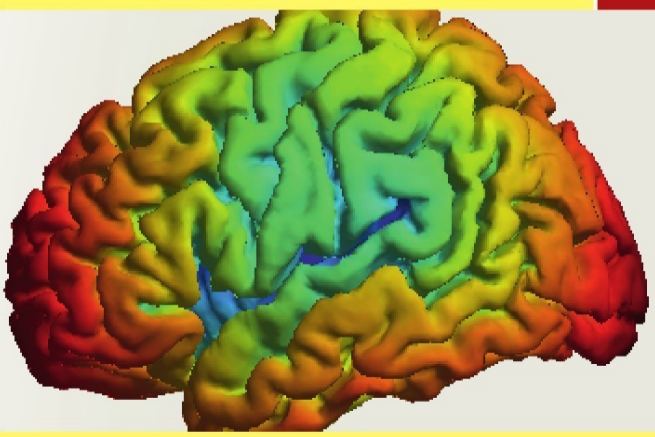
Computational Imaging and Vision 41



Luc Florack Marie-Colette van Lieshout
Remco Duits Laurie Davies *Eds.*
Geurt Jongbloed

Mathematical Methods for Signal and Image Analysis and Representation

Springer

Mathematical Methods for Signal and Image Analysis and Representation

Computational Imaging and Vision

This comprehensive book series embraces state-of-the-art expository works and advanced research monographs on any aspect of this interdisciplinary field.

Topics covered by the series fall in the following four main categories:

- Imaging Systems and Image Processing
- Computer Vision and Image Understanding
- Visualization
- Applications of Imaging Technologies

Only monographs or multi-authored books that have a distinct subject area, that is where each chapter has been invited in order to fulfill this purpose, will be considered for the series.

Volume 41

For further volumes:
www.springer.com/series/5754

Luc Florack • Remco Duits • Geurt Jongbloed •
Marie-Colette van Lieshout • Laurie Davies
Editors

Mathematical Methods for Signal and Image Analysis and Representation

Springer

Editors
Prof. Luc Florack
Dept. Mathematics & Computer Science
Eindhoven University of Technology
Eindhoven
The Netherlands

Dr. Remco Duits
Dept. Mathematics & Computer Science
Eindhoven University of Technology
Eindhoven
The Netherlands

Prof. Geurt Jongbloed
Dept. Applied Mathematics
Delft University of Technology
Delft
The Netherlands

Dr. Marie-Colette van Lieshout
Probability & Stochastic Networks (PNA)
Centrum Wiskunde & Informatica
Amsterdam
The Netherlands

Prof. Laurie Davies
Fakultät für Mathematik
Universität Duisburg-Essen
Essen
Germany

ISSN 1381-6446 Computational Imaging and Vision
ISBN 978-1-4471-2352-1 e-ISBN 978-1-4471-2353-8
DOI 10.1007/978-1-4471-2353-8
Springer London Dordrecht Heidelberg New York

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Control Number: 2012930491

Mathematics Subject Classification: 62H35, 68U10

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

The research institute EURANDOM (European Institute for Statistics, Probability, Stochastic Operations Research and its Applications) was established in 1997 on the campus of Eindhoven University of Technology, The Netherlands. Its mission is to foster research in the area of stochastics and its applications. It achieves this mission by recruiting and training talented young researchers and helping them to find their way to tenured positions in academia and industry, and by carrying out and facilitating research through postdoctoral and graduate appointments, visitor exchange and workshops. Its chief mission statement has been given nationwide support in The Netherlands by a recently installed national cluster called STAR (Stochastics—Theoretical and Applied Research), for which EURANDOM acts as coordinating and facilitating node.

As part of its workshop programme, EURANDOM organized a series of international workshops on image processing and analysis. The third one in this series was the workshop on *Locally Adaptive Filters in Signal and Image Processing*, November 24–26, 2008, focusing specifically on locally adaptive methods. The ability of a system to adapt to the local state is important in many problems in image analysis. Many renowned young experts were invited to give overview talks on this theme covering state-of-the-art and novel research.

Despite the high quality of contributions, no proceedings of this workshop have been issued. Instead, the workshop initiated a collaborative effort, focusing more generally on mathematical methods for signal and image analysis and representation. The results of this effort are described in this book.

Contributions have been carefully selected to be representative for a variety of generic approaches as well as to illustrate formal connections among these. Roughly speaking deterministic methods are central to the first half of the book, whereas the second half considers mainly statistical methods. However, some chapters in the middle of the book clearly encompass both approaches, and more than a hundred cross-references throughout the book emphasize the many formal connections and analogies that exist between seemingly different paradigms.

This book differs from most existing books on medical signal and image analysis or computer vision to the extent that it does not focus on specific applications (al-

though some are detailed for the sake of illustration), but on *methodological frameworks* on which such applications may be built. This book should therefore be of interest to all those in search of a suitable methodological basis for specific applications, as well as to those who are interested in fundamental methodologies per se.

Eindhoven, Netherlands Luc Florack

Acknowledgements

The institute EURANDOM of Eindhoven University of Technology has funded and facilitated the visitor exchange programme and workshop for the international collaboration that has resulted in this edited book. Further funding was obtained from the Netherlands Organisation for Scientific Research (NWO) through the programme for Incidental Financial Support, through a personal grant in the Innovational Research Incentives Scheme to Luc Florack, and through a grant from the Royal Netherlands Academy of Arts and Sciences (KNAW).

Special thanks go to Lucienne Coolen-van Will, Marèse Wolfs-van de Hurk and Enna van Dijk for administrative support, the Department of Mathematics & Computer Science, the Department of Biomedical Engineering, and the Executive Board of Eindhoven University of Technology for supporting the cross-divisional Imaging Science & Technology Eindhoven High Potential Research Program (IST/e).

Contents

Contributors

Maxime Boucher School of Computer Science, McGill University, Montreal, Canada

Bernhard Burgeth Mathematical Image Analysis Group, Faculty of Mathematics and Computer Science, Saarland University, Saarbruecken, Germany

Laurent Demaret HelmholtzZentrum München, Institut für Biomathematik und Biometrie (IBB), Neuherberg, Germany

Xavier Descombes Laboratoire d'Informatique, Signaux et Systèmes de Sophia-Antipolis I3S, UMR6070, UNS CNRS 2000, Sophia Antipolis Cedex, France

Stephan Didas Abteilung Bildverarbeitung, Fraunhofer-Institut für Techno- und Wirtschaftsmathematik, Kaiserslautern, Germany

Remco Duits Department of Mathematics and Computer Science & Department of Biomedical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands

Alan Evans McConnell Brain Imaging Center, Montreal Neurological Institute, McGill University, Montreal, Canada

Michael Felsberg Computer Vision Laboratory, Department of Electrical Engineering, Linköping University, Linköping, Sweden

Luc Florack Department of Mathematics and Computer Science & Department of Biomedical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands

Hartmut Führ Lehrstuhl A für Mathematik, RWTH Aachen, Aachen, Germany

Yaniv Gur SCI Institute, University of Utah, Salt Lake City, UT, USA

Michal Haindl Institute of Information Theory and Automation of the ASCR, Prague, Czech Republic

Armin Iske Department of Mathematics, University of Hamburg, Hamburg, Germany

Bart Janssen Department of Biomedical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands

Wahid Khachabi Department of Mathematics, University of Hamburg, Hamburg, Germany

Kai Krajsek Institute for Chemistry and Dynamics of the Geosphere, ICG-3, Forschungszentrum Jülich GmbH, Jülich, Germany

Jean-Michel Loubes Institut de Mathématiques de Toulouse, UMR 5219, Université Toulouse 3, Toulouse cedex 9, France

Michal Matuszak Faculty of Mathematics & Computer Science, Nicolaus Copernicus University, Toruń, Poland

