \text{if } p < 2. \end{cases}$$

3. Use the convergence proof of Theorem 6.141 to estimate the norm of the iterates (u^n, w^n) of the primal-dual method from Table 6.3.

Exercise 6.45 Implement the primal-dual inpainting method from Table 6.3.

Add-on: Use the results from Exercise 6.44, to derive a modified duality gap $\tilde{\mathcal{G}}$ according to Example 6.144. Prove that $\tilde{\mathcal{G}}(u^n, w^n) \rightarrow 0$ for the iterates (u^n, w^n) and modify your program such that it terminates if $\tilde{\mathcal{G}}$ falls below a certain threshold.

Exercise 6.46 Let $1 \leq p < \infty$, $N_1, N_2, M_1, M_2 \in \mathbf{N}$ positive, let the map $A_h : \mathbf{R}^{N_1 \times M_1} \rightarrow \mathbf{R}^{N_2 \times M_2}$ be linear and surjective, and let $A_h \mathbf{1} \neq 0$. Consider the minimization problem

$$\min_{u \in \mathbf{R}^{N_1 \times M_1}} \frac{\|\nabla_h u\|_p^p}{p} + I_{\{A_h v = U^0\}}(u)$$

for some $U^0 \in \mathbf{R}^{N_2 \times M_2}$. Define $X = \mathbf{R}^{N_1 \times N_2}$, $Z = \mathbf{R}^{N_2 \times M_2} \times \mathbf{R}^{N_1 \times M_1 \times 2}$ and

$$F_1 : X \rightarrow \mathbf{R}, \quad F_1(u) = 0, \quad F_2 : Z \rightarrow \mathbf{R}, \quad F_2(v, w) = I_{\{0\}}(v - U^0) + \frac{\|w\|_p^p}{p}.$$

Prove that the minimization problem is equivalent to the saddle point problem for

$$L(u, v, w) = (A_h u, v) + (\nabla_h u, w) + F_1(u) - F_2^*(v, w).$$

Derive an alternative method for the minimization of the discrete Sobolev and total variation semi-norm, respectively, that in contrast to the method from Table 6.4, does not use the projection onto $\{A_h u = U^0\}$ and hence does not need to solve a linear system.

Exercise 6.47 Let $N, M \in \mathbf{N}$ be positive, $U^0 \in \mathbf{R}^{N \times M}$ and $K \in \mathbf{N}$ with $K \geq 1$. For $1 \leq p < \infty$ consider the interpolation problem

$$\min_{u \in \mathbf{R}^{KN \times KM}} \frac{\|\nabla_h u\|_p^p}{p} + I_{\{A_h u = U^0\}}(u)$$

with

$$(A_h u)_{i,j} = \frac{1}{K^2} \sum_{k=1}^K \sum_{l=1}^K u_{((i-1)K+k), ((j-1)K+l)}.$$

Use the algorithm from Table 6.4 to implement a numerical method.

Add-on: Implement and test the alternative method from Exercise 6.46. How do the methods differ in practice?

References

1. R. Acar, C.R. Vogel, Analysis of bounded variation penalty methods for ill-posed problems. *Inverse Probl.* **10**(6), 1217–1229 (1994)
2. R.A. Adams, J.J.F. Fournier, *Sobolev Spaces*. Pure and Applied Mathematics, vol. 140, 2nd edn. (Elsevier, Amsterdam, 2003)
3. L. Alvarez, F. Guichard, P.-L. Lions, J.-M. Morel, Axioms and fundamental equations in image processing. *Arch. Ration. Mech. Anal.* **123**, 199–257 (1993)
4. H. Amann, Time-delayed Perona-Malik type problems. *Acta Math. Univ. Comenian. N. Ser.* **76**(1), 15–38 (2007)
5. L. Ambrosio, N. Fusco, D. Pallara, *Functions of Bounded Variation and Free Discontinuity Problems*. Oxford Mathematical Monographs (Oxford University Press, Oxford, 2000)
6. L. Ambrosio, N. Fusco, J.E. Hutchinson, Higher integrability of the gradient and dimension of the singular set for minimisers of the Mumford-Shah functional. *Calc. Var. Partial Differ. Equ.* **16**(2), 187–215 (2003)
7. K.J. Arrow, L. Hurwicz, H. Uzawa, *Studies in Linear and Non-linear Programming*. Stanford Mathematical Studies in the Social Sciences, 1st edn. (Stanford University Press, Palo Alto, 1958)
8. G. Aubert, P. Kornprobst, *Mathematical Problems in Image Processing* (Springer, New York, 2002)
9. G. Aubert, L. Blanc-Féraud, R. March, An approximation of the Mumford-Shah energy by a family of discrete edge-preserving functionals. *Nonlinear Anal. Theory Methods Appl. Int. Multidiscip. J. Ser. A Theory Methods* **64**(9), 1908–1930 (2006)
10. J.-F. Aujol, A. Chambolle, Dual norms and image decomposition models. *Int. J. Comput. Vis.* **63**(1), 85–104 (2005)
11. V. Aurich, J. Weule, Non-linear gaussian filters performing edge preserving diffusion, in *Proceedings 17. DAGM-Symposium, Bielefeld* (Springer, Heidelberg, 1995), pp. 538–545
12. C. Bär, *Elementary Differential Geometry* (Cambridge University Press, Cambridge, 2010). Translated from the 2001 German original by P. Meerkamp
13. M. Bertalmio, G. Sapiro, V. Caselles, C. Ballester, Image inpainting, in *Proceedings of SIGGRAPH 2000*, New Orleans (2000), pp. 417–424
14. M. Bertero, P. Boccacci, *Introduction to Inverse Problems in Imaging* (Institute of Physics, London, 1998)
15. F. Bornemann, T. März, Fast image inpainting based on coherence transport. *J. Math. Imaging Vis.* **28**(3), 259–278 (2007)

16. J.M. Borwein, A.S. Lewis, *Convex Analysis and Nonlinear Optimization: Theory and Examples*. CMS Books in Mathematics, vol. 3, 2nd edn. (Springer, New York, 2006)
17. A. Borzì, K. Ito, K. Kunisch, Optimal control formulation for determining optical flow. *SIAM J. Sci. Comput.* **24**, 818–847 (2002)
18. B. Bourdin, A. Chambolle, Implementation of an adaptive finite-element approximation of the Mumford-Shah functional. *Numer. Math.* **85**(4), 609–646 (2000)
19. K. Bredies, K. Kunisch, T. Pock, Total generalized variation. *SIAM J. Imaging Sci.* **3**(3), 492–526 (2010)
20. M. Breuß, J. Weickert, A shock-capturing algorithm for the differential equations of dilation and erosion. *J. Math. Imaging Vis.* **25**(2), 187–201 (2006)
21. H. Brézis, *Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert*. North-Holland Mathematics Studies, vol. 5. Notas de Matemática (50) (North-Holland, Amsterdam; Elsevier, New York, 1973).
22. H. Brézis, *Analyse fonctionnelle - Théorie et applications*. Collection Mathématiques Appliquées pour la Maîtrise (Masson, Paris, 1983)
23. T. Brox, O. Kleinschmidt, D. Cremers, Efficient nonlocal means for denoising of textural patterns. *IEEE Trans. Image Process.* **17**(7), 1083–1092 (2008)
24. A. Bruhn, J. Weickert, C. Schnörr, Lucas/Kanade meets Horn/Schunck: combining local and global optical flow methods. *Int. J. Comput. Vis.* **61**(3), 211–231 (2005)
25. A. Buades, J.-M. Coll, B. Morel, A review of image denoising algorithms, with a new one. *Multiscale Model. Simul.* **4**(2), 490–530 (2005)
26. M. Burger, O. Scherzer, Regularization methods for blind deconvolution and blind source separation problems. *Math. Control Signals Syst.* **14**, 358–383 (2001)
27. E.J. Candès, D.L. Donoho, New tight frames of curvelets and optimal representations of objects with piecewise c^2 singularities. *Commun. Pure Appl. Math.* **57**(2), 219–266 (2004)
28. E. Candès, L. Demanet, D. Donoho, L. Ying, Fast discrete curvelet transforms. *Multiscale Model. Simul.* **5**(3), 861–899 (2006)
29. J. Canny, A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **8**(6), 679–698 (1986)
30. F. Catté, P.-L. Lions, J.-M. Morel, T. Coll, Image selective smoothing and edge detection by nonlinear diffusion. *SIAM J. Numer. Anal.* **29**(1), 182–193 (1992)
31. A. Chambolle, P.-L. Lions, Image recovery via Total Variation minimization and related problems. *Numer. Math.* **76**, 167–188 (1997)
32. A. Chambolle, B.J. Lucier, Interpreting translation-invariant wavelet shrinkage as a new image smoothing scale space. *IEEE Trans. Image Process.* **10**, 993–1000 (2001)
33. A. Chambolle, G.D. Maso, Discrete approximation of the Mumford-Shah functional in dimension two. *Math. Model. Numer. Anal.* **33**(4), 651–672 (1999)
34. A. Chambolle, T. Pock, A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vis.* **40**(1), 120–145 (2011)
35. A. Chambolle, R.A. DeVore, N. Lee, B.J. Lucier, Nonlinear wavelet image processing: variational problems, compression and noise removal through wavelet shrinkage. *IEEE Trans. Image Process.* **7**, 319–335 (1998)
36. T.F. Chan, S. Esedoglu, Aspects of total variation regularized L^1 function approximation. *SIAM J. Appl. Math.* **65**, 1817 (2005)
37. T.F. Chan, J. Shen, *Image Processing And Analysis: Variational, PDE, Wavelet, and Stochastic Methods* (Society for Industrial and Applied Mathematics, Philadelphia, 2005)
38. T.F. Chan, L.A. Vese, Active contours without edges. *IEEE Trans. Image Process.* **10**(2), 266–277 (2001)
39. T.F. Chan, C. Wong, Total variation blind deconvolution. *IEEE Trans. Image Process.* **7**, 370–375 (1998)
40. T.F. Chan, A. Marquina, P. Mulet, High-order total variation-based image restoration. *SIAM J. Sci. Comput.* **22**(2), 503–516 (2000)

