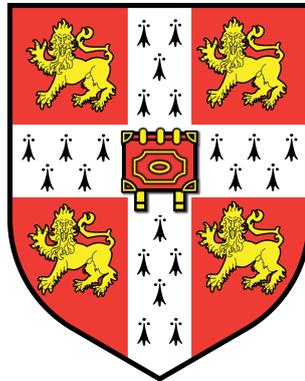


New PDE models for imaging problems and applications



Luca Calatroni

Hughes Hall College,

Cambridge Centre for Analysis (CCA), Faculty of Mathematics

University of Cambridge

A thesis submitted for the degree of

Doctor of Philosophy

September 2015

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text. In particular, in Section 1.5 a detailed description of the original contribution of this thesis and of the work which has been done in collaboration is given.

Luca Calatroni

Al mio Angelo

Acknowledgements

In the following, I would like to thank all the people who made my last four years exceptional and unique. The path towards the completion of my PhD has been very intense and memorable from both a professional and a personal point of view. Here below I will list some of the main protagonists of this incredible experience.

First among anyone else, an enormous thank you goes to Carola Schönlieb, my PhD supervisor, friend and guiding light in all these years. My thanks are for all the passion, motivation and intelligence that you have put in your work with me and for having been the leading figure of my professional growth during my PhD. You have carefully followed and supported my research experiences, giving me the extraordinary opportunities to present my work everywhere in the world and to meet leading experts in our field at several international conferences. But not only that, Carola. You have helped me in so many ways: from not leaving me alone in an Austrian hospital till being an incredible host when needed. Much more than that, you have kept my spirits up in a very difficult moment when my motivations and my concentration were lost. Thanks for believing in me and for having always followed my progresses, giving me always the chance to improve and reinforce my skills. My future career and (hopefully!) successes will always be a reflection of your guidance and help.

During my PhD I had the extraordinary opportunity to go to Ecuador to work with Juan Carlos De Los Reyes, the director of the Centro de Modelización Matemática (ModeMat) and to initiate with him a fruitful and stimulating collaboration in the wonderful scenario of Quito. To you, Juan Carlos, I want to devote a huge thank you for being such an incredible person, mathematician, leader and friend. You helped me a lot with your kindness in making me feel home so far away. Alongside with you I want to thank Estefy, Sofía, Myri, Majo, Pedro, Sergio and Luis Miguel for being my on-site family and for making me feel part of the group since the very first day.

A special thank you goes also to all my other collaborators for their dedication and help in our work together. In particular I want to thank Bertram Düring and Martin Benning for our nice and friendly collaboration and to Yves van Gennip for his commitment and competence in our last research project. From

you all I have learnt a lot and I hope there will be chances in the future to work together one more time.

I deeply want to thank also Prof. John Aston and Prof. Martin Burger for the extremely useful suggestions and discussion we had during my viva examination which, despite its name, was probably one of the most rewarding day of my entire academic life.

I would also like to thank all the members of the Cambridge Image Analysis group for creating such an incredibly stimulating work environment where I could broaden my knowledge in the field.

I am also deeply grateful to the Cambridge Centre for Analysis crew. The path towards the PhD is very long and, often, very frustrating. Spending these years of work in Cambridge with my wonderful colleagues has been fundamental (whenever I was there). In particular, Eoin, I will miss a lot our afternoon discussions about our work and our plans for the future and I will miss our lunch breaks in the very far common room of Pavilion D with you my friends Gil, Milana, Lukas, Clarice and Kevin. Having you as part of my professional and personal growth has been incredibly helpful especially in our hectic first year as well in our last one where we pushed each other towards a brilliant conclusion of our period together and the achievement of even brighter successes. I wish you all a wonderful career and a life that could give you all the professional and personal achievements you deserve.

A deep thank you to my Mediterranean connections in CMS. Thanks Kostas for playing the part of my older brother during these years, giving me always wise advices which helped me a lot in keeping my Cambridge life on track. Similarly, thanks to you Sara, Amalio, Nayia and Vaggelis to filling these grey UK years with the colour of your lives and genuine southern emotions at every cigarette break, dinner and party we have been together.

Nothing of what I have mentioned above would have been possible without the support and the unconditional love of my family. I want then to express an immense thank you to my mother, Ornella, for having been in every single moment the person I could rely on. Without knowing any mathematics, any English, any of my everyday struggle about research and work life, you have always known the right words to tell me to raise me up in my downs. During these years, your strength has been unimaginable, and my success in this path is largely your success since without your help all this would not have been achieved. Thanks to my brother Stefano who followed my path and always believed in me. Thanks to all the rest of my family who has always been watching me as ‘the genius’ making me feel always happy with their genuine and authen-

tic love whenever I was back. Among you, let me express a particular thank you to my cousin, Elena, for making me understand how to prioritise my life from a completely different point of view after her extraordinary example of strength and determination she gave me in these months: you will get back the piece of life you have been stolen and you will shine and smile even more than you used to do!

My experience in Cambridge has been made special and unforgettable by a number of people with whom I have shared my worries, successes and, mainly, I have enjoyed my free time. Accidentally, they are all Italians and they are all building up their brilliant career far from our country which I hope one day will welcome them back. Among all of them, a special thanks goes to Barbara for being my confessor and talking cricket. Thanks to Stefania for her random life which has made me think that things can always be worse, to Enrico for being always on my same wavelength, to Giulia for her wisdom and sensitivity and to Alice and Elena for their constant help and warm hospitality. Further, thanks to Olivia, Zuni, Enrico, Federico, Alessandra, Francesco, Andrea, Antonio, Rocco, Mauro, Carlo, Alessandro, Anna and all the others. Each one of you left a deep sign in my life and showed me how interesting and ambitious people can be. You are all models for me, my friends, and I wish we can all achieve what we have been fighting for in all these years.

Thanks to my emigrated friends Paolo, Milly and Ciccio for being the most solid friendships in my life. Wherever and whenever needed, you have always been close to me, understanding my weaknesses and strengthening my personality with your intelligence. I am so proud of you my friends! Thanks to Teresa for being the light of my professional and personal path in these years and for compensating my anxiety and stresses with her smile. Thanks to Mara, Eleonora and Alessandra, my mathematician friends, for having followed and supported all the changes I have been through during these years and for still being able to make me laugh by inventing terrible mathematical songs. Thanks to my whole old college crew for making me feel part of a squad still after so many years and to Federico, Marco and all the others for always waiting for me whenever I am back and for tolerating my no-shows. Thanks to Sabina, my favourite flute player, for let me think again about my future as a musician and for motivating me with her passion and incredible talent.

And finally, thanks to you, Dad, for what you have taught me with your smile, your patience, your curious eyes and your tenderness. Even though you are the only absent in this moment, your presence is stronger than the one of anyone else and your hand will be holding mine wherever I will go.

Abstract

Variational methods and Partial Differential Equations (PDEs) have been extensively employed for the mathematical formulation of a myriad of problems describing physical phenomena such as heat propagation, thermodynamic transformations and many more. In imaging, PDEs following variational principles are often considered. In their general form these models combine a regularisation and a data fitting term, balancing one against the other appropriately. Total variation (TV) regularisation is often used due to its edge-preserving and smoothing properties. In this thesis, we focus on the design of TV-based models for several different applications.

We start considering PDE models encoding higher-order derivatives to overcome well-known TV reconstruction drawbacks. Due to their high differential order and nonlinear nature, the computation of the numerical solution of these equations is often challenging. In this thesis, we propose directional splitting techniques and use Newton-type methods that despite these numerical hurdles render reliable and efficient computational schemes.

Next, we discuss the problem of choosing the appropriate data fitting term in the case when multiple noise statistics in the data are present due, for instance, to different acquisition and transmission problems. We propose a novel variational model which encodes appropriately and consistently the different noise distributions in this case.

Balancing the effect of the regularisation against the data fitting is also crucial. For this sake, we consider a learning approach which estimates the optimal ratio between the two by using training sets of examples via bilevel optimisation. Numerically, we use a combination of SemiSmooth (SSN) and quasi-Newton methods to solve the problem efficiently.

Finally, we consider TV-based models in the framework of graphs for image segmentation problems. Here, spectral properties combined with matrix completion techniques are needed to overcome the computational limitations due to the large amount of image data. Further, a semi-supervised technique for the measurement of the segmented region by means of the Hough transform is proposed.

Keywords: Total variation, higher-order PDEs, directional splitting, quasi-Newton methods, image denoising, image inpainting, mixed noise distribution, bilevel optimisation, parameter learning, SemiSmooth Newton methods, image segmentation, graph clustering, matrix completion, Hough transform.

Contents

List of Notation	17
1 Introduction	21
1.1 What is image processing?	21
1.1.1 The mathematics of digital images	22
1.1.2 Discrete VS. continuous modelling	25
1.1.3 Some imaging tasks	27
1.2 Mathematical image reconstruction techniques	28
1.2.1 The Bayesian framework	29
1.2.2 The variational approach	30
1.2.3 The PDE approach	31
1.3 Total variation models	32
1.3.1 Motivation	32
1.3.2 Nonsmooth regularisation	33
1.3.3 Previous work	35
1.4 Choosing the ingredients for a good model	36
1.4.1 Improved reconstructions with higher-order models	37
1.4.2 Fitting the data	39
1.4.3 Regularisation VS. data fitting: how to balance?	39
1.5 Contribution	41
1.5.1 Efficient solution of higher-order TV models	42
1.5.2 The TV-IC variational model for image denoising	43
1.5.3 Learning from examples	44
1.5.4 Image segmentation via graph-clustering	46
1.6 Organisation of the thesis	47
2 Mathematical preliminaries	49
2.1 Functional analysis	50
2.1.1 Basic properties of Radon measures	50
2.1.2 Functions of bounded variation	53

2.1.3	Relaxed functionals and Γ -convergence	55
2.2	Derivatives and subgradients	58
2.3	Computational approaches to TV models	59
2.3.1	Primal formulation	60
2.3.2	Dual formulation	62
2.3.3	Primal-dual formulation	63
2.4	PDE-constrained optimisation	65
2.4.1	Characterisations of solutions	66
2.4.2	Numerical methods for PDE-constrained problems	68
3	Numerical solution of higher-order TV flows	77
3.1	Two non-standard TV gradient flows	78
3.1.1	The TV gradient flow in the space H^{-1}	78
3.1.2	The L^2 -Wasserstein TV gradient flow	79
3.1.3	Related work	81
3.1.4	Preliminaries	82
3.2	Alternating Directional Implicit (ADI) splitting schemes	84
3.2.1	The Peaceman-Rachford scheme	85
3.2.2	The Douglas-Hundsdorfer scheme	85
3.2.3	Additive-Multiplicative Operator Splitting (AMOS) schemes	86
3.3	Solving $(\text{TV}-H^{-1})$ by ADI splitting	88
3.3.1	The linear case: the biharmonic equation	88
3.3.2	The nonlinear case: the $\text{TV}-H^{-1}$ model	90
3.4	Primal-dual formulation of $(\text{TV}-\text{Wass})$	99
3.4.1	Formulation of the problem	99
3.4.2	A damped Newton system	100
3.5	Numerical results	101
3.5.1	Numerical solution of $(\text{TV}-H^{-1})$ via ADI splitting	101
3.5.2	Numerical solution of $(\text{TV}-\text{Wass})$	108
4	A novel variational model for mixed noise removal	115
4.1	Heuristics	117
4.1.1	The reference models	117
4.2	Statistical motivation	118
4.2.1	The impulse and Gaussian case	119
4.2.2	The Poisson and Gaussian case	121
4.2.3	Difference with existing approaches	123
4.3	The variational model	124
4.3.1	Impulse-Gaussian fidelity	124
4.3.2	Gaussian-Poisson fidelity	125

4.3.3	Well-posedness	127
4.4	Recovery of the pure noise models	131
4.4.1	Recovery of TV- L^1 and TV- L^2 solutions	131
4.4.2	Recovery of TV- L^2 and TV-KL solutions	133
4.5	Numerical results	134
4.5.1	Impulse-Gaussian case	136
4.5.2	Gaussian-Poisson case	140
5	Learning the optimal setup by bilevel optimisation	145
5.1	A Review of learning methods in imaging	146
5.1.1	Learning from the model	146
5.1.2	Learning from the data	146
5.1.3	Bilevel optimisation approaches in function spaces	147
5.2	The General learning model	147
5.2.1	Assessing the quality: the choice of the cost functional	148
5.3	Learning the image model	149
5.3.1	ICTV and TGV regularisations	149
5.4	Learning the data model	153
5.4.1	Theoretical results	155
5.5	Numerical optimisation	158
5.5.1	Adjoint-based methods	159
5.5.2	Optimality system	161
5.5.3	Lagrangian derivation of the optimality system for TV-IC model	162
5.6	Robust estimation for large training sets	165
5.6.1	Dynamic sampling	167
5.7	Numerical results	172
5.7.1	Single noise estimation	173
5.7.2	Multiple noise estimation	181
6	Graph clustering and object measurement in images	185
6.1	A review on classical image segmentation models	186
6.1.1	Canny edge detection	186
6.1.2	The Mumford-Shah functional	186
6.1.3	The Chan-Vese model	187
6.2	Image segmentation as graph clustering	189
6.2.1	The Ginzburg-Landau functional as approximation of TV	189
6.2.2	Towards the modelling: the graph framework	191
6.2.3	The Nyström extension for matrix completion	195
6.2.4	Convex splitting	199
6.2.5	Numerical realisation and algorithm	200

CONTENTS

6.3	The Hough transform	203
6.3.1	Line detection	203
6.3.2	Circle detection	204
6.3.3	Pseudocode	206
6.4	Applications to object measurement in images	207
6.4.1	Male pied flycatcher's blaze segmentation	207
6.4.2	Moles monitoring for melanoma staging	216
6.4.3	Other applications: animal tracks and archaeological finds	217
7	Conclusions	219
	Bibliography	221

List of notation

For each entry the page at which the notation is introduced for the first time is reported.

Sets, Measures and Function Spaces:

\mathbb{R}_+	51	The set of nonnegative real numbers.
$\overline{\mathbb{R}}$	55	The extended real line, $\overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$.
Ω	49	An open subset of \mathbb{R}^d .
$\mathcal{B}(X)$	49	The Borel σ -algebra of X .
$ \mu $	51	The total variation measure of the finite Radon measure μ .
$\mathcal{M}(X, \mathbb{R}^\ell)$	51	Space of \mathbb{R}^ℓ -valued finite Radon measures on X with the norm $\ \mu\ _{\mathcal{M}(X)} = \mu (X)$.
$\mathcal{M}(X)$	51	Short version of $\mathcal{M}(X, \mathbb{R})$.
$\mathcal{M}^+(X)$	52	The set of finite positive Radon measures on X .
$\ \mathcal{T}\ _{\mathcal{M}}$	52	The Radon norm of the distribution \mathcal{T} on a set Ω : $\ \mathcal{T}\ _{\mathcal{M}} = \sup\{\langle \mathcal{T}, v \rangle : v \in \mathcal{C}_c^\infty(\Omega), \ v\ _\infty \leq 1\}$.
\mathcal{L}^d	50	The Lebesgue measure on \mathbb{R}^d .
$\frac{\mu}{\nu}$	52	The Radon–Nikodym density of μ with respect to ν .
μ^{ac}	52	The absolutely continuous part of μ with respect to Lebesgue measure, i.e., $\mu^{ac} = \frac{\mu}{\mathcal{L}^d}$.
μ^s	52	The singular part of μ with respect to Lebesgue measure.
$\mathcal{C}_c(X, \mathbb{R}^\ell)$	49	The space of \mathbb{R}^ℓ -valued continuous functions of compact support in X with norm $\ u\ _\infty = \sup_{x \in X} u(x) $.
$\mathcal{C}_c(X)$	49	Short version of $\mathcal{C}_c(X, \mathbb{R})$.
$\mathcal{C}_0(X, \mathbb{R}^\ell)$	49	The completion of $\mathcal{C}_c(X, \mathbb{R}^\ell)$ under the supremum norm.
$\mathcal{C}_0(X)$	49	Short version of $\mathcal{C}_0(X, \mathbb{R})$.
$\mathcal{C}_c^k(X, \mathbb{R}^\ell)$	49	The space of \mathbb{R}^ℓ -valued, k -times continuously differentiable functions with compact support in X .
$\mathcal{C}_c^k(X)$	49	Short version of $\mathcal{C}_c^k(X, \mathbb{R})$.
$\mathcal{C}_c^\infty(X, \mathbb{R}^\ell)$	49	The space of infinitely many times differentiable \mathbb{R}^ℓ -valued functions with compact support in X .
$\mathcal{C}_c^\infty(X)$	49	Short version of $\mathcal{C}_c^\infty(X, \mathbb{R})$.

$L^p(X, \mathbb{R}^\ell; \mu)$	50	Space of \mathbb{R}^ℓ -valued, μ -measurable functions with $\int_X u ^p d\mu < \infty$, where $1 \leq p < \infty$ and μ is a positive measure. The corresponding norm is $\ u\ _{L^p(X, \mathbb{R}^\ell; \mu)} = (\int_X u ^p d\mu)^{1/p}$.
$L^p(\Omega, \mathbb{R}^\ell)$	50	Short version of $L^p(\Omega, \mathbb{R}^\ell; \mathcal{L}^d)$.
$L^p(\Omega)$	50	Short version of $L^p(\Omega, \mathbb{R})$.
$L^\infty(X, \mathbb{R}^\ell; \mu)$	50	Space of μ -essentially bounded measurable functions, where μ is a positive measure. The corresponding norm is $\ u\ _{L^\infty(X, \mathbb{R}^\ell; \mu)} = \operatorname{ess\,sup}_{x \in X} u(x) $.
$L^\infty(\Omega, \mathbb{R}^\ell)$	50	Short version of $L^\infty(X, \mathbb{R}^\ell; \mathcal{L}^d)$.
$L^\infty(\Omega)$	50	Short version of $L^\infty(\Omega, \mathbb{R})$.
Du	53	The distributional derivative of the function u .
$W^{k,p}(\Omega)$	50	Sobolev space of functions $u \in L^p(\Omega)$ with weak derivatives up to order k also in $L^p(\Omega)$. The corresponding norm is $\ u\ _{W^{k,p}(\Omega)} = \left(\sum_{ \alpha \leq k} \int_\Omega D^\alpha u ^p dx \right)^{1/p}$. D^α is the α -th distributional derivative of u and $ \alpha = \alpha_1 + \dots + \alpha_d$ is the order of the multi index $\alpha = (\alpha_1, \dots, \alpha_d)$.
$H^k(\Omega)$	50	The Sobolev space $W^{k,2}(\Omega)$.
$H_0^k(\Omega)$	50	The completion of $\mathcal{C}_c^\infty(\Omega)$ under the $\ \cdot\ _{H^k(\Omega)}$ norm.
$\operatorname{BV}(\Omega)$	53	The space of functions of bounded variation on Ω endowed with the norm is $\ u\ _{\operatorname{BV}(\Omega)} = \ u\ _{L^1(\Omega)} + Du (\Omega)$.
$\operatorname{TV}(u)$	34	The total variation of $u \in L^1(\Omega)$.
∇u	53	Short version of $(Du)^\alpha = \frac{Du}{\mathcal{L}^d}$, the absolutely continuous part of Du with respect to Lebesgue measure.
$D^s u$	53	Short version of $(Du)^s$, the singular part of Du with respect to Lebesgue measure.
J_u	55	The jump set of u .

Miscellaneous notation:

\mathcal{X}_A	29	The characteristic function of the set A , i.e., $\mathcal{X}_A(x) = \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{if } x \notin A. \end{cases}$
$\operatorname{sc}_\tau F$	56	The lower semicontinuous envelope of $F : X \rightarrow \mathbb{R}$ with respect to the topology τ .
X^*	58	The analytic dual of the space X .
T^*	32	The adjoint operator $T^* : Y^* \rightarrow X^*$ of the linear and bounded functional $T : X \rightarrow Y$, where X, Y are Banach spaces.
∂F	59	The subdifferential of F .
$\delta F(w)[z]$	58	The directional derivative of F at w in the direction z .

Chapter 1

Introduction

This chapter addresses the non-expert audience and intends to give some background and a rather general overview of the main concepts of mathematical image processing. Starting with a review of standard notions in image digitalisation, we then describe some of the main problems faced in imaging applications and focus on nonsmooth image reconstruction models which have become largely popular over the last thirty years. For these models, we highlight the main challenges and open questions that represent the starting points of this thesis.

In particular, in the following Section 1.1 we start our exposition with a revision of the main concepts in digital image processing. In Section 1.2 we discuss how a generic imaging task can be mathematically formalised within a variational model, focusing our attention on total variation regularisation models in Section 1.3. In Section 1.4 we investigate the choice of the different ingredients of the modelling in view of obtaining a tailored approach with respect to the application considered. A summary of the main contributions of this thesis in this respect is finally presented in Section 1.5.

1.1 What is image processing?

Images are nowadays one of the most powerful means of communication. They surround us, in every aspect of our life. We see thousands of images when looking at the news in the newspaper or checking the updates on our favourite social network. Images taken by professional photographers with modern cameras can capture details we thought could not be caught. Furthermore, several other fields benefit from the use and the analysis of images. For instance, images acquired by means of Closed-Circuit Television (CCTV) cameras are often employed for security purposes. Another example is the use of images in the field of medical imaging, which is one of the main non-invasive practice used nowadays by medical doctors for the study and treatment of serious diseases such as cancer. Several medical imaging techniques like Computed Tomography (CT), Magnetic Resonance

Imaging (MRI), Positron Emission Tomography (PET), only to name a few, are indeed fundamental tools for the diagnosis, the treatment-planning and the follow-up of patients. Images are also essential in life sciences. In the study of earthquakes, for instance, images describing the density of seismic events are often used for predictions. In biological sciences such as zoology or botany, databases of images are used for species classification. Finally, astronomical images are employed for the study of the morphology of stars and planets.

As a result of different acquisition and transmission factors, very often images cannot be manipulated as raw data. On the contrary, they need to be processed in order to remove interferences corrupting the signal and/or to fill in missing or occluded regions in the image by some type of interpolation. Image processing (or Imaging) is the field of mathematics that enhances the quality of the imperfect image acquired by its rigorous and appropriate modelling.

Several post-processing tasks can be realised as well. The easiest example is the detection (segmentation) of objects in the image. This is a typical practice, for instance, in medical applications where doctors need to identify cancer regions out of the tumour-free ones and can be similarly used in many other disciplines.

In the following, we start from a revision on the mathematical modelling of digital images. For further readings we refer to [AK06, CS05a, GW06].

1.1.1 The mathematics of digital images

The first examples of photographs date back to the work of Nicéphore Niépce in 1826-1827. The technique used to acquire those images is often referred to as *heliography*. Essentially, it consists in the use of a plate made up by an expanding material which hardens when exposed to sunlight for a long time (around eight hours). By cleaning the plate with a suitable oil, the acquired scene is left imprinted on the plate, see Figure 1.1a¹.

In 1833, Louis Daguerre reduced the long exposition times required by Niépce's techniques to ten minutes and designed a new image acquisition tool (called *daguerreotype* after his name) which can be thought as a prototype of the modern Polaroid cameras, see Figure 1.1b².

In subsequent years, image acquisition instruments similar in essence to daguerreotypes were used and refined to improve the quality of the acquired images. The need of using a more compact modelling of images became stronger in the early 1920s when a prototype of digital images started to be used in the newspaper industry where pictures had to be sent by a submarine cable between London and New York with transmission times longer than

¹Image from: <http://www.hrc.utexas.edu/exhibitions/permanent/firstphotograph/process/#top>

²Image from: https://en.wikipedia.org/wiki/Louis_Daguerre#/media/File:Boulevard_du_Temple_by_Daguerre.jpg.