Ofer Pasternak Department of Psychiatry, Brigham and Women's Hospital Harvard Medical School, Boston, MA, USA

Luis Pizarro Department of Computing, Imperial College London, London, UK

Jörg Polzehl Weierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany

Markus Reiß Institute of Mathematics, Humboldt University Berlin, Berlin, Germany

Paul Rochet Institut de Mathématiques de Toulouse, UMR 5219, Université Toulouse 3, Toulouse cedex 9, France

Yves Rozenholc MAP5, UMR CNRS 8145, University Paris Descartes, Paris, France

Hanno Scharr Institute for Chemistry and Dynamics of the Geosphere, ICG-3, Forschungszentrum Jülich GmbH, Jülich, Germany

Tomasz Schreiber Faculty of Mathematics & Computer Science, Nicolaus Copernicus University, Toruń, Poland

Nir Sochen Department of Applied Mathematics, Tel Aviv University, Tel Aviv, Israel

Karsten Tabelow Weierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany

Markus van Almsick Department of Biomedical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands

Joachim Weickert Mathematical Image Analysis Group, Faculty of Mathematics and Computer Science, Saarland University, Saarbruecken, Germany

Chapter 1
A Short Introduction to Diffusion-Like Methods

Hanno Scharr and Kai Krajsek

Abstract This contribution aims to give a basic introduction to diffusion-like methods. There are many different methods commonly used for regularization tasks. Some of them will be briefly introduced and their connection to diffusion shown. In addition to this we will go into some detail for diffusion-like methods in a narrower sense, i.e. methods based on PDEs similar to diffusion PDEs known from physics. Main issues highlighted here are which PDE to use, how diffusivities in such a PDE are constructed, and which discretization is suitable for a given task.

1.1 Introduction

There are quite a few methods for regularization tasks like noise reduction, inpainting, super-resolution, or interpolation described in literature. Many of them, if not all, can somehow be brought into connection with diffusion. Obviously we cannot visit all of them in this paper making this introduction incomplete. Our focus here will be on nonlinear averaging, mainly used for noise reduction, even though currently best performing denoising algorithms on natural grey value images are not diffusions in a narrower sense (see e.g. [346]). Nevertheless one goal beneath basic introduction is to mention at least some of the major contributions to this field.

We will nearly completely ignore the fact that diffusion can be used to build a scale-space. First discovered in Japan [231, 442] scale-space filtering is a topic of its own (see e.g. [290, 450]). A linear scale space is built by applying linear diffusion (see Sect. 1.2.1) to a signal in short time steps and recording the more and more smoothed signal. Other scale spaces can be derived by applying other diffusion-like schemes, cf. Chaps. 7 and 9.

Regularization schemes are represented in literature from different view-points. Diffusion schemes typically start with the formulation of a continuous partial dif-

H. Scharr (✉) · K. Krajsek
Institute for Chemistry and Dynamics of the Geosphere, ICG-3, Forschungszentrum Jülich GmbH, 52425 Jülich, Germany
e-mail: h.scharr@fz-juelich.de

K. Krajsek
e-mail: k.krajsek@fz-juelich.de

L. Florack et al. (eds.), *Mathematical Methods for Signal and Image Analysis and Representation*, Computational Imaging and Vision 41,
DOI 10.1007/978-1-4471-2353-8_1,

ferential equation (PDE) describing a process which changes data over time [334, 433]. This PDE is then discretized yielding an iterative update scheme. The classical diffusion defined by the heat equation known from physics involves linear filtering of the input data by derivatives

$$\partial_t s(\mathbf{x}, t) = \operatorname{div}(\mathbf{D} \nabla s(\mathbf{x}, t)), \tag{1.1}$$

where $s(\mathbf{x}, t)$ is the evolving signal or image with $s(\mathbf{x}, 0) = r(\mathbf{x})$ and the initially observed data $r(\mathbf{x})$; $\nabla = (\partial_{x_1}, \ldots, \partial_{x_N})$ is the vector of spatial derivatives. Diffusion tensor $\mathbf{D}$ is a symmetric, positive definite tensor which may vary with space and evolution time and may depend on local data. Adaptivity of a diffusion scheme is achieved via adaptation of the diffusion tensor. This makes the scheme nonlinear. In computational physics diffusion is typically simulated using e.g. finite differences on a sampling grid. This grid is refined when the result is not accurate enough, making discretization simple. In image processing no such refinement is typically applied, giving more influence to discretization details. How to discretize an anisotropic nonlinear diffusion process will be subject of Sect. 1.5.

There are different naming conventions in the literature for diffusion schemes. Especially the term *anisotropic diffusion* is inconsistently used. Following [433] we use the term isotropic diffusion, when $\mathbf{D}$ is proportional to the identity matrix $\mathbf{D} = c\mathbb{1}$, i.e. when Eq. (1.1) collapses to

$$\partial_t s(\mathbf{x}, t) = \operatorname{div}(c \nabla s(\mathbf{x}, t)). \tag{1.2}$$

If *diffusivity* or *edge stopping function* c depends on the image $s(\mathbf{x}, t)$, we call a diffusion scheme *isotropic nonlinear diffusion*. We call diffusions with general $\mathbf{D}$, not proportional to $\mathbb{1}$, *anisotropic diffusion*, in contrast to several publications using this term for isotropic nonlinear diffusion (e.g. [41, 334]). This naming inconsistency originates from the fact that the overall effect of isotropic nonlinear diffusion on the evolved data is anisotropic. We call a diffusion scheme *linear* if $\mathbf{D}$ does not depend on the evolving image. The simplest case is *linear homogenous diffusion* with a constant edge stopping function c, where Eq. (1.2) simplifies to

$$\partial_t s(\mathbf{x}, t) = c \Delta s(\mathbf{x}, t), \tag{1.3}$$

where $\Delta = \sum_i^N \partial_{x_i}^2$ is the spatial Laplacian. A time step applying linear diffusion to s on the unbounded domain is solved by convolution with a Gaussian kernel (see Sect. 1.2.1).

Diffusion equations with many different edge stopping functions c and diffusion tensors D have been proposed in literature (see e.g. [41, 364, 369, 433]). They depend on gray value or color gradients, curvatures, or other image features (see [376, 440] for possible dependencies when regularizing optical flow). There is rich ongoing work on diffusions for vector, matrix or tensor-valued data (see e.g. [114, 266, 337, 381, 414], and elsewhere in this volume, cf. Chaps. 3, 4, and 5, where the main problem is to select a useful metric and discretize operators respecting it. We will show an example in Sect. 1.5.2. Diffusion methods explicitly focusing on metrics

induced by known group structures are shown e.g. in this volume, Chaps. 5, 7, 8, 9, and 10, and also elsewhere [114].

A framework for diffusions in generalized image spaces has been defined using the Beltrami operator, a natural generalization of the Laplacian on non-flat manifolds [250]. E.g. an RGB-color image is a 2D manifold the 5D space (x, y, I^r, I^g, I^b) spanned by 2 spatial dimensions x and y and 3 color intensities I^r, I^g, and I^b. The Beltrami framework will be introduced in Sect. 1.2.3.

Equation (1.1) can be seen as a first order Taylor approximation describing a dynamic process as often used in physics. This is sufficient in continuous time and space as used in physics, but in discrete time and/or space richer representations of a process are sometimes advantageous. First order derivatives ∇ may therefore be exchanged by other and/or more operators [361, 366, 465]. We introduce the isotropic nonlinear case in Sect. 1.3.2 and give an example for the anisotropic nonlinear case in Sect. 1.6.2.