41. T.F. Chan, S. Esedoglu, F.E. Park, A fourth order dual method for staircase reduction in texture extraction and image restoration problems. Technical report, UCLA CAM Report 05-28 (2005)
42. K. Chen, D.A. Lorenz, Image sequence interpolation using optimal control. *J. Math. Imaging Vis.* **41**(3), 222–238 (2011)
43. K. Chen, D.A. Lorenz, Image sequence interpolation based on optical flow, segmentation, and optimal control. *IEEE Trans. Image Process.* **21**(3), 1020–1030 (2012)
44. Y. Chen, K. Zhang, Young measure solutions of the two-dimensional Perona-Malik equation in image processing. *Commun. Pure Appl. Anal.* **5**(3), 615–635 (2006)
45. U. Clarenz, U. Diewald, M. Rumpf, Processing textured surfaces via anisotropic geometric diffusion. *IEEE Trans. Image Process.* **13**(2), 248–261 (2004)
46. P.L. Combettes, V.R. Wajs, Signal recovery by proximal forward-backward splitting. *Multiscale Model. Simul.* **4**(4), 1168–1200 (2005)
47. R. Courant, K.O. Friedrichs, H. Lewy, Über die partiellen Differenzengleichungen der mathematischen Physik. *Math. Ann.* **100**(1), 32–74 (1928)
48. I. Daubechies, Orthonormal bases of compactly supported wavelets. *Commun. Pure Appl. Math.* **41**(7), 909–996 (1988)
49. G. David, *Singular Sets of Minimizers for the Mumford-Shah Functional*. Progress in Mathematics, vol. 233 (Birkhäuser, Basel, 2005)
50. J. Diestel, J.J. Uhl Jr., *Vector Measures*. Mathematical Surveys and Monographs, vol. 15 (American Mathematical Society, Providence, 1977)
51. J. Dieudonné, *Foundations of Modern Analysis*. Pure and Applied Mathematics, vol. 10-I (Academic, New York, 1969). Enlarged and corrected printing
52. U. Diewald, T. Preußer, M. Rumpf, Anisotropic diffusion in vector field visualization on Euclidean domains and surfaces. *IEEE Trans. Visual. Comput. Graph.* **6**(2), 139–149 (2000)
53. N. Dinculeanu, *Vector Measures*. Hochschulbücher für Mathematik, vol. 64 (WEB Deutscher Verlag der Wissenschaften, Berlin, 1967)
54. D.L. Donoho, Denoising via soft thresholding. *IEEE Trans. Inf. Theory* **41**(3), 613–627 (1995)
55. V. Duval, J.-F. Aujol, Y. Gousseau, The TVL1 model: a geometric point of view. *Multiscale Model. Simul.* **8**(1), 154–189 (2009)
56. J. Eckstein, D.P. Bertsekas, On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Math. Program.* **55**, 293–318 (1992)
57. I. Ekeland, R. Temam, *Convex Analysis and Variational Problems*. Studies in Mathematics and Its Applications, vol. 1 (North-Holland, Amsterdam, 1976)
58. H.W. Engl, M. Hanke, A. Neubauer, *Regularization of Inverse Problems*. Mathematics and Its Applications, vol. 375, 1st edn. (Kluwer Academic, Dordrecht, 1996)
59. S. Esedoğlu, Stability properties of the Perona-Malik scheme. *SIAM J. Numer. Anal.* **44**(3), 1297–1313 (2006)
60. L.C. Evans, A new proof of local $C^{1,\alpha}$ regularity for solutions of certain degenerate elliptic P.D.E. *J. Differ. Equ.* **45**, 356–373 (1982)
61. L.C. Evans, R.F. Gariepy, *Measure Theory and Fine Properties of Functions* (CRC Press, Boca Raton, 1992)
62. H. Federer, *Geometric Measure Theory* (Springer, Berlin, 1969)
63. B. Fischer, J. Modersitzki, Ill-posed medicine — an introduction to image registration. *Inverse Probl.* **24**(3), 034008 (2008)
64. I. Galić, J. Weickert, M. Welk, A. Bruhn, A. Belyaev, H.-P. Seidel, Image compression with anisotropic diffusion. *J. Math. Imaging Vis.* **31**, 255–269 (2008)
65. E. Giusti, *Minimal Surfaces and Functions of Bounded Variation*. Monographs in Mathematics, vol. 80 (Birkhäuser, Boston, 1984)
66. G.H. Golub, C.F. Van Loan, *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences, 4th edn. (Johns Hopkins University Press, Baltimore, 2013)
67. R.C. Gonzalez, P.A. Wintz, *Digital Image Processing* (Addison-Wesley, Reading, 1977)
68. K. Gröchenig, *Foundations of Time-Frequency Analysis* (Birkhäuser, Boston, 2001)

69. F. Guichard, J.-M. Morel, Partial differential equations and image iterative filtering, in *The State of the Art in Numerical Analysis*, ed. by I.S. Duff, G.A. Watson. IMA Conference Series (New Series), vol. 63 (Oxford University Press, Oxford, 1997)
70. A. Haddad, Texture separation $BV - G$ and $BV - L^1$ models. Multiscale Model. Simul. **6**(1), 273–286 (electronic) (2007)
71. P.R. Halmos, *Measure Theory* (D. Van Nostrand, New York, 1950)
72. M. Hanke-Bourgeois, *Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens*, 3rd edn. (Vieweg+Teubner, Wiesbaden, 2009)
73. L. He, S.J. Osher, Solving the Chan-Vese model by a multiphase level set algorithm based on the topological derivative, in *Scale Space and Variational Methods in Computer Vision*, ed. by F. Sgallari, A. Murli, N. Paragios. Lecture Notes in Computer Science, vol. 4485 (Springer, Berlin, 2010), pp. 777–788
74. W. Hinterberger, O. Scherzer, Variational methods on the space of functions of bounded Hessian for convexification and denoising. Computing **76**, 109–133 (2006)
75. W. Hinterberger, O. Scherzer, C. Schnörr, J. Weickert, Analysis of optical flow models in the framework of the calculus of variations. Numer. Funct. Anal. Optim. **23**(1), 69–89 (2002)
76. M. Hintermüller, W. Ring, An inexact Newton-CG-type active contour approach for the minimization of the Mumford-Shah functional. J. Math. Imaging Vis. **20**(1–2), 19–42 (2004). Special issue on mathematics and image analysis
77. M. Holler, Theory and numerics for variational imaging — artifact-free JPEG decompression and DCT based zooming. Master’s thesis, Universität Graz (2010)
78. B.K.P. Horn, B.G. Schunck, Determining optical flow. Artif. Intell. **17**, 185–203 (1981)
79. G. Huisken, Flow by mean curvature of convex surfaces into spheres. J. Differ. Geom. **20**(1), 237–266 (1984)
80. J. Jost, *Partial Differential Equations*. Graduate Texts in Mathematics, vol. 214 (Springer, New York, 2002). Translated and revised from the 1998 German original by the author
81. L.A. Justen, R. Ramlau, A non-iterative regularization approach to blind deconvolution. Inverse Probl. **22**, 771–800 (2006)
82. S. Kakutani, Concrete representation of abstract (m)-spaces (a characterization of the space of continuous functions). Ann. Math. Second Ser. **42**(4), 994–1024 (1941)
83. B. Kawohl, N. Kutev, Maximum and comparison principle for one-dimensional anisotropic diffusion. Math. Ann. **311**, 107–123 (1998)
84. S.L. Keeling, W. Ring, Medical image registration and interpolation by optical flow with maximal rigidity. J. Math. Imaging Vis. **23**, 47–65 (2005)
85. S.L. Keeling, R. Stollberger, Nonlinear anisotropic diffusion filtering for multiscale edge enhancement. Inverse Probl. **18**(1), 175–190 (2002)
86. S. Kichenassamy, The Perona-Malik paradox. SIAM J. Appl. Math. **57**, 1328–1342 (1997)
87. S. Kindermann, S.J. Osher, J. Xu, Denoising by BV-duality. J. Sci. Comput. **28**(2–3), 411–444 (2006)
88. J.J. Koenderink, The structure of images. Biol. Cybern. **50**(5), 363–370 (1984)
89. G.M. Korpelevič, An extragradient method for finding saddle points and for other problems. Ékonomika i Matematicheskie Metody **12**(4), 747–756 (1976)
90. G. Kutyniok, D. Labate, Construction of regular and irregular shearlets. J. Wavelet Theory Appl. **1**, 1–10 (2007)
91. E.H. Lieb, M. Loss, *Analysis*. Graduate Studies in Mathematics, vol. 14, 2nd edn. (American Mathematical Society, Providence, 2001)
92. L.H. Lieu, L. Vese, Image restoration and decomposition via bounded Total Variation and negative Hilbert-Sobolev spaces. Appl. Math. Optim. **58**, 167–193 (2008)
93. P.-L. Lions, B. Mercier, Splitting algorithms for the sum of two nonlinear operators. SIAM J. Numer. Anal. **16**(6), 964–979 (1979)
94. A.K. Louis, P. Maass, A. Rieder, *Wavelets: Theory and Applications* (Wiley, Chichester, 1997)