(a) N. Niépce: photo etching of an engraving of Cardinal Georges D'Amboise, 1827.



(b) L. Daguerre: "Boulevard du Temple", 1838.

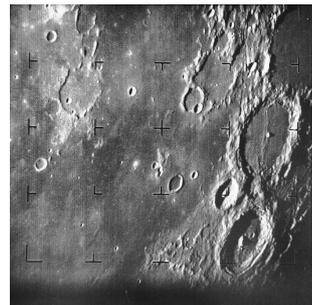
Figure 1.1: First examples of photographs.

a week. For this purpose, Mr. Bartholomew and Captain MacFarlane, both of the Daily Mirror in London, invented a new cable image transmission system (called *Bartlane*) which made the image transmission much faster (less than three hours). In this system, images were first digitised into five different levels of grey (increased to fifteen in the following years) using special printing devices and then standard telegraphic typewriting was used to transmit images to the receivers, see Figure 1.2a³

The idea of reducing image descriptors to a truncated, finite set of values became the basis of modern approaches to image digitalisation. Such approaches developed in parallel to digital computers due to the computational power and storage required for image display and transmission. The first computers powerful enough to handle image processing tasks appeared in the early 1960s and were used at the Jet Propulsion Laboratory in Pasadena to process and enhance images of the moon taken and transmitted by *Ranger 7* spacecraft and distorted by some interferences, see Figure 1.2b⁴



(a) A digital image produced from a coded tape by a telegraph printer with special type faces using Bartlane cable image transmission system, 1921.



(b) The first image of the moon taken by a U.S. spacecraft digitalised and enhanced in Jet Propulsion Laboratory, Pasadena, USA.

Figure 1.2: First examples of digital images.

Let us now be a more precise and explain in the following how a digital image is

³Image from: <http://www.jmcvey.net/cable/elements/pictures2.htm>.

⁴Image from: https://en.wikipedia.org/wiki/Ranger_7#/media/File:Ranger7_PIA02975.jpg.

created. For simplicity, let us consider in the following the case of grey scale images. The objective is to convert the real-world two-dimensional continuous image $u(x, y)$ into a digital form. In order to make the storage and the representation of the image feasible on computers, only a discrete number of points in the image can be used. Practically, in both the x - (horizontal) and the y - (vertical) directions a sample of points is extracted. This procedure is called *sampling*. The amplitude of the continuous signal (e.g. its colour intensity) presents also continuous variations. Therefore, similarly as above, a sampling is also required for its representation. In this case, the process is called *quantisation*. We make more clear our explanation by looking at Figure 1.4. By sampling, a regular grid of squared elements (pixels) is superimposed to the continuous grey scale image (left image). This is what happens in practice using cameras, scanners or any other sensor acquiring image signals. By quantisation, only one single numerical value is then assigned to every element of the grid (right image) such as, for instance, the average grey value of the element. In this way, the image has been digitalised.

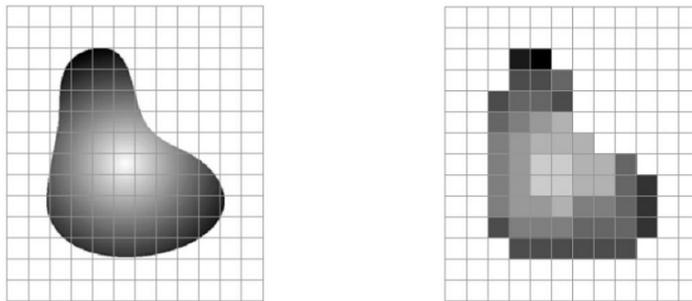


Figure 1.3: Left: continuous image projected into a sensor array. Right: Result after sampling and quantisation. Image from [GW06].

Mathematically, a digital image can then be represented as follows: let Ω the rectangular image grid, i.e. $\Omega = \{1, \dots, M\} \times \{1, \dots, N\}$, where M and N are positive integer values indicating the number of rows and columns of the grid, respectively. This grid describes the sampling of the image, i.e. its spatial discretisation into a finite number of MN elements. In order to describe the grey scale intensity of the image at each element of the grid, a numerical value is then assigned. Due to storage and hardware considerations, the number of grey levels allowed for each pixels is normally taken to be a power of 2. In practice, the number of allowed grey levels is $2^8 = 256$ and in particular their values are the integer values ranging from 0 (black) to 255 (white). The range of values spanned by the grey scale is called *dynamic range* of the image. The digital representation of a grey scale image u is then the function:

$$u : \Omega = \{1, \dots, M\} \times \{1, \dots, N\} \rightarrow \{0, \dots, 255\},$$

where, clearly, the finer the grid, the closer the representation of the real-world continuous

image (see Section 1.1.2 for more details). In Figure 1.4 we observe how this modelling corresponds to consider the image u to be a matrix having as entries $u_{i,j}$ the grey values evaluated in the pixel (i, j) . Alternatively, the image can be represented as a surface when interpreted as a function defined on a subset of the plane taking values into a subset of the real line.

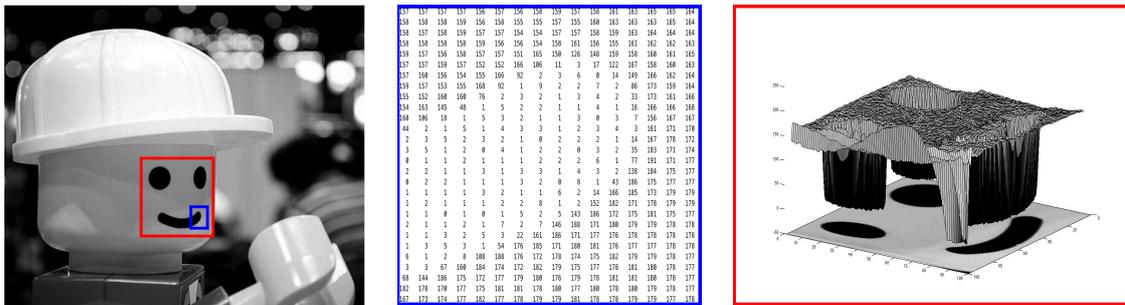


Figure 1.4: Matrix and surface representation of details of the image on the left.

For colour images, a triplet of intensity values ranging each from 0 to 255 is associated to the red (R), green (G) and blue (B) channel at every pixel: their combination is the RGB colour descriptor. In this case the image u is therefore a function:

$$u : \Omega = \{1, \dots, M\} \times \{1, \dots, N\} \rightarrow \{0, \dots, 255\}^3.$$

Digital images are therefore discretised representations of real-world continuous objects. Two questions arise naturally: are we losing some important information in this reduction approach? How close is this modelling to the realistic representation of what we actually see?

1.1.2 Discrete VS. continuous modelling

In the imaging community, there are two main attitudes to deal with image processing tasks. One follows directly from the modelling introduced in the previous section. There, an image is considered as a map defined on a discrete set taking values in a discrete range of intensity variations. On the other hand, in the last years significant improvements have been made in the field of image acquisition and representation (see Figure 1.5⁵) with the purpose of refining the representation of real-world images more and more, trying to approximate their original continuous nature by increasing their spatial resolution. Modern cameras are nowadays of the order of 36 megapixels and the ‘gigapixel race’ has already started.

Aiming to a finer and finer discretisation of the image representing its actual continuous nature is important for a number of reasons. Firstly, more and more details are visible

⁵Image from: https://en.wikipedia.org/wiki/Display_resolution

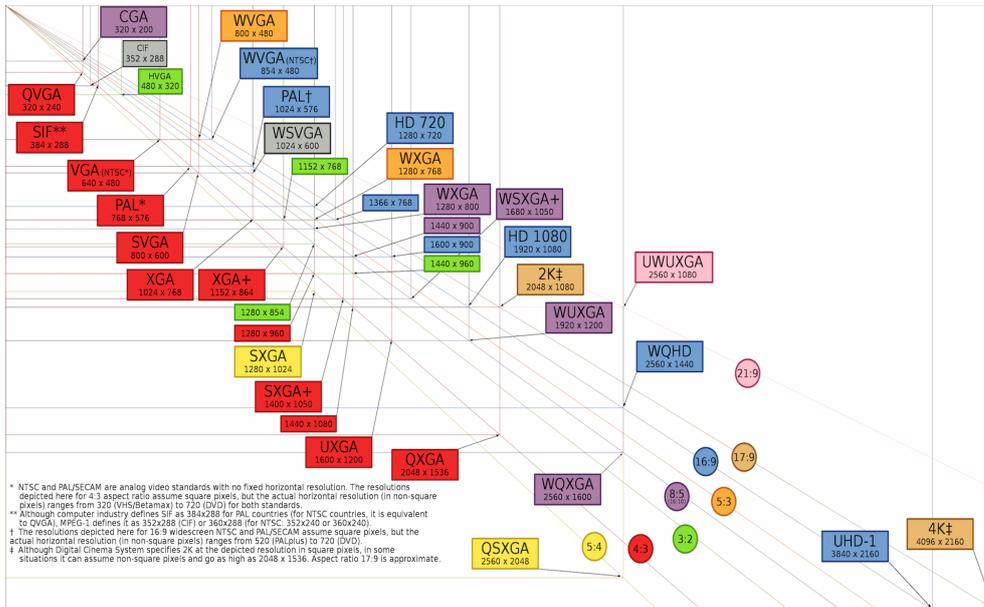


Figure 1.5: The most common display resolutions.

using finer spatial resolutions. This is important, for instance, in the case of CCTV images where fine details can reveal useful details, see Figure 1.6⁶. Secondly, in order to represent

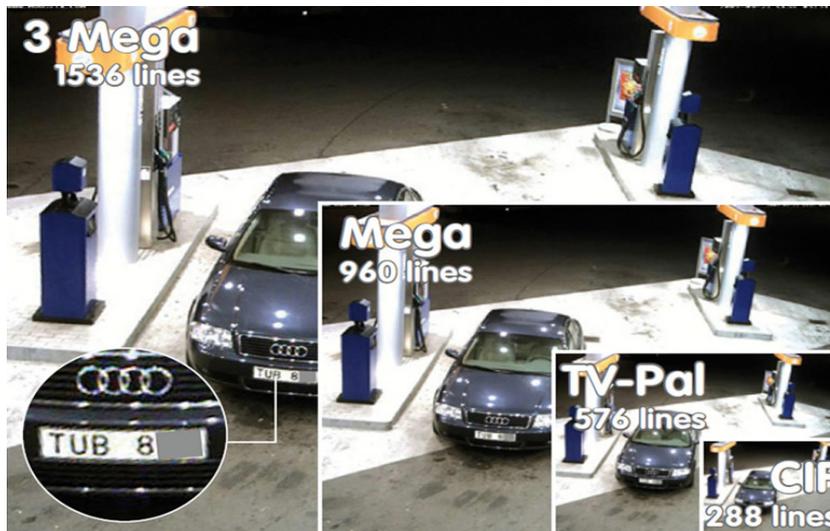


Figure 1.6: Increasing resolution reveals fine details which are fundamental, for instance, in CCTV imaging applications.

image information around us as our human vision system actually perceives it, a high-resolution representation of images accounting for infinitesimal variations is required. In fact, the human-eye resolution is of the order of 576 megapixels⁷, a much bigger number

⁶Image from: <http://www.idealcctv.co.uk/mobotix-security-cameras-mobotix-systems-registered-partner>

⁷Source: <http://www.clarkvision.com/articles/eye-resolution.html>.

than the state-of-the-art resolution of professional cameras. As a matter of fact, our fovea sight range is actually more of the order of 7 megapixels. This is because our eye cannot be considered as a single frame snapshot camera, but rather more like a video stream. When perceiving information around, the eye moves rapidly and detects also small angular variations in very short time intervals of the order of milliseconds. Then, the brain combines all the information coming from both eyes to increase resolution of the perceived image even further. In other words, even though our eye can be thought as a 7 megapixel-sensor, the actual number of information combined by the brain and transmitted to the retina is much higher. Therefore, aiming to a modelling of image processing problems which could encode appropriately the continuous nature of the image in terms of its properties and infinitesimal variations is desirable. Also, certain natural structures in images, such as edges, are meaningful in the continuum world only and can be appropriately modelled as discontinuities. From a mathematical point of view, rather than considering images as functions defined on discrete domains, the approach here consists in looking at images as elements of suitable function spaces defined over a continuous planar domain $\Omega \subset \mathbb{R}^2$. In this modelling, problems can then be formulated in terms of energy-minimisation and PDE models, used traditionally to describe continuous physical phenomena such as, for instance, heat conduction, wave propagation and many more.

1.1.3 Some imaging tasks

Before giving more details on the mathematical approaches used in this thesis, we present in the following some of the image processing tasks frequently encountered in applications.

- **Image denoising:** the task here consists in eliminating the interferences (noise) in the data due to faults or defects in the process of image acquisition and transmission, see Figure 1.7.



(a) Brain CT scan



(b) Camera image



(c) Astronomical image

Figure 1.7: The task of image denoising is frequent in several applications such as medical and astronomical imaging as well as in image retouching. Qualitatively, the noise in the image may appear different, depending on the tools used for acquisition and transmission.

- **Image inpainting:** the problem here is to fill in parts of the image where information has been lost or occluded. This is the case, for instance, of text and object removal, retouching of old photographs with scratches or folding creases, restoration of ancient frescoes and many more, see Figure 1.8.



Figure 1.8: The task of image inpainting consists in recovering piece of image information which has been lost or occluded.

- **Image segmentation:** here the problem is to partition the digital image into its constituent parts by extracting regions of interest characterised by similar features (colour intensity, texture, illumination. . .), see Figure 1.9.

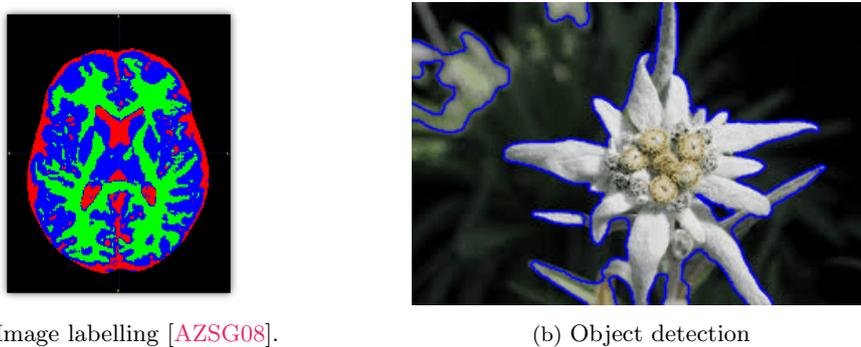


Figure 1.9: Through image segmentation, relevant objects in the image are extracted from the background and classified.

1.2 Mathematical image reconstruction techniques

A generic image processing task can be described as follows: given the measured image f which has been deteriorated by a degradation process T , we want to recover the original, uncorrupted image u . Assuming an additional additive noise component n in the measurements, the problem reads:

$$\text{find } u \text{ such that } f = T(u) + n. \quad (1.1)$$

The problem above fits the definition of **inverse problem** given by Keller in [Kel76] by which “We call two problems *inverses* of one another if the formulation of each involves all or part of the solution of the other.” Inverse problems are ubiquitous in many fields such as engineering, physics and medicine. Their solution requires the inversion of the operator T , which is very often not trivial. Consequently, the well-posedness of the problem (1.1) is often not guaranteed especially in terms of uniqueness and stability.

Standard models in imaging simplify (1.1) by assuming that the operator T is continuous and linear. In the following, we will proceed analogously. In the case of image denoising, for instance, the operator T is assumed to be the identity operator. Convolution operators may be considered as well to model the additional presence of blur. In the case of image inpainting, if $D \subseteq \Omega$ is the domain to inpaint, the operator T is chosen as $T = \chi_{\Omega \setminus D}$, the characteristic function of the complement of D . In this case, T is clearly not invertible and the additional presence of noise could make the solution problem even harder. One standard way to go around these issues consists in formulating a different problem whose solution is not exactly the u we seek, but one approximation of it which is found by exploiting prior knowledge on the desired reconstruction u in terms, for instance, of its likelihood conditioned to the observation of the data f .

1.2.1 The Bayesian framework

Among all stochastic and statistical approaches to solve (1.1), the Bayesian one is probably the easiest to be understood. In this framework, the problem (1.1) is reformulated as a maximisation problem of the posterior probability

$$P(u|f) = \frac{P(f|u)P(u)}{P(f)}, \quad (1.2)$$

where $P(u|f)$ is the probability of observing u whenever f is observed and the equality (1.2) is simply the Bayes’ theorem. The probability $P(u)$ is called the *prior* model since it specifies *a priori* information on the target quantity u which is independent on the data f . The conditional probability $P(f|u)$ is the *data* model since it specifies how information f is distributed once u is known. Therefore, thanks to (1.2), the solution \hat{u} of the MAP estimation problem satisfies

$$\hat{u} = \arg \max_u P(u|f) = \arg \max_u P(f|u)P(u), \quad (1.3)$$

where the quantity $P(f)$ does not affect the minimisation over u and therefore has been neglected. The technique described is generally called Maximum A Posteriori (MAP) estimation. Standard reference books for the use of such Bayesian statistical framework to solve inverse problems are [Stu10, Idi13].

1.2.2 The variational approach

The MAP estimation Bayesian approach described above presents connections with other, deterministic models through the use of Gibbs' models for images as random fields, see [GG84], by which the prior model is assumed to be of the form:

$$P(u) \sim \exp\left(-\frac{1}{\lambda}R(u)\right)$$

for a convex function $R(u)$ and a scalar parameter $\lambda > 0$. Assuming a similar structure for $P(f|u) \sim \exp(-\Phi(Tu, f))$ and considering in (1.2) an equivalent formulation in terms of the negative log-likelihood that reads

$$\hat{u} = \arg \min_u -\left(\log P(u|f)\right) = \arg \min_u \left(-\log P(f|u) - \log P(u)\right),$$

we have that the maximisation problem (1.3) can be equivalently expressed as the minimisation problem

$$\hat{u} = \arg \min_u R(u) + \lambda \Phi(Tu, f),$$

where we have dropped the normalisation constants not depending on u . Of course, in one side the modelling above appears quite natural and sensible in the general framework, the tailored and specific choice of the single terms depends on the application considered: in particular, as we will detail more in the following, the likelihood functional $P(u)$ encodes information on the desired type of reconstruction one aims fore, whereas the choice of the data model $P(f|u)$ can be driven by more structural information on the data available (i.e. distributed, concentrated noise with signal dependent/independent properties).

Whenever the modelling is done in a function space V , the minimisation problem above naturally leads to a variational approach. The theoretical foundation for the use of variational approaches to the regularisation of ill-posed inverse problems of the form (1.1) has been studied by Tikhonov [Tik63] and is traditionally known in literature as *Tikhonov regularisation* of inverse problems. In our imaging framework, the regularisation of the inverse problem (1.1) reads:

$$\min_{u \in V} J(u) := R(u) + \lambda \Phi(Tu, f), \tag{1.4}$$

where, given the measured image f , its reconstructed version is computed by minimising the functional J in the Banach space V . We can observe that the energy functional J is the combination of different terms:

- The **regularisation term** $R(u)$, which encodes *a priori* information on the reconstruction u we seek by imposing extra-regularity on it. Typically in imaging applications, this information is incorporated through some norm of the image gradient.

The regularising term drives the reconstruction process. In the case of denoising, for instance, it smooths out the noise in the measured image f , while in inpainting tasks it transfers information inside the inpainting domain.

- The **fidelity term** $\Phi(Tu, f)$, which measures in an appropriate sense the distance between the data f and its reconstruction u after the action of the operator T .
- The **weighting parameter** $\lambda > 0$, which balances the effect of the regularisation against the fitting of the data.

The choice of the different ingredients of the model is crucial for obtaining a good reconstruction of the image. We give more details on the modelling of (1.4) in the forthcoming Section 1.4.

Some pioneering works in variational methods for image reconstruction are the ones of Rudin, Osher and Fatemi [ROF92] for image denoising and the approach presented by Mumford and Shah in [MS89] for image segmentation. Let us point out that several variational formulations of imaging tasks have been proposed also in the context of graphs, see [ELB08, Gra06, SM00, BF12]. Typically, variational models are handled here using elements of spectral graph theory (see, e.g., [Chu97]) in order to exploit clustering properties of the discrete differential operators involved.

1.2.3 The PDE approach

Other approaches used to solve (1.1) rely on the use of Partial Differential Equations (PDEs). PDE models have been widely used to study several physical problems such as heat conduction, phase transition, wave propagation, electromagnetic transformations, thermodynamic processes and many more. Standard reference books for the study of PDEs are, for instance, [Eva10] and [Fol95]. PDEs describe changes with respect to continuous spatial and temporal variables in a compact form. Consistently with our motivation given in Section 1.1.2, it is therefore appropriate to consider this framework to model imaging tasks. Some specific readings on the use of the PDE framework in Image Processing are [GMmM01, AK06, PWO89].

PDE approaches to imaging problems can be connected to the variational approaches described in Section 1.2.2 as follows: minimisers of (1.4) are stationary points of the first-variation, i.e. the Fréchet derivative of the functional J in the Banach space V . In other words, they are solutions of the corresponding Euler-Lagrange equations. Therefore, imposing the vanishing of the Fréchet derivative of the energy J in (1.4) (under the assumption that all the terms are differentiable and the functional J is convex) we obtain that the following PDE holds in correspondence with a minimiser \hat{u} of J :

$$\nabla_V R(\hat{u}) + \lambda T^* \nabla_V \Phi(T\hat{u}, f) = 0, \tag{1.5}$$

where we have denoted by ∇_V the Fréchet derivative in the space V and by T^* the adjoint operator of T .

A common practice used to solve (1.5) consists in introducing an artificial time variable mimicking the evolution of the reconstruction process starting from the initial image f towards the solution \hat{u} . In other words this corresponds to consider the parabolic PDE

$$\frac{\partial u}{\partial t} = -\nabla_V R(u) - \lambda T^* \nabla_V \Phi(Tu, f). \quad (1.6)$$

The evolution process computes a family of images representing progressive regularised versions of f . When no more time variations occur, the solution of (1.5) is then recovered. In literature this approach is called *steepest descent* and is very often used in imaging applications. Equations of the form (1.6) are called *gradient flows* of the functional J with respect to the metric of the Banach space V .

Several nonlinear PDE models used in imaging do not follow a variational principle, i.e. they cannot be derived directly from the energy functional J by considering its gradient flow, see, e.g., [PM90, Wei98, CKS02, CS01, SB11]. Therefore, many PDE models are formulated only in terms of a PDE, without thinking of any upper energy functional.

1.3 Total variation models

Let us focus for a moment on the choice of the regularisation term $R(u)$ we shall consider in (1.4). As mentioned above, such term encodes *a priori* information and knowledge on the reconstruction we seek. In this section we take a closer look at the consequences of this choice on the qualitative properties of the reconstructed image u for the particular case of image denoising ($T = Id$) and image inpainting ($T = \chi_{\Omega \setminus D}$, where D is the region to inpaint).