Some, but not all of these PDEs may be derived from suitable energy functions via calculus of variations. The PDE then changes the data such that the energy is minimized. Consequently a diffusion scheme optimizes a property our image data is assumed to fulfill. Energy functions may be designed assuming data models motivated from physics underlying the imaging process, e.g. modeling step edges explicitely via a line process [42, 316]. Such energy functions also occur in robust statistics and can be expressed in terms of probabilities via Gibbs distributions or Markov Random Fields (MRFs), cf. [183] and Chaps. 13, 14 and 15. The edge stopping function then corresponds to a robust error function and diffusion is then related to M-estimation [41]. Probability distributions forming potential functions in the energy can be learned from training data using this relation to image statistics [371, 465] (cf. also Chap. 11). Robust error functions are derived from histograms of filtered images. In the classical diffusion case filter kernels are spatial derivatives, but other kernels may be used as well. The kernels may even be learned from training data [361]. The histograms are treated as observed statistics or empirical marginal distributions defining a probability to observe a certain image. Maximizing this probability means minimizing an energy. We will go into more detail in Sect. 1.3.1.

There are algorithmic approaches presenting and evaluating *discrete schemes* used for filtering, e.g. nonlinear Gaussian Filtering [17, 187, 445], Mean-Shift Filtering [81] or Bilateral Filtering [409]. Typically they may also be formulated in terms of minimization of a *cost functional* corresponding to the energy functions formulated in continuous time and space. We show some prominent examples and their relation to diffusion in Sect. 1.4. Here it is important to note that diffusion in a strict sense is only defined in continuous space and time. Therefore a discrete scheme is called a *consistent* diffusion scheme, if it becomes the diffusion equation in the limit $h \to 0$ and $\tau \to 0$ for spatial sampling step h and temporal sampling step τ. This is the definition for consistency known from numerics. Associating some discrete scheme with a discrete scheme for diffusion does *not* show that the first scheme is a diffusion. We will elaborate this for the case of bilateral filtering in Sect. 1.4.1.

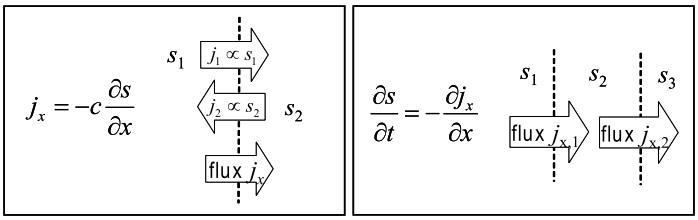


Fig. 1.1 Fick's Law. *Left*: Number of particles crossing a given border is proportional to the number of particles s per unit volume, i.e. proportional to the density. Overall flux j depends on the density difference (or continuously: gradient). *Right*: Change of particle number in a given volume equals in-flux minus out-flux (continuously: divergence). Combining the two laws yields the heat equation (by means of the Gauss divergence theorem) with diffusion coefficient or diffusivity c

1.2 Diffusion in a Narrow Sense

In this section, we introduce different types of diffusion. We start with the simplest case well-known as Fick's law from physics since 1855 [142], continue with formulation of isotropic nonlinear diffusion (sometimes called Perona-Malik Diffusion [334]) and anisotropic diffusion (sometimes called Coherence [433] or Edge-Enhancing Diffusion). Finally we show the currently most general formulation, the Beltrami framework [250].

1.2.1 Diffusion in Physics, Basic Solution and Numerics

Diffusion occurs in statistical physics and thermodynamics where random 'Brownian' motion of particles leads e.g. to heat transport or mixing of liquids or gases. Densities or temperature then evolve with time as described by the heat equation (1.3): $\partial_t s(\mathbf{x}, t) = c\Delta s(\mathbf{x}, t)$ (cf. Fig. 1.1), in which $s(\mathbf{x}, t)$ is the evolving density and c a diffusion constant or *diffusivity*. The evolution of s by linear isotropic diffusion, i.e. diffusion with $c = \text{const.}$ is given by convolution of s with a Gaussian kernel with variance $\sigma^2 = 2ct$.

The nonlinear heat equation (1.2) may be solved by finite differences. In the simplest case we exchange the time derivative on the left hand side by a forward difference (Euler forward), and derivatives by neighbor differences

$$\frac{s^{t+\tau} - s^t}{\tau} = \left(\begin{bmatrix}1\\-1\end{bmatrix}\right)^T * \left(c\left(\begin{bmatrix}1\\-1\end{bmatrix}\right) * s^t\right) \tag{1.4}$$

and get an update scheme

$$s^{t+\tau} = \begin{bmatrix} 0 & \tau c_{x,y+\frac{h}{2}} & 0 \\ \tau c_{x+\frac{h}{2},y} & 1 - \tau(c_{x+\frac{h}{2},y} + c_{x-\frac{h}{2},y} + c_{x,y+\frac{h}{2}} + c_{x,y-\frac{h}{2}}) & \tau c_{x-\frac{h}{2},y} \\ 0 & \tau c_{x,y-\frac{h}{2}} & 0 \end{bmatrix} * s^t \tag{1.5}$$

or $s^{t+\tau} = (1 + \tau A_{x,y}) * s^t$. This is called an explicit scheme. It boils down to convolution of the signal with a spatially (and temporally) varying kernel $(1 + \tau A_{x,y})$. In the case of spatially constant c this simplifies to

$$s^{t+\tau} = \begin{pmatrix} 0 & \tau c & 0 \\ \tau c & 1 - 4\tau c & \tau c \\ 0 & \tau c & 0 \end{pmatrix} * s^t, \tag{1.6}$$

where A is the so-called 5-point-star times c. This scheme has positive entries only, i.e. is a convex regularizer and features absolute stability, if and only if $\tau c < 0.25$. It becomes instable if $\tau c > 0.5$, as then frequencies at the Nyquist border are amplified by a factor < -1. For small τc the convolution kernel applied to s^t is a reasonable discretization of a Gaussian. The same bounds can be derived by application of the Gershgorin circle theorem on the spectrum which is supposed to be contained within $(-1, 1)$.

Discretizing the left hand side by a backward difference quotient (Euler backward) we get an implicit scheme $(s^t - s^{t-\tau})/\tau = A_{x,y} * s^t$ or equivalently $s^{t+\tau} = (1 - \tau A_{x,y})^{-1} * s^t$ boiling down to a recursive filter applied to s^t.

1.2.2 Anisotropic Diffusion

Anisotropic diffusion typically acts along measured local image orientations (cf. Fig. 1.2). They are described by the structure tensor $\mathbf{J}_\rho$ [36]

$$\mathbf{J}_\rho = \int w_\rho(\mathbf{x}) \nabla s(\mathbf{x}) \nabla^T s(\mathbf{x})\, d\mathbf{x}, \tag{1.7}$$

where $w_\rho(x)$ are Gaussian weights with standard deviation ρ. Being symmetric $\mathbf{J}_\rho$ can be diagonalized, i.e. $\mathbf{M}$ is a diagonal matrix with eigenvalues $\mathbf{M}_{ii} = \mu_i$ and

$$\mathbf{J}_\rho = (e_1, \ldots, e_N)\mathbf{M}(e_1, \ldots, e_N)^T. \tag{1.8}$$