95. M. Lysaker, A. Lundervold, X.-C. Tai, Noise removal using fourth-order partial differential equation with applications to medical magnetic resonance images in space and time. *IEEE Trans. Image Process.* **12**(12), 1579–1590 (2003)
96. J. Ma, G. Plonka, The curvelet transform: a review of recent applications. *IEEE Signal Process. Mag.* **27**(2), 118–133 (2010)
97. S. Mallat, *A Wavelet Tour of Signal Processing - The Sparse Way, with Contributions from Gabriel Peyré*, 3rd edn. (Elsevier/Academic, Amsterdam, 2009)
98. D. Marr, E. Hildreth, Theory of edge detection. *Proc. R. Soc. Lond.* **207**, 187–217 (1980)
99. Y. Meyer, *Oscillating Patterns in Image Processing and Nonlinear Evolution Equations*. University Lecture Series, vol. 22 (American Mathematical Society, Providence, 2001). The fifteenth Dean Jacqueline B. Lewis memorial lectures
100. J. Modersitzki, *FAIR: Flexible Algorithms for Image Registration*. Fundamentals of Algorithms, vol. 6 (Society for Industrial and Applied Mathematics, Philadelphia, 2009)
101. D. Mumford, J. Shah, Optimal approximations by piecewise smooth functions and variational problems. *Commun. Pure Appl. Math.* **42**(5), 577–685 (1989)
102. H.J. Muthsam, *Lineare Algebra und ihre Anwendungen*, 1st edn. (Spektrum Akademischer Verlag, Heidelberg, 2006)
103. F. Natterer, F. Wuebbeling, *Mathematical Methods in Image Reconstruction* (Society for Industrial and Applied Mathematics, Philadelphia, 2001)
104. J. Nečas, *Les méthodes directes en théorie des équations elliptiques* (Masson, Paris, 1967)
105. S.J. Osher, R. Fedkiw, *Level Set Methods and Dynamic Implicit Surfaces*. Applied Mathematical Sciences, vol. 153 (Springer, Berlin, 2003)
106. S.J. Osher, J.A. Sethian, Fronts propagating with curvature-dependent speed: algorithms based on Hamilton-Jacobi formulations. *J. Comput. Phys.* **79**, 12–49 (1988)
107. S.J. Osher, A. Sole, L. Vese, Image decomposition and restoration using Total Variation minimization and the h^{-1} norm. *Multiscale Model. Simul.* **1**(3), 349–370 (2003)
108. S. Paris, P. Kornprobst, J. Tumblin, F. Durand, Bilateral filtering: theory and applications. *Found. Trends Comput. Graph. Vis.* **4**(1), 1–73 (2009)
109. W.B. Pennebaker, J.L. Mitchell, *JPEG: Still Image Data Compression Standard* (Springer, New York, 1992)
110. P. Perona, J. Malik, Scale-space and edge detection using anisotropic diffusion. *IEEE Trans. Pattern Anal. Mach. Intell.* **12**(7), 629–639 (1990)
111. G. Plonka, G. Steidl, A multiscale wavelet-inspired scheme for nonlinear diffusion. *Int. J. Wavelets Multiresolut. Inf. Process.* **4**(1), 1–21 (2006)
112. T. Pock, D. Cremers, H. Bischof, A. Chambolle, An algorithm for minimizing the Mumford-Shah functional, in *2009 IEEE 12th International Conference on Computer Vision* (2009), pp. 1133–1140
113. L.D. Popov, A modification of the Arrow-Hurwicz method for search of saddle points. *Math. Notes* **28**, 845–848 (1980)
114. W.K. Pratt, *Digital Image Processing* (Wiley, New York, 1978)
115. T. Preußer, M. Rumpf, An adaptive finite element method for large scale image processing. *J. Vis. Commun. Image Represent.* **11**(2), 183–195 (2000)
116. J.M.S. Prewitt, Object enhancement and extraction, in *Picture Processing and Psychopictorics*, ed. by B.S. Lipkin, A. Rosenfeld (Academic, New York, 1970)
117. T.W. Ridler, S. Calvard, Picture thresholding using an iterative selection method. *IEEE Trans. Syst. Man Cybern.* **8**(8), 630–632 (1978)
118. R.T. Rockafellar, *Convex Analysis*. Princeton Mathematical Series (Princeton University Press, Princeton, 1970)
119. A. Rosenfeld, A.C. Kak, *Digital Picture Processing* (Academic, New York, 1976)
120. E. Rouy, A. Tourin, A viscosity solutions approach to shape-from-shading. *SIAM J. Numer. Anal.* **29**(3), 867–884 (1992)
121. L.I. Rudin, S.J. Osher, E. Fatemi, Nonlinear total variation based noise removal algorithms. *Phys. D Nonlinear Phenom.* **60**(1–4), 259–268 (1992)

122. W. Rudin, *Functional Analysis*. McGraw-Hill Series in Higher Mathematics (McGraw-Hill, New York, 1973)
123. W. Rudin, *Principles of Mathematical Analysis*. International Series in Pure and Applied Mathematics, 3rd edn. (McGraw-Hill, New York, 1976)
124. G. Sapiro, Color snakes. *Comput. Vis. Image Underst.* **68**(2), 247–253 (1997)
125. O. Scherzer, Denoising with higher order derivatives of bounded variation and an application to parameter estimation. *Computing* **60**(1), 1–27 (1998)
126. O. Scherzer, M. Grasmair, H. Grossauer, M. Haltmeier, F. Lenzen, *Variational Methods in Imaging* (Springer, New York, 2009)
127. L. Schwartz, *Théorie des distributions*, vol. 1, 3rd edn. (Hermann, Paris, 1966)
128. J.A. Sethian, *Level Set Methods and Fast Marching Methods*, 2nd edn. (Cambridge University Press, Cambridge, 1999)
129. S. Setzer, G. Steidl, Variational methods with higher order derivatives in image processing, in *Approximation XII*, ed. by M. Neamtu, L.L. Schumaker (Nashboro Press, Brentwood, 2008), pp. 360–386
130. R.E. Showalter, *Monotone Operators in Banach Space and Nonlinear Partial Differential Equations*. Mathematical Surveys and Monographs, vol. 49 (American Mathematical Society, Providence, 1997)
131. J. Simon, Régularité de la solution d'une équation non linéaire dans \mathbf{R}^N , in *Journées d'Analyse Non Linéaire*, ed. by P. Bénilan, J. Robert. Lecture Notes in Mathematics, vol. 665 (Springer, Heidelberg, 1978)
132. S.M. Smith, J.M. Brady, SUSAN—A new approach to low level image processing. *Int. J. Comput. Vis.* **23**(1), 45–78 (1997)
133. I.E. Sobel, Camera models and machine perception. PhD thesis, Stanford University, Palo Alto (1970)
134. P. Soille, *Morphological Image Analysis - Principles and Applications* (Springer, Berlin, 1999)
135. J. Stoer, R. Bulirsch, *Introduction to Numerical Analysis*. Texts in Applied Mathematics, vol. 12, 3rd edn. (Springer, New York, 2002). Translated from the German by R. Bartels, W. Gautschi and C. Witzgall
136. C. Tomasi, R. Manduchi, Bilateral filtering for gray and color images, in *International Conference of Computer Vision* (1998), pp. 839–846
137. F. Tröltzsch, *Optimal Control of Partial Differential Equations*. Graduate Studies in Mathematics, vol. 112 (American Mathematical Society, Providence, 2010). Theory, methods and applications, Translated from the 2005 German original by Jürgen Sprekels
138. G.N. Watson, *A Treatise on the Theory of Bessel Functions*. Cambridge Mathematical Library, 2nd edn. (Cambridge University Press, Cambridge, 1995). Revised edition
139. J.B. Weaver, Y. Xu, D.M. Healy Jr., L.D. Cromwell, Filtering noise from images with wavelet transforms. *Magn. Reson. Med.* **21**, 288–295 (1991)
140. J. Weickert, *Anisotropic Diffusion in Image Processing*. European Consortium for Mathematics in Industry (B. G. Teubner, Stuttgart, 1998)
141. J. Weidmann, *Linear Operators in Hilbert Spaces*. Graduate Texts in Mathematics, vol. 68. (Springer, New York, 1980). Translated from the German by Joseph Szűcs
142. M. Welk, J. Weickert, G. Steidl, A four-pixel scheme for singular differential equations, in *Scale-Space and PDE Methods in Computer Vision*, ed. by R. Kimmel, N. Sochen, J. Weickert. Lecture Notes in Computer Science, vol. 3459 (Springer, Berlin, 2005), pp. 610–621
143. M. Welk, J. Weickert, F. Becker, C. Schnörr, C. Feddern, B. Burgeth, Median and related local filters for tensor-valued images. *Signal Process.* **87**(2), 291–308 (2007)
144. A.P. Witkin, Scale-space filtering, in *Proceedings of the International Joint Conference on Artificial Intelligence* (1983), pp. 1019–1021
145. L.P. Yaroslavsky, *Digital Picture Processing, An Introduction* (Springer, Berlin, 1985)
146. J. Yeh, *Real Analysis*, 3rd edn. (World Scientific, Hackensack, 2014). Theory of measure and integration

147. K. Yosida, *Functional Analysis*. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 123, 6th edn. (Springer, Berlin, 1980)
148. E. Zeidler, *Nonlinear Functional Analysis and Its Applications. II/A - Linear Monotone Operators* (Springer, New York, 1990)
149. W.P. Ziemer, *Weakly Differentiable Functions* (Springer, New York, 1989)

Picture Credits

All figures not listed in these picture credits only contain our own pictures. Only the first figure that features a certain image is listed.