1.3.1 Motivation

Let us start from the problem of image denoising. In [Tik63], Tikhonov considered the following quadratic regularisation term

$$R(u) = \int_{\Omega} |\nabla u|^2 dx,$$

where ∇u is the weak gradient of u . Combined with a squared L^2 data fidelity term Φ of the difference between u and the data f , compare [ROF92] and see Section 1.4.2 for more details, the variational model for Gaussian noise removal reads

$$\min_{u \in H^1(\Omega)} \int_{\Omega} |\nabla u|^2 dx + \frac{\lambda}{2} \int_{\Omega} (u - f)^2 dx. \quad (1.7)$$

The minimisation of the energy above enforces the reduction of high frequency noise oscillations in the reconstruction process as well as the fitting of the given data f . Standard tools in calculus of variations ensure that the problem above is well-posed. When taking the gradient flow with respect to the L^2 norm, the following heat-type PDE is obtained:

$$\frac{\partial u}{\partial t} = \Delta u + \lambda(f - u).$$

Smoothing properties of such equations have been widely studied in the mathematical literature of parabolic PDEs. Therefore, their use seems appropriate in the context of noise reduction. However, the resulting Laplace diffusion is isotropic and can not distinguish between high oscillations in the image due to noise and image distinctive structures such as edges, compare Figure 1.10.

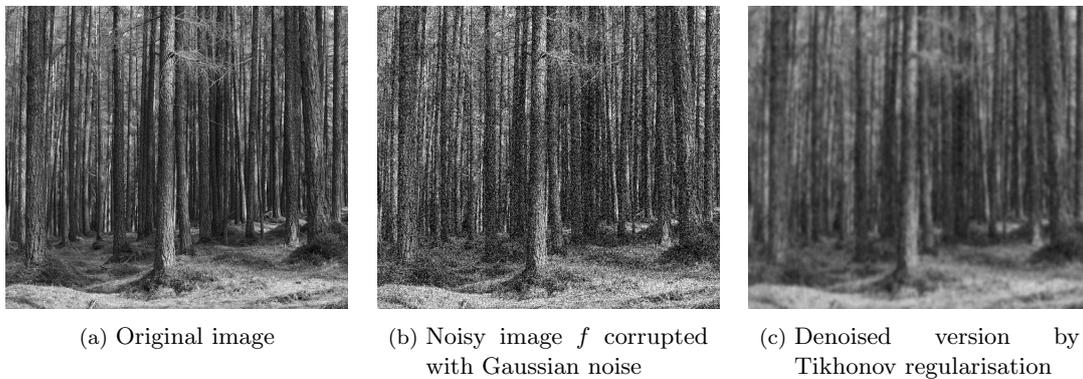


Figure 1.10: Image denoising in presence of Gaussian noise using Tikhonov regularisation model (1.7).

A different type of regularisation shall then be considered. Its action should eliminate the noise while preserving at the same time the characteristic and structural features in the image such as its discontinuities corresponding to image edges.

1.3.2 Nonsmooth regularisation

In order to formulate a regularisation model which could encode the existence of jumps, while still being mathematically tractable, in their seminal work of 1992 [ROF92], Rudin, Osher and Fatemi proposed the use of the Total Variation (TV) regularisation and the framework of functions of bounded variation on Ω ($BV(\Omega)$) to solve the task of image denoising. Previous studies on BV functions date back to the work by Ambrosio in [Amb89, AT90] and Giusti in [Giu84] and provide the mathematical foundations necessary for the rigorous study of the related regularisation models. An exhaustive monograph on BV functions can be found in [AFP00].

The main advantage of the use of BV functions is that in this framework the gradient of a discontinuous function has still some meaning. In particular, it is interpreted as a

measure. Namely, for a regular domain $\Omega \subset \mathbb{R}^2$ and a function $u \in L^1(\Omega)$, the total variation of u is defined as

$$\text{TV}(u) = \sup \left\{ \int_{\Omega} u \nabla \cdot v \, dx : v \in \mathcal{C}_c^1(\Omega, \mathbb{R}^2), \|v\|_{\infty} \leq 1 \right\},$$

where $\mathcal{C}_c^1(\Omega, \mathbb{R}^2)$ is the space of continuously differentiable functions with compact support in Ω . A function with $\text{TV}(u) < \infty$ is said to be a function of bounded variation, i.e. an element of $\text{BV}(\Omega)$. In this case, the distributional gradient of u can be represented by a finite Radon measure Du . Moreover, denoting by $|Du|(\Omega)$ the total variation measure of Du , if $u \in \text{BV}(\Omega)$ the following equality holds:

$$\text{TV}(u) = |Du|(\Omega).$$

Therefore, the total variation of a function with bounded variation is nothing but the total variation measure of its distributional gradient interpreted as a Radon measure (see Chapter 2 for more details). The interpretation of the total variation clarifies in the case of smooth functions $u \in W^{1,1}(\Omega)$. In this case one can show that $\text{TV}(u) = \int_{\Omega} |\nabla u| \, dx$. Total variation can then be interpreted as a generalisation of the L^1 norm of Du in the nonsmooth case.

The capability of TV of describing discontinuities in images was shown rigorously by Ambrosio in [Amb89], where a decomposition of Du into a Lebesgue and a singular component accommodating jumps is studied.

The corresponding Rudin, Osher and Fatemi (ROF) TV denoising model for Gaussian noise removal [ROF92] reads :

$$\min_{u \in \text{BV}(\Omega)} \text{TV}(u) + \frac{\lambda}{2} \int_{\Omega} (u - f)^2 \, dx. \quad (1.8)$$

The well-posedness of the problem above can be proved using compactness and embedding properties of the space $\text{BV}(\Omega)$. Differently from (1.7), the ROF model accomplishes at the same time the removal of the noise and the preservation of image structures through the use of a nonsmooth regularisation, see Figure 1.11.

TV regularisation models have been applied also to other imaging tasks. In image inpainting, the use of total variation models has been considered for instance in [SC02, CKS02, CS05b]. The minimisation problem in this case reads:

$$\min_{u \in \text{BV}(\Omega)} \text{TV}(u) + \frac{\lambda}{2} \int_{\Omega \setminus D} (u - f)^2 \, dx, \quad (1.9)$$

where $D \subseteq \Omega$ is the region in the image where information has been lost, i.e. the inpainting domain. In this case, the feature of total variation to preserve discontinuities is used to propagate image information inside D . The reconstruction result is sharp and almost

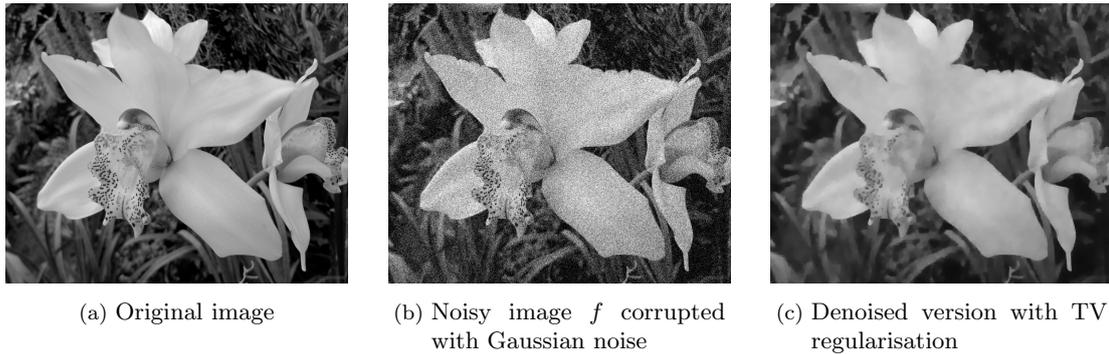


Figure 1.11: Image denoising in presence of Gaussian noise using ROF TV regularisation model (1.7). The noise is successfully removed and structures are preserved.

perfect in the case when the missing regions are small enough, see Figure 1.12 and compare Section 1.4.1 for more details.

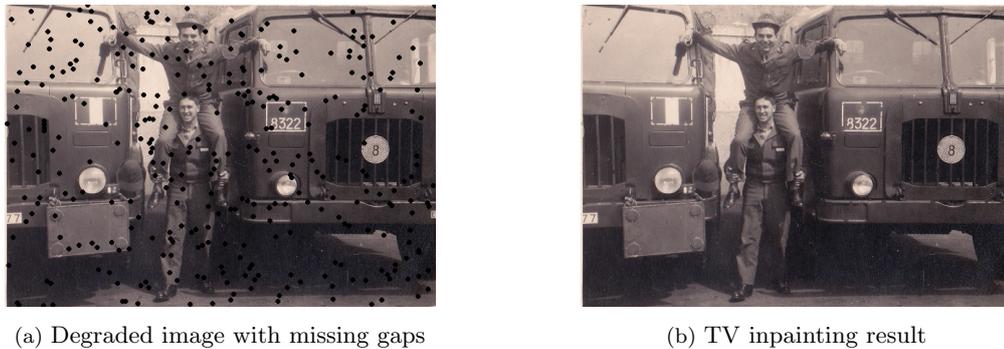


Figure 1.12: Image inpainting using total variation model (1.9).

Some work on the use of total variation for image segmentation purposes also exists in the literature. Typically, a smoothing of TV depending on a small parameter $\varepsilon > 0$ is considered [KS89, ET06, Ese04, ES02, BF12]. Such smoothing often relates to approximations of TV (in suitable sense) which have been used to model physical transformations of two-phase materials, see [MM77a, MM77b], and is therefore often employed for binary segmentation where the desired region of interest is extracted from the background.

1.3.3 Previous work

The literature on the use of TV models for imaging problems is very vast. From a theoretical point of view, the well-posedness of TV models and on the structure of their solutions have been studied in several works [AV94, Mey01, Ves01, Nik02, CS03, DAG09, CCN07, All08a, All08b].

Since its first use in 1992 in [ROF92], total variation denoising has further been considered for [CL97, CKS01, AV97, CM99] for Gaussian noise removal, whereas in [Nik04] it is used in combination a combination with an L^1 fidelity term for the modelling of a

different type of noise that switches the intensity value of the image to either the maximum/minimum value of its dynamic range or to a random value within it, with positive probability. This type of noise is called *impulse* or “salt & pepper”, respectively. Similarly, in [LCA07, SBMB09] TV regularisation is used for the design of variational models describing the presence of Poisson noise in the data. In [MBBS14] TV regularisation is applied to the RGB channels separately for several image reconstruction problems. TV-based denoising models have been considered in the context of MRI and PET reconstruction in [Saw11, Mü13, BGH⁺14, BMPS14] and for image inpainting in [SC02, CS05b]. In [CSZ06] an inpainting strategy based on total variation regularisation is used to fill in missing regions in the wavelet domain.

Due the nonlinear nature of total variation regularisation, the design of algorithms computing the numerical solution of TV image reconstruction models efficiently is a research field on its own regard. In the case of image denoising, the first numerical approach considered by Rudin, Osher and Fatemi [ROF92] solved the resulting steepest descent PDE of the minimisation problem (1.8). Further, algorithms based on duality [ZWC10], Newton-type methods [NQYH07], graph cuts [CD09] have been considered as well. A celebrated algorithm solving TV denoising is the projection algorithm described by Chambolle in [Cha04]. There, by using Legendre-Fenchel duality, the TV minimisation problem is shown to correspond to the computation of the nonlinear projection of the given noisy data f onto a closed and convex set. Numerically, the problem is solved by using a convergent fixed point algorithm. Chambolle’s algorithm has also been used to solve image deblurring [BBFAC04] and image inpainting and colorisation [LBLZ] problems. In [CP11] an efficient hybrid primal-dual algorithm is studied for Gaussian and impulse noise removal and applications to image deconvolution, zooming and inpainting are considered. Operator splitting techniques have also been employed. In particular, in [GO09b] a splitting technique is combined with Bregman iteration [OBG⁺05] to solve efficiently nonsmooth image processing problems. The resulting split Bregman algorithm decomposes the costly original problem into a sequence of computationally cheap subproblems and solve them alternately.

1.4 Choosing the ingredients for a good model

In Section 1.2.2 we have discussed the role of the ingredients of the variational model (1.4). Depending on the particular application considered, very often a customised choice of the regulariser, data term and weighting parameter is essential. In this Section we briefly comment on the main extensions of the reference ROF model (1.8) studied in the last years for improving upon standard, often unsatisfying properties of TV models with L^2 data fidelities.

1.4.1 Improved reconstructions with higher-order models

Despite being appropriate prototypes for imaging problems as discussed in Section 1.3.2, variational approaches based on the use of pure TV regularisation present some drawbacks and limitations. For instance, in the case of image denoising, it can be observed and proved analytically that TV regularisation promotes piecewise constant reconstructions, see Figure 1.13. In the case of natural (non-piecewise constant) images, this often results in watercolour-like reconstructions characterised by the presence of blocky regions. This drawback is called *staircasing effect* and has been studied analytically, e.g., in [CCN07, BKV13, DS96].

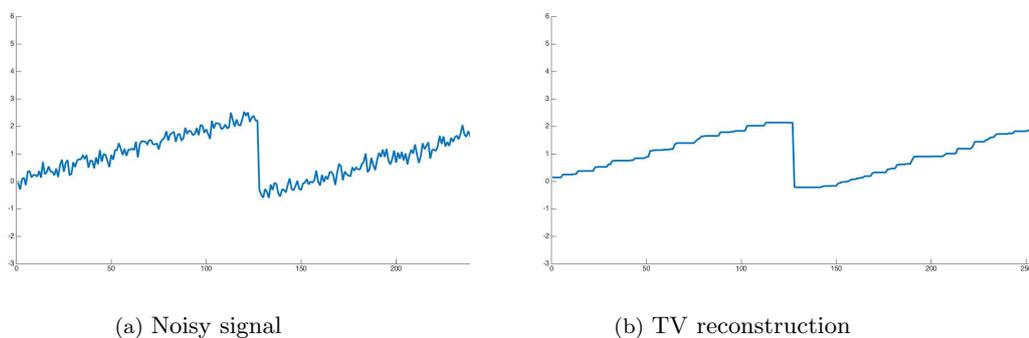


Figure 1.13: TV denoising promotes piecewise constant reconstructions.

In order to counteract this undesired drawback, several extensions of TV models have been considered. Despite their differences, these models share the common idea of incorporating higher-order derivatives in the regularisation term. In this way, more complex structures such as piecewise linear regions can be reconstructed, improving upon the limitation of pure TV models. In [CL97], for instance, this is achieved by introducing a combination of first and second order regularisation which decomposes the image into its piecewise constant and piecewise affine part. Further higher-order models being a combination of first and second order models [PS14] or pure higher-order models have also been considered in [Sch98, BP10, PS08]. A celebrated and extensively studied higher-order regulariser is the Total Generalised Variation (TGV) regularisation introduced in [BKP10]. Compared to standard TV regularisation, TGV models promote smooth reconstructions, allowing the preservation of linear structures, see Figure 1.14. Theoretical aspects of TGV models have been studied in a vast number of papers [BV11, Mü13, Bre14, BKV13, PB15] and applications of TGV to MRI and diffusion tensor imaging [BGH⁺14, VBK13] and image decomposition [Hol13] have also been considered.

In the case of image inpainting, TV regularisation exhibits a different problem. Through the coarea formula, it can be shown that minimising the total variation energy corresponds to penalise the length of the level lines over the inpainting domain. During the reconstruction process, information is then interpolated via the shortest path, [CKS02, Sch09a]. This

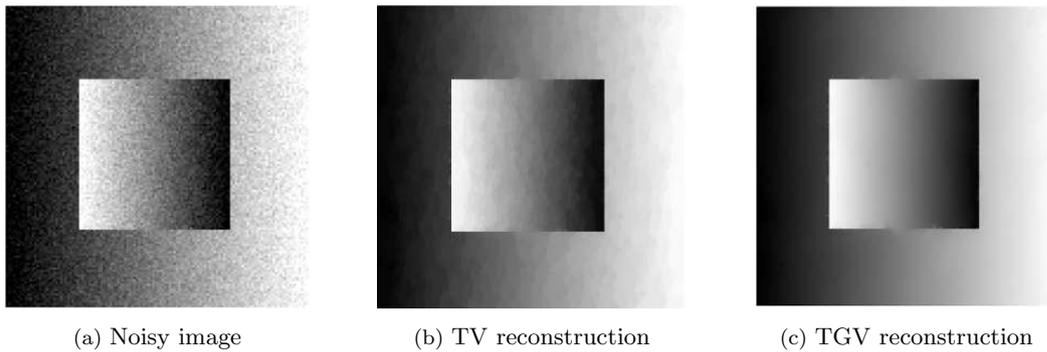


Figure 1.14: Comparison between TV and TGV reconstruction. Image from [BKP10].

is of course problematic when the missing gap is large. In this case, TV inpainting fails completely the reconstruction, i.e. it does not satisfy the so-called *connectivity principle*, that is the property of interpolating over large gaps. Similarly, such models show drawbacks in the smooth propagation of the level lines of regions with high curvature, i.e. do not satisfy the *curvature preservation* principle. By incorporating higher-order derivatives, more information on the directional derivatives and curvature is encoded in the regularisation term and good reconstructions over large gaps and in presence of high curvature regions can be obtained, see Figure 1.15 and Figure 1.18. Among the vast literature of higher-order inpainting approaches we recall here Euler’s elastica models introduced in [NMS93], Cahn–Hilliard models for binary inpainting [BEG08] and its generalisation to grey value images in the Sobolev space H^{-1} introduced by Burger et al. in [BHS09] and studied further in [SBBH09, SB11] from both an analytical and numerical point of view.

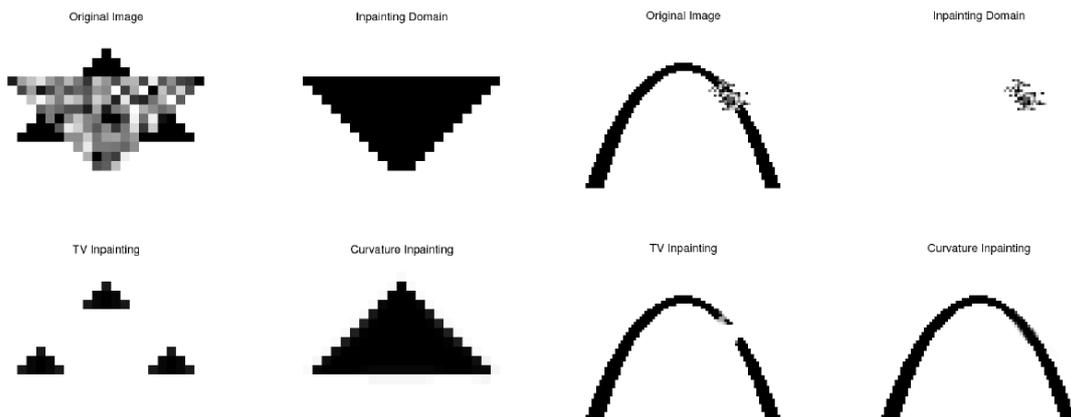


Figure 1.15: TV inpainting fails in presence of large gaps. By using higher-order models, information is propagated in the missing region correctly. Image from [CS05b].

1.4.2 Fitting the data

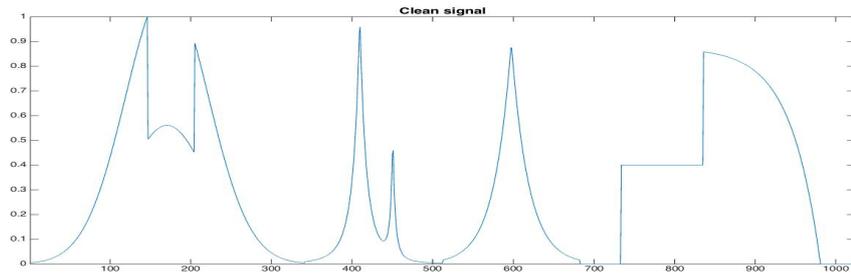
If on one side the choice of a proper regularisation is fundamental for an artefact-free reconstruction of the desired image, on the other one the selection of an appropriate data fidelity term in (1.4) is essential for the design of a model which is consistent with the given data. As mentioned briefly in Section 1.2.2, typically the choice of Φ reflects specific, *a priori* known properties of the noise corrupting the data, i.e. to the likelihood function modelling the probability density of observing f given u . In the examples considered so far, the noise has been assumed to be additive and Gaussian distributed. In this case L^2 squared data fidelities are considered, see, e.g., [ROF92, CL97]. In the case when the data are corrupted by noise with different statistical properties, different data fidelities are used. This is somehow natural as signals corrupted by different noise distributions appear different, see Figure 1.16⁸.

In the case of impulse noise, i.e. an instance of the noise consisting in sparse random occurrences of ‘spikes’ affecting only a percentage of the pixels in the image, the L^1 norm of the difference $u - f$, i.e. $\Phi(u, f) = \|u - f\|_{L^1(\Omega)}$ has been shown to be a more appropriate data model in [Nik04, Nik02, DAG09, CE05]. In the case of uniform noise, an L^∞ modelling $\Phi(u, f) = \|u - f\|_{L^\infty(\Omega)}$ has been considered in [Cla12]. When Poisson noise is assumed, variational models making use of (some variations of) the Kullback-Leibler functional of the type $\Phi(u, f) = \int_{\Omega} (u - f \log(u)) dx$, have been employed in [SBMB09, Saw11, BSW⁺13, LCA07]. Some work exists also in the variational modelling of signal-dependent noise distributions whose intensity depends in a multiplicative manner on the underlying signal, which generalise the Poisson case. One example of this type of noise is speckle noise, which can be observed, for instance, in Synthetic Aperture Radar (SAR) images. In this case, the statistical model assumed corresponds to a Gamma density function. In [AA08] a TV restoration model combining appropriate data fidelity term is proposed for this type of problems. In the case when multiple noise distributions are assumed, combination of data fidelities have been considered in [CCN08, HL13, LBU11, JCPT12, BLCT⁺08, DLRS13], but very often such approaches are not statistically consistent with the assumptions on the combined distribution of the noise.

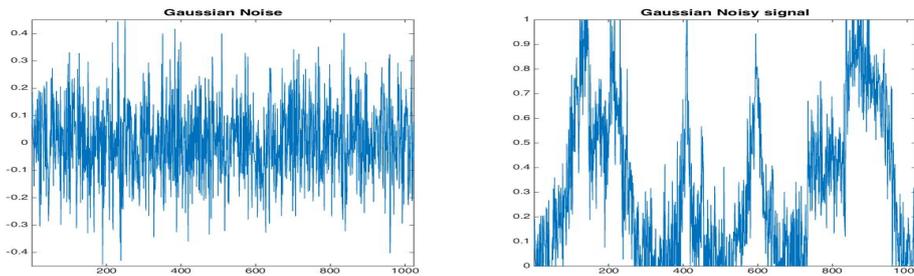
1.4.3 Regularisation VS. data fitting: how to balance?

Another crucial modelling aspect is the choice of the weighting term $\lambda > 0$ appearing in (1.4). Its size balances the effect of the regularisation against the fitting of the data in the reconstruction process. Figure 1.17 illustrates the problematics of choosing a parameter which is either too large or too small. In the former case, a strong fitting with the data is imposed: therefore, in the case of noisy data, a poor reduction of the noise is observed.

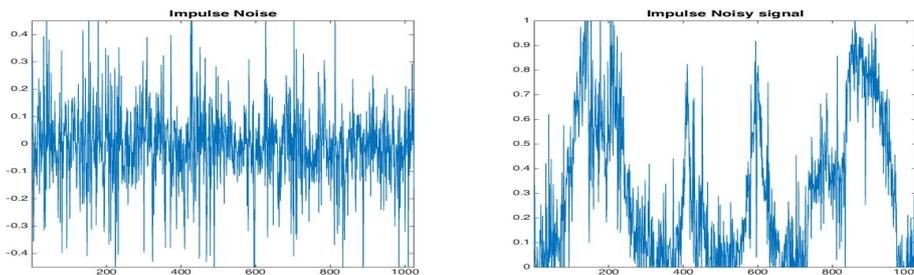
⁸Images from http://nbviewer.ipython.org/github/gpeyre/numerical-tours/blob/master/matlab/denoisingsimp_1_noise_models.ipynb, see [Pey11].