Anisotropic diffusion filtering evolves the initial noisy image $s(x, 0)$ via

$$\partial_t s = \nabla \cdot (\mathbf{D} \nabla s) \tag{1.9}$$

(cf. Eq. (1.1)). $\mathbf{D}$ is the diffusion tensor, a positive definite symmetric matrix, and $s(\mathbf{x}, t)$ is the evolving spatio-temporal image. Diffusion time t is sometimes used as the scale parameter in a scale-space. It should not be confused with the time coordinate x_3 of a 2D image sequence. The diffusion tensor $\mathbf{D}$ usually applied in anisotropic diffusion uses the same eigenvectors e_i as the structure tensor $\mathbf{J}_\rho$ (see Eq. (1.7)). Thus smoothing is applied according to the spatio-temporal image structure. Smoothing strengths along these structures are given by eigenvalues λ_i of $\mathbf{D}$. Given a diagonal matrix $\mathbf{L}$ with $\mathbf{L}_{ii} = \lambda_i$, the diffusion tensor $\mathbf{D}$ is transformed into the image coordinate system given by the eigenvectors e_i:

$$\mathbf{D} = (e_1, \ldots, e_N)\mathbf{L}(e_1, \ldots, e_N)^T. \tag{1.10}$$

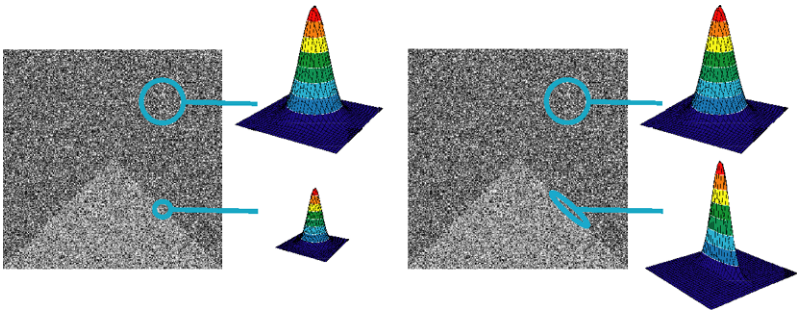


Fig. 1.2 Isotropic nonlinear versus anisotropic diffusion. *Left*: Isotropic nonlinear diffusion reduces diffusivity in all directions when an image structure is present. *Right*: Anisotropic diffusion reduces diffusivities across edges only

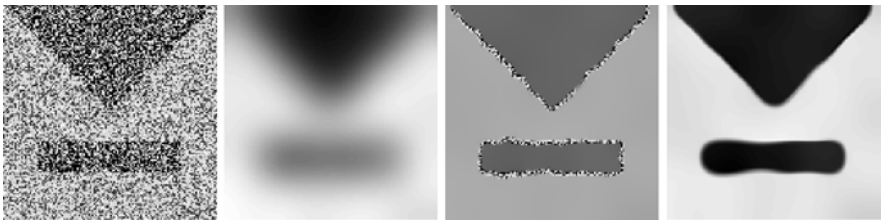


Fig. 1.3 Effect of diffusion filtering. Illustrative examples exaggerating dominating smoothing effects. *From left to right*: Noisy input image, smoothing result with isotropic linear, isotropic nonlinear and anisotropic diffusion

The directional diffusivities $\lambda_i, i \in \{1, \ldots, N\}$ determine the behavior of the diffusion. For image enhancing they shall be high for low values of μ_i and vice versa.

Different possible choices for $\mathbf{L}$ include isotropic linear, isotropic nonlinear and anisotropic processes. There are anisotropic choices with fixed smallest or fixed largest directional diffusivities λ_i and choices where all diffusivities vary. Results demonstrating the different smoothing effects are shown in Fig. 1.3.

Common choices for directional diffusivities are

Isotropic-Linear The standard linear diffusion using a Gaussian kernel corresponds to $D = \alpha \mathbb{1}$, with $\alpha \geq 0$.

Isotropic Non-linear Perona-Malik [334] type diffusion seeks to adapt the smoothing strength to the absolute value of the gray value gradient. Tensor D is given by $D = f(\nabla g)\mathbb{1}$. Among the choices for the diffusivity f the following is given:

$$f(\nabla g) = \exp(-\|\nabla g\|/K). \tag{1.11}$$

The considered diffusivities have in common that they decrease with increasing gradient magnitude. Thus smoothing across edges is prevented.

Edge Enhancing This is basically an anisotropic version of the previous. Following [164] we extend the original 2D formulation [433, 434] to nD as follows:

$$\begin{aligned} \lambda_i = \lambda_2 &:= f(\nabla g) \quad \text{for } i \neq N, \\ \lambda_N &:= 1. \end{aligned} \tag{1.12}$$

The largest diffusivity fixed to 1 enforces strong smoothing in the direction of the corresponding eigenvector even when there is no clear linear structure present.

Coherence Enhancing In this type of diffusion the eigenvalues of the diffusion tensor are chosen as $\lambda_i = \alpha$ for $i \neq N$ and [435]:

$$\lambda_N := \begin{cases} \alpha & \text{if } \kappa = 0, \\ \alpha + (1-\alpha)\exp(\frac{-C}{\kappa}) & \text{else.} \end{cases} \tag{1.13}$$

With a small positive parameter α and the coherence κ measured by:

$$\kappa = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (\mu_i - \mu_j)^2. \tag{1.14}$$

This process is designed to smooth only when there is a large spread in the eigenvalues, enhancing line-like, coherent structures.

Orientation-Enhancing In order to fully exploit the information provided by the structure tensor and to facilitate orientation estimation (= optical flow in image sequences) the eigenvalues of the diffusion tensor can be chosen [369, 431]:

$$\lambda_i := \begin{cases} 1 & \text{if } \mu_i \leq \sigma^2, \\ 1 - \exp(-\frac{c}{(\mu_i - \sigma^2)^2}) & \text{else,} \end{cases} \tag{1.15}$$

where $c > 0$ regulates the transition and σ is related to the noise variance of the image derivatives. This exponential function has been used in [334, 433].

The functions f used to calculate each λ_i are often chosen ad hoc, e.g. selected from robust error functions [41]. As we will see in the next section (Sect. 1.3.1) they may in principle be learned from image statistics. However no energy functional exists for anisotropic nonlinear diffusion, where $\mathbf{D}$ depends on signal s. Therefore in this case generative learning is not possible in a strict sense (cf. [371]). An energy function for anisotropic nonlinear diffusion can be given, if $\mathbf{D}$ in principle should not depend on s, even though s is used to calculate an approximation of the true $\mathbf{D}$ [366]. We derive and use this energy function in Sect. 1.6.1.

1.2.3 Beltrami Framework

The basic concept behind Beltrami flow is to consider an image as a (curved) surface embedded into a higher dimensional space [250]. This concept is frequently used in

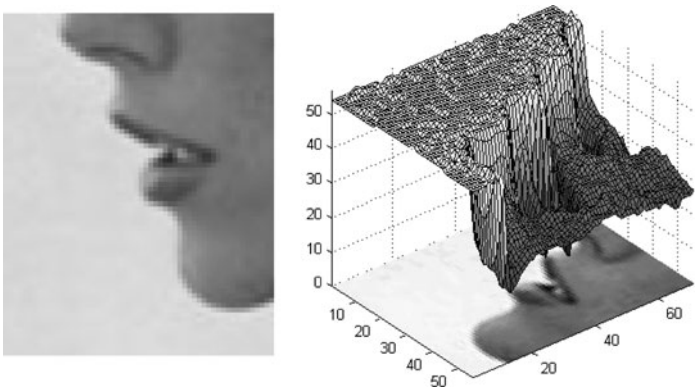


Fig. 1.4 Embedding of a gray value image in a feature space

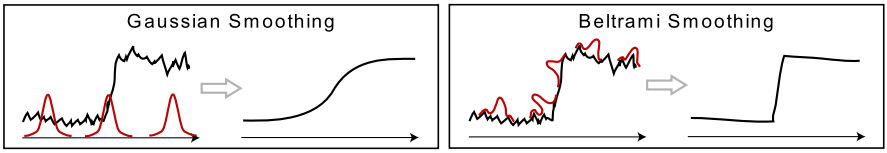


Fig. 1.5 Linear diffusion (i.e. Gaussian smoothing) compared to Beltrami smoothing. Linear diffusion averages a signal respecting spatial distances only. Beltrami smoothing averages a signal respecting distances along the curved manifold