Fig. 1.1	Matthew Mendoza@Flickr http://www.flickr.com/photos/mattmendoza/2421196777/ (License: http://creativecommons.org/licenses/by-sa/2.0/legalcode)PerkinElmer (http://www.cellularimaging.com/assays/receptor_activation),CNRS/Université de St-Etienne (France), Labor für Mikrozerspanung (Universität Bremen)	3
Fig. 1.5	Last Hero@Flickr http://www.flickr.com/photos/uwe_schubert/4594327195/ (License: http://creativecommons.org/licenses/by-sa/2.0/legalcode), Kai Schreiber@Flickr http://www.flickr.com/photos/genista/1249056653/ (License: http://creativecommons.org/licenses/by-sa/2.0/legalcode), http://grin.hq.nasa.gov/ABSTRACTS/GPN-2002-000064.html	8
Fig. 3.14	huangjiahui@Flickr http://www.flickr.com/photos/huangjiahui/3128463578/ (license: http://creativecommons.org/licenses/by-sa/2.0/legalcode)	96
Fig. 5.17	Mike Baird@Flickr http://www.flickr.com/photos/mikebaird/4533794674/ (License: http://creativecommons.org/licenses/by/2.0/legalcode)	223
Fig. 5.19	Benson Kua@Flickr http://www.flickr.com/photos/bensonkua/3301838191/ (License: http://creativecommons.org/licenses/by-sa/2.0/legalcode)	227

- Fig. 6.4 Grzegorz Łobiński@Flickr
<http://www.flickr.com/photos/gregloby/3073384793/>
(License: <http://creativecommons.org/licenses/by/2.0/legalcode>),
damo1977@Flickr <http://www.flickr.com/photos/damo1977/3944418313>
(License: <http://creativecommons.org/licenses/by/2.0/legalcode>) .. 261
- Fig. 6.11 John Lambert Pearson@Flickr <http://www.flickr.com/photos/orphanjones/419121524/> (License: <http://creativecommons.org/licenses/by/2.0/legalcode>) 336
- Fig. 6.13 Paul Mannix@Flickr <http://www.flickr.com/photos/paulmannix/552264573/> (License: <http://creativecommons.org/licenses/by/2.0/legalcode>) 339
- Fig. 6.14 jphilipg@Flickr <http://www.flickr.com/photos/15708236@N07/2754478731/> (License: <http://creativecommons.org/licenses/by/2.0/legalcode>) 345
- Fig. 6.26 jphilipg@Flickr <http://www.flickr.com/photos/15708236@N07/3317203380/> (License: <http://creativecommons.org/licenses/by/2.0/legalcode>) 388
- Fig. 6.27 Jerry Ferguson@Flickr <http://www.flickr.com/photos/fergusonphotography/3056953388/> (License: <http://creativecommons.org/licenses/by/2.0/legalcode>) 389
- Fig. 6.29 D. Sharon Pruitt@Flickr <http://www.flickr.com/photos/pinksherbet/2242046686/> (License: <http://creativecommons.org/licenses/by/2.0/legalcode>) 390
- Fig. 6.30 John Morgan@Flickr <http://www.flickr.com/photos/aidanmorgan/2574179254/> (License: <http://creativecommons.org/licenses/by/2.0/legalcode>) 391

Notation

Abbreviations

$\mathcal{B}(\mathbf{R}^d)$	space of bounded functions, page 89
$\mathfrak{B}(\Omega)$	Borel algebra over Ω , page 32
$\mathcal{BC}(\mathbf{R}^d)$	space of bounded and continuous functions, page 173
$\mathcal{BUC}(\mathbf{R}^d)$	space of bounded and uniformly continuous functions, page 187
$B_r(x)$	ball around x with radius r , page 17
$\mathcal{C}(\overline{U}, Y)$	space of bounded and uniformly continuous mapping on U with values in Y , page 19
$\mathcal{C}(U, Y)$	space of continuous mappings on U with values in Y , page 19
\mathbf{C}	set of complex numbers, page 16
$C_{\alpha, \beta}(u)$	semi-norms on the Schwartz space, page 112
$\mathcal{C}_c(\Omega, X)$	space of continuous function with compact support, page 42
$\mathcal{C}_0(\Omega, X)$	closure of $\mathcal{C}_c(\Omega, X)$ in $\mathcal{C}(\Omega, X)$, page 44
$\mathcal{C}^k(\Omega)$	space of k -times continuously differentiable functions, page 49
$\mathcal{C}^\infty(\Omega)$	space of infinitely differentiable functions, page 49
$\mathcal{C}_b^\infty(\mathbf{R}^d)$	space of infinitely differentiable functions with bounded derivatives, page 173
c_ψ	constant in the admissibility condition for a wavelet ψ , page 147
$DF(x)$	Fréchet derivative of F at the point x , page 20
$D^k F$	k th derivative of F , page 21
$\mathcal{D}(\Omega)$	space of test functions, page 49
$\mathcal{D}(\Omega)^*$	space of distributions, page 50
DAu	linear coordinate transformation of $u : \mathbf{R}^d \rightarrow \mathbf{K}$ using $A \in \mathbf{R}^{d \times d}$, i.e., $DAu(x) = u(Ax)$, page 56
$DCT(u)$	discrete cosine transform of u , page 140
$\mathcal{D}_{\text{div}}^m$	space of vector fields with m th weak divergence and vanishing trace on the boundary, page 329

$\mathcal{D}_{\text{div}, \infty}$	space of L^∞ vector fields with weak divergence and vanishing normal trace on the boundary, page 366
$\text{div } F$	divergence of F , page 22
$\text{div}^m F$	m th divergence of F , page 328
$\text{dom}(F)$	domain of a mapping $F : X \supset \text{dom}(F) \rightarrow Y$, page 18
$\text{dom } F$	effective domain of definition of a functional $F : X \rightarrow \mathbf{R}_\infty$, page 263
$\text{epi } F$	epigraph of a functional $F : X \rightarrow \mathbf{R}_\infty$, page 263
\mathcal{F}	Fourier transform, page 110
$\mathcal{G}_g u$	windowed Fourier transform of u w.r.t. the window g , page 142
$\text{graph}(F)$	graph of a mapping $F : X \rightarrow Y$, page 18
G_σ	Gaussian function with variance σ , page 75
\mathfrak{H}^k	k -dimensional Hausdorff measure, page 34
$H^{m,p}(\Omega)$	Sobolev space, page 51
$H_0^{m,p}(\Omega)$	closure of $\mathcal{D}(\Omega)$ in $H^{m,p}(\Omega)$, page 51
$H^m(\Omega)$	abbreviation for $H^{m,2}(\Omega)$, page 52
$H^s(\mathbf{R}^d)$	fractional Sobolev space, page 124
H_u	histogram of u , page 63
I_K	indicator functional of the set K , page 273
i	imaginary unit, page 16
id	identity map, page 20
$\text{IDCT}(u)$	inverse discrete cosine transform of u , page 141
$\text{Im } z$	imaginary part of a complex number $z \in \mathbf{C}$, page 16
$\text{int}(U)$	interior of the set U , page 17
$J_0(\nabla u_\sigma)$	structure tensor of u and noise level σ , page 222
$J_\rho(\nabla u_\sigma)$	structure tensor to u , noise level σ , and spatial scale ρ , page 224
\mathbf{K}	set of real or complex numbers, page 16
$\ker(F)$	kernel of mapping F , page 20
$L_{\text{loc}}^1(\Omega)$	space of locally integrable functions, page 49
\mathfrak{L}^d	d -dimensional Lebesgue measure, page 34
$\mathcal{L}(X, Y)$	space of linear continuous mappings from X to Y , page 19
$\mathcal{L}^k(X, Y)$	space of k -linear continuous mappings from X to Y , page 21
$L_\mu^p(\Omega, X)$	space of p -integrable functions w.r.t. the measure μ , page 38
$L^p(\Omega)$	standard Lebesgue space, page 39
$L_\psi u$	continuous wavelet transform of u w.r.t. the wavelet ψ , page 145
$\mathfrak{M}(\Omega, X)$	space of vector-valued finite Radon measures on Ω , page 43
\mathcal{M}_n	sequence of operators for the \mathcal{C}^∞ -approximation on bounded Lipschitz domains, page 317
$\text{MSE}(u, v)$	mean squared error of the images u and v , page 61
$M_y u$	modulation of $u : \mathbf{R}^d \rightarrow \mathbf{K}$ with $y \in \mathbf{R}^d$, i.e., $M_y u(x) = \exp(i x \cdot y) u(x)$, page 111
P_m	projection onto the complement of Π^m , page 322
$\text{PSNR}(u, v)$	Peak-Signal-to-Noise-Ratio of images u and v , page 61
Q_m	projection onto Π^m , page 322

Q_p	projection matrix onto the subspace perpendicular to $p \in \mathbf{R}^d$, i.e., $Q_p = (\text{id} - \frac{p \otimes p}{ p ^2})$, page 200
\mathbf{R}	set of real numbers, page 16
\mathbf{R}_∞	extended real numbers, page 263
$\operatorname{Re} z$	real part of a complex number $z \in \mathbf{C}$, page 16
$\operatorname{rg}(F)$	range of mapping F , page 20
$\mathcal{S}(\mathbf{R}^d)$	Schwartz space of rapidly decreasing functions, page 112
$\mathcal{S}(\mathbf{R}^d)^*$	space of tempered distributions, page 120
$S^{d \times d}$	space of symmetric $d \times d$ matrices, page 189
sinc	sinus cardinalis, page 58
$\operatorname{spt}(F)$	set of affine linear supporting functionals for F , page 304
$\operatorname{supp} f$	support of f , page 42
$T_u^d \sigma$	trace of σ w.r.t. ∇u of a BV-function u , page 373
T_f	distribution induced by the function f , page 50
$T_u^v \sigma$	normal trace of σ w.r.t. ∇u of a BV-function u , page 370
$T_{u^0} u$	translation of $u \in X$ by $u_0 \in X$, i.e., $T_{u^0} u = u + u^0$, page 294
$T_y u$	shifting $u : \mathbf{R}^d \rightarrow \mathbf{K}$ by $y \in \mathbf{R}^d$, i.e., $T_y u(x) = u(x + y)$, page 56
$\operatorname{TV}(u)$	total variation of u , page 354
V_j	subspace in a multiscale analysis, also approximation space to scale j , page 150
W_j	detail space to scale j , page 152