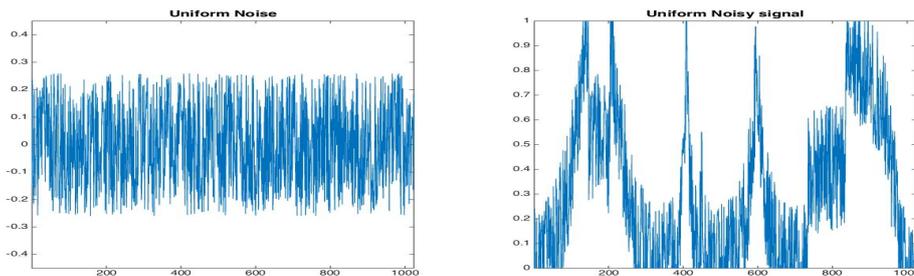
(a) Noise-free signal



(b) Gaussian noise



(c) Impulse noise



(d) Uniform noise

Figure 1.16: Different noise distributions result in different noisy signals.

On the other hand, in the latter case a stronger regularisation effect is promoted: while the noise is reduced, the distinctive structures we desire to preserve are over-smoothed and destroyed.

Trial-error parameter selection methods aside, several techniques for the estimation of

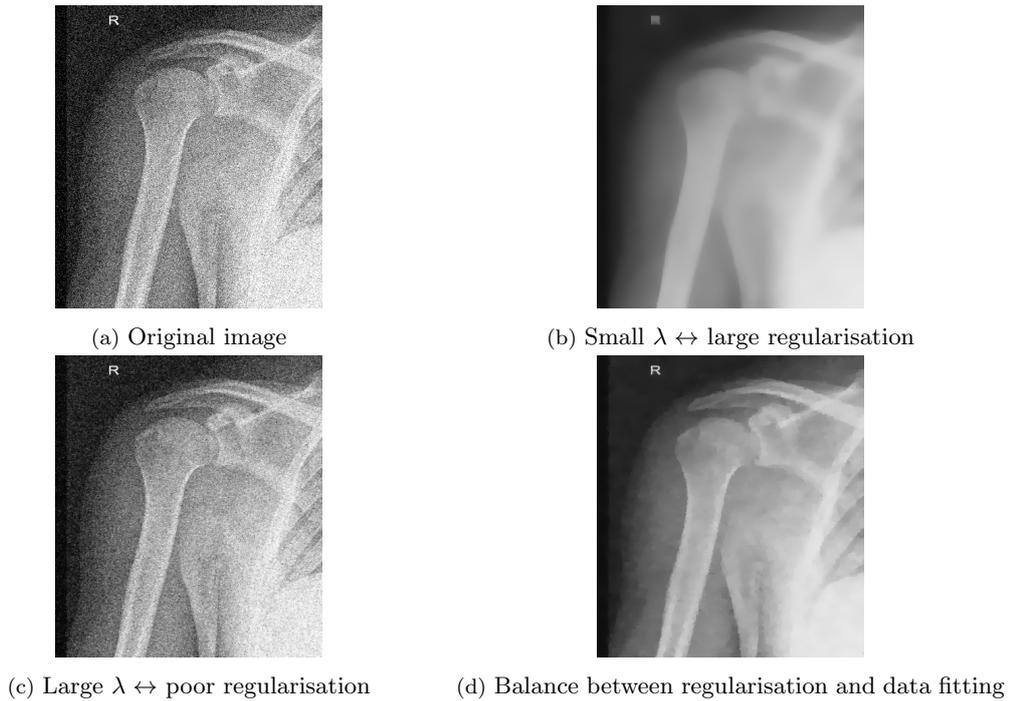


Figure 1.17: Different TV-denoising reconstructions obtained for different choices of λ in (1.8).

the optimal λ have been proposed in the literature. In the case of image denoising, when the level of noise is assumed to be known, a common practice consists in estimating the optimal λ with a precision level equal to the noise intensity. This approach is historically called *Morozov discrepancy principle* and has been firstly considered in [Mor66] and adapted in [TSC13] for the case of non- L^2 fidelities. In the case when the noise level is not known, another approach has been proposed by Lawson in [LH95]. There, the plot of the noise residual against the solution of the problem after a transformation through appropriate functions is drawn and the optimal value of λ is found by looking at the minimum of the function obtained. Due to the shape of the resulting curve considered this approach is called “L-curve” approach and has been further considered for inverse problems regularisation in [EHN96, CMRS00]. Cross-validation [Wah77] and variational Bayes’ models [BMK08] have also been considered for parameter identification, but a robust and efficient estimation of the optimal parameter not depending on *a priori* assumptions on the noise level is still an open question.

1.5 Contribution

The main body of this thesis consists of four different chapters corresponding to different topics of research which has been investigated by the author during his Ph.D. studies. These investigations resulted in the publication of a number of scientific papers which

are, at the time of the submission of this thesis, either published in mathematical journals [CDS14, BCDS13, CDLRS14], accepted for publication [CCDLR⁺15] or at the final revision stage before submission [CvGS⁺15, CDLRS15]. In the following, we give a short summary of the main contributions given. They will be presented in more detail in Chapters 3, 4, 5 and 6.

1.5.1 Efficient solution of higher-order TV models

In [CDS14, BCDS13] we focus on the numerical solution of higher-order nonlinear PDE models for imaging. These papers are the result of a collaboration with Martin Benning, Bertram Düring and Carola-Bibiane Schönlieb.

In particular, in [CDS14] we present directional operator splitting schemes for the numerical solution of the H^{-1} -gradient flow of the total variation energy. This reads:

$$\begin{aligned} u_t &= \Delta q, & q &\in \partial|Du|(\Omega) & \text{in } \Omega \times (0, \infty), \\ u(t=0) &= u_0 & & & \text{in } \Omega, \end{aligned}$$

and has been previously considered by Burger, Schönlieb and others in [BHS09, SBBH09, Sch09a, SB11] as a higher-order extension of classical second-order TV models for inpainting problems due to the better connectivity properties over large gaps. The efficient numerical solution of this equation is very challenging due to the stiffness of most numerical schemes. We propose a combination of directional splitting schemes with implicit time-stepping providing a stable and computationally cheap numerical realisation of the equation, see Figure 1.18

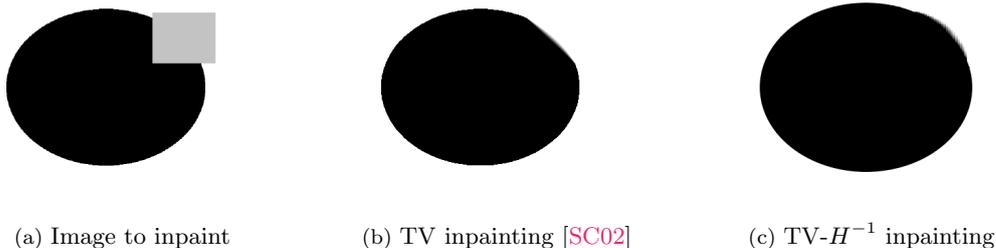


Figure 1.18: Comparison between TV and TV- H^{-1} inpainting. The solution is computed efficiently using directional splitting.

In [BCDS13] we consider a similar nonlinear fourth-order diffusion equation that arises in denoising of image densities that reads

$$\begin{aligned} u_t &= \nabla \cdot (u \nabla q), & q &\in \partial|Du|(\Omega) & \text{in } \Omega \times (0, \infty), \\ u(t=0) &= u_0 \geq 0 & & & \text{in } \Omega, \end{aligned}$$

with normalised mass $\int_{\Omega} u_0 = 1$. Such equation has been considered by Burger, Schönlieb

et al. in [DS12, BFS12]. Such PDE is the gradient flow of the total variation energy with respect to the L^2 -Wasserstein metric in the space of probability densities. We focus here on the application of this PDE to the problem of image denoising. In particular, we show improvements upon classical drawbacks of second-order TV solutions such as stair-casing. Furthermore, the density framework considered ensures that natural properties (such as mass and positivity) are still preserved. For this model, we propose an implicit time stepping scheme that employs a primal-dual method for computing the generalised derivative of the total variation functional. From a numerical point of view, dealing with the dual formulation of TV introduces a constraint on the dual variable which we relax by adding a penalty term, weighted by a parameter that determines the strength of the penalisation. The resulting primal-dual system is solved efficiently by means of a damped Newton's method. We report numerical examples showing the denoising properties of the model considered as well as the preservation of the gradient flow structure.

1.5.2 The TV-IC variational model for image denoising

In [CDLRS15] we concentrate on the choice of appropriate data fidelities Φ in (1.4) for images corrupted by a combination of noise distributions frequent in applications (such as impulse-Gaussian and Gaussian-Poisson). The results in this paper have been obtained in collaboration with Juan Carlos De Los Reyes and Carola-Bibiane Schönlieb.

The variational model considered is derived using the statistical assumptions on the noise corrupting the data and combines in an infimal convolution fashion standard L^1 [Nik04, DAG09], L^2 [ROF92] and Kullback-Leibler-type [LCA07, SBMB09] fidelity terms used in literature for single-noise denoising modelling.

Previous methods dealing with the combined case have been considered in a number of papers. Typically, the combination of noise distributions is encoded in the variational model considered as a linear combination of data fidelities. In particular, in [HL13] a TV- L^1 - L^2 model is considered for the combination of Gaussian and impulse noise. For the same type of problems, in a previous work of Cai et al. [CCN08] a two-phase approach is considered and a sequential smoothing with L^1 and L^2 fidelity is performed to remove the impulse and the Gaussian component of the noise, respectively. The Gaussian and Poisson mixture of noise distributions has been addressed, for instance, in [JCPT12, JCPT15] where an exact, though complex expression for the negative log-likelihood is considered and in [DLRS13, CDLRS14, CCDLR⁺15] where the authors considered a linear combination modelling of a general class of data fidelities.

The model we consider in [CDLRS15] introduces an additional variable v in the minimisation problem (1.4) which corresponds to one of the two noise components in the data. From a variational point of view, an appropriate data fidelity Φ_1 is associated to this variable and, similarly, a suitable second fidelity term Φ_2 is associated to the residual quantity (i.e. the difference between the given noisy image and the extra-variable introduced). In

its general form the model reads:

$$\min_{\substack{u \in V \subset \text{BV}(\Omega) \\ v \in W \subset L^2(\Omega)}} \left\{ |Du|(\Omega) + \lambda_1 \Phi_1(v) + \lambda_2 \Phi_2(u, f - v) \right\},$$

where V and W are suitable subspaces of $\text{BV}(\Omega)$ and $L^2(\Omega)$, respectively, where Φ_1 and Φ_2 are well-defined. The infimal convolution model considered (denoted by TV-IC) is a sound approximation to the exact statistical model accommodating mixed noise statistics. The numerical solution of this model is computed using semismooth Newton-type methods. Consistently with our modelling assumptions, the noise is decomposed into its constituting components (see Figure 1.19) and classical single-noise models are recovered asymptotically as the weighting parameters λ_1 and λ_2 of the model go to infinity.

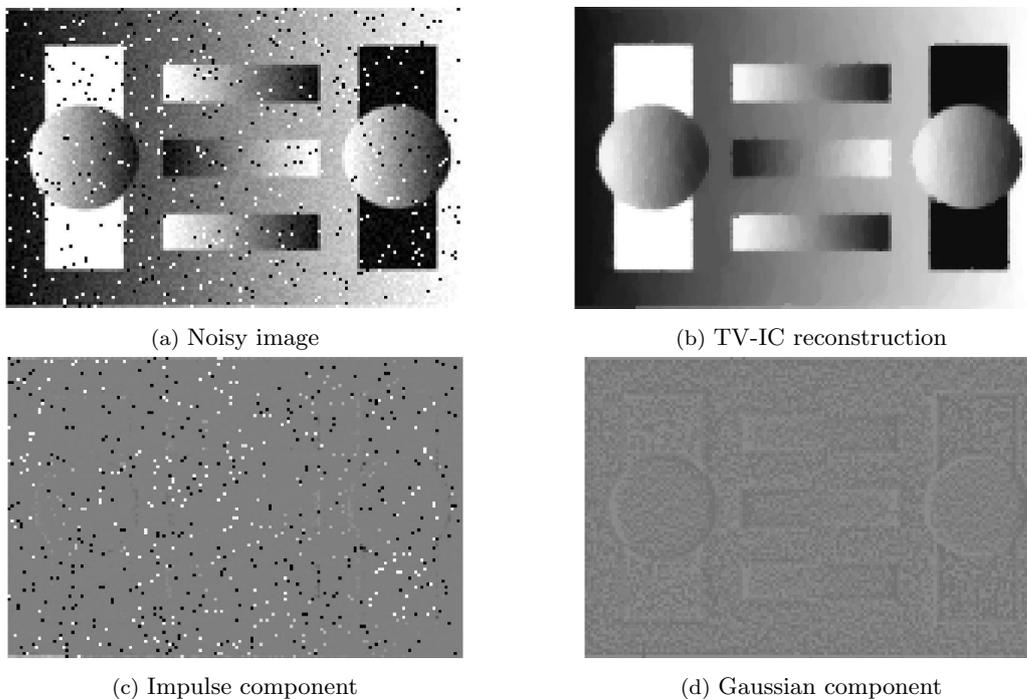


Figure 1.19: TV-IC reconstruction of a synthetic image corrupted with a mixture of impulse and Gaussian noise and noise decomposition property illustrated.

1.5.3 Learning from examples

In [CCDLR⁺15] we review some recent learning approaches in variational imaging based on bilevel optimisation and emphasise the importance of their treatment in function spaces. This survey paper is the result of a collaboration with Cao Chung, Juan Carlos De Los Reyes, Carola-Bibiane Schönlieb and Tuomo Valkonen.

The PDE-constrained framework has been considered for the purpose of *learning* the optimal ingredients of the general model (1.4) (i.e. the regulariser R , the fidelity term Φ

and the balancing parameter λ) in a number of papers, see, e.g. [HHH12, DLRS13, KP13, BNS14, SR14, CYP15, DLRSV15a, DLRSV15b]. Depending on the modelling assumptions considered, learning approaches can be regarded from two main points of view. The former uses *a priori* knowledge on the data and the type of reconstruction we seek (i.e. which noise distribution is affecting the given signals and which solution is expected), whereas in the latter one looks directly at the given data and learns the optimal setup without any further modelling assumption. Both approaches presents advantages and disadvantages in terms of adaptivity for real data and reconstruction guarantees. Hence, in order to design a reliable, adaptable and effective model (1.4) a unified modelling combining example- and data-driven learning methods is crucial.

In our work, we focus on the approach introduced in [DLRS13] for noise learning from training sets of examples, a practice that is not uncommon in real-world applications such as MRI, for instance. From an analytical point of view, results on the existence and structure of minimisers, as well as optimality conditions for their characterisation have been derived and thoroughly analysed in [DLRS13, DLRSV15b]. Based on this information, efficient second-order (Newton-type) methods are studied for the numerical solution of the problems considered. In applications, the use of databases (dictionaries) allows an accurate estimation of the noise model, but reflects in high computational costs due to the size of the databases and to the nonsmooth nature of the TV-type PDE constraints. To overcome this computational barrier we consider in [CDLRS14] an optimisation algorithm that by sampling dynamically from the set of constraints and using a quasi-Newton method, solves the problem accurately and efficiently.

Our numerical results show a robust and efficient estimation of the noise in the image and of the optimal balance between data fitting and image regularisation, see Figure 1.20. We use this framework for different single and mixed noise estimation, assessing optimality with respect to different quality measures.

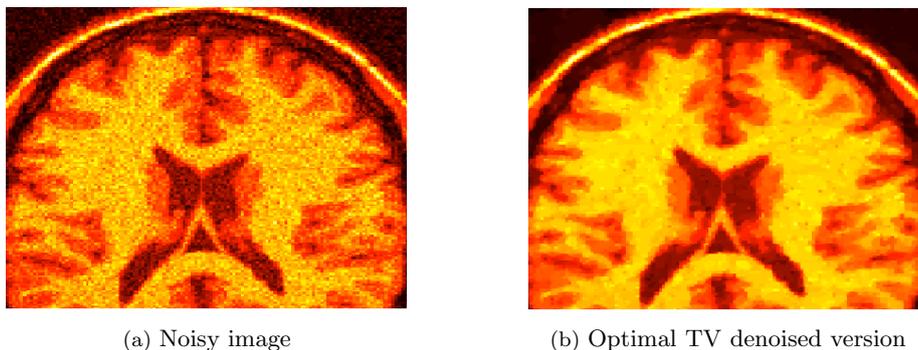
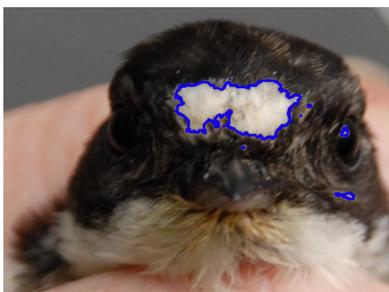


Figure 1.20: TV Gaussian denoising with optimal λ computed by means of bilevel optimisation.

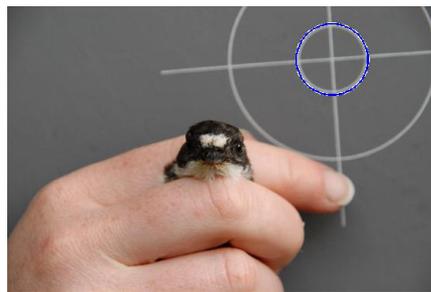
1.5.4 Image segmentation via graph-clustering

In [CvGS⁺15] we focus on a completely different framework for imaging problems. Namely, we consider the task of image segmentation and present it as a clustering problem defined on a graph (the pixel image). Furthermore, we study the problem of scale detection in images where the region of interest we desire to extract is present together with a measurement tool (e.g. a ruler). Our research finds its motivation in a behavioural study investigated by Hannah Rowland from the Department of Zoology of the University of Cambridge on a particular species of birds. The segmentation results have been obtained in collaboration with Yves van Gennip, Carola-Bibiane Schönlieb and Arjuna Flenner.

The graph clustering method considered has been presented in [BF12] and further developed in [GMB⁺14b, GMB⁺14a, MBBT15]. It reinterprets classical continuous Ginzburg-Landau minimisation models in a totally discrete framework. Nonetheless, a suitable gradient flow can be still derived in this case, its solution being the segmentation we seek. To overcome the numerical difficulties due to the large size of the images considered we use matrix completion and splitting techniques. In particular, we consider the Nyström technique presented in [Nys29] for computing the similarity matrix of the image graph through its reduced computation only in correspondence with a subset of pixels. In order to solve efficiently the Ginzburg-Landau minimisation we split the functional into the difference of two convex terms and design a time stepping scheme which simplifies the numerical solution of the resulting gradient flow by distributing the linear/nonlinear components in the implicit/explicit terms, respectively. The scale in the image is detected via a Hough transform based algorithm which identifies objects with a specific *a priori* known shape in the image by mapping them into an auxiliary space where geometrical properties (such as linear alignment or circularity) are easily identifiable. The two main steps of this method are reported in Figure 1.21.



(a) Forehead blaze segmentation



(b) Scale detection

Figure 1.21: Segmentation of male pied flycatcher forehead blaze by minimisation of the discrete Ginzburg-Landau functional defined on the image graph. The image scale is found by detecting the inner circle (with radius 1 cm) using the circular Hough transform.

The combined method is then applied to some measurement tasks arising in real-world applications such as zoology, medicine and archaeology.

The application of this method to the behavioural study involving populations of male pied flycatchers is an ongoing research project in collaboration with the Zoology department of the University of Cambridge and will be the topic of a future interdisciplinary paper [CvGS⁺].

1.6 Organisation of the thesis

In Chapter 2 we review the main mathematical ingredients used in the thesis. In particular, we follow [Eva10, AFP00, Giu84, Rud87, DM93] for a review on the main tools and definitions on functional calculus and functions of bounded variation. Then, we give a brief account of the classical computational approaches for solving TV models via smoothing. We conclude the chapter with a condensed description of the PDE-constrained optimisation analytical and numerical framework, following essentially [DLR15, Trö10].

In Chapter 3 we report the results published in [CDS14] and [BCDS13]. We start from describing the general form of the fourth-order PDEs considered, which can be seen as non-standard gradient flows of the TV functional with respect to norms defined on appropriate function spaces. Following mainly [HV03, PR55, vdHV79] we then review a particular class of operator splitting methods and show how these can be applied to solve numerically higher-order flows. As an alternative numerical strategy, we then proceed similarly as in [Mül08] and reformulate these problems in a primal-dual form and design a quasi-Newton scheme for their numerical solution. We conclude the chapter presenting some results for image inpainting and denoising problems.

Chapter 4 describes the TV-infimal convolution approach presented in [CDLRS15]. After a brief review on related variational approaches for mixed noise denoising problems, we derive the model through a detailed *maximum a posteriori* (MAP) estimation [CS05a, Stu10]. We then show well-posedness results for the model together with asymptotical properties and report the numerical results obtained via the application of a SemiSmooth Newton's (SSN) method.

In Chapter 5 we give a general review on bilevel optimisation models in the context of learning approaches for imaging based on our work in [CCDLR⁺15]. In particular, we consider the tasks of choosing an optimal regularisation [DLRSV15b], data fidelity and balance between the two [DLRS13]. In the case when a robust estimation of the noise is required by means of training databases, we provide the details on the dynamical sampling algorithm considered in [CDLRS14] for an efficient computation through SSN methods.

In Chapter 6 we revise the graph segmentation framework introduced in [BF12] and used in our paper [CvGS⁺15]. After a quick review of the main notions of graph theory, we focus on the numerical realisation of the model considered which is challenging due to the large size of the images in the database as well as their characterisation in terms of non-standard features (such as texture). The computational difficulties are overcome by means

of matrix completion [Nys29] and convex splitting [SB11] techniques. In order to detect in the image objects characterised by specific geometrical features (such as rulers), we design algorithms based on the use of the Hough transform [Hou62] for identifying those features in an auxiliary space where such properties are found as peaks of appropriate functions.

We conclude the thesis with some final comments and open questions in Chapter 7.

Chapter 2

Mathematical preliminaries

In this chapter we present the main mathematical tools and concepts which will be used in the rest of the thesis. After fixing some notation, we start from revising some definitions and properties on Radon measures in order to define the space of function of bounded variation, the main framework of our work. We then recall some basic concepts on lower continuous envelopes and convergence of functionals. Basic knowledge in real and functional analysis and on Sobolev space is assumed. We refer the reader to [Bre83, LL01, Eva10] for further readings on these topics. Moreover, we briefly review some standard concepts in numerical PDE (constrained) optimisation such as gradient-descent and Newton's methods. For these subjects we refer the reader to [Kel99, NW99, DLR15].

Notation. Let $(X, \mathcal{B}(X))$ be a metric space where $\mathcal{B}(X)$ denotes the σ -algebra of X . The space of continuous functions with compact support $u : X \rightarrow \mathbb{R}^\ell$ will be denoted by $\mathcal{C}_c(X, \mathbb{R}^\ell)$ and endowed with the supremum norm $\|u\|_\infty = \sup_{x \in X} |u(x)|$. Its completion under this norm will be denoted by $\mathcal{C}_0(X, \mathbb{R}^\ell)$. The space of \mathbb{R}^ℓ -valued, k -times continuously differentiable functions defined on an open subset Ω of \mathbb{R}^d , will be denoted by $\mathcal{C}^k(\Omega, \mathbb{R}^\ell)$, where with $k = \infty$ we will denote the space $\mathcal{C}^\infty(\Omega, \mathbb{R}^\ell)$ of infinitely differentiable functions. When considering functions with compact support, similarly as above, we will write $\mathcal{C}_c^k(\Omega, \mathbb{R}^\ell)$ and $\mathcal{C}_c^\infty(\Omega, \mathbb{R}^\ell)$, respectively.