literature, especially when working on spaces with underlying non-flat group structures, cf. Chaps. 5, 8 and 9, and elsewhere [114, 157]. Results from differential geometry are then used to process this surface. We will illustrate this concept by means of denoising gray valued images, however this concept can be generalized to other image types (e.g. color images or tensor valued images) in a straightforward manner. Instead of considering a gray valued image as a function $s(\mathbf{x})$ from the image domain $\Omega \subset \mathbb{R}^2$ into a one dimensional feature space $I \subset \mathbb{R}$, an image is considered as a surface M embedded in the product space (trivial fiber bundle) $E = \Omega \times I$ (cf. Figs. 1.4 and 1.5). The embedding is described by the map $X(\mathbf{x}) = (\mathbf{x}, s(\mathbf{x}))$. In order to be able to define energies on the image, M is considered as a Riemannian manifold: at each tangent space $T_{\mathbf{x}}(M)$ on a manifold M a (positive definite) inner product $g(\mathbf{x}) : TM \times TM \to \mathbb{R}$ is given by $g(\mathbf{x}) = g_{ij}(\mathbf{x})\, dx^i \otimes dx^j$, in which $\otimes$ denotes tensor outer product. The metric defines the length of an "infinitesimal line element" via $ds^2 = g_{ij}(\mathbf{x})\, dx^i\, dx^j$ as well as the "infinitesimal volume element" $\sqrt{|g|}\, dx\, dy$, where $|g(\mathbf{x})| = \det[g_{ij}(\mathbf{x})]$. The geometrical framework allows to relate the metric of the embedding space E with the metric of the image surface M via the so-called *pullback metric* $g_{ij} = h_{uv} \partial_i X^u \partial_j X^v$ assuring that infinitesimal distances defined by h_{uv} in E equal infinitesimal distances in M. If we consider the usual Euclidean metric in Ω and I, the metric in the embedding space reads $d\ell^2 = dx^2 + dy^2 + ds^2(x, y)$ leading to following pullback metric on M: $g_{11} = 1 + (\partial_x s)^2$, $g_{12} = g_{21} = \partial_x s \partial_y s$ and $g_{22} = 1 + (\partial_y s)^2$. Based on this

mathematical structure an image can now be characterized by an energy based on distances on the image surface M. For instance, a denoised image may have minimal surface, i.e. the denoised image minimizes the energy

$$S(M) = \int \sqrt{|g|}\, dx\, dy. \tag{1.16}$$

The corresponding diffusion equation can then be obtained with calculus of variation, i.e. setting the negative functional derivative of the energy functional equal to the temporal derivative of the signal (cf. Eq. (1.22))

$$\partial_t s = \operatorname{div}\left(\frac{1}{\sqrt{|g|}}\nabla s\right) \tag{1.17}$$

which is isotropic nonlinear diffusion (or Perona Malik diffusion [334]) with the edge stopping function $\phi(x) = 1/\sqrt{|g|}$. The geometrical framework thus allows to relate an ad hoc chosen edge stopping function with a metric of the image surface M. However the choice of a suitable edge stopping function has only been shifted to the choice of a suitable metric (in the embedding space). A more general energy functional has been proposed [250]:

$$S(X^u, g_{ij}, h_{uv}) = \int \sqrt{|g|} g^{ij} h_{uv} \partial_i X^u \partial_j X^v\, d^n x, \tag{1.18}$$

where n denotes the dimension of the image domain, h_{uv} the embedding space metric and g_{ij} the metric of the image manifold. Depending on the interpretation of the different entities, different well known diffusion schemes can be reproduced by this energy functional. These include the reparameterization invariant linear scale-space by Florack et al. [155], Perona Malik anisotropic diffusion [334], Mumford-Shah segmentation models [316], Rudin-Osher-Fatemi total variation TV method for image enhancement based on the L1 norm [364], and the different Blake-Zisserman membrane models [42]. Also diffusion schemes for vector valued images or images whose feature space constitute itself a nonlinear manifold arise in a natural way, cf. Sect. 1.5.2 as well as Chaps. 3, 5, 8, 9, as well as [114].

1.3 Diffusion and Image Statistics

Diffusion and diffusion-like methods can be derived from image statistics, probability distribution functions (PDFs), or energy functionals. This section introduces the main statistical concepts needed, pathways to diffusion and extensions to diffusion.

A considerable advantage of a statistical point of view is an airtight justification of noise reduction. In scientific applications changing measured data e.g. by denoising it or by rejecting outliers is only allowed if such a change improves the data. Obviously, improving means to optimize some criterion which needs to be given explicitly. What is more, the criterion must be the right one for the data-set at hand.

However, what is the right criterion and in what sense is it right? From a statistical point of view the denoised or otherwise reconstructed data should be the most likely one given the data and prior knowledge! Consequently we formulate suitable probabilities and show in which way optimization schemes correspond to diffusion.

1.3.1 From Probability Distributions to Diffusion

An isotropic nonlinear diffusion process can be derived from an energy function, that itself may be derived from a probability distribution. The smooth image s is the maximizer of the posterior probability distribution $p(s|r)$, i.e. the probability that $s := s(\cdot, t) : \mathbb{R}^N \to \mathbb{R}$ for some fixed $t > 0$ is the desired smooth image when $r : \mathbb{R}^N \to \mathbb{R}$ has been observed

$$\hat{s} = \arg\max_s p(s|r) \quad \text{with } p(s|r) \propto \prod_i (p(r_i|s_i)\, p(\|\nabla s_i\|)). \tag{1.19}$$

The sampling distribution (likelihood function for fixed r_i) $p(r_i|s_i)$ at every pixel i, with $s_i = s(\mathbf{x}_i)$, may be defined by a measured image statistics, i.e. a normalized histogram of observed noise. In image processing it typically is modeled to only depend on intensity differences $\varepsilon_i = r_i - s_i$, i.e. on measurement noise.[1] The spatial term $p(\|\nabla s_i\|)$ formulates prior knowledge about the solution s. It exploits a Markov Random Field (MRF) assumption [183], which defines the prior in terms of local neighbor properties. For a 1D signal the assumption used here is that if we know a signal value s_i at a position i, then we can give a probability to observe a certain s_{i+1} at neighbor position $i + 1$, and vice versa.

Please note that the likelihood term depends on measured data r and is therefore often called *data term*. The prior term only depends on the sought for smooth solution s and is therefore often called *smoothness term*.

We may interpret $p(s|r)$ to be a Gibbs distribution

$$p(s) = \frac{1}{Z} e^{-E(s)},$$

where E denotes the energy corresponding to p and Z is the partition function normalizing the integral of p over all s. Maximizing $p(s|r)$ is equivalent to minimizing its energy, i.e. its negative logarithm

$$\hat{s} = \arg\min_s E(s) \quad \text{with } E(s) = -\sum_i (\rho_0(s_i - r_i) + \lambda \rho_1(\|\nabla s_i\|)), \tag{1.20}$$

where we used the notation $\rho(x) = -\log p(x)$, added indices to stress that sampling and prior are different distributions, and introduced weight λ which accounts for the

[1] Considering e.g. Poisson or shot noise and low intensities, this is not a good approximation.