Symbols

$ x $	Euclidean norm of vectors $x \in \mathbf{K}^n$, page 16
$ x _p$	p -norm of vector $x \in \mathbf{K}^n$, page 16
$ \alpha $	order of a multi-index $\alpha \in \mathbf{N}^d$, page 22
$ \mu $	total variation measure to μ , page 43
$ \Omega $	Lebesgue measure of a set $\Omega \in \mathfrak{B}(\mathbf{R}^d)$, page 34
\angle_x	direction of vector $x \in \mathbf{R}^2$, page 78
χ_B	characteristic function of the set B , page 57
δ_x	Dirac measure in x , page 33
$\langle x^*, x \rangle_{X^* \times X}$	duality pairing of x^* and x , page 25
η, ξ	local image coordinates, page 204
$\frac{\partial^\alpha}{\partial x^\alpha}$	α th derivative, page 22
$\lfloor y \rfloor$	ceiling function applied to y , largest integer that is smaller than y , page 56
$\operatorname{med}_B(u)$	median filter with B applied to u , page 101
$\mu \llcorner \Omega'$	restriction of μ to Ω' , page 34
∇F	gradient of F , page 21
$\nabla^2 F$	Hessian matrix of F , page 21
$\nabla^m F$	m th derivative of F organized as an m -tensor, page 321
$\ \mu\ _{\mathfrak{M}}$	norm of the Radon measure μ , page 43
$\ f\ _{m,p}$	norm in the Sobolev space $H^{m,p}(\Omega)$, page 51
$\ f\ _p$	norm in the Banach space $L^p(\Omega, X)$, page 39

$\frac{\ x\ _X}{U}$	norm in a Banach space X , page 16 closure of the set U , page 17
∂F	subdifferential of the convex functional F , page 286
∂U	boundary of the set U , page 17
∂^α	α th derivative, page 22
$\partial_\eta u, \partial_\xi u$	derivatives of u in local coordinates, page 204
$(x, y)_X$	inner product in the Hilbert space X , page 29
(x_n)	sequence (x_1, x_2, \dots) , page 17
Δu	Laplace operator applied to u , page 22
Γ	gamma function, page 34
$\Gamma_0(X)$	space of pointwise suprema of continuous affine linear functionals, page 305
$\Pi^m(\Omega)$	space of polynomials of degree up to $m - 1$ on Ω , page 322
\check{u}	inverse Fourier transform of u , page 115
\check{T}	inverse Fourier transform of a distribution T , page 121
\widehat{T}	Fourier transform of a distribution T , page 121
\widehat{u}	Fourier transform of u , page 110
\widehat{u}	discrete Fourier transform of u , page 135
$F \triangle G$	infimal convolution of F and G , page 311
F^*	Fenchel conjugate of F , page 305
F^*	adjoint mapping of F , page 27
F^*	Hilbert space adjoint of F , page 32
K^\perp	cone normal to K , page 290
$p \otimes q$	tensor product of two \mathbf{R}^d vectors, i.e., $p \otimes q = pq^T$, page 199
p^*	dual exponent to $p \in [1, \infty[$, i.e., $\frac{1}{p} + \frac{1}{p^*} = 1$, page 41
$u \ominus B$	erosion of u with B , page 88
$u \oplus B$	dilation of u with B , page 88
$u \sqcup B$	black-top-hat operator with B applied to u , page 95
$u \odot (B, C)$	hit-or-miss operator with B and C applied to u , page 94
$u * h$	convolution of u with h , page 69
$U \boxtimes H$	U filtered with H , page 83
$u \bullet B$	closing of u with B , page 93
$u \circ B$	opening of u with B , page 93
$u \diamond_m B$	m th rank-order filter, page 101
$u \sqcap B$	white-top-hat operator with B applied to u , page 95
$U \subset\subset X$	U is a compact subset of X , page 17
$u \circledast v$	periodic convolution of u and v , page 138
U^\perp	annihilator of the set U , page 24
U^\perp	orthogonal complement of U , page 30
$V_j \otimes V_j$	tensor product of the approximation spaces V_j with itself, page 161
$x \cdot y$	Euclidean inner product of x and y in \mathbf{K}^N , page 30
$x \perp y$	x is orthogonal to y , i.e., $(x, y)_X = 0$, page 30
$X \hookrightarrow Y$	X is continuously embedded in Y , page 20
$X \succcurlyeq Y$	matrix inequality for symmetric matrices X and Y : $X - Y$ is positive semi-definite, page 189

$x \vee y$	maximum of x and y , page 88
$x \wedge y$	minimum of x and y , page 88
X^*	dual space of X , page 24
$x_n \rightarrow x$	x is the limit of the sequence (x_n) , page 17
$x_n \rightharpoonup x$	x is the weak limit of the sequence (x_n) , page 25
$x_n \stackrel{*}{\rightharpoonup} x$	x is the weak*-limit of the sequence (x_n) , page 26
$\psi_{j,k}$	shifted and scaled function ψ , i.e., $\psi_j, k(x) = 2^{-j/2}\psi(2^{-j}x - k)$, page 149

Index

- Aberration
 - chromatic, 222
- Adjoint, 27
 - Hilbert space \sim , 32
 - of unbounded mappings, 27
 - of the weak gradient, 329
- Admissibility condition, 147
- Algorithm
 - Arrow-Hurwicz, 408
 - edge detection \sim according to Canny, 96
 - extra gradient \sim , 408
 - forward-backward splitting, 397
 - isodata \sim , 67
 - primal-dual, 316
- Alias effect, 60
- Aliasing, 125, 128, 133, 137
- Almost everywhere, 35
- Annihilator, 24, 308
- Anti-extensionality
 - of opening, 93
- Aperture problem, 11
- Approximation space, 152
- Artifact, 1, 2, 340
 - color, 387
 - compression \sim , 61
 - staircasing \sim , 376
- Average
 - moving, 68, 84, 247
 - nonlocal, 103
- Averaging filter
 - non-local, 104
- Axioms
 - \sim of a scale space, 173
- Ball
 - closed, 17
 - open, 16
- Banach space, 23
- Bandwidth, 128
- Bessel function
 - modified, 254
- Bessel's inequality, 31
- Bidual space, 25
- Bilateral filter, 101
- Binomial filter, 84
- Black top-hat operator, 95
- Borel algebra, 32
- Boundary
 - topological, 17
- Boundary extension, 82
 - constant, 82
 - periodical, 82
 - symmetrical, 82
 - zero- \sim , 82
- Boundary initial value problem, 210
- Boundary treatment, 82
- Boundedness, 18
- Caccioppoli set, 355
- Calculus of variations, 263
 - direct method of, 263
 - fundamental lemma of the, 50
- Cauchy problem, 184, 191
- CFL condition, 245
- Characteristic
 - method of \sim s, 240
 - of a transport equation, 240, 241

- Closing, 93
- Closure
 - topological, 17
- CMYK space, 4
- Coercivity, 265
- Coherence, 5, 224, 226
- Color channel, 4
- Color space, 2
 - CMYK~, 4
 - discrete, 3
 - HSV~, 4
 - RGB~, 4
- Comparison principle
 - of a scale space, 174
- Complement
 - orthogonal, 30
- Completely continuous, 28
- Completion
 - of a normed space, 25
 - of a σ -algebra, 35
- Compression, 13, 61
 - with the DCT, 141
 - by inpainting, 260
 - by means of the wavelet transform, 164
- Cone
 - convex, 290, 308
 - dual, 308
 - normal, 290
- Conjugate
 - Fenchel, 305
- Continuity, 18
 - Lipschitz \sim , 18
 - at a point, 18
 - sequential- \sim , 18
 - uniform, 18
 - weakly sequentially \sim , 27
 - weak*-sequentially \sim , 27
- Contrast invariance, 92
 - of erosion and dilation, 92
- Contrast invariance of a scale space, 175
- Convergence
 - in $\mathcal{D}(\Omega)$, 50
 - in $\mathcal{D}(\Omega)^*$, 50
 - of sequences, 17
 - strict, 360
 - weak, 25
 - weak*- \sim , 26
- Convex, 28, 270
 - \sim analysis, 270
 - strictly \sim , 270
- Convolution, 69
 - discrete, 81, 82
 - \sim kernel, 69
 - infimal, 311
- periodic, 138
- \sim theorem, 111, 122, 138
- Coordinates
 - local \sim , 204
- Coordinate transformation
 - linear, 56, 176
- Corner, 225
- Correspondence problem, 10
- Cosine transform
 - discrete, 140
- Counting measure, 33
- Curvature motion, 205, 241
- Curvelet, 165
 - \sim transform, 165
- Deblurring, 7, 79, 119
 - variational, 254
- Deconvolution, 119
 - blind, 258
 - Sobolev, 338
 - total variation, 377
 - variational, 254, 420
- Definiteness
 - positive, 16, 29
- Delta comb, 59, 129
- Denoising, 6, 102, 104, 117, 418
 - with the heat equation, 198
 - with median filter, 101
 - with the moving average, 68
 - of objects, 86
 - with the Perona-Malik equation, 219
 - Sobolev, 334
 - total variation, 374
 - variational, 252
 - by wavelet soft thresholding, 184
- Derivative, 20
 - of a distribution, 51
 - distributional, 51
 - Fréchet- \sim , 20
 - Gâteaux- \sim , 23
 - at a point, 20
 - weak, 51
- Detail space, 152
- Differentiability, 20
 - continuous, 20
 - Fréchet, 20
 - Gâteaux, 23
 - weak, 51
- Differential equation
 - numerical solution, 229
 - partial, 184, 196
- Differential operator
 - elliptic, 189