Let now μ be a positive measure on $(X, \mathcal{B}(X))$. For $1 \leq p < \infty$, we will denote by $L^p(X, \mathbb{R}^\ell; \mu)$ the space of μ -measurable functions $u : X \rightarrow \mathbb{R}^\ell$, such that $\int_X |u|^p d\mu < \infty$ endowed with the norm:

$$\|u\|_{L^p(X, \mathbb{R}^\ell; \mu)} = \left(\int_X |u|^p d\mu \right)^{1/p}, \quad 1 \leq p < \infty,$$

while, for $p = \infty$ we will denote by $L^\infty(X, \mathbb{R}^\ell; \mu)$ the space of μ -essentially bounded

functions endowed with the norm:

$$\|u\|_{L^\infty(X, \mathbb{R}^\ell; \mu)} = \operatorname{ess\,sup}_{x \in X} |u(x)|.$$

The spaces $L^p(X, \mathbb{R}^\ell; \mu)$ are Banach spaces for every $1 \leq p \leq \infty$. For simplicity of notation, we will omit to write μ in the case when $\mu = \mathcal{L}^d$, the d -dimensional Lebesgue measure in \mathbb{R}^d , for $\Omega \subset \mathbb{R}^d$. Consequently, we will intend $L^p(\Omega, \mathbb{R}^\ell)$ as a short version of $L^p(\Omega, \mathbb{R}^\ell; \mathcal{L}^d)$. Also, we will omit to write the range of functions in the spaces above whenever $\ell = 1$. We will do similarly for $L^\infty(\Omega, \mathbb{R}^\ell)$ and $L^\infty(\Omega)$.

Sobolev spaces of real-valued functions u whose distributional derivatives up to order $k \geq 1$ are functions in $L^p(\Omega)$ will be denoted by $W^{k,p}(\Omega)$ and endowed with the norm

$$\|u\|_{W^{k,p}(\Omega)} = \left(\sum_{|\alpha| \leq k} \int_{\Omega} |D^\alpha u|^p dx \right)^{1/p}, \quad 1 \leq p < \infty,$$

$$\|u\|_{W^{k,\infty}(\Omega)} = \sum_{|\alpha| \leq k} \operatorname{ess\,sup}_{x \in \Omega} |D^\alpha u(x)|, \quad p = \infty,$$

where $D^\alpha u$ is the α -th distributional derivative of u and $|\alpha| = \alpha_1 + \dots + \alpha_d$ is the order of the multi index $\alpha = (\alpha_1, \dots, \alpha_d)$. When $p = 2$ we will denote by $H^k(\Omega)$ the Hilbert space $W^{k,2}(\Omega)$ and $H_0^k(\Omega)$ is the completion of $\mathcal{C}_c^\infty(\Omega)$ under the $\|\cdot\|_{H^k(\Omega)}$ norm. The corresponding \mathbb{R}^ℓ -valued Sobolev spaces will be denoted similarly by $W^{k,p}(\Omega, \mathbb{R}^\ell)$, $H^k(\Omega, \mathbb{R}^\ell)$ and $H_0^k(\Omega, \mathbb{R}^\ell)$.

2.1 Functional analysis

The mathematical description of an image starts from the analytical modelling of its discontinuities, i.e. the image edges. Their representation is fundamental in most of the imaging and computer-vision problems considered and the choice of a framework where discontinuous functions can be analysed is a crucial starting point for our work. Classical Sobolev spaces do not allow this modelling since the gradient of a Sobolev function is still a function. The gradient of an image has rather to be understood as a measure. In the space of function of bounded variation this is possible and a rigorous mathematical analysis can be done. In this section we recall some basic definitions about measures and define such space, giving a brief description of its main properties. Our description follows essentially [AFP00, Giu84, Rud87].

2.1.1 Basic properties of Radon measures

We summarise the main definitions and properties of a special class of measures which will serve us for the description of image discontinuities. Here and in the following Ω will

denote an open domain in \mathbb{R}^d .

Definition 2.1.1 (Finite Radon measures). *Let X be a locally compact, Hausdorff space and let $\mathcal{B}(X)$ denote the Borel σ -algebra of X . We say that a measure $\mu : \mathcal{B}(X) \rightarrow \mathbb{R}^\ell$ is a \mathbb{R}^ℓ -valued finite Radon measure if the following three conditions hold:*

- $\mu(\emptyset) = 0$,
- for any sequence $(E_n)_{n \in \mathbb{N}}$ of pairwise disjoint elements of $\mathcal{B}(X)$ we have

$$\mu \left(\bigcup_{n=0}^{\infty} E_n \right) = \sum_{n=0}^{\infty} \mu(E_n),$$

- for any compact set $K \subset \Omega$ $\mu(K) < \infty$.

We denote by $\mathcal{M}(X, \mathbb{R}^\ell)$ the space of \mathbb{R}^ℓ -valued finite Radon measures and simply write $\mathcal{M}(X)$ in the case $\ell = 1$ (real valued measures). Every $\mu \in \mathcal{M}(X, \mathbb{R}^\ell)$ can be written as $\mu = (\mu_1, \dots, \mu_\ell)$ where $\mu_i \in \mathcal{M}(X)$ for every $i = 1, \dots, \ell$.

Definition 2.1.2 (Total variation measure). *For every $\mu \in \mathcal{M}(X, \mathbb{R}^\ell)$, we define the total variation measure $|\mu| : \mathcal{B}(X) \rightarrow \mathbb{R}_+$ for every set $E \in \mathcal{B}(X)$ as follows*

$$|\mu|(E) := \sup \left\{ \sum_{n=0}^{\infty} |\mu(E_n)| : E_n \in \mathcal{B}(X) \text{ pairwise disjoint, } E = \bigcup_{n=0}^{\infty} E_n \right\}.$$

It can be shown that the measure $|\mu|$ is finite and positive.

In order to define a norm on $\mathcal{M}(X, \mathbb{R}^\ell)$, we state a representation theorem which provides another interpretation of $|\mu|(X)$.

Theorem 2.1.3 (Riesz representation theorem). *Let X be a locally compact, Hausdorff metric space and let T be a bounded linear functional on $(\mathcal{C}_0(X, \mathbb{R}^\ell), \|\cdot\|_\infty)$. Then there exists a unique element $\mu \in \mathcal{M}(X, \mathbb{R}^\ell)$ such that*

$$T(u) = \sum_{i=1}^{\ell} \int_X u_i d\mu_i, \quad \forall u \in \mathcal{C}_0(X, \mathbb{R}^\ell).$$

Moreover, the isomorphism is isometric with respect to the total variation measure, i.e.

$$\|T\|_* = |\mu|(X),$$

where $\|T\|_*$ is the operator norm of T in the dual space of $(\mathcal{C}_0(X, \mathbb{R}^\ell), \|\cdot\|_\infty)$.

Therefore, the space $\mathcal{M}(X, \mathbb{R}^\ell)$ can be viewed as the dual space of $\mathcal{C}_0(X, \mathbb{R}^\ell)$ with respect to the infinity norm. Hence, the space $\mathcal{M}(X, \mathbb{R}^\ell)$ is a Banach space under the norm $\|\mu\|_{\mathcal{M}(X, \mathbb{R}^\ell)} = |\mu|(X)$.

Recalling that $C_c^\infty(X, \mathbb{R}^\ell)$ is dense in $C_0(X, \mathbb{R}^\ell)$ under the infinity norm, we have that by definition of operator norm the following holds for every $\mu \in \mathcal{M}(X, \mathbb{R}^\ell)$:

$$\|\mu\|_{\mathcal{M}(X, \mathbb{R}^\ell)} = |\mu|(X) = \sup \left\{ \langle \mu, v \rangle : v \in C_c^\infty(X, \mathbb{R}^\ell), \|v\|_\infty \leq 1 \right\}, \quad (2.1)$$

where the duality pairing $\langle \mu, v \rangle$ is intended here as $\langle \mu, v \rangle = \sum_{i=1}^\ell \int_X v_i d\mu_i$.

Remark 2.1.4 (Radon norm of distributions). *The formula (2.1) gives an alternative definition for $\|\mu\|_{\mathcal{M}(X, \mathbb{R}^\ell)}$ and can be used in the framework of distributions as follows. Given an \mathbb{R}^ℓ -valued distribution \mathcal{T} in Ω we can define the Radon norm of \mathcal{T} as*

$$\|\mathcal{T}\|_{\mathcal{M}(\Omega, \mathbb{R}^\ell)} = \sup \left\{ \langle \mathcal{T}, v \rangle : v \in C_c^\infty(\Omega, \mathbb{R}^\ell), \|v\|_\infty \leq 1 \right\}.$$

Theorem 2.1.3 ensures that $\|\mathcal{T}\|_{\mathcal{M}(\Omega, \mathbb{R}^\ell)} < \infty$ if and only if \mathcal{T} is a finite Radon measure μ and in that case $\|\mathcal{T}\|_{\mathcal{M}(\Omega, \mathbb{R}^\ell)} = |\mu|(\Omega)$.

We recall now one of the most important tool in integration and measure theory: the Lebesgue decomposition theorem. In the following, we will use it in the context of functions of bounded variation. Hereafter, we denote by $\mathcal{M}^+(X)$ the set of finite positive Radon measures on X .

Definition 2.1.5. *Let $\mu \in \mathcal{M}^+(X)$ and $\nu \in \mathcal{M}(X, \mathbb{R}^\ell)$.*

- (i) *We say that ν is absolutely continuous with respect to μ and we write $\nu \ll \mu$, if whenever $\mu(E) = 0$, then $|\nu|(E) = 0$.*
- (ii) *We say that μ and ν are mutually singular and we write $\mu \perp \nu$ if there exists a set $E \in \mathcal{B}(X)$ such that $\mu(X \setminus E) = \nu(E) = 0$.*

Theorem 2.1.6 (Lebesgue decomposition). *Let $\mu \in \mathcal{M}^+(X)$ and $\nu \in \mathcal{M}(X, \mathbb{R}^\ell)$. Then there exists a unique pair $\nu^{ac}, \nu^s \in \mathcal{M}(X, \mathbb{R}^\ell)$ such that*

$$\nu = \nu^{ac} + \nu^s, \quad \nu^{ac} \ll \mu, \quad \nu^s \perp \mu.$$

Moreover, there exists a unique $f \in L^1(X, \mathbb{R}^\ell; \mu)$ satisfying $\nu^{ac} = f\mu$. The element f is called the Radon-Nikodym derivative of ν^{ac} with respect to μ and it is denoted by $f = \frac{\nu^{ac}}{\mu}$. Moreover, the following properties hold:

$$\frac{\nu}{\mu} = \frac{\nu^{ac}}{\mu}, \quad \text{and} \quad \frac{\nu^s}{\mu} = 0, \quad \mu - a.e. \quad \text{and} \quad \nu(E) = \int_E \frac{\nu^{ac}}{\mu} d\mu + \nu^s(E),$$

for every $E \in \mathcal{B}(X)$.

All the following Lebesgue decompositions will be considered with respect to the d -dimensional Lebesgue measure \mathcal{L}^d , which will not be explicitly written unless needed.

Hence, with the notation μ^{ac} and μ^s we will denote the absolutely continuous and singular part of μ with respect to the Lebesgue measure, respectively.

We now recall the classical notion of weak* convergence for measures in $\mathcal{M}(X, \mathbb{R}^\ell)$.

Definition 2.1.7 (Weak* convergence of measures). *We say that the sequence $(\mu_n)_{n \in \mathbb{N}} \subseteq \mathcal{M}(X, \mathbb{R}^\ell)$ converges weakly* to $\mu \in \mathcal{M}(X, \mathbb{R}^\ell)$ if for every $u \in \mathcal{C}_0(X)$ we have*

$$\lim_{n \rightarrow \infty} \int_X u d\mu_n = \int_X u d\mu.$$

The above notion of weak* convergence is the usual notion of weak* convergence in dual spaces of Banach spaces since from the Riesz representation Theorem 2.1.3 $\mathcal{M}(X, \mathbb{R}^\ell) = (\mathcal{C}_0(X, \mathbb{R}^\ell))^*$.

2.1.2 Functions of bounded variation

We now define the space of functions of bounded variation, the principal framework of the mathematical imaging models we will consider in this work. Dealing with functions with discontinuities in this space corresponds to consider edges in the modelling of images. Here, we examine the main properties of this space.

Definition 2.1.8 (Functions of bounded variation). *A function $u \in L^1(\Omega)$ is a function of bounded variation or, alternatively, $u \in \text{BV}(\Omega)$ if its distributional derivative Du can be represented by a \mathbb{R}^d -valued finite Radon measure. For simplicity, such measure will be still denoted by Du , but interpreted as an element of $\mathcal{M}(\Omega, \mathbb{R}^d)$. In other words, $u \in \text{BV}(\Omega)$ if there exists $Du \in \mathcal{M}(\Omega, \mathbb{R}^d)$ such that*

$$\int_\Omega u \frac{\partial v}{\partial x_i} dx = - \int_\Omega v dD_i u, \quad \forall v \in \mathcal{C}_c^\infty(\Omega), \quad i = 1, \dots, d.$$

Thanks to Lebesgue decomposition Theorem 2.1.6, Du can be decomposed into its absolutely continuous and singular part with respect to the Lebesgue measure as follows:

$$Du = \nabla u + D^s u, \tag{2.2}$$

where, for simplicity, we have denoted by ∇u the absolutely continuous part $(Du)^{ac}$ and by $D^s u$ the singular part $(Du)^s$ of Du , respectively.

Remark 2.1.9. *From Definition 2.1.8 we observe that $W^{1,1}(\Omega) \subseteq \text{BV}(\Omega)$ since if $u \in W^{1,1}(\Omega)$ then $Du = \nabla u$. Hence, $\text{BV}(\Omega)$ is a more general space than the standard Sobolev space $W^{1,1}(\Omega)$.*

Definition 2.1.10 (Total Variation). *For a function $u \in L^1(\Omega)$ the total variation of u ,*

$\text{TV}(u)$, is defined as

$$\text{TV}(u) = \sup \left\{ \int_{\Omega} u \nabla \cdot v \, dx : v \in \mathcal{C}_c^1(\Omega, \mathbb{R}^d), \|v\|_{\infty} \leq 1 \right\}. \quad (2.3)$$

It can be proved that $\text{TV}(u) < \infty$ if and only if $u \in \text{BV}(\Omega)$ and in that case $\text{TV}(u) = |Du|(\Omega)$, see Remark 2.1.4. Hence, in the following we will indifferently use the notations $\text{TV}(u)$ and $|Du|(\Omega)$ to indicate the total variation of functions $u \in \text{BV}(\Omega)$.

Furthermore, if $u \in W^{1,1}(\Omega, \cdot)$ then $|Du|(\Omega)$ is simply $\|\nabla u\|_{L^1(\Omega)} = \int_{\Omega} |\nabla u| \, dx$. One side of the equality is a simple application of the Gauss-Green divergence theorem, whereas the other inequality can be shown using mollification techniques [Fol99].

Finally, we note that by definition, the total variation is lower semicontinuous with respect to the strong L^1 convergence being the supremum of functions with such property.

Properties of $\text{BV}(\Omega)$

The space $\text{BV}(\Omega)$ endowed with the norm

$$\|u\|_{\text{BV}(\Omega)} = \|u\|_{L^1(\Omega)} + |Du|(\Omega), \quad (2.4)$$

is a Banach space with bad compactness properties (for instance, smooth functions are not dense with respect to the topology induced by the norm above). Hence, a weaker notion of convergence is generally considered.

Definition 2.1.11 (Weak* convergence in BV). *We say that the sequence $(u_n)_{n \in \mathbb{N}} \subseteq \text{BV}(\Omega)$ converges to u weakly* in $\text{BV}(\Omega)$ if*

$$\lim_{n \rightarrow \infty} \|u_n - u\|_{L^1(\Omega)} = 0 \quad \text{and} \quad \lim_{n \rightarrow \infty} \int_{\Omega} v \, dDu_n = \int_{\Omega} v \, dDu, \quad \forall v \in \mathcal{C}_0(\Omega).$$

In other words, (u_n) converges weakly to u in $\text{BV}(\Omega)$ if (u_n) converges to u in $L^1(\Omega)$ and (Du_n) converges to Du weakly* in $\mathcal{M}(\Omega, \mathbb{R}^d)$.*

The norm in the space of functions of bounded variation introduced in (2.4) shows better compactness properties with respect to weak* convergence introduced above. This appears clear in the following compactness theorem which is often used to prove existence results through the direct method of calculus of variations in order to overcome the non-reflexivity of the Sobolev space $W^{1,1}(\Omega)$.

Theorem 2.1.12 (Compactness in BV). *Let $(u_n)_{n \in \mathbb{N}}$ be a sequence in $\text{BV}(\Omega)$ such that*

$$\sup_{n \in \mathbb{N}} \|u_n\|_{\text{BV}(\Omega)} < \infty.$$

Then, there exists a subsequence $(u_{n_k})_{k \in \mathbb{N}}$ and some $u \in \text{BV}(\Omega)$, such that $u_n \rightarrow u$ weakly in $\text{BV}(\Omega)$.*

Similar definitions and properties can be given also in the vectorial case, that is when $u \in L^1(\Omega, \mathbb{R}^\ell)$ with $\ell > 1$. In our exposition, we limit ourselves to the scalar case for better clarity.

We recall now an embedding theorem for the space BV.

Theorem 2.1.13. (BV embedding theorem) *Let Ω an open bounded domain in \mathbb{R}^d with Lipschitz boundary. Then,*

$$\text{BV}(\Omega) \subseteq L^{d^*}(\Omega)$$

with continuous embedding. In the inclusion above $d^ = d/(d - 1)$ if $d > 1$ and $d^* = \infty$ if $d = 1$.*

Recalling the Lebesgue decomposition (2.2) for functions $u \in \text{BV}(\Omega)$ we can now say more for the singular part $D^s u$ of the distributional gradient Du . More precisely, in [Amb89] Ambrosio showed that $D^s u$ can be further decomposed into a “jump” part J_u and a singular “Cantor” part C_u such that (2.2) actually becomes:

$$Du = \nabla u + \overbrace{J_u + C_u}^{D^s u}. \quad (2.5)$$

Heuristically, the presence of the jump part J_u in (2.5) clarifies the extension to the BV space introduced above. Jump discontinuities are in fact encoded in this term. The analysis of problems formulated in this space is often referred to as “nonsmooth” analysis, due to the low regularity of the functions considered. Nonetheless, the tools and the properties of functions in this space, still ensure a mathematically interesting analysis where notions from functional analysis and calculus of variations can be still used to study the models in a rigorous way.

2.1.3 Relaxed functionals and Γ -convergence

Recalling the variational approach for imaging problems introduced in Section 1.2.2, we now follow [DM93, Bra02] and recall some notions in the case when J is not weakly lower semicontinuous. In this case one cannot expect in general to guarantee the existence of a minimum for the problem (1.4). However, an auxiliary functional can be associated to J such that the minima of the auxiliary functional are weak cluster points of minimising sequences for J . Similarly, we recall the concept of Γ -convergence which gives a meaning to the convergence of sequence of functionals.

In the following we denote by $\overline{\mathbb{R}}$ the extended real line, i.e., $\overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$ and consider for simplicity X to be a separable Banach space endowed with a topology τ .

Let us start by recalling the definition of lower semicontinuity with respect to the topology τ .

Definition 2.1.14. *The functional J is called τ -lower semicontinuous (l.s.c.) at a point $x \in X$ if for all sequences $(x_n)_{n \in \mathbb{N}}$ converging to x in the topology τ , we have*

$$J(x) \leq \liminf_{n \rightarrow \infty} J(x_n).$$

J is l.s.c. if it is l.s.c. at every point.

In the following, we shall think at τ to be either the strong or the weak topology of X . In the direct method of calculus of variations the notion of weak l.s.c. emerges very naturally, but, in general, it is a hard property to prove. A sufficient condition that implies weak l.s.c. is convexity.

Theorem 2.1.15. *Let $J : X \rightarrow \overline{\mathbb{R}}$ be convex. Then J is weakly l.s.c. if and only if J is strongly l.s.c.*

The theorem above is useful, since proving strong l.s.c. is not very hard in general.

Let us now consider the following definition which will be useful in the following to deal with problems which are not τ -lower semicontinuous.

Definition 2.1.16 (Lower semicontinuous envelope). *Let $J : X \rightarrow \overline{\mathbb{R}}$. The τ -lower semicontinuous envelope (also called the relaxed functional) of J with respect to the topology τ , $\text{sc}_\tau J : X \rightarrow \overline{\mathbb{R}}$ is defined for every $x \in X$ by*

$$\text{sc}_\tau J(x) := \sup \{ G(x) : G : X \rightarrow \overline{\mathbb{R}}, \tau\text{-l.s.c.}, \text{ and } G(y) \leq J(y), \forall y \in X \}.$$

In words, $\text{sc}_\tau J$ is the greatest τ -lower semicontinuous functional which is smaller or equal than J .

Theorem 2.1.17. *The τ -lower semicontinuous envelope $\text{sc}_\tau J$ of J with respect to the topology τ is characterised by the following two properties:*

(i) *For every sequence $(x_n)_{n \in \mathbb{N}}$ τ -converging to x , we have*

$$\text{sc}_\tau J(x) \leq \liminf_{n \rightarrow +\infty} J(x_n).$$

(ii) *For every $x \in X$ there exists a sequence $(x_n)_{n \in \mathbb{N}}$ τ -converging to x , such that*

$$\text{sc}_\tau J(x) \geq \limsup_{n \rightarrow +\infty} J(x_n).$$

In the following Proposition we consider the relation between the original and the relaxed problem.

Proposition 2.1.18. *Let $J : X \rightarrow \overline{\mathbb{R}}$ and suppose that $\inf_{x \in X} J(x) \in \mathbb{R}$. Then if $\text{sc}_\tau J$ has a minimum point we have that*

- (i) $\min_{x \in X} \text{sc}_\tau J(x) = \inf_{x \in X} J(x)$,
- (ii) Every cluster point of a minimising sequence for J is a minimum point for $\text{sc}_\tau J$,
- (iii) Every minimum point for $\text{sc}_\tau J$ is the limit of a minimising sequence for J .

Hence, starting with a minimisation problem for which existence of a solution is not guaranteed, a relaxed, ‘close’ problem can be defined. The connections with the original one are the ones stated in the Proposition 2.1.18.

Another notion which turns out to be particularly useful for approximating non-convex problems is the notion of Γ -convergence of sequence of functionals.

Definition 2.1.19 (Γ -convergence). *Let $(J_n)_{n \in \mathbb{N}}$ a sequence of functionals such that $J_n : X \rightarrow \overline{\mathbb{R}}$ for every n . We say that J_n Γ -converges to J for the topology τ if the following two conditions hold:*

- (i) For every $x \in X$ and every sequence $(x_n)_{n \in \mathbb{N}}$ τ -converging to x in X ,

$$J(x) \leq \liminf_{n \rightarrow +\infty} J_n(x_n).$$

- (ii) For every $x \in X$ there exists a sequence $(x_n)_{n \in \mathbb{N}}$ τ -converging to x in X such that:

$$J(x) \geq \limsup_{n \rightarrow +\infty} J_n(x_n).$$

Definition 2.1.20 (Equicoercive sequence of functionals). *A sequence of functionals $(J_n)_{n \in \mathbb{N}}$ is said to be equicoercive if for every $t \geq 0$ there exists a compact subset K_t of X such that $\{x \in X : J_n(x) \leq t\} \subset K_t$ for every $n \in \mathbb{N}$.*

We can finally condensate the main properties of Γ -convergence and relaxed functionals in the following theorem.