confidence one has in the different model terms. The smoothness term in (1.20) can be interpreted as nonlinear isotropic diffusion [41, 375]. As diffusion is defined in continuous domain, we rewrite the smoothness term as energy functional

$$E(s) = \int \rho(\|\nabla s\|)\, dx. \tag{1.21}$$

We denote with δE the functional derivative of E if the differential

$$\langle \delta E(u), w \rangle := \lim_{\varepsilon \to 0} \frac{E(u + \varepsilon w) - E(u)}{\varepsilon}, \tag{1.22}$$

exists for all test functions $w : \mathbb{R}^N \to \mathbb{R}$. The functional derivative can be seen as a generalization of the gradient of a multivariate function in vector calculus to a functional defined on a function space. Consequently, $\delta E = 0$ is a necessary condition for a minimizer of E, known as the Euler-Lagrange equation. If we embed the signal into a 1-parameter family $s : \mathbb{R}^N \times \mathbb{R}^+ \to \mathbb{R}$, then the stationary point can be interpreted as the steady state solution, if it exists, of the following evolution equation:

$$\langle \partial_t s(\cdot, t), w \rangle = -\langle \delta E(s(\cdot, t)), w \rangle, \tag{1.23}$$

which, in a weak sense, amounts to the following gradient flow PDE:

$$\partial_t s = \operatorname{div}(\psi(\|\nabla s\|)\nabla s) \quad \text{with } \psi(\alpha) = \rho'(\alpha)/\alpha. \tag{1.24}$$

This means that nonlinear isotropic diffusion can be interpreted in terms of classical or Bayesian statistics. The term ρ is denoted as an *error norm* in the context of classical robust statistics and *potential function* in a Bayesian interpretation. In a robust statistical approach, the main data is assumed to follow a certain distribution, e.g. a Gaussian distribution and there are a few outliers whose distribution is not explicitly known. The challenge is to choose an error norm such that outliers do not influence the estimate. An energy function in a robust statistical approach does not necessarily belong to a valid probability distribution. In contrast to this in the Bayesian approach the complete probability distribution of main data and outliers is modeled by a probabilistic distribution, i.e. the robust error norm directly follows from the statistical properties of the complete data. A collection of well-known error norms and the corresponding diffusivities[2] are depicted in Fig. 1.6.

1.3.2 Gibbs Reaction-Diffusion

The prior term above is based on the absolute value of the image gradient $|\nabla s|$. This choice is ad hoc and the partial derivatives in $|\nabla s|$ may be exchanged by a set of

[2]Diffusivities calculated by e.g. Tuckey or Cup functions may become 0 and thus a diffusion tensor based on them is not guaranteed to be positive definite, but positive semi-definite.

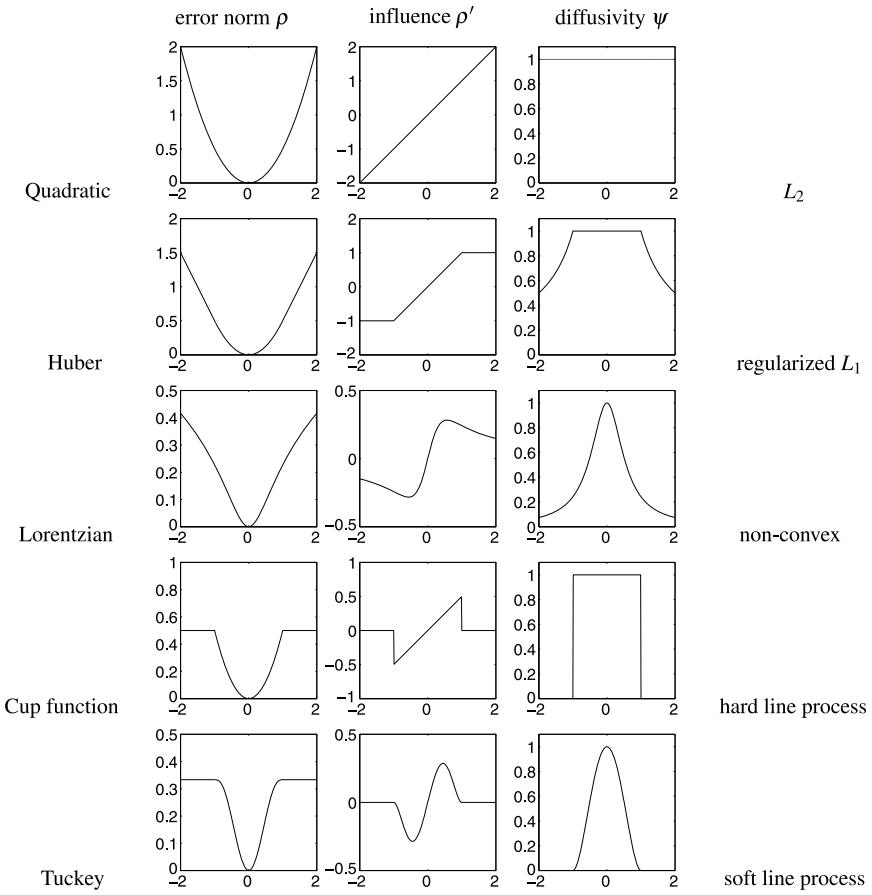


Fig. 1.6 Different (robust) error norms

linear filters. Zhu and Mumford [465] proposed to use Gibbs distributions of the form $p(s, R, F) = \frac{1}{Z} e^{-E(s,R,F)}$, where $R = \rho_1, \ldots, \rho_J$ and F is the set of filters applied, $F = \{F_1, \ldots, F_J\}$, Z a normalization factor making p integrate to 1 and

$$E(s, R, F) = \sum_i \sum_{j=1}^{J} \rho_j (F_j * s_i), \tag{1.25}$$

where $*$ denotes convolution. Filters F_j on different scales are used and the respective ρ_j are learned from training data. Zhu and Mumford observe that ρ-functions are well modeled by $\rho(\xi) = a(1 - (1 + (|\xi - \xi_0|/b)^{-\gamma})^{-1})$ for different parameters a, b, ξ_0, and γ. When $a > 0$ we get a typical potential function with minimum at ξ_0. In most cases one gets $a > 0$ for filters like ∇ or Δ which capture the general smoothness of an image (cf. Fig. 1.7, left). Interestingly for filters characterizing prominent features, e.g. Gabor filters at various orientations and scales one gets

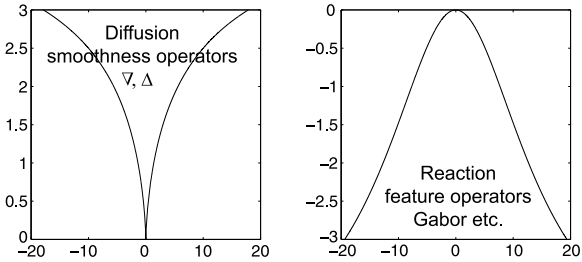


Fig. 1.7 Typical ρ-functions as derived by Zhu and Mumford for diffusion and reaction terms

$a < 0$, i.e. destabilizing behavior, resulting e.g. in edge-enhancement (cf. Fig. 1.7, right). Zhu and Mumford call these terms *reaction*, while the smoothing terms are called *diffusion*. All these terms are generalized isotropic nonlinear diffusions in our nomenclature. Roth and Black [361] propose a framework for also learning the filters F.