- Diffusion
 - anisotropic, 206, 222
 - coherence enhancing, 226
 - edge-enhancing, 226
 - equation, 207
 - ~ equation
 - isotropic, 206
 - numerical solution, 234
 - ~ tensor, 206
- Dilation
 - of binary images, 88
 - of grayscale images, 89
 - multiscale ~, 181
- Discrepancy functional, 252
- Discrepancy term, 252
- Distribution, 50
 - delta ~, 50, 121
 - Dirac ~, 50
 - regular, 50
 - tempered, 120
- Distribution function, 63
- Distributivity
 - of erosion and dilation, 90
- Divergence, 22
 - weak, 328, 343
- Domain, 47
 - bounded, 47
 - of a linear mapping, 19
 - Lipschitz ~, 47
- Domain of definition
 - effective, 263
- Duality
 - of Banach spaces, 23
 - of erosion and dilation, 90
 - Fenchel, 301
 - Fenchel-Rockafellar, 312
 - of opening and closing, 93
- Duality gap, 414
- Duality pairing, 25
- Dual space, 24
 - of Lebesgue spaces, 41
 - of spaces of continuous functions, 44
- Edge, 5, 213, 224
 - detection, 7
 - ~detection according to Canny, 76
- Elliptic, 189
- Embedded, 20
- Embedding
 - compact, 28
- Epigraph, 263
- Erosion
 - of binary images, 88
- of grayscale images, 90
 - multiscale ~, 181
- Error
 - mean squared ~, 61
- Error measure
 - mean squared error, 61
 - peak signal-to-Noise ratio, 61
- Euler-Lagrange equation, 259
 - for Sobolev deconvolution, 340
 - for Sobolev denoising, 335
 - of Sobolev inpainting, 344
 - for Sobolev interpolation, 348
 - for total variation deconvolution, 378
 - for total variation-denoising, 375
 - for total variation inpainting, 381
 - for total variation interpolation, 383
- Exponent
 - dual, 41
- Extensionality
 - closing, 93
- Fenchel conjugate, 305
 - of indicator functionals, 308
 - of norm functionals, 307
 - of positively homogeneous functionals, 307
 - of sums, 311
 - of suprema, 310
- Fenchel conjugation
 - calculus for, 309
- Field, 16
- Filter
 - bilateral, 101
 - binomial, 84
 - discrete, 83
 - efficient implementation, 85
 - Gaussian, 84
 - high-pass , 117
 - linear, 68, 75
 - low-pass , 117
 - median , 101
 - morphological, 86
 - moving average, 84
 - rank-order ~, 100
- Filter function, 68
- Filter mask, 83
 - separable, 85
- Finite differences, 234
- Fourier coefficients, 126
- Fourier series, 125
- Fourier transform, 109
 - discrete, 135
 - on $L^1(\mathbf{R}^d)$, 110
 - on $L^2(\mathbf{R}^d)$, 116

- on $\mathcal{S}(\mathbf{R}^d)^*$, 121
- windowed, 142
- Fréchet derivative, 20
- Frequency representation, 117
- Function
 - harmonic, 259
 - integrable, 36
 - interpolation \sim , 58
 - measurable, 36
 - p -integrable, 38
 - weakly harmonic, 259
- Functional
 - affine linear supporting, 304
 - bidual, 305
 - coercive, 265
 - convex, 270
 - discrepancy, 252
 - dual, 305
 - indicator, 273
 - Lagrange, 315
 - objective, 252
 - penalty, 252
 - proper, 263
 - Rudin-Osher-Fatemi- \sim , 375
 - supporting, 304
 - Tikhonov, 279, 420
- Gabor transform, 143
- Gamma function, 34
- Gaussian filter, 84
- Gaussian function, 75, 78, 167, 177
- Generator
 - infinitesimal, 187
 - \sim of a multiscale analysis, 150
- Gradient, 21
- Gradient flow, 395
- Graph, 285
- Grayscale invariance
 - of a scale space, 175
- Gray-value-shift invariance
 - of a scale space, 175
- Hat function, 57
- Heat conduction, 78
- Heat equation, 196, 206, 231
- Hessian matrix, 21
- High-level methods, 4
- Hilbert space, 29
- Hilbert space adjoint, 32
- Histogram, 63
- Hit-or-miss operator, 94
- Homogeneity
 - positive, 16, 308
 - HSV space, 4
- Idempotence
 - of opening and closing, 93
- Image, 2
 - binary \sim , 3
 - continuous, 2
 - \sim decomposition, 6, 117, 427
 - discrete, 2
 - grayscale- \sim , 3
 - \sim processing
 - high-level method, 5
 - low-level method, 5
- Image domain, 2
- Inequality
 - Cauchy-Schwarz \sim , 29
 - Fenchel, 306
 - Hölder's \sim , 41
 - Minkowski- \sim , 16
 - Minkowski \sim for integrals, 39
 - Poincaré-Wirtinger \sim , 323, 360
 - subgradient, 286
 - Young's \sim , 69
 - Young's \sim for products, 307
- Infimal convolution, 311
 - exact, 436
- Infimum
 - essential, 39
- Injection
 - canonical, 25
- Inner product, 29
 - Euclidean, 30
- Inpainting, 12, 246, 258
 - harmonic \sim , 260
 - Sobolev \sim , 341
 - total variation \sim , 379
- Integral, 36
 - Bochner \sim , 37
- Integration theory, 32
- Interior
 - topological, 17
- Interior point, 17
- Interpolation, 55, 75
 - bilinear \sim , 58
 - nearest-neighbor \sim , 56, 58
 - piecewise constant \sim , 56
 - piecewise linear \sim , 57
 - separable \sim , 56
 - Sobolev, 346
 - tensor product \sim , 58
 - total variation, 383
- Interpolation function, 58, 75

- Inverse problem
 - ill-posed, 257
- Isometry invariance
 - of a scale space, 175
- Isomorphism
 - isometric, 20
 - linear, 20
- Jacobian matrix, 22
- Kernel
 - of a linear mapping, 20
- Lagrange multiplier, 282, 299
- Laplace
 - \sim filter, 85
 - \sim operator, 22, 80
 - \sim sharpening, 79
- Lebesgue space, 38
 - standard \sim , 39
- Leibniz rule, 22
- Lemma
 - Fatou's, 40
 - fundamental \sim of the calculus of variations, 50
 - Weyl's, 260
- Level-set
 - method, 426
 - sub- \sim , 275
- Limit point, 17
- Lipschitz
 - \sim constant, 19
 - \sim continuity, 18
 - \sim domain, 47
 - \sim property, 47
- Locality
 - of a scale space, 174
- Lower semicontinuity
 - sequential, 263
- Low-level methods, 5
- Low-pass filter
 - perfect, 118, 134, 346
- Map
 - affine linear, 433
 - p -integrable, 38
- Mapping
 - adjoint, 27
 - affine linear, 271
 - bilinear, 21
- compact, 28
- densely defined linear, 19
- differentiable, 20
- essentially bounded, 39
- k -linear and continuous, 21
- linear and continuous, 19
- multilinear and continuous, 21
- nonexpansive, 396
- Riesz \sim , 32
- self-adjoint, 32
- set-valued, 285
- unbounded, 19
- weakly closed, 26
- weakly*-closed, 26
- weakly continuous, 26
- weakly*-continuous, 26
- Maximum principle
 - discrete, 236
 - for harmonic functions, 260
 - of the Perona-Malik equation, 212, 250
 - of Sobolev denoising, 334
 - for Sobolev inpainting, 341
 - for total variation inpainting, 440
 - of total variation denoising, 375
- Mean curvature motion, 203
- Mean squared error, 61
- Mean-value property, 260
- Measurability, 32
 - of a function, 36
 - Lebesgue \sim , 35
 - of vector-valued mappings, 36
 - w.r.t. a measure, 35
- Measure, 33
 - Borel \sim , 33
 - counting \sim , 59
 - Dirac \sim , 33
 - finite, 33
 - Hausdorff \sim , 34
 - Lebesgue \sim , 34
 - positive Radon \sim , 33
 - product \sim , 45
 - restriction of a \sim , 34
 - σ -finite, 33
 - signed, 42
 - \sim theory, 32
 - total variation, 43, 364
 - vector-valued, 42
- Measure space, 33
- Median, 101
- Median filter, 104
- Method
 - \sim of characteristics, 240
 - direct \sim in the calculus of variations, 263
 - level-set, 426

- splitting, 397
 - See also* Algorithm
- Mexican hat function, 148
- Minimal surface, 381
- Modulation, 110
- Mollifier, 52, 73
- Monotonicity
 - of erosion and dilation, 90
 - of the integral, 37
 - of opening and closing, 93
 - of a scale space, 174
- Morphology, 86
- Motion blur, 75
- Motion blurring, 119, 120
- Multi-index, 22
- Multiscale analysis, 150, 172

- Neighborhood, 17
- Noise, 1, 6, 61
 - impulsive, 101
- Norm, 16
 - equivalent, 16
 - Euclidean vector \sim , 16
 - Frobenius, 385
 - operator \sim , 19
 - spectral, 386
 - strictly convex, 272
- Normal
 - outer \sim , 48
 - \sim trace, 331
- Null set, 35
- Numbers
 - complex, 16
 - extended real, 263
 - real, 16
- Nyquist rate, 128

- Opening, 93
- Operator
 - divergence \sim , 22
 - forward, 278
 - Laplace \sim , 22
 - normal trace \sim , 371
 - p -Laplace, 333
 - trace \sim , 52
- Operator norm, 19
- Optical flow, 9, 427
- Order
 - of a multi-index, 22
- Orthogonal, 30
- Orthogonality, 30
- Orthonormal basis, 31