Theorem 2.1.21. *Let $(J_n)_{n \in \mathbb{N}}$ be a sequence of equicoercive functionals from X to $\overline{\mathbb{R}}$. Then:*

- (i) If the Γ -limit J of $(J_n)_{n \in \mathbb{N}}$ exists, J is unique and l.s.c.
- (ii) There exists a subsequence $(J_{n_k})_{k \in \mathbb{N}} \subset (J_n)_{n \in \mathbb{N}}$ and J such that $J = \Gamma\text{-}\lim_{k \rightarrow +\infty} J_{n_k}$.
- (iii) If $J = \Gamma\text{-}\lim_{n \rightarrow +\infty} J_n$, then $J + G = \Gamma\text{-}\lim_{h \rightarrow +\infty} (J_n + G)$ for every continuous function $G : X \rightarrow \mathbb{R}$.
- (iv) If $(J_n)_{n \in \mathbb{N}}$ converges to J uniformly, then $(J_n)_{n \in \mathbb{N}}$ Γ -converges to J .
- (v) If $(J_n)_{n \in \mathbb{N}}$ is an increasing sequence of l.s.c. functionals converging to J pointwise, then J is l.s.c. and $(J_n)_{n \in \mathbb{N}}$ Γ -converges to $\text{sc}_\tau J$.
- (vi) If $(J_n)_{n \in \mathbb{N}}$ is a decreasing sequence converging to J pointwise, then $(J_n)_{n \in \mathbb{N}}$ Γ -converges to $\text{sc}_\tau J$, the τ -lower semicontinuous envelope of J defined in 2.1.16.

2.2 Derivatives and subgradients

Let now V be a Banach space endowed with norm $\|\cdot\|_V$ and $F : V \rightarrow \mathbb{R}$. Following [Eva10], we list here different notions of differentiability for the functional F . In the following we will use the standard notation V^* to indicate the analytical dual space of V , i.e. the space of linear and continuous functionals $T : V \rightarrow \mathbb{R}$. The duality pairing between V^* and V will be denoted by $\langle \cdot, \cdot \rangle_{V^*, V}$.

We start this section from the simplest notion of differentiability in one single direction.

Definition 2.2.1 (Directional derivative). *Let $F : V \rightarrow \mathbb{R}$. If, for given $v, h \in V$ the limit:*

$$\delta F(v)[h] := \lim_{t \rightarrow 0^+} \frac{F(v + th) - F(v)}{t}$$

exists, then $\delta F(v)[h]$ is called the directional derivative of F at v in the direction h . If the limit exists for every $h \in V$, F is called directionally differentiable at v .

In the case where F is everywhere directionally differentiable and its derivative is a linear and continuous operator, we have a stronger notion of differentiability.

Definition 2.2.2 (Gâteaux differentiability). *Let $F : V \rightarrow \mathbb{R}$. If for a given $v \in V$ and every $h \in V$ the limit:*

$$\delta F(v)[h] = \lim_{t \rightarrow 0^+} \frac{F(v + th) - F(v)}{t}$$

exists and $\delta F \in V^$, then $\delta F(v)$ is called the Gâteaux derivative of F at v and F is called Gâteaux differentiable at v .*

Finally, we introduce a notion of differentiability in norm. The following definition can be easily generalised to the case when $F : V \rightarrow W$, where W is a normed space.

Definition 2.2.3 (Fréchet differentiability). *Let $F : V \rightarrow \mathbb{R}$. If F is Gâteaux differentiable at $v \in V$ and satisfies in addition*

$$\lim_{\|h\|_V \rightarrow 0^+} \frac{|F(v + h) - F(v) - \delta F(v)h|}{\|h\|_V} = 0,$$

then F is said to be Fréchet differentiable at v . In this case the operator $\delta F(v)$ is denoted by $\nabla F(v)$ or, equivalently, with $F'(v)$ and called the Fréchet derivative of F in v .

Higher-order directional, Gâteaux and Fréchet derivatives can be defined similarly. We list in the following the main properties of these derivatives.

Proposition 2.2.4 (Properties of Fréchet derivative). *Let $F : V \rightarrow \mathbb{R}$ be Fréchet differentiable at $v \in V$. Then:*

- (i) F is continuous at v ;

(ii) $F'(v) = \left[\frac{d}{dt} F(v + tw) \right]_{t=0}$ for every $w \in V$.

Let V , W and Z be three Banach spaces and let $F : V \rightarrow W$ and $G : W \rightarrow Z$ be Fréchet differentiable at $v \in V$ and $F(v)$, respectively. Then the function $H : V \rightarrow Z$ defined by:

$$H(v) := G \circ (F(v))$$

is also Fréchet differentiable and its Fréchet derivative is given by the chain rule:

$$H'(v) = G'(F(v))F'(v).$$

Remark 2.2.5. Expressed in the form (ii) of the proposition above, $F'(v)$ is also called functional derivative of F in v .

Finally, we introduce now a weaker notion of differentiability which will be used in the following for the design of second-order numerical methods in the case of nonsmooth functions.

Definition 2.2.6 (Newton differentiability). Let D an open subset of V . The mapping $F : V \rightarrow \mathbb{R}$ is called Newton differentiable on D if there exists a generalised derivative $G \in V^*$ such that

$$\lim_{\|h\|_V \rightarrow 0^+} \frac{|F(v+h) - F(v) - G(v+h)h|}{\|h\|_V} = 0,$$

for every $v \in D$. In this case the operator $G(v)$ is called the Newton derivative of F at v .

We introduce now a more general notion of differentiability which can be used in the case when the limits above do not exist.

Definition 2.2.7 (Subdifferential and subgradients). Let $F : V \rightarrow \mathbb{R}$. The subdifferential of F at $v \in V$ is defined as the set:

$$\partial F(v) := \{p \in V^* : \langle p, w - v \rangle_{V^*, V} \leq F(w) - F(v), \quad \forall w \in V\}.$$

If $\partial F(v) \neq \emptyset$ we say that F is subdifferentiable at v and we call each element of $\partial F(v)$ a subgradient of F in v .

Of course, in the case when F is Fréchet differentiable in $v \in V$ we have $\partial F(v) = \{F'(v)\}$. However, subdifferentials turn out to be more interesting in the (convex) nonsmooth case where the subdifferential is a multivalued operator.

2.3 Computational approaches to TV models

We now discuss on some computational aspects regarding the numerical solutions of TV-based models. Due to the nonsmoothness of the TV regularisation functional (2.3), the

numerical solution of the the minimisation problem

$$\min_{u \in \text{BV}(\Omega)} J(u) := \text{TV}(u) + \lambda \Phi(u, u_0) \quad (2.6)$$

is a challenging task which has been addressed in several works, see, e.g., [ROF92, AV94, VO96, CL97, CM99, Vog02, CS05a]. In this section we focus on the regularisations of TV classically used in literature to solve the problem (2.6) numerically. Therefore, we assume for simplicity in this section that the fidelity term Φ is convex and differentiable, i.e. it does not add computational difficulties to our problem. The case of nonsmooth fidelities will be addressed later on.

In the following, we list three numerical regularisations of TV which are used in this thesis. The first two techniques aim to regularise the TV functional in order to make it differentiable so that the corresponding Euler-Lagrange equations can be considered and their numerical solutions can be computed either via some fixed-point or steepest descent methods with appropriate boundary conditions and a suitable initial guess. They can be seen as numerical strategies solving the *primal* formulation of TV models where the image u is the *primal* variable of both a regularised version of the minimisation problem (2.6) and the associated Euler-Lagrange equations. An alternative choice is represented by the *dual* formulation of the problem (2.6) where one takes into account also the dual variable $v \in \mathcal{C}_c^1(\Omega, \mathbb{R}^d)$ with the constraint $\|v\|_\infty \leq 1$ appearing in the definition of TV (2.3). These model have been addressed in a number of papers due to the resulting efficient computational algorithms they lead to, see for instance [Cha04]. Combinations of the two models also exist in literature: there, the primal and the dual variable u and v are used at the same time. These methods are known in literature under the name of *primal-dual* methods and have been considered, for instance, in [CGM99, CP11].

2.3.1 Primal formulation

We assume that u is a smooth function in the space $W^{1,1}(\Omega)$ and compute the Euler-Lagrange equation of (2.6) by taking the first variation of J and get:

$$0 = -\nabla \cdot \left(\frac{\nabla u}{|\nabla u|} \right) + \lambda \Phi'(u, u_0), \quad (2.7)$$

where the divergence term is intended in the distributional sense and suitable boundary conditions (such as homogeneous Neumann) are considered.

ε -regularisation

Let us consider now the following regularised TV-functional defined for $0 \leq \varepsilon \ll 1$:

$$\text{TV}_\varepsilon(u) := \int_{\Omega} \sqrt{|\nabla u|^2 + \varepsilon} \, dx. \quad (2.8)$$

Whenever $\varepsilon = 0$, TV_ε reduces simply to the TV functional (2.3) for the smooth case, whereas taking $\varepsilon > 0$ and replacing the TV term with TV_ε in (2.6), i.e. considering the regularised functional

$$J_\varepsilon(u) := \text{TV}_\varepsilon(u) + \lambda \Phi(u, u_0),$$

offers a number of advantages, like, for instance the property of differentiability. Using convex analysis tools, Acar and Vogel showed in [AV94] that TV_ε is actually well-defined for every $L^1(\Omega)$ and that its effective domain is, in fact, $\text{BV}(\Omega)$. Moreover, the authors showed that the nonsmooth TV functional can be interpreted as the pointwise limit of TV_ε as $\varepsilon \rightarrow 0$ and that, for every $\varepsilon \geq 0$, TV_ε is convex and lower semicontinuous with respect to the weak topologies of $L^p(\Omega)$ for $1 \leq p < \infty$. Under some additional restrictions on p , such properties ensure the existence of minimisers of J_ε in $L^p(\Omega)$ which are shown to converge to the minimiser of J in (2.6) as $\varepsilon \rightarrow 0$.

The regularised version of the Euler-Lagrange equation (2.7) reads:

$$0 = -\nabla \cdot \left(\frac{\nabla u}{\sqrt{|\nabla u|^2 + \varepsilon}} \right) + \lambda \Phi'(u, u_0), \quad (2.9)$$

which is intended in the distributional sense and for suitable boundary conditions on $\partial\Omega$. Alternatively, as in [ROF92], one can write the steepest descent for (2.9) with artificial time t as:

$$\frac{\partial u}{\partial t} = \nabla \cdot \left(\frac{\nabla u}{\sqrt{|\nabla u|^2 + \varepsilon}} \right) - \lambda \Phi'(u, u_0) \quad (2.10)$$

with the same choice of boundary conditions on $\partial\Omega$ and initial guess $u(x, 0) = u_0(x)$. Equation (2.10) can be solved numerically by using explicit Euler [ROF92], lagged diffusivity [CM99] or Newton's [BCM⁺00] methods.

Huber regularisation

Alternatively, let us consider the continuous Huber regularisation of the absolute value of the distributional gradient of $u \in W^{1,1}(\Omega)$ depending on a parameter $\gamma \gg 1$:

$$|\nabla u|_\gamma := \begin{cases} |\nabla u| - \frac{1}{2\gamma} & \text{if } |\nabla u| \geq \frac{1}{\gamma}, \\ \frac{\gamma}{2} |\nabla u|^2 & \text{if } |\nabla u| < \frac{1}{\gamma}. \end{cases} \quad (2.11)$$

The corresponding regularised version of the nonsmooth problem 2.6 is then:

$$\min_{u \in W^{1,1}(\Omega)} J_\gamma(u) := \text{TV}_\gamma(u) + \lambda \Phi(u, u_0),$$

with

$$\text{TV}_\gamma(u) := \int_{\Omega} |\nabla u|_\gamma \, dx$$

be the Huber-TV functional. According to the (2.11), regions in the image with large gradients (i.e. edges) are treated in a L^1 fashion while the numerical difficulties arising in homogeneous regions (compare (2.7)) are overcome by an L^2 -smoothing. The corresponding regularised Euler-Lagrange equation reads in this case:

$$0 = -\nabla \cdot \left(\frac{\gamma \nabla u}{\max(\gamma |\nabla u|, 1)} \right) + \lambda \Phi'(u, u_0). \quad (2.12)$$

Such equation can be solved, for instance, using SemiSmooth Newton's methods as in [DLRS13].

As detailed in [Val04, DLRSV15b], we remark that the functional TV_γ can be defined in general for functions $u \in L^1(\Omega)$ using duality as:

$$\text{TV}_\gamma(u) = \sup \left\{ \int_{\Omega} u \nabla \cdot v \, dx - \frac{\gamma}{2} \|v\|_{L^2(\Omega, \mathbb{R}^d)}^2 : v \in \mathcal{C}_c^1(\Omega, \mathbb{R}^d), \|v\|_\infty \leq 1 \right\}.$$

Moreover, in [BPPS15], the authors showed that the regularised TV_γ functional can be equivalently expressed as the infimal convolution of the TV semi-norm and the L^2 norm.

Huber-regularisation is useful especially in the context of second-order optimisation methods [DLRS13, DLRLM14]. For this type of problems, smoother (namely, C^2) Huber-type regularisations are considered in order to guarantee further differentiability of the regularisation term.

2.3.2 Dual formulation

In order to overcome the non-differentiability of the TV regulariser, one other approach consists in reformulating the minimisation problem (2.6) in terms of the dual variable $v \in \mathcal{C}_c^1(\Omega, \mathbb{R}^d)$ with the side constraint $\|v\|_\infty \leq 1$.

Namely, writing down explicitly definition of TV given in (2.3), we have that (2.6) becomes:

$$\min_{u \in \text{BV}(\Omega)} \left\{ \sup_{v \in \mathcal{C}_c^1(\Omega, \mathbb{R}^d), \|v\|_\infty \leq 1} \int_{\Omega} u \nabla \cdot v \, dx + \lambda \Phi(u, u_0) \right\}.$$

Since the fidelity term does not involve v , the problem above can be re-written as the

saddle-point problem:

$$\min_{u \in \text{BV}(\Omega)} \sup_{v \in \mathcal{C}_c^1(\Omega, \mathbb{R}^d), \|v\|_\infty \leq 1} \left\{ \int_{\Omega} u \nabla \cdot v \, dx + \lambda \Phi(u, u_0) \right\}. \quad (2.13)$$

By exchanging the min and the sup operator due to classical convex optimisation results (see, e.g., [BV04, ET76, HUL93]), we get:

$$\sup_{v \in \mathcal{C}_c^1(\Omega, \mathbb{R}^d), \|v\|_\infty \leq 1} \min_{u \in \text{BV}(\Omega)} \left\{ \int_{\Omega} u \nabla \cdot v \, dx + \lambda \Phi(u, u_0) \right\}. \quad (2.14)$$

The strategy now consists in finding an expression of u in terms of the dual variable v by computing the optimality conditions in (2.14) and use such expression to formulate an optimisation problem in terms of the sole variable v which can be solved easily. To have a better understanding of this, we consider the simple case of a squared data fidelity.

Example (L^2 -data fidelity) In the case of quadratic data fidelity term $\Phi = \frac{1}{2} \|u - u_0\|_2^2$ (such as in the classical ROF denoising model [ROF92]), it is easy to compute the optimal u by setting the gradient of the objective function in brackets in (2.14) to zero, thus obtaining

$$u = u_0 - \frac{1}{\lambda} \nabla \cdot v, \quad (2.15)$$

which can now be inserted into (2.13), yielding in this case:

$$\sup_{v \in \mathcal{C}_c^1(\Omega, \mathbb{R}^d), \|v\|_\infty \leq 1} \left\{ \int_{\Omega} u_0 \nabla \cdot v \, dx - \frac{1}{2\lambda} \int_{\Omega} (\nabla \cdot v)^2 \, dx \right\}. \quad (2.16)$$

Therefore, by computing the optimal v and substituting in (2.15), the optimal primal variable u can be easily found.

In synthesis, duality methods translate the computation of the solution of the minimisation problem (2.6) in the computation of the solution of dual problems similar to (2.16). For more general fidelities Φ , a dual formulation of the problem (2.6) can be derived by using classical tools from convex analysis such as subgradients and convex conjugates, see, e.g., [Cha04]. Compared to primal methods, dual methods allow for the design of efficient algorithms, the main difficulty in their design being the constraint on the dual variable v .

2.3.3 Primal-dual formulation

Another approach for solving (2.6) consists in solving directly the saddle problem (2.13)

$$\min_{u \in \text{BV}(\Omega)} \sup_{v \in \mathcal{C}_c^1(\Omega, \mathbb{R}^d), \|v\|_\infty \leq 1} L(u, v), \quad (2.17a)$$

where

$$L(u, v) := \int_{\Omega} u \nabla \cdot v \, dx + \lambda \Phi(u, u_0). \quad (2.17b)$$

We get the following necessary saddle-point conditions:

$$\frac{\partial L}{\partial u} = \nabla \cdot v + \lambda \Phi'(u, u_0) = 0,$$

and

$$L(u, v) \geq L(u, w), \quad \text{for every } w \in \mathcal{C}_c^1(\Omega, \mathbb{R}^d), \quad \text{s.t. } \|w\|_{\infty} \leq 1. \quad (2.18)$$

Condition (2.18) can be equivalently written as:

$$\int_{\Omega} u \nabla \cdot (v - w) \, dx \geq 0 \quad \text{for every } w \in \mathcal{C}_c^1(\Omega, \mathbb{R}^d), \quad \text{s.t. } \|w\|_{\infty} \leq 1,$$

which can be shown to be actually a sufficient condition to guarantee that u is a solution of (2.6) (see, for instance, [Mül08]). An example of a celebrated primal-dual algorithm is the one that has been proposed by Chambolle and Pock in [CP11] where applications to several image denoising, inpainting and segmentation problems are shown.

Relaxation of the dual constraint

In order to rewrite (2.17) in an unconstrained form, the constraint on the variable v can be relaxed by adding a standard penalty term which penalises the functional whenever the constraint is not satisfied. In particular, one possibility is to replace $L(u, v)$ in (2.17b) by:

$$L_{\eta}(u, v) := L(u, v) - \frac{1}{2\eta} F(|v|)$$

where $0 < \eta \ll 1$ is a weighting parameter and F penalises whenever $|v| > 1$. A classical example for F is:

$$F(|v|) = \frac{1}{2} \int_{\Omega} \max(|v|, 1)^2 \, dx. \quad (2.19)$$

We observe that with this choice violations of the constraint for v are allowed, but at the points where they occur the functional L_{η} is penalised. Stronger relaxations can be used, preventing the constraint from being violated. These terms have usually logarithmic structure and are known under the name of *barrier* terms or, in alternative, as *interior-point* methods, see [Mül08].

In the case of smooth functions $u \in H^1(\Omega)$, the relaxation on the constraint of the dual variable v given by (2.19) can be shown to correspond to the choice

$$\text{TV}_{\eta}(u) := \frac{\eta}{2} \int_{\Omega} |\nabla u|^2 \, dx + \int_{\Omega} |\nabla u| \, dx \quad (2.20)$$

as primal functional. This can be interpreted as a combination of a nonsmooth L^1 regular-

isation smoothed by a small diffusion term whose strength is modulated by the weighting parameter η .

Acknowledgement. The interpretation of the primal functional (2.20) for the case of smooth functions described above is the result of a nice discussion the author of the thesis had with Dr. Michael Möller, Department of Computer Science, Technische Universität München, Germany.

2.4 PDE-constrained optimisation

We conclude this chapter with a condensed revision of the classical numerical approaches used in optimisation to find solutions of PDE-constrained minimisation problems (which will also be referred to later on in this thesis as *bilevel optimisation* problems). In our exposition, we follow mainly [DLR15, Trö10]. In Chapter 5 we will consider extensions of these methods to the nonsmooth case. For the sake of completeness, we start our exposition with a brief review on the underlying theoretical framework of these models.

A standard PDE-constrained bilevel optimisation problem has the form:

$$\min_{(w,v) \in W \times V} \mathcal{F}(w, v) \quad (2.21a)$$

subject to the PDE constraint:

$$e(w, v) = 0, \quad (2.21b)$$

where $\mathcal{F} : W \times V \rightarrow \mathbb{R}$, $e : W \times V \rightarrow B$ and W, V, B are reflexive Banach spaces. Hereafter, we assume that the PDE (2.21b) has a unique solution $w(v)$ and refer to the operator $\mathcal{S} : V \rightarrow W$ which assigns to each $v \in V$ the solution $w(v)$ of (2.21b) as the *control-to-state* or *solution* operator. Using this operator, the optimisation problem (2.21) can be written in reduced form as:

$$\min_{v \in V} F(v) := \mathcal{F}(\mathcal{S}(v), v) = \mathcal{F}(w(v), v). \quad (2.22)$$

For simplicity, we also assume in the following that $F : V \rightarrow \mathbb{R}$ is bounded from below and sufficiently smooth, specifying the regularity required as long as we get on. Under standard coercivity and weak lower semicontinuity assumptions, problem (2.22) has a global solution. Under the additional assumption of strictly convexity, the solution is unique.

We recall now a standard property satisfied by every solution $\bar{v} \in V$ of (2.22) which will be used in the following.

Theorem 2.4.1. *Let $\bar{v} \in V$ a local minimum in (2.22). If F is directionally differentiable at \bar{v} in direction $\eta - \bar{v} \in V$, then:*

$$\delta F(\bar{v})[\eta - \bar{v}] \geq 0.$$

Additionally, if F is Fréchet differentiable at \bar{v} there holds:

$$F'(\bar{v})[u] = 0, \quad \text{for every } u \in V. \quad (2.23)$$

2.4.1 Characterisations of solutions

We now derive the necessary conditions used in practice to characterise solutions of the minimisation problem (2.22) in terms of their first order information.

First order necessary conditions and adjoint states

Let us assume that $\mathcal{F} : W \times V \rightarrow \mathbb{R}$ and $e : W \times V \rightarrow B$ introduced in (2.21) are continuously Fréchet differentiable. Furthermore, denoting for brevity the optimal state with $\bar{w} := w(\bar{v})$ let us assume that, given a solution (\bar{w}, \bar{v}) of (2.21), the partial derivative of the map e with respect to w is a linear and continuous isomorphism between W and B , i.e. satisfies:

$$e_w(\bar{w}, \bar{v}) \in \mathcal{L}(W, B) \text{ is a bijection.}$$

This condition guarantees that in a neighbourhood of (\bar{w}, \bar{v}) a unique solution $w(v)$ to the state equation exists, via the implicit function theorem. Moreover, the control-to-state operator is also continuously Fréchet differentiable in such neighbourhood.

Under these assumptions, the necessary optimality condition (2.23) for $\bar{v} \in V$ to be a local optimal solution, can equivalently be interpreted in terms of the *adjoint state* associated to the problem (2.22) defined by the following Theorem.