1.3.3 Steerable Random Fields and Anisotropic Diffusion

Following [371] anisotropic diffusion with a diffusion tensor can be derived using Zhu and Mumford's [465] approach (see Sect. 1.3.2). With the special filter choice $F = \mathbf{n}_1\nabla, \ldots, \mathbf{n}_J\nabla$, i.e. directional derivatives along the normalized vectors $\mathbf{n}_j$ we get the posterior probability (cf. Eqs. (1.19) and (1.25))

$$\hat{s} = \arg\max_s p(s|r) \quad \text{with } p(s|r) \propto \prod_i \left(p(r_i|s_i) \prod_{j=1}^{J} p_j(\mathbf{n}_j \nabla s_i) \right). \tag{1.26}$$

Maximizing $p(s|r)$ is equivalent to minimizing its negative logarithm, i.e. the energy

$$\hat{s} = \arg\min_s E(s) \quad \text{with } E(s) = -\sum_i \left(\rho_0(s_i - r_i) + \lambda \sum_{j=1}^{J} \rho_j(\mathbf{n}_j \nabla s_i) \right), \tag{1.27}$$

where we used $\rho_0(s_i - r_i) = -\log p(r_i|s_i)$ and $\rho_j(\mathbf{n}_j\nabla s_i) = -\log p_j(\mathbf{n}_j\nabla s_i)$, added indices to stress that likelihood and prior are different distributions, and introduced weight λ which accounts for the confidence one has in the smoothness terms (as before in Sect. 1.3.1). We set up a gradient descent minimization scheme using the functional derivative of E (cf. Eq. (1.22))

$$\partial_t s = -\rho_0'(s_i - r_i) + \lambda \nabla^T \sum_i \sum_{j=1}^{J} \psi_j(\mathbf{n}_j \nabla s_i) \mathbf{n}_j \mathbf{n}_j^T \nabla s_i \quad \text{for all } i \tag{1.28}$$

with $\psi_j(\alpha) = \rho_j'(\alpha)/\alpha$. Comparing the second term in Eq. (1.28) with anisotropic diffusion (Eq. (1.1)) reveals the relation between the diffusion tensor

$$\mathbf{D} = \sum_{i,j} \psi_j(\mathbf{n}_j \nabla s_i)\mathbf{n}_j\mathbf{n}_j^T \tag{1.29}$$

and the derivatives of the potential functions $\psi_j(\mathbf{n}_j \nabla s_i)$. Consequently, the diffusion tensor can be learned from training data.

Unfortunately the diffusion tensor from Eq. (1.28) is not constructed from a structure tensor as commonly done (cf. Sect. 1.2.2). Defining the prior as $\prod_j p(\mu_j)$, where μ_j are the eigenvalues of the structure tensor sorted by size yields a scheme similar to structure-tensor-based anisotropic diffusion [371]. However, as eigenvalues of the structure tensor are *smoothed* squared directional derivatives, the gradient in the spatial term of the diffusion operator (rightmost ∇ in Eqs. (1.1) and (1.28)) is also smoothed. If this smoothing is taken out of the structure tensor, the resulting scheme reduces to isotropic nonlinear diffusion [371]. This problem can be circumvented, when only the orientation of the diffusion is derived via the structure tensor, but directional diffusivities depend on unsmoothed directional derivatives [362]. Again this is not structure-tensor-based anisotropic diffusion, but called steerable random fields. An energy functional yielding structure-tensor-based anisotropic diffusion can be derived in a strict sense [366], however the structure tensor then is only used as a proxy for an orientation tensor coming from a linear model. This tensor does not depend on the signal s. We show this approach in Sect. 1.6.1.

1.3.4 Robust Statistics and Kernel Estimation

In the last section, we discussed the relation between probabilistic and diffusion based denoising methods. In particular we showed that the prior term in a Bayesian approach is directly linked to the energy of a robust estimator which then leads to classical diffusion schemes. Examining the likelihood term in (1.19) in the same manner reveals similar relations between different denoising methods as shown next. Let us consider several observations r_j in a local neighborhood. Assuming that all these observations belong to the same signal value[3] corrupted with identical independently distributed Gaussian noise, the corresponding likelihood function reads

$$p(\{r_j\}|s) \propto \exp\left\{-\frac{1}{2\sigma^2}\sum_j (s - r_j)^2\right\} \tag{1.30}$$

with the corresponding energy

$$E(s) = \sum_j (s - r_j)^2. \tag{1.31}$$

[3] In the case of an image, where r_j are spatially distributed on the pixel grid this is of course a smoothness assumption on the underlying signal s.

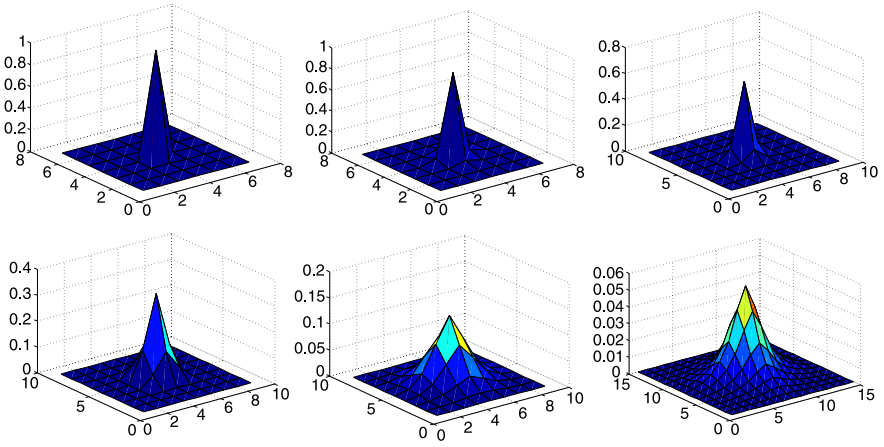


Fig. 1.8 Optimal spatial filter masks in a data term depending on an assumed noise level. *From left to right and top to bottom*: Noise level 1, 5, 10, 20, 50, 100

The maximum likelihood estimator is thus given by the *linear least squares estimator*, i.e. the signal is estimated by the mean of all observations. The assumption of a constant signal model might be too restrictive and one is attempted to give pixel values closer to a position x_k of interest more weight leading to the *weighted least squares estimator*

$$E(s_k) = \sum_j w_x(x_k - x_j)(s_k - r_j)^2 \quad \Rightarrow \quad s_k = \frac{\sum_j w_x(x_k - x_j) r_j}{\sum_j w_x(x_k - x_j)}, \tag{1.32}$$

where x_j and x_k denote the (pixel) position of r_j and s_k, respectively, and w_x is a weight function. The weighted least squares scheme here is usually implemented by a convolution with a (normalized) smoothing kernel w_x, equivalent to linear diffusion if a Gaussian kernel is used (cf. Sect. 1.2.1). Optimal weights depending on pixel position can be chosen from statistical characteristics of the model error (a constant signal here) and are typically not of Gaussian shape (cf. [267] and Fig. 1.8).

If the observed signal contains outliers, i.e. either the signal model or the noise model is severely violated, the estimator can be made robust. Such a *robust estimator* is obtained by exchanging the quadratic energy function by a robust error metric, i.e. the corresponding estimator is the minimum of the energy

$$E(s_k) = \sum_j \rho(s_k - r_j). \tag{1.33}$$

It may be minimized by iterating

$$s_k^{t+1} = \frac{\sum_j w_s(s_k^t - r_j) r_j}{\sum_j w_s(s_k^t - r_j)}, \tag{1.34}$$

Likelihood (data term)	**Prior (smoothness term)**
• May be learned from data.	• May be learned from data.
• Smoothness assumption on signal only if r spatially distributed.	• Smoothness assumption on signal.
• Term contains measured data r and solution s.	• Term contains only solution s.
• Initialized with mean or r (typically).	• Initialized with r, if only one term.
• Solution non-constant.	• Solution (piece-wise) constant.
• Scheme iterated till convergence.	• Only few iterations, then stopped.
• Variance of spatial kernel constant wrt. $h \Rightarrow 0$, $\tau \Rightarrow 0$.	• Variance of spatial Gaussian decreases with $h \Rightarrow 0$, $\tau \Rightarrow 0$ ($\sigma^2 = 2c\tau$).
• Consistency check: No diffusion.	• Consistent diffusion.

Fig. 1.9 Summary: Likelihood versus prior. First iteration may be identical in both cases

where $w_s(u) = \rho'(u)/u$, and the upper index t is the iteration number. Please note that if ρ is a negative Gaussian $\rho(u) = -\exp(-u^2)/2$, then $\rho'(u) = u\exp(-u^2)$ and $w_s(u) = \exp(-u^2)$ is a Gaussian as well.