- Orthonormal system, 31
 - complete, 31
- Outer normal, 48
- Oversampling, 132

- Parseval
 - \sim identity, 31
- Partition of unity, 48
- Peak signal-to-noise-ratio, 61
- Penalty functional, 252
- Perimeter
 - of a measurable set, 355
- Perona-Malik equation, 207
 - in local coordinates, 212
 - modified, 216
 - one-dimensional, 213
- Photograph, 1
- Pixel, 1
- Plancherel's formula, 117
- Plateau function, 57
- Poisson formula, 128
- Polar decomposition, 365
 - of measures, 43
- Pre-dual space, 25
- Pre-Hilbert space, 29
- Prewitt filter, 85
- Probability measure, 33
- Problem
 - dual, 301
 - minimization \sim , 252
 - saddle point, 315
- Projection
 - orthogonal, 30
- Pushforward measure, 46

- Range
 - of a linear mapping, 20
- Recursivity
 - of a scale space, 173
- Reflexive, 25
- Registration, 11, 429
- Regularity
 - of a scale space, 174
- Relative topology, 17
- Resolvent, 397
- RGB space, 4
- Riesz mapping, 32

- Saddle point, 315
 - problem, 315
- Sampling, 59, 125, 127, 129, 131

- of mean values, 59
- over~, 132
- point ~, 59
- under~, 132, 134
- Sampling rate, 126
- Sampling theorem, 127, 137
- Scale, 173
- Scale invariance
 - of a scale space, 175
- Scale space, 171, 173, 392
 - linear, 196
 - morphological, 199
- Scaling equation, 151
- Scaling function
 - ~ of a multiscale analysis, 150
- Schwartz function, 112
- Schwartz space, 112
- Segmentation, 8, 66
- Semi-group property
 - of a scale space, 173
- Semi linear, 31
- Seminorm
 - admissible, 322
- Sequence, 17
 - Cauchy ~, 23
 - convergent, 17
 - strictly convergent, 360
 - weak* convergent, 26
 - weakly convergent, 25
- Sequence space, 40
- Sharpening, 79, 119
- Shearlet, 165
- Shearlet transform, 166
- σ -algebra, 32
- Singular value decomposition, 386
- Sobel filter, 85
- Sobolev space, 51
 - fractional, 124
- Soft thresholding, 405
 - Fourier, 184
 - wavelet~, 184
- Space
 - Banach ~, 23
 - bidual, 25
 - of bounded and continuous functions, 173
 - of bounded functions, 89
 - of the continuous mappings, 19
 - of differentiable functions, 49
 - of distributions, 50
 - dual, 24
 - embedded, 20
 - of functions of bounded total variation, 356
 - of k -linear and continuous mappings, 21
 - Lebesgue ~, 38
- of linear and continuous mappings, 19
- of locally integrable functions, 49
- measurable, 32
- measure ~, 33
- normed, 16
- of p -integrable functions, 38
- of polynomials of degree up to $m - 1$, 322
- pre-dual, 25
- pre-Hilbert ~, 29
- quotient~, 18
- of Radon measures, 43
- reflexive, 25
- Schwartz ~, 112
- separable, 17
- Sobolev ~, 51
- of special functions of bounded total variation, 426
- standard Lebesgue ~, 39
- of tempered distributions, 120
- of test functions, 49
- Staircasing effect, 215, 376, 429
- Step function, 36
- Step-size restriction, 236
- Structure element, 88, 89
- Structure tensor, 222, 224
- Subdifferential, 286
 - chain rule, 294, 434
 - chain rule for the, 435
 - for convex functions of integrands, 293
 - of indicator functionals, 290
 - of norm functionals, 291
 - of the Sobolev seminorm, 332
 - sum rule, 294, 434
 - of the total variation, 373
- Subgradient, 286
 - existence of a ~, 297
 - inequality, 286
- Sublevel-set, 275
- Subset
 - closed, 17
 - compact, 17
 - dense, 17
 - open, 17
 - sequentially closed, 17
 - sequentially compact, 17
 - weakly sequentially compact, 26
 - weak*-sequentially compact, 26
- Support
 - compact, 42
- Support functional
 - affine linear, 281
- Supremum
 - essential, 39
- Symlets, 155

- Symmetry
 - Hermitian, 29
- Tensor product, 161
- Test functions, 50
- Texture, 5, 119
- Theorem
 - Baire category~, 23
 - Banach-Alaoglu, 44
 - Banach-Alaoglu \sim , 26
 - Banach-Steinhaus, 24
 - closed range, 28
 - convolution \sim , 111, 122, 138
 - divergence \sim , 49
 - Eberlein-Šmulyan \sim , 26, 42
 - Edelheit's \sim , 287
 - Fischer-Riesz, 40
 - Fubini's \sim , 45
 - Gauss' integral \sim on $BV(\Omega)$, 371
 - Gauss's \sim , 49
 - Hahn-Banach extension \sim , 25
 - Hahn-Banach separation \sim , 28
 - of the inverse mapping, 24
 - Lebesgue's dominated convergence \sim , 40
 - of the open mapping, 24
 - Pythagorean, 30
 - Riesz-Markov representation \sim , 44
 - Riesz representation \sim , 31, 41
 - weak Gauss' \sim , 53
- Threshold, 66
- Topology, 16, 264
 - norm- \sim , 16
 - strong, 16, 264
 - weak, 26, 265
 - weak*- \sim , 26
- Total variation, 354
 - bounded, 356
 - vectorial, 386
- Trace, 52
 - of a $BV(\Omega)$ function, 362
 - of $\mathcal{D}_{\text{div}, \infty}$ -vector field, 372
 - normal, 331, 371
- of a Sobolev function, 52
- Transfer function, 117, 138
- Translation, 56
- Translation invariance
 - of the convolution, 69
 - of erosion and dilation, 90
 - of opening and closing, 93
 - of a scale space, 175
- Transport equation, 203, 205, 234
 - numerical solution, 240
- Triangle inequality, 16
- Undersampling, 132, 134
- Upwind scheme, 245
- Variation
 - \sim al method, 252
 - \sim al problem, 252
 - \sim measure, 43
 - total, 354
- Vector field visualization, 228
- Viscosity
 - numerical, 245
- Viscosity solution, 192, 194
- Viscosity sub-solution, 193
- Viscosity super-solution, 194
- Wavelet, 145, 147
 - Daubechies \sim s, 155
 - fast \sim reconstruction, 159
 - fast \sim transform, 158
 - Haar \sim , 149, 153
 - inverse \sim transform, 147
 - \sim space, 152
- Weak formulation, 219
- Weak solution, 217, 218
- Weyl's Lemma, 260
- White top-hat operator, 95
- Window function, 141

Applied and Numerical Harmonic Analysis (81 Volumes)

1. A. I. Saichev and W.A. Woyczyński: *Distributions in the Physical and Engineering Sciences* (ISBN 978-0-8176-3924-2)
2. C.E. D'Attellis and E.M. Fernandez-Berdaguer: *Wavelet Theory and Harmonic Analysis in Applied Sciences* (ISBN 978-0-8176-3953-2)
3. H.G. Feichtinger and T. Strohmer: *Gabor Analysis and Algorithms* (ISBN 978-0-8176-3959-4)
4. R. Tolimieri and M. An: *Time-Frequency Representations* (ISBN 978-0-8176-3918-1)
5. T.M. Peters and J.C. Williams: *The Fourier Transform in Biomedical Engineering* (ISBN 978-0-8176-3941-9)
6. G.T. Herman: *Geometry of Digital Spaces* (ISBN 978-0-8176-3897-9)
7. Teolis: *Computational Signal Processing with Wavelets* (ISBN 978-0-8176-3909-9)
8. J. Ramanathan: *Methods of Applied Fourier Analysis* (ISBN 978-0-8176-3963-1)
9. J.M. Cooper: *Introduction to Partial Differential Equations with MATLAB* (ISBN 978-0-8176-3967-9)
10. Procházka, N.G. Kingsbury, P.J. Payner, and J. Uhlir: *Signal Analysis and Prediction* (ISBN 978-0-8176-4042-2)
11. W. Bray and C. Stanojevic: *Analysis of Divergence* (ISBN 978-1-4612-7467-4)
12. G.T. Herman and A. Kuba: *Discrete Tomography* (ISBN 978-0-8176-4101-6)
13. K. Gröchenig: *Foundations of Time-Frequency Analysis* (ISBN 978-0-8176-4022-4)
14. L. Debnath: *Wavelet Transforms and Time-Frequency Signal Analysis* (ISBN 978-0-8176-4104-7)
15. J.J. Benedetto and P.J.S.G. Ferreira: *Modern Sampling Theory* (ISBN 978-0-8176-4023-1)
16. D.F. Walnut: *An Introduction to Wavelet Analysis* (ISBN 978-0-8176-3962-4)