Theorem 2.4.2 (Adjoint state and optimality system). *Under the assumptions above, if $\bar{v} \in V$ is a local optimal solution of (2.22) with associated state $\bar{w} \in W$, then there exists an element $p \in B^*$ such that the triplet (\bar{v}, \bar{w}, p) satisfies the following system of equations:*

$$e(\bar{w}, \bar{v}) = 0, \quad (2.24a)$$

$$e_w(\bar{w}, \bar{v})^* p = \mathcal{F}_w(\bar{w}, \bar{v}), \quad (2.24b)$$

$$e_v(\bar{w}, \bar{v})^* p = \mathcal{F}_v(\bar{w}, \bar{v}). \quad (2.24c)$$

The element $p \in B^$ is called the adjoint state of the problem (2.21) and the system (2.24) above is called the optimality system for \bar{v} .*

In other words, Theorem 2.4.2 translates the necessary optimality condition (2.23) into the existence of an adjoint state solving the optimality system (2.24).

Lagrangian formalism

The derivation of the optimality system (2.24) may be very hard in practice. In particular, it may be not intuitive determining *a priori* its structure and the regularity of the adjoint state p in B^* . We briefly revise here the Lagrangian approach, which is *formally* used in practice to provide a hint on how the adjoint equations should look like.

Definition 2.4.3 (Lagrangian functional). *The Lagrangian functional associated to the reduced problem (2.22) is defined as the following functional $\mathcal{L} : W \times V \times B^* \rightarrow \mathbb{R}$:*

$$\mathcal{L}(w, v, p) := \mathcal{F}(w, v) - \langle p, e(w, v) \rangle_{B^*, B}. \quad (2.25)$$

By differentiating the Lagrangian functional with respect to the variable w and v one can recover the optimality system (2.24). In fact, by taking the derivative of the Lagrangian functional with respect to w in the direction $y \in W$ one obtains:

$$\mathcal{L}_w(w, v, p)[y] = \mathcal{F}_w(w, v)[y] - \langle p, e_w(w, v)y \rangle_{B^*, B} = \mathcal{F}_w(w, v)[y] - \langle e_w(w, v)^* p, y \rangle_{W^*, W}.$$

Similarly, taking the derivative with respect to the variable v in the direction $h \in V$, we obtain:

$$\mathcal{L}_v(w, v, p)[h] = \mathcal{F}_v(w, v)[h] - \langle p, e_v(w, v)h \rangle_{B^*, B} = \mathcal{F}_v(w, v)[h] - \langle e_v(w, v)^* p, h \rangle_{V^*, V}.$$

Therefore, the optimality system (2.24) evaluated in correspondence with an optimal solution $(\bar{w}, \bar{v}) \in W \times H$ can be written in terms of the Lagrangian functional \mathcal{L} as follows:

$$e(\bar{w}, \bar{v}) = 0, \quad (2.26a)$$

$$\mathcal{L}_w(\bar{w}, \bar{v}, p) = 0, \quad (2.26b)$$

$$\mathcal{L}_v(\bar{w}, \bar{v}, p) = 0. \quad (2.26c)$$

Second order sufficient optimality conditions

The description above shows how to find *necessary* conditions for an element $\bar{v} \in V$ to be a minimum of the reduced functional F defined in (2.22). Unfortunately, first order information is generally not enough to prove they are also sufficient. One has to look at second order optimality conditions which for the problem considered translate

basically into a positivity condition of the second Fréchet derivative. Without any further assumptions on F , this guarantees the property of \bar{v} to be a local minimum. We do not include here the corresponding theory and results, but we refer the reader to [Trö10] where a detailed analysis for several different PDE problems is considered. Nonetheless, we present in the following some insight on the numerical methods for solving (2.22), pointing out the second-order-type conditions required for their convergence to a global minimum of the problem.

2.4.2 Numerical methods for PDE-constrained problems

We give now an account of classical numerical optimisation methods, with a particular focus on the case of PDE constrained problems. The general philosophy of the methods presented in this section is an iterative procedure which, starting from an initial iterate v_0 , decreases the value of the cost functional F in (2.22) along the iterations by exploiting first and second order differential information. Depending on the method, different descent directions can be considered as well as different procedures determining how far iterates need to move along them. In what follows, we restrict to the case when V is a Hilbert space and denote by (\cdot, \cdot) the scalar product in V .

Optimise-then-discretise approach Hereafter in the thesis, we consider *infinite* dimensional optimisation methods for the solution of the problem 2.21b. In fact, we follow the *optimise-then-discretise* approach where first the numerical algorithm is designed and then the PDE is discretised. This differs from the *discretise-then-optimise* approach where, after the discretisation of both the cost functional and the PDE is carried out, the optimisation is performed via large-scale optimisation tools. In our work we prefer the former approach mainly for two different reasons: firstly, as previously discussed in Section 1.1.2, because the variational imaging models we consider are set up in infinite dimensional function spaces. Secondly, such approach guarantees the key property of *mesh independence*, that is the property of an infinite dimensional method to reflect the corresponding properties of discretised problems with sufficiently small step sizes. This property is crucial in the design of an optimisation method independent on the size of the grid used for its discretisation.

Gradient-descent methods

In the case when the reduced cost functional F in (2.22) is assumed to be continuously Fréchet differentiable, a standard descent method approach consists in exploiting first order information at each iteration $k \geq 0$ and define a *descent* direction $d_k \in V$ such that, for $\alpha > 0$

$$F(v_k + \alpha d_k) < F(v_k), \quad (2.27)$$

is satisfied for every k . By linearising the left hand side of (2.27), we approximate F as:

$$F(v_k + \alpha d_k) \approx F(v_k) + \alpha(\nabla F(v_k), d_k).$$

Ideally, we would like then to choose the descent direction d_k as:

$$d_k = \arg \min_{\|d\|_V=1} (\nabla F(v_k), d). \quad (2.28)$$

The following theorem defines the direction d_k along which F decreases most rapidly.

Theorem 2.4.4. *Let $v_k \in V$ such that $\nabla F(v_k) \neq 0$ for every $k \geq 0$. Then, the problem (2.28) has a unique solution given by*

$$\bar{d}_k = -\frac{\nabla F(v_k)}{\|\nabla F(v_k)\|_V}, \quad \text{for every } k \geq 0.$$

The Theorem 2.4.4 above, justifies the classical choice for the direction d_k to be chosen as:

$$d_k = -\nabla F(v_k), \quad k \geq 0, \quad (2.29)$$

which corresponds to the standard *gradient-descent* (or *steepest-descent*) optimisation method solving (2.22).

The following global convergence result holds:

Theorem 2.4.5 (Global convergence of gradient-descent methods). *Let F in (2.22) be Fréchet differentiable and bounded from below. If $(v_k)_{k \in \mathbb{N}}$, $(d_k)_{k \in \mathbb{N}}$ are sequence of iterates and gradient-descent directions provided by Theorem 2.4.4 and $\alpha > 0$ is small enough, then:*

$$\lim_{k \rightarrow +\infty} \nabla F(v_k) = 0,$$

and every accumulation point of $(v_k)_{k \in \mathbb{N}}$ is a stationary point of F .

In the theorem above, we intentionally do not make precise how to choose the parameter α . In fact, this can be allowed to change along the iterations, so that $\alpha = \alpha_k$, $k \geq 0$, under some conditions. We will comment more on this choice in the following, providing some extra conditions for the sequence $(\alpha_k)_{k \in \mathbb{N}}$ which ensure the validity of Theorem 2.4.5 in that case as well.

Remark 2.4.6. *Theorem 2.4.5 ensures global convergence of the algorithm, but local convergence of gradient-descent methods is typically slow, see [Kel99].*

Theorem 2.4.5 holds actually for a class of more general descent conditions which is generally referred to as *angle condition*. They read:

$$-(F(v_k), d_k) \geq \theta \|\nabla F(v_k)\|_V \|d_k\|_V, \quad \theta \in (0, 1), \quad (2.30)$$

and needs to be satisfied at each iteration $k \geq 1$. In this general framework, several choices of directions d_k can be made. A special class is given by directions d_k defined as:

$$d_k = -D_k^{-1} \nabla F(v_k), \quad (2.31)$$

where the family of matrices $(D_k)_{k \in \mathbb{N}}$ satisfies:

$$\alpha_1 \|v\|_V^2 \leq (D_k v, v) \leq \alpha_2 \|v\|_V^2, \quad \text{for every } v \in V \text{ and every } k \geq 1, \quad (2.32)$$

for some constants $0 < \alpha_1 < \alpha_2$ independent of k . It can be easily shown that inequality (2.30) is satisfied for this choice and, consequently, the global convergence result guaranteed in Theorem 2.4.5 holds.

Recalling the PDE-constrained optimisation framework (2.21a)- (2.21b) as well as Theorem 2.4.2, we have that in correspondence with an optimal solution $\bar{v} \in V$ and its correspondent state $\bar{w} \in W$ the following characterisation of the cost functional F holds:

$$(\nabla F(\bar{v}), h) = -\langle e_v(\bar{w}, \bar{v})^* p, h \rangle_{V^*, V} + \mathcal{F}_v(\bar{w}, \bar{v})[h], \quad (2.33)$$

where p is the solution of the adjoint system (2.24). Equation (2.33) is the one which is used in practice to choose the descent direction d_k as in (2.29) at each iteration of the optimisation algorithm. Once d_k has been defined, the state and adjoint equations are solved simultaneously to compute the corresponding new iterates w_{k+1} and p_{k+1} , respectively.

Newton's methods

We consider now a second-order Taylor's expansion of the functional F at the point $v_k + \alpha d$ with $\alpha > 0$. We can write

$$F(v_k + \alpha d) \approx F(v_k) + \alpha (\nabla F(v_k), d) + \frac{1}{2} (\nabla^2 F(v_k) d, d),$$

where $\nabla^2 F(v_k) d$ denotes the Riesz representative of $F''(v_k) d$. Analogously as before, a descent-direction d_k can be computed at each iteration $k \geq 1$ by finding the minimiser of the expression above with respect to d . In particular, such d_k satisfies the following first-order optimality condition:

$$\alpha \nabla F(v_k) + \nabla^2 F(v_k) d_k = 0,$$

which can be equivalently written in the more usual form:

$$\nabla^2 F(v_k) d_k = -\alpha \nabla F(v_k). \quad (2.34)$$

Equation (2.34) is classically known in literature under the name of *Newton's* iteration for the solution of the optimality condition:

$$\nabla F(v) = 0. \quad (2.35)$$

Such a class of methods is used for computing iteratively local minima of the function through the optimality condition $\nabla F(v) = 0$. In (2.34), if the second derivative is invertible, the Newton's iteration reads:

$$v_{k+1} = v_k - \alpha(\nabla^2 F(v_k))^{-1} \nabla F(v_k). \quad (2.36)$$

Moreover, if the second derivative satisfies condition (2.32) and $\alpha > 0$ is small enough, global convergence of the iterates again holds, as guaranteed by Theorem 2.4.5. Such condition, though, is quite difficult to check at each Newton iteration. Generally, some alternative conditions are preferred such as Lipschitz continuity of $\nabla^2 F$ in a neighbourhood of a local optimal point. We state the result in the framework of the constrained-optimisation (2.21a)- (2.21b).

Theorem 2.4.7 (Local convergence of Newton method). *Let $(\bar{w}, \bar{v}) \in W \times V$ a local optimal solution of (2.21a)- (2.21b) and assume that the maps \mathcal{F} and e are twice continuously Fréchet differentiable with Lipschitz continuous second derivatives. Further, assume that $e_w(w, v)$ is locally bijective in a neighbourhood of the solution pair (\bar{w}, \bar{v}) . If there is a constant $\mu > 0$ such that the Lagrangian functional \mathcal{L} defined in (2.25) satisfies:*

$$\mathcal{L}''_{w,v}(\bar{w}, \bar{v}, p)[(y, h)]^2 \geq \mu \|h\|_2^2,$$

for every pair $(y, h) \in W \times V$ such that

$$e_w(\bar{w}, \bar{v})y + e_v(\bar{w}, \bar{v})h = 0,$$

then the Newton iterates (2.36) converge locally quadratically, that is there is a constant $C > 0$ such that:

$$\|v_{k+1} - \bar{v}\|_V \leq C \|v_k - \bar{v}\|_V^2,$$

for every $k \geq 1$.

To summarise: when considering $D_k^{-1} = I$ in (2.31), gradient descent iterations are obtained. Theorem 2.4.5 ensures global convergence to the minimum, but the rate of convergence is often slow (typically linear). On the other hand, the choice $D_k^{-1} = (\nabla^2 F(v_k))^{-1}$ in (2.31) gives rise to Newton's methods. For those methods quadratic local convergence is obtained, but global convergence is not generally guaranteed without any further assumptions on the positivity of $\nabla^2 F$ at least in a neighbourhood of a solution. Furthermore,

the computation of the Hessian requires typically more computational efforts. To circumvent this problem, a common strategy in the design of second-order optimisation methods consists in the approximation of the Hessian through first-order information in order to reduce the computational costs, while ensuring rapid local convergence rate and preserving the positivity of the approximated Hessian, this leading to a globally convergent method. These methods go under the name of quasi-Newton methods.

Quasi-Newton methods

More precisely, at each Newton's iteration $k + 1$ the second derivative can be approximated by considering a linear and continuous operator B_{k+1} which can be expressed in terms of first order quantities evaluated in the previous iteration k and satisfying, for instance, the secant equation, compare Figure 2.1:

$$B_{k+1}(v_{k+1} - v_k) = \nabla F(v_{k+1}) - \nabla F(v_k).$$

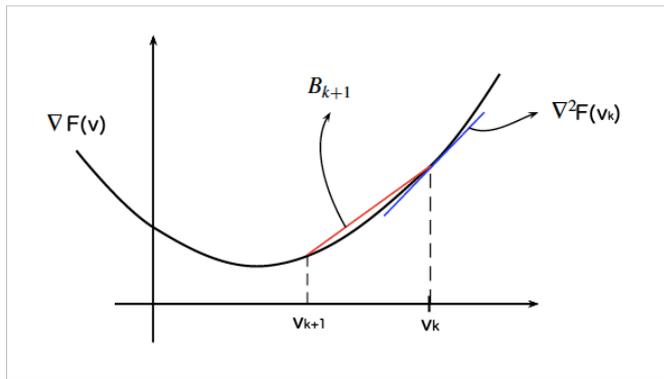


Figure 2.1: Secant equation. Image from [DLR15].

Let us now define for every $k \geq 1$ the quantities $s_k := v_{k+1} - v_k$ and $z_k := \nabla F(v_{k+1}) - \nabla F(v_k)$. In order to uniquely determine the operator B_{k+1} additional conditions are needed. One possibility consists in choosing B_{k+1} as the solution of:

$$\begin{cases} \min_B \|W(B^{-1} - B_k^{-1})W\|_F \\ \text{subject to :} \\ B = B^\top, \quad B s_k = z_k. \end{cases} \quad (2.37)$$

Here $\|\cdot\|_F$ stands for the matrix Frobenius norm and W is a positive definite matrix satisfying $W^2 s_k = z_k$. Therefore, the solution of the problem (2.37) can be expressed as:

$$B_{k+1} = B_k - \frac{B_k s_k \otimes B_k s_k}{(B_k s_k, s_k)} + \frac{z_k \otimes z_k}{(z_k, s_k)}, \quad (2.38)$$

where the for every $w_1, w_2 \in V$ the operation $w_1 \otimes w_2$ is defined as:

$$(w_1 \otimes w_2)(v) := (w_2, v)w_1, \quad v \in V.$$

The update (2.38) is known in literature under the name of Broyden-Fletcher- Goldfarb-Shanno (BFGS) method to compute an approximation of $\nabla^2 F(v_k)$. Note that such approximation exploits first-order information only. The following Theorem guarantees that this approximation can be used in practice.

Theorem 2.4.8 (Convergence of BFGS). *Let F be twice Fréchet differentiable and let $\nabla^2 F$ be Lipschitz continuous in a neighbourhood of an optimal solution \bar{v} of (2.22), with bounded inverse. Let B_0 an initial positive operator such that $B_0 - \nabla^2 F(\bar{v})$ is compact and let $v_0 \in V$ be an initial guess. Then, the BFGS iteration*

$$v_{k+1} = v_k - B_k^{-1}F(v_k)$$

converges superlinearly to \bar{v} provided that

$$\|v_0 - \bar{v}\|_V \quad \text{and} \quad \|B_0 - \nabla^2 F(\bar{v})\|_{\mathcal{L}(V)}$$

are sufficiently small.

As with Newton's methods, condition of boundedness of the inverse of $\nabla^2 F$ can be replaced by a convexity condition.

SemiSmoothNewton (SSN) methods

A different Newton-type method can be designed in order to solve (2.35) in the case when the function F is not twice Fréchet differentiable. Recalling Definition 2.2.6 of Newton differentiability, we have the following theorem.

Theorem 2.4.9 (Convergence of SSN). *Let F be Fréchet differentiable and \bar{v} be a solution to (2.22). Let $v_0 \in V$ be an initial guess. If ∇F is Newton differentiable with Newton derivative H in a neighbourhood D of \bar{v} and H has bounded inverse in D , then the SemiSmooth Newton (SSN) iteration*

$$v_{k+1} = v_k - H(v_k)^{-1}F(v_k)$$

converges superlinearly to \bar{v} provided that $\|v_0 - \bar{v}\|_V$ is sufficiently small.

SSN methods will turn out to be particularly useful in the design of second-order optimisation methods for nonsmooth max-type functions appearing in the Euler-Lagrange equation of the Huber-regularised TV functional in (2.12).

Line search

As briefly discussed before, another important choice in (2.27) is the one of the parameter $\alpha > 0$. Once having fixed the descent direction d_k in each iteration $k \geq 1$, such a choice determines how far the current iteration needs to move along d_k . As such, the parameter α can let vary along the iterations by defining:

$$\alpha_k := \arg \min_{\alpha > 0} F(v_k + \alpha d_k), \quad \text{for every } k \geq 1, \quad (2.39)$$

or, alternatively, one can choose α_k as the smallest positive root of the corresponding necessary optimality condition for (2.39), whenever (2.39) is easy to solve. However, in practice, this may be too complicated. In general, the following two properties need to be satisfied by the sequence $(\alpha_k)_{k \in \mathbb{N}}$:

$$F(v_k + \alpha_k d_k) < F(v_k), \quad \text{for every } k \geq 1, \quad (2.40a)$$

$$F(v_k + \alpha_k d_k) - F(v_k) \rightarrow 0, \quad \text{as } k \rightarrow +\infty. \quad (2.40b)$$

Assuming that the sequence $(\alpha_k)_{k \in \mathbb{N}}$ satisfies the properties (2.40a)-(2.40b) above, a global convergence result for gradient-descent methods similar to Theorem 2.4.5 holds. Similarly, provided that the line search parameter guarantees that the curvature condition

$$(z_k, s_k) > 0 \quad (2.41)$$

is satisfied in (2.38), global convergence of BFGS method is obtained.

Armijo rule One popular example of line search strategy is the *Armijo rule* with backtracking. The selection of α_k is performed in each iteration $k \geq 1$ as follows: given a descent direction d_k for F and the current iterate v_k , α_k is chosen as the largest element in $\{1, \frac{1}{2}, \frac{1}{4}, \dots\}$ such that:

$$F(v_k + \alpha_k d_k) - F(v_k) \leq \theta \alpha_k (\nabla F(v_k), d_k), \quad (2.42)$$

where $0 < \theta \ll 1$ is a constant. Defining $\phi(\alpha) := F(v_k + \alpha d_k)$, (2.42) can be rewritten as

$$\phi(\alpha) \leq \phi(0) + \theta \phi'(0).$$

Such sequence satisfies properties (2.40a)-(2.40b) and is a good choice for the design of a descent algorithm, compare Figure 2.2.

Despite improving the general local convergence behaviour of the optimisation algorithm used, line search strategies may be very costly in practice, especially for PDE-

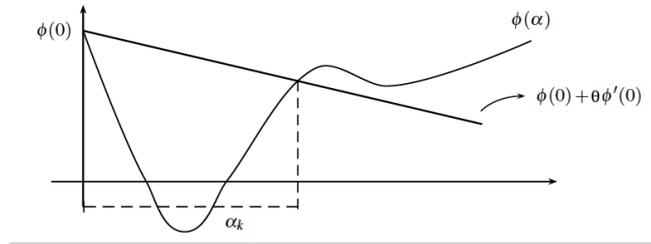


Figure 2.2: Armijo line search rule. Image from [DLR15].

constrained optimisation problems due to the evaluation of the cost functional in each outer iteration of the optimisation loop.

Remark 2.4.10 (Computational complexity). *The use of BFGS (or, in general, of other Quasi-Newton) superlinearly-convergent optimisation algorithms applied to compute the approximation of the $n \times n$ Hessian matrix for a problem with n unknowns is $O(n^2)$. In order to reduce such factor some low-complexity alternatives can be used. For instance, limited-memory BFGS methods (L-BFGS) store in each iteration of the optimisation only few vectors of the Hessian matrix to compute implicitly its approximation, whereas in \mathcal{LQN} methods the approximation of the Hessian matrix (corresponding to B_{k+1} in (2.38)) is found in a low-complexity space. Using these alternatives, the number of operations can be reduced up to $O(n \log n)$.*

Chapter 3

Numerical solution of higher-order TV flows

In this chapter we start our investigation considering two non-standard gradient flows of the TV energy which have been considered in recent works to overcome some of the TV reconstruction drawbacks discussed in Section 1.4.1. The nonlinear PDEs considered have a differential order typically of order four and, as such, encode more differential information which enhance the reconstruction of the image.

From a computational point of view, finding numerical schemes that solve such type of equations is a challenging problem. Dealing with an evolutionary nonlinear higher-order PDE, an easily applicable method is desirable. On one hand, a naive explicit discretisation in time may restrict the choice of the time-step size Δt up to an order $(\Delta x)^k$, where Δx denotes the step size of the spatial grid and k is the differential order, compare [Sme03]. Moreover, the strong nonlinearity of subgradients of the total variation adds additionally constraints to the stability condition of a discrete time stepping scheme, compare [CM99, DWB09]. In particular, when approximating the subgradient of the total variation by regularising it – either by a square root ε -regularisation as in (2.8) or by a regularisation of its dual formulation (2.13) – the size of the regularisation parameter encodes the accuracy of this approximation, that is the strength of the nonlinearity in the approximated subgradient. The presence of this nonlinearity together with the higher differential order of the equation may result then in restrictive stability conditions on numerical time stepping for small values of this regularisation parameter.

In the following we consider two prototypes of these higher-order flows and propose efficient numerical methods for the computation of their solution based either on a dimensional splitting strategy or on quasi-Newton schemes solving a relaxed primal-dual formulation of the problem.

3.1 Two non-standard TV gradient flows

We consider higher-order nonlinear PDEs of the type

$$\begin{aligned} \frac{\partial u}{\partial t} &= \nabla \cdot (h(u)\nabla q) & q \in \partial|Du|(\Omega) & \text{ in } \Omega \times (0, \infty), \\ u(t=0) &= u_0 & & \text{ in } \Omega, \end{aligned} \tag{3.1}$$

where $h : \mathbb{R} \rightarrow \mathbb{R}$ is a modelling function, u_0 is a sufficiently regular initial condition and $\partial|Du|(\Omega)$ denotes the subdifferential of the total variation functional, see Definition 2.2.7. Equations of the form (3.1) typically arise in the study of tension driven flow of thin liquid films [Mye98, ODB97] and have recently found applications in image processing, see, e.g., [OSV03, BHS09, SBBH09]. We limit ourselves to the consideration of two prototypes of (3.1). In particular, we focus on the two choices $h(u) \equiv 1$ and $h(u) = u$ and apply the two resulting models to solve image inpainting and image denoising tasks, respectively.

3.1.1 The TV gradient flow in the space H^{-1}

In the case when $h(u) \equiv 1$ in (3.1), we receive

$$\begin{aligned} \frac{\partial u}{\partial t} &= \Delta q, & q \in \partial|Du|(\Omega), & \text{ in } \Omega \times (0, \infty), \\ u(t=0) &= u_0 & & \text{ in } \Omega. \end{aligned} \tag{TV- H^{-1} }$$

Elements q of the subdifferential $\partial|Du|(\Omega)$ have the property that, if $q \in \partial|Du|(\Omega)$, then (see [Ves01, Proposition 4.1]), for every $x \in \Omega$ s.t. $|\nabla u(x)| \neq 0$

$$q = -\nabla \cdot \left(\frac{\nabla u}{|\nabla u|} \right), \tag{3.2}$$

and $\partial|Du|(\Omega)$ is multivalued otherwise. This characterisation motivates the fourth differential order and the strong nonlinearity on the gradient in equation (3.1) in this case.

Equation (TV- H^{-1}) constitutes a gradient flow of the total variation seminorm in the space H^{-1} . This is rather different from classical approaches to image processing task where the L^2 -gradient flows of the TV functional is generally considered and results in a nonlinear second-order PDE. In [Mey01, OSV03, LV08, DW05] higher-order approaches like (TV- H^{-1}) are proposed for image denoising and image decomposition. When applied to image denoising, a given noisy image u_0 is decomposed into its piecewise smooth part $u = u(T)$, solution of (TV- H^{-1}) at time T , and its oscillatory, noisy part n , i.e., $u_0 = u + n$. Similarly, in image decomposition the piecewise smooth part u represents the structure/cartoon part of the image, and the oscillatory part n its texture part. The advantage of using an H^{-1} subgradient of the total variation rather than an L^2 subgradient is that the application of the model (TV- H^{-1}) shows better separability of the data into

oscillatory and piecewise constant components and it reduces unwanted artefacts of the total variation filter such as staircasing, cf. [OSV03, LV08, LT06].

In [BHS09, SB11] equation like (TV- H^{-1}) is used for image inpainting. As discussed in Section 1.4.1, the higher-order subgradient is necessary for improving upon the ability of total variation-type inpainting to smoothly connect image structures also over large distances, cf. for instance [SC02, CS01, CKS02]. In [BHS09] the inpainted image u is reconstructed from a given image $f \in L^2(\Omega)$ which is damaged inside the inpainting domain $D \subset \Omega$ by evolving it via the flow

$$\begin{aligned} \frac{\partial u}{\partial t} &= \Delta q, & q \in \overline{\partial TV}(u), & \text{ in } D, \\ u &= f & & \text{ in } \Omega \setminus D, \end{aligned} \tag{3.3}$$

where

$$\overline{TV}(u) := \begin{cases} |Du|(\Omega) & \text{if } \|u\|_{L^\infty(\Omega)} \leq 1 \quad \text{a.e. in } \Omega, \\ +\infty & \text{otherwise.} \end{cases}$$

Here, the L^∞ -bound on solutions of (3.3) is a technical assumption needed for the existence analysis in [BHS09] and does not present a restriction when dealing with greyvalue images whose values are always bounded within a fixed interval.

3.1.2 The L^2 -Wasserstein TV gradient flow

Considering (3.1) with the choice $h(u) = u$, we get the following equation for u :

$$\begin{aligned} \frac{\partial u}{\partial t} &= \nabla \cdot (u \nabla q), & q \in \partial |Du|(\Omega), & \text{ in } \Omega \times (0, \infty), \\ u(t=0) &= u_0 \geq 0 & & \text{ in } \Omega, \end{aligned} \tag{TV-Wass}$$

where we assume the given initial condition u_0 to be nonnegative and with unitary mass, i.e. $\int_{\Omega} u_0(x) dx = 1$.

Wasserstein spaces of probability densities and abstract gradient flows

Before proceeding, we now follow [AGS08, Vil03, ACB⁺03] and briefly revise the definition of the L^2 -Wasserstein distance in the space of probability measures defined on a general metric space (Ω, d) . This will facilitate and make more clear the following discussion.

Let $\mathbb{P}(\Omega)$ be the space of probability measures on Ω and let $\mu \in \mathbb{P}(\Omega)$. Let us denote by $\mathbb{P}_2(\Omega)$ the space of all probability measures on Ω with integrable second moment.

Definition 3.1.1. *Let $\mu_1, \mu_2 \in \mathbb{P}_2(\Omega)$. The L^2 -Wasserstein distance between μ_1 and μ_2 is defined by:*

$$W_2(\mu_1, \mu_2)^2 := \min_{\Pi \in \Gamma(\mu_1, \mu_2)} \int_{\Omega \times \Omega} d(x, y)^2 d\Pi(x, y).$$

Here, $\Gamma(\mu_1, \mu_2)$ denotes the class of all transport maps $\gamma \in \mathbb{P}(\Omega \times \Omega)$ such that

$$\pi_{\#}^1 \gamma = \mu_1, \quad \pi_{\#}^2 \gamma = \mu_2,$$

where for $i = 1, 2$, $\pi^i : \Omega \times \Omega \rightarrow \Omega$ are the canonical projections on Ω and $\pi_{\#}^i \gamma \in \mathbb{P}(\Omega)$ is the push-forward of γ through π^i .

Equation (TV-Wass) can be derived as the L^2 -Wasserstein gradient flow of the TV functional $|Du|(\Omega)$ in the abstract framework of Riemannian manifolds proceeding similarly as in [Ott99]. In this abstract framework, let (\mathcal{M}, g) a Riemannian manifold and let $\mathcal{E} : \mathcal{M} \rightarrow \mathbb{R}$ an energy functional defined on \mathcal{M} . The abstract gradient flow of \mathcal{E} on (\mathcal{M}, g) can be expressed as

$$\frac{\partial u}{\partial t} = -\text{grad } \mathcal{E}|_u, \quad (3.4)$$

where $\text{grad } \mathcal{E}|_u$ is a tangent vector field to \mathcal{M} in u . We note here that the metric tensor g is a necessary ingredient to the notion of gradient flow since it converts the differential $\text{diff } \mathcal{E}$ of \mathcal{E} , a cotangent vector field, into the gradient $\text{grad } \mathcal{E}$ through the relation:

$$g(\text{grad } \mathcal{E}, s) = \text{diff } \mathcal{E}.s, \quad \text{for all vector fields } s \text{ on } \mathcal{M}.$$

The abstract gradient flow (3.4) can then be equivalently expressed as

$$g\left(\frac{\partial u}{\partial t}, s\right) + \text{diff } \mathcal{E}|_u.s = 0, \quad \text{for all vector fields } s \text{ on } \mathcal{M}. \quad (3.5)$$

Taking $s = \frac{\partial u}{\partial t}$ we have that:

$$\frac{d}{dt} \mathcal{E}(u) = \text{diff } \mathcal{E}|_u \cdot \frac{\partial u}{\partial t} = -g\left(\frac{\partial u}{\partial t}, \frac{\partial u}{\partial t}\right) \leq 0, \quad (3.6)$$

which is nothing but the energy decreasing property along $\frac{\partial u}{\partial t}$. Let us fix now:

- $\mathcal{M} = \{u : \Omega \rightarrow \mathbb{R} : u \geq 0, \int_{\Omega} u \, dx = 1\}$.
- $g(s_1, s_2) = \int_{\Omega} u \nabla p_1 \cdot \nabla p_2 \, dx$, where for $i = 1, 2$ $s_i \in \mathcal{T}_u \mathcal{M}$, the tangent space to \mathcal{M} in u , and is defined as

$$s_i = -\nabla \cdot (u \nabla p_i). \quad (3.7)$$

- $\mathcal{E}(u) = |Du|(\Omega)$, the TV functional.

We note that thanks to (3.7) we have

$$g(s_1, s_2) = \int_{\Omega} s_1 p_2 \, dx.$$

For $q \in \partial|Du|(\Omega)$ and every p related to s through (3.7) we have then that the flow

(TV-Wass) can be formally derived as:

$$\begin{aligned} \int_{\Omega} u \nabla \left(\frac{\partial u}{\partial t} \right) \cdot \nabla s \, dx + \int_{\Omega} q s \, dx &= \int_{\Omega} \frac{\partial u}{\partial t} p \, dx - \int_{\Omega} q \nabla \cdot (u \nabla p) \, dx \\ &= \int_{\Omega} \frac{\partial u}{\partial t} p \, dx - \int_{\Omega} \nabla \cdot (u \nabla q) p \, dx = 0, \end{aligned} \quad (3.8)$$

where we have used (3.5)-(3.5), integration by parts and the divergence rule.

In the following we will consider equation (TV-Wass) as a higher-order regularising procedure for u_0 being a noisy image. This approach has been proposed in [BFS12] for density estimation and smoothing. Therein, the authors propose to compute a smoothed version u of a given probability density u_0 as a minimiser of

$$\frac{1}{2} W_2(u_0 \mathcal{L}^d, u \mathcal{L}^d)^2 + \alpha |Du|(\Omega). \quad (3.9)$$

This minimisation problem can be interpreted as a discrete approximation of a solution of the gradient flow (TV-Wass) of $|Du|(\Omega)$ with respect to the L^2 -Wasserstein metric. More precisely, the minimisation of (3.9) represents one timestep of De Giorgi's minimising movement scheme (see, e.g. [AGS08, JKO98]) to the functional $|Du|(\Omega)$ with timestep α . By construction, the regularisation method (3.9) proposed in [BFS12] is nonsmooth, i.e., edge preserving, and conserves mass, i.e., is density preserving.

Equation (TV-Wass) has been further investigated in [DS12], where the authors numerically studied the scale space properties and high-contrasting effects of the equation by solving it with an *alternating direction implicit* (ADI) operator splitting approach similar to the one we want to apply to the (TV- H^{-1}) model. There, the subgradient q of the TV seminorm in (TV-Wass) is approximated by an ε -regularisation using the characterisation (3.2) above.

3.1.3 Related work

Some numerical methods have been proposed to solve equations of the type (3.1). Lieu and Vese [LV08] proposed a numerical method to solve TV- H^{-1} denoising/decomposition by using the Fourier representation of the H^{-1} norm on the whole \mathbb{R}^d , $d \geq 1$. This leads to a second-order PDE defined in Fourier space. In [ES07] and [ES09] the authors propose an algorithm using a finite element method to solve such equation, while in [AC05, Sch09b] a dual approach similar to the one described in [Cha04] is presented with interesting applications both to denoising and inpainting. In [SB11] the authors present results of convergence and stability for a particular numerical splitting method solving equation (TV- H^{-1}), called *convex splitting*, see Section 6.2.4. Therein, the equation is modelled as the gradient flow in H^{-1} of the difference of two convex energies. The result is the presence of a linear diffusion term in the numerical scheme which balances the unstable

behaviours coming from the nonlinear terms.

The use of the L^2 -Wasserstein distance function in (3.9) may appear unusual. However, in the last years, several connections between the world of imaging and the one of optimal transport have been considered in the literature. For instance Wasserstein distances have been used for image segmentation in [CEN07], in [Mem11] for image matching and in [LLSV14] for texture-cartoon decomposition. In [BFS12] the numerical solution of (3.9) has been accomplished by a combination of the Benamou-Brenier formulation [BB00] for the Wasserstein distance, an augmented Lagrangian method and a split Bregman technique [GO09b, Bru10]. This numerical procedure is in the flavour of several recently proposed numerical schemes for equations with gradient flow structure, cf., e.g., [CM09, BCC08, DMM10, BCW10] and references therein. In a recent work of Peyré [Pey] an entropic smoothing of the Wasserstein term resulting in a Kullback-Leibler approximation is considered and an efficient algorithm solving the modified problem is proposed.

We consider in the following two numerical strategies proposed equations of the type (3.1). In particular, we propose a directional operator splitting strategy for solving the (TV- H^{-1}) flow and a damped Newton's method solving the primal-dual formulation of (TV-Wass).

3.1.4 Preliminaries

In order to make the following exposition more clear, we specify here some useful definitions and notation details.

Notation

Since we are going to discuss the numerical solution of evolutionary differential equations, we need to distinguish between the exact solution u of the continuous equation (3.1) and the approximate semi-discrete and fully discrete solutions of the numerical schemes. Therefore, in the following we will denote by $u(x, y, t)$ the exact solution at point $(x, y) \in \Omega = [a, b] \times [c, d]$ and $t \geq 0$. We will write $u_n(x, y)$ to indicate the approximation in time of the function $u(x, y, t_n)$, where $(x, y) \in \Omega$, $t_n = n\Delta t$, $n \geq 0$ having fixed Δt as the time step size. Further, we will approximate Ω by a finite grid $\{a = x_1 < \dots < x_N = b\} \times \{c = y_1 < \dots < y_M = d\}$ with equidistant spatial size $h = (b - a)/N = (d - c)/M$ and then denote by $u_{i,j}(t)$ the approximation of $u(x_i, y_j, t)$ in the node (x_i, y_j) at time t and by $U(t) = (u_{i,j}(t))_{i,j}$ the semi-discretised solution vector. Finally, the fully discrete approximation of u will be denoted by $U_n = (u_{i,j}(n\Delta t))_{i,j}$. Also, when dealing with vectors, we will indicate their components using the superscripts notation: $\mathbf{Y} = (Y^1 Y^2)^\top$. For the discretised operators, we will indicate by D_* the differential operator acting in the direction $*$ and we will use the notations D_∇ , D_{div} and $D_\Delta U$ to indicate the discrete gradient, divergence and Laplace operator of the approximating solution U , respectively.

The discretised operators

In the following we define all the discrete operators that we use for approximating the spatial derivatives that appear in the design of our numerical schemes. Throughout the section we will make use of homogeneous Neumann boundary conditions. Alternatively, periodic boundary conditions can also be used. In our numerical simulations we did not notice any significant change in this latter case.

We will approximate the first derivatives $u_x(x_i, y_j, \cdot)$ in the x - directions using forward differences $(D_x^+ U)_{i,j}$ and backward differences $(D_x^- U)_{i,j}$, where:

$$(D_x^+ U)_{i,j} = \frac{1}{h}(u_{i+1,j} - u_{i,j}) \quad \text{and} \quad (D_x^- U)_{i,j} = \frac{1}{h}(u_{i,j} - u_{i-1,j}).$$

The first derivatives of u with respect to y will be approximated analogously by $(D_y^+ U)_{i,j}$ and by $(D_y^- U)_{i,j}$. The pure second derivatives will be approximated by using the five-point formula, this means the Laplace operator $\Delta u = u_{xx} + u_{yy}$ will be approximated by:

$$(D_{\Delta} U)_{i,j} = (D_{xx} U)_{i,j} + (D_{yy} U)_{i,j} = \frac{1}{h^2}(u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{i,j}). \quad (3.10)$$

The mixed derivatives $u_{xy}(x_i, y_j, \cdot)$ will be approximated by:

$$(D_{xy} U)_{i,j} = \frac{1}{4h^2}(u_{i+1,j+1} + u_{i-1,j-1} - u_{i-1,j+1} - u_{i+1,j-1}).$$

Other discretisations of the mixed derivative terms can be considered as well. For instance, in [itHW07] a more general discretisation weighting also the other points in the 3×3 is considered. Nonetheless, for the sake of simplicity we stick in the following with the discretisation given as above. An appropriate and maybe adaptive choice of such operator could be further investigated.

For the discretisation of the pure and mixed fourth-order derivatives appearing in Section 3.3, we will use a 5×5 stencil and approximate u_{xxxx} , u_{yyyy} and u_{xxyy} by:

$$\begin{aligned} (D_{xxxx} U)_{i,j} &= (D_{xx}(D_{xx} U))_{i,j} = \frac{1}{h^4}(u_{i+2,j} - 4u_{i+1,j} + 6u_{i,j} - 4u_{i-1,j} + u_{i-2,j}), \\ (D_{yyyy} U)_{i,j} &= (D_{yy}(D_{yy} U))_{i,j} = \frac{1}{h^4}(u_{i,j+2} - 4u_{i,j+1} + 6u_{i,j} - 4u_{i,j-1} + u_{i,j-2}), \\ (D_{xxyy} U)_{i,j} &= (D_{xy}(D_{xy} U))_{i,j} = \frac{1}{16h^4}(u_{i+2,j+2} + u_{i-2,j-2} + 4u_{i,j} + u_{i-2,j+2} + u_{i+2,j-2}) \\ &\quad - \frac{1}{8h^4}(u_{i,j+2} + u_{i+2,j} + u_{i-2,j} + u_{i,j-2}). \end{aligned}$$

In order to preserve the adjointness property between the discrete gradient and divergence operators, we will use forward differences for the discretisation of ∇ and backward

differences for $\nabla \cdot$, see, e.g., [Cha04]. One can check that this indeed implies

$$-D_{\text{div}}^-(\mathbf{P}) \cdot U = \mathbf{P} \cdot D_{\nabla}^+ U, \quad \forall U \in \mathbb{R}^{N \times M}, \mathbf{P} \in (\mathbb{R}^{N \times M})^2. \quad (3.11)$$

Under the choice of the boundary conditions specified above, this formula corresponds essentially to the discrete analogue of integration by parts. Whenever alternative (such as upwind-type) discretisations are used, they will be specified in the text.

Definition of stability

We will often refer in the following to the property of *unconditional stability* of time stepping methods solving equations of the type (3.1). In the following, we make precise what we mean with such a property.

Definition 3.1.2. *Let u be an element of a suitable function space \mathcal{H} defined on $\Omega \times [0, T]$, with $\Omega \subset \mathbb{R}^2$ open and bounded and $T > 0$. Let G be a real valued function and $\frac{\partial u}{\partial t} = G(u, D^\alpha u)$ a partial differential equation with all spatial D^α with order $|\alpha| \leq 4$. A corresponding time stepping method*

$$u_{n+1} = u_n + \Delta t G_n(u_n, u_{n+1}, D^\alpha u_n, D^\alpha u_{n+1}) \quad (3.12)$$

where G_n is a suitable approximation of G in u_n and u_{n+1} is **unconditionally stable** if all the solutions of (3.12) are bounded for all $\Delta t > 0$ and all n such that $n\Delta t \leq T$ and **stable** if, for a given $\Delta t > 0$, the solutions of (3.12) are bounded for all n such that $n\Delta t \leq T$.

3.2 Alternating Directional Implicit (ADI) splitting schemes

The general idea behind splitting methods is breaking down a complicated problem into smaller (and, typically, easier to approach) parts, such that these smaller problems can be solved efficiently. Different splitting methods have been presented both to solve ODEs and PDEs. The main idea is the splitting of the operator defining the problem into different components. We focus on an operator splitting method called *directional splitting*, see [HV03]. In particular, we discuss three types of directional splitting: the Peaceman-Rachford scheme [PR55], the Douglas-Hundsdorfer scheme [vdHV79, Hun02], and an additive multiplicative operator splitting (AMOS) scheme [BIK01].

In what follows we consider large systems of generic ordinary differential equations (ODEs) arising from a semi-discretisation of initial boundary value problems such as

$$\begin{cases} U'(t) = F(t, U(t)) & \text{for } t \geq 0, \\ U(0) = U_0, \end{cases}$$

for a given function $F : (0, T) \times \mathbb{R} \rightarrow \mathbb{R}$, $U : (0, T) \rightarrow \mathbb{R}$ and initial condition $U_0 \in \mathbb{R}$. The explicit dependence on time of F may not appear, thus considering *autonomous* systems.

Directional splitting methods simplify the problem above by decomposing the differential operator F into the sum:

$$F(t, U) = F_0(t, U) + F_1(t, U) + \cdots + F_s(t, U) \quad \text{for some } s \geq 1, \quad (3.13)$$

where s is the spatial dimension of the problem, i.e. $\Omega \subset \mathbb{R}^s$. In (3.13), the components F_j , $j = 1, \dots, s$, encode the linear action of F along the space direction $j = 1, \dots, s$, respectively, and F_0 may contain additional contributes coming from mixed directions and non-stiff nonlinear terms. In their general form, *alternating directional implicit* (ADI) schemes are time-stepping methods that treat the unidirectional components F_j , $j \geq 1$ implicitly and the F_0 component, if present, explicitly in time. Having in mind a standard finite difference space discretisation of the differential operators involved (see Section 3.5 for the details), this splitting idea translates numerically into considering tridiagonal, easily invertible operators acting each on one direction only, which is a desirable feature in the design of (semi-)implicit numerical schemes.

3.2.1 The Peaceman-Rachford scheme

The first method we consider in this framework is the second-order accurate *Peaceman-Rachford* ADI method (see [PR55] and [HV03]) where the operator F is simply splitted into $F = F_1 + F_2$, i.e. no mixed derivative or nonlinear terms are present. For every $n \geq 0$ the approximation U_{n+1} is computed through the following time-stepping:

$$\begin{cases} U_{n+1/2} = U_n + \frac{\Delta t}{2} F_1(t_n, U_n) + \frac{\Delta t}{2} F_2(t_{n+1/2}, U_{n+1/2}), \\ U_{n+1} = U_{n+1/2} + \frac{\Delta t}{2} F_1(t_{n+1}, U_{n+1}) + \frac{\Delta t}{2} F_2(t_{n+1/2}, U_{n+1/2}). \end{cases} \quad (3.14)$$

In (3.14) we observe that forward and backward Euler are applied alternatively in a symmetrical fashion, thus resulting in second-order accuracy. The scheme (3.14), however, does not have a natural extension for the case when $F_0 \neq 0$, so more general ADI methods have been proposed in the literature to deal with the more general case.

3.2.2 The Douglas-Hundsdorfer scheme

Another ADI numerical method that allows a more general decomposition like the one in (3.13) is the so-called *Douglas* method (see [vdHV79] and [Hum02]). In it, the numerical approximation in each time step is computed by applying at first a forward Euler predictor and then it is stabilised by intermediate s steps where just the unidirectional components F_j of the splitting (3.13) appear, weighted by a parameter $\theta \in [0, 1]$. The size of θ balances

the implicit/explicit behaviour of these steps. In other words, the unidirectional operators are applied to the convex combination $\theta U_{n+1} + (1 - \theta)U_n$, thus considering fully implicit steps for $\theta = 1$, explicit ones for $\theta = 0$ and a Crank-Nicolson type scheme for $\theta = 1/2$. The time-consistency order of the scheme is equal to two whenever $F_0 = 0$ and $\theta = 1/2$ and it is of order one otherwise. However, in many applications $F_0 \neq 0$ (for instance, when considering contributions coming from mixed derivative operators) for any given θ . In these cases, one would like to still preserve second-order time accuracy. Consequently, some extensions of the Douglas scheme have been proposed. Denoting by Y_j and $\tilde{Y}_j, j = 1, \dots, s$ the intermediate approximations of U_{n+1} computed in the unidirectional steps, the following scheme proposed in [Hun02] by Hundsdorfer is an extension of the Douglas method where a second stabilising parameter $\sigma > 0$ appears:

$$\left\{ \begin{array}{l} Y_0 = U_n + \Delta t F(t_n, U_n) \\ Y_j = Y_{j-1} + \theta \Delta t (F_j(t_{n+1}, Y_j) - F_j(t_n, U_n)), \quad j = 1, 2, \dots, s \\ \tilde{Y}_0 = Y_0 + \sigma \Delta t (F(t_{n+1}, Y_s) - F(t_n, U_n)) \\ \tilde{Y}_j = \tilde{Y}_{j-1} + \theta \Delta t (F_j(t_{n+1}, \tilde{Y}_j) - F_j(t_{n+1}, Y_s)), \quad j = 1, 2, \dots, s \\ U_{n+1} = \tilde{Y}_s. \end{array} \right. \quad (3.15)$$

In this extension, the approximation Y_s to U_{n+1} obtained from the application of the pure Douglas method is used to introduce a correction with respect to whole operator F as well. The advantage