In a Bayesian framework, the error metric is interpreted as a potential function that directly encodes the statistical properties of signal model and noise. As for the weighted least squares estimator, we may introduce a further weight w_x reducing the influence of estimates being further away from the central position x_k. The minimizer of the corresponding energy

$$E(s_k) = \sum_j w_x(\mathbf{x}_k - \mathbf{x}_j)\rho(s_k - r_j) \tag{1.35}$$

is denoted as the robust M-smoother in literature (cf. e.g. [75, 449]). This energy may be minimized by iterating

$$s_k^{t+1} = \frac{\sum_j w_x(\mathbf{x}_k - \mathbf{x}_j) w_s(s_k^t - r_j) r_j}{\sum_j w_x(\mathbf{x}_k - \mathbf{x}_j) w_s(s_k^t - r_j)}. \tag{1.36}$$

The equivalence to nonlinear diffusion becomes apparent, when we select a spatial neighborhood consisting of nearest neighbors only for w_x and start with $s_k^0 = r_k$. The first iteration step then is equivalent to a diffusion step, where w_s models diffusivities. This equivalence can also be shown for larger neighborhoods [20, 21].

Please note When we start with a smoothness or constancy assumption formulated on the measured data r, we construct a *data term*. In the respective estimation schemes r is *never updated*. When we start with a similar assumption on the underlying signal s as in Sect. 1.3.1, we get diffusion schemes, where r_j stands for the initial value s_j^0 and is updated. Here, we introduced spatial weights ad hoc, in Sect. 1.3.1 they come from an MRF assumption.

In Fig. 1.9 we summarize properties of likelihood-based data terms and prior-based smoothness terms.

1.4 Some Diffusion-Like Nonlinear Regularization Schemes

There are several regularization techniques using averaging kernels, usually Gaussians, together with a nonlinearity (for overviews see e.g. [55, 75, 449] as well Chap. 4 in this volume). Local M-smoothing [75, 449] and related robust statistics-based methods have already been shown in Sect. 1.3.4. We only show two prominent examples here: bilateral filtering [409] also called cascaded nonlinear Gaussian filtering [17, 187, 445], and channel smoothing [139], cf. Chap. 2. They are closely connected to nonlinear isotropic diffusion. A false friend in the list of diffusion-like methods is mean shift filtering [81], a mode-seeking method. Iterations occurring in this approach are similar to diffusion, but only the first iteration really is equivalent to a diffusion step.

1.4.1 Bilateral Filtering and Nonlinear Gaussian Filtering

Bilateral Filtering [409] as well as nonlinear Gaussian filtering [17, 187, 445] operate on the input data given by $r(\mathbf{x})$ and filter it via

$$\hat{s}(\mathbf{x}) = k^{-1}(\mathbf{x}) \sum_{\boldsymbol{\xi}} w_1(\mathbf{x} - \boldsymbol{\xi}) w_2(r(\mathbf{x}) - r(\boldsymbol{\xi})) r(\mathbf{x}), \tag{1.37}$$

$\mathbf{x} \in \mathbb{R}^N$, where $k(\mathbf{x}) = \sum_{\boldsymbol{\xi}} w_1(\mathbf{x} - \boldsymbol{\xi}) w_2(r(\mathbf{x}) - r(\boldsymbol{\xi}))$ is a normalization, and $\hat{s}(\mathbf{x}) : \mathbb{R}^N \to \mathbb{R}$ is the filtered image. The filter weights w_1 and w_2 may be Gaussians as suggested in [17, 187, 409, 445] but other filter weights may also be applied. Positions $\mathbf{x}_i$ may be restricted to a spatial local neighborhood with size depending on the standard deviation σ_1 of w_1, typically $3\sigma_1$.

Equation (1.37) is simple linear Gaussian smoothing if the second kernel $w_2 \equiv 1$ and thus a direct solution of the heat equation (1.3) for one given time step, cf. Sect. 1.2.1. The second kernel $w_2(r(\mathbf{x}) - r(\xi))$ down-weights the contribution of a value $r(\boldsymbol{\xi})$ if it differs from the value $r(\mathbf{x})$ at the current position $\mathbf{x}$. This is equivalent to reducing the diffusivity between the points $\mathbf{x}$ and $\boldsymbol{\xi}$. Iterating Bilateral Filtering by applying it to the filtered data is therefore similar to isotropic nonlinear diffusion in relatively coarse time steps. An investigation based on a detailed analysis of discrete schemes also shows this connection between diffusion and bilateral filtering [20, 21].

So-called cascaded nonlinear Gaussian filtering changes standard deviations σ_1 and σ_2 of w_1 and w_2, respectively, in every iteration step. Typically one starts with small σ_1 (space) and large σ_2 (range) and doubles σ_1 while halving σ_2 in every iteration step.

We will now check for numerical consistency of bilateral filtering with isotropic non-linear diffusion. It is not sufficient to compare discrete schemes in order to decide whether or not bilateral filtering is a consistent numerical scheme for isotropic nonlinear diffusion. We need to known how it behaves in the limit of continuous

signals s and r. It is clear that if we only go to continuous domain, without associating r with the initial (i.e. $t = 0$) signal $s(x,t)|_{t=0}$ this scheme cannot become diffusion—the diffusion equation contains no r. In this case the energy associated with the respective likelihood term reads

$$E(s_k) = \sum_j w(\mathbf{x}_k - \mathbf{x}_j)\rho(r_k - r_j)(s_k - r_j)^2 \tag{1.38}$$

with $r_j = r(\mathbf{x}_j)$, which can be interpreted as the energy function of a Gaussian distribution with the precision matrix $\Lambda_{kj} = w(\mathbf{x}_k - \mathbf{x}_j)\rho(r_k - r_j)$. Rewriting Eq. (1.37) as

$$\hat{s}(\mathbf{x}, t+\tau) = k^{-1}(\mathbf{x}) \sum_{\boldsymbol{\xi}} w_1(\mathbf{x} - \boldsymbol{\xi}) w_2(s(\mathbf{x}, t) - r(\boldsymbol{\xi}))s(\mathbf{x}, t) \tag{1.39}$$

we do not end up with diffusion either, but perform robust averaging of multiple measurements r. This update scheme corresponds to a robust M-smoother.

Only if we exchange also $r(\xi)$ by $s(\xi, 0)$, we may end up with diffusion if we do the limiting process right

$$\hat{s}(\mathbf{x}, t+\tau) = k^{-1}(\mathbf{x}) \sum_{\boldsymbol{\xi}} w_1(\mathbf{x} - \boldsymbol{\xi}) w_2(s(\mathbf{x}, t) - s(\boldsymbol{\xi}))s(\mathbf{x}, t). \tag{1.40}$$

This update scheme is proposed in [409]. If the variance of kernel w_1 decreases with selected time step τ and τ is small enough, only nearest neighbors need to be addressed. Comparing Eq. (1.40) with the update scheme in Eq. (1.5) reveals that w_2 may be interpreted as diffusivity c in that scheme. This interpretation of bilateral filtering is consistent with isotropic nonlinear diffusion.

1.4.2 Mean Shift Filtering

Noise reduction may be regarded as estimation of the most likely measurement value (or other feature) at a given position. Density of features in feature space may be regarded as empirical probability function (PDF) of the represented parameter. The modes of a feature space, i.e. maximal dense regions, may therefore be identified as local maxima of the unknown PDF. Mean shift filtering locates maxima, i.e. stationary density points by gradient ascent without estimating the density. How does this work?

A kernel density estimator in a flat space is given by

$$\hat{f}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} K(\mathbf{x} - \mathbf{x}_i), \tag{1.41}$$

where $\hat{f}$ is the estimated density, n data points $\mathbf{x}_i \in \mathbb{R}^N$, $i \in \{1, \ldots, n\}$, are given and K is some normalized smoothing kernel typically radially symmetric $K(\mathbf{x}) =$