17. Abbate, C. DeCusatis, and P.K. Das: *Wavelets and Subbands* (ISBN 978-0-8176-4136-8)
18. O. Bratteli, P. Jorgensen, and B. Treadway: *Wavelets Through a Looking Glass* (ISBN 978-0-8176-4280-80)
19. H.G. Feichtinger and T. Strohmer: *Advances in Gabor Analysis* (ISBN 978-0-8176-4239-6)
20. O. Christensen: *An Introduction to Frames and Riesz Bases* (ISBN 978-0-8176-4295-2)
21. L. Debnath: *Wavelets and Signal Processing* (ISBN 978-0-8176-4235-8)
22. G. Bi and Y. Zeng: *Transforms and Fast Algorithms for Signal Analysis and Representations* (ISBN 978-0-8176-4279-2)
23. J.H. Davis: *Methods of Applied Mathematics with a MATLAB Overview* (ISBN 978-0-8176-4331-7)
24. J.J. Benedetto and A.I. Zayed: *Sampling, Wavelets, and Tomography* (ISBN 978-0-8176-4304-1)
25. E. Prestini: *The Evolution of Applied Harmonic Analysis* (ISBN 978-0-8176-4125-2)
26. L. Brandolini, L. Colzani, A. Iosevich, and G. Travaglini: *Fourier Analysis and Convexity* (ISBN 978-0-8176-3263-2)
27. W. Freeden and V. Michel: *Multiscale Potential Theory* (ISBN 978-0-8176-4105-4)
28. O. Christensen and K.L. Christensen: *Approximation Theory* (ISBN 978-0-8176-3600-5)
29. O. Calin and D.-C. Chang: *Geometric Mechanics on Riemannian Manifolds* (ISBN 978-0-8176-4354-6)
30. J.A. Hogan: *Time–Frequency and Time–Scale Methods* (ISBN 978-0-8176-4276-1)
31. Heil: *Harmonic Analysis and Applications* (ISBN 978-0-8176-3778-1)
32. K. Borre, D.M. Akos, N. Bertelsen, P. Rinder, and S.H. Jensen: *A Software-Defined GPS and Galileo Receiver* (ISBN 978-0-8176-4390-4)
33. T. Qian, M.I. Vai, and Y. Xu: *Wavelet Analysis and Applications* (ISBN 978-3-7643-7777-9)
34. G.T. Herman and A. Kuba: *Advances in Discrete Tomography and Its Applications* (ISBN 978-0-8176-3614-2)
35. M.C. Fu, R.A. Jarrow, J.-Y. Yen, and R.J. Elliott: *Advances in Mathematical Finance* (ISBN 978-0-8176-4544-1)
36. O. Christensen: *Frames and Bases* (ISBN 978-0-8176-4677-6)
37. P.E.T. Jorgensen, J.D. Merrill, and J.A. Packer: *Representations, Wavelets, and Frames* (ISBN 978-0-8176-4682-0)
38. M. An, A.K. Brodzik, and R. Tolimieri: *Ideal Sequence Design in Time-Frequency Space* (ISBN 978-0-8176-4737-7)
39. S.G. Krantz: *Explorations in Harmonic Analysis* (ISBN 978-0-8176-4668-4)
40. Luong: *Fourier Analysis on Finite Abelian Groups* (ISBN 978-0-8176-4915-9)
41. G.S. Chirikjian: *Stochastic Models, Information Theory, and Lie Groups, Volume 1* (ISBN 978-0-8176-4802-2)

42. Cabrelli and J.L. Torrea: *Recent Developments in Real and Harmonic Analysis* (ISBN 978-0-8176-4531-1)
43. M.V. Wickerhauser: *Mathematics for Multimedia* (ISBN 978-0-8176-4879-4)
44. B. Forster, P. Massopust, O. Christensen, K. Gröchenig, D. Labate, P. Vandergheynst, G. Weiss, and Y. Wiaux: *Four Short Courses on Harmonic Analysis* (ISBN 978-0-8176-4890-9)
45. O. Christensen: *Functions, Spaces, and Expansions* (ISBN 978-0-8176-4979-1)
46. J. Barral and S. Seuret: *Recent Developments in Fractals and Related Fields* (ISBN 978-0-8176-4887-9)
47. O. Calin, D.-C. Chang, and K. Furutani, and C. Iwasaki: *Heat Kernels for Elliptic and Sub-elliptic Operators* (ISBN 978-0-8176-4994-4)
48. C. Heil: *A Basis Theory Primer* (ISBN 978-0-8176-4686-8)
49. J.R. Klauder: *A Modern Approach to Functional Integration* (ISBN 978-0-8176-4790-2)
50. J. Cohen and A.I. Zayed: *Wavelets and Multiscale Analysis* (ISBN 978-0-8176-8094-7)
51. Joyner and J.-L. Kim: *Selected Unsolved Problems in Coding Theory* (ISBN 978-0-8176-8255-2)
52. G.S. Chirikjian: *Stochastic Models, Information Theory, and Lie Groups, Volume 2* (ISBN 978-0-8176-4943-2)
53. J.A. Hogan and J.D. Lakey: *Duration and Bandwidth Limiting* (ISBN 978-0-8176-8306-1)
54. Kutyniok and D. Labate: *Shearlets* (ISBN 978-0-8176-8315-3)
55. P.G. Casazza and P. Kutyniok: *Finite Frames* (ISBN 978-0-8176-8372-6)
56. V. Michel: *Lectures on Constructive Approximation* (ISBN 978-0-8176-8402-0)
57. D. Mitrea, I. Mitrea, M. Mitrea, and S. Monniaux: *Groupoid Metrization Theory* (ISBN 978-0-8176-8396-2)
58. T.D. Andrews, R. Balan, J.J. Benedetto, W. Czaja, and K.A. Okoudjou: *Excursions in Harmonic Analysis, Volume 1* (ISBN 978-0-8176-8375-7)
59. T.D. Andrews, R. Balan, J.J. Benedetto, W. Czaja, and K.A. Okoudjou: *Excursions in Harmonic Analysis, Volume 2* (ISBN 978-0-8176-8378-8)
60. D.V. Cruz-Uribe and A. Fiorenza: *Variable Lebesgue Spaces* (ISBN 978-3-0348-0547-6)
61. W. Freeden and M. Gutting: *Special Functions of Mathematical (Geo-)Physics* (ISBN 978-3-0348-0562-9)
62. A. I. Saichev and W.A. Woyczyński: *Distributions in the Physical and Engineering Sciences, Volume 2: Linear and Nonlinear Dynamics of Continuous Media* (ISBN 978-0-8176-3942-6)
63. S. Foucart and H. Rauhut: *A Mathematical Introduction to Compressive Sensing* (ISBN 978-0-8176-4947-0)
64. Herman and J. Frank: *Computational Methods for Three-Dimensional Microscopy Reconstruction* (ISBN 978-1-4614-9520-8)

65. Paprotny and M. Thess: *Realtime Data Mining: Self-Learning Techniques for Recommendation Engines* (ISBN 978-3-319-01320-6)
66. Zayed and G. Schmeisser: *New Perspectives on Approximation and Sampling Theory: Festschrift in Honor of Paul Butzer's 85th Birthday* (ISBN 978-3-319-08800-6)
67. R. Balan, M. Begue, J. Benedetto, W. Czaja, and K.A Okoudjou: *Excursions in Harmonic Analysis, Volume 3* (ISBN 978-3-319-13229-7)
68. Boche, R. Calderbank, G. Kutyniok, J. Vyiral: *Compressed Sensing and its Applications* (ISBN 978-3-319-16041-2)
69. S. Dahlke, F. De Mari, P. Grohs, and D. Labate: *Harmonic and Applied Analysis: From Groups to Signals* (ISBN 978-3-319-18862-1)
70. Aldroubi, *New Trends in Applied Harmonic Analysis* (ISBN 978-3-319-27871-1)
71. M. Ruzhansky: *Methods of Fourier Analysis and Approximation Theory* (ISBN 978-3-319-27465-2)
72. G. Pfander: *Sampling Theory, a Renaissance* (ISBN 978-3-319-19748-7)
73. R. Balan, M. Begue, J. Benedetto, W. Czaja, and K.A Okoudjou: *Excursions in Harmonic Analysis, Volume 4* (ISBN 978-3-319-20187-0)
74. O. Christensen: *An Introduction to Frames and Riesz Bases, Second Edition* (ISBN 978-3-319-25611-5)
75. E. Prestini: *The Evolution of Applied Harmonic Analysis: Models of the Real World, Second Edition* (ISBN 978-1-4899-7987-2)
76. J.H. Davis: *Methods of Applied Mathematics with a Software Overview, Second Edition* (ISBN 978-3-319-43369-1)
77. M. Gilman, E. M. Smith, S. M. Tsynkov: *Transitionospheric Synthetic Aperture Imaging* (ISBN 978-3-319-52125-1)
78. S. Chanillo, B. Franchi, G. Lu, C. Perez, E.T. Sawyer: *Harmonic Analysis, Partial Differential Equations and Applications* (ISBN 978-3-319-52741-3)
79. R. Balan, J. Benedetto, W. Czaja, M. Dellatorre, and K.A Okoudjou: *Excursions in Harmonic Analysis, Volume 5* (ISBN 978-3-319-54710-7)
80. Pesenson, Q.T. Le Gia, A. Mayeli, H. Mhaskar, D.X. Zhou: *Frames and Other Bases in Abstract and Function Spaces: Novel Methods in Harmonic Analysis, Volume 1* (ISBN 978-3-319-55549-2)
81. Pesenson, Q.T. Le Gia, A. Mayeli, H. Mhaskar, D.X. Zhou: *Recent Applications of Harmonic Analysis to Function Spaces, Differential Equations, and Data Science: Novel Methods in Harmonic Analysis, Volume 2* (ISBN 978-3-319-55555-3)
82. F. Weisz: *Convergence and Summability of Fourier Transforms and Hardy Spaces* (ISBN 978-3-319-56813-3)
83. Heil: *Metrics, Norms, Inner Products, and Operator Theory* (ISBN 978-3-319-65321-1)
84. S. Waldron: *An Introduction to Finite Tight Frames: Theory and Applications.* (ISBN: 978-0-8176-4814-5)
85. Joyner and C.G. Melles: *Adventures in Graph Theory: A Bridge to Advanced Mathematics.* (ISBN: 978-3-319-68381-2)

86. B. Han: *Framelets and Wavelets: Algorithms, Analysis, and Applications* (ISBN: 978-3-319-68529-8)
87. H. Boche, G. Caire, R. Calderbank, M. März, G. Kutyniok, R. Mathar: *Compressed Sensing and Its Applications* (ISBN: 978-3-319-69801-4)
88. N. Minh Chong: *Pseudodifferential Operators and Wavelets over Real and p -adic Fields* (ISBN: 978-3-319-77472-5)
89. A. I. Saichev and W.A. Woyczyński: *Distributions in the Physical and Engineering Sciences, Volume 3: Random and Fractal Signals and Fields* (ISBN: 978-3-319-92584-4)

For an up-to-date list of ANHA titles, please visit <http://www.springer.com/series/4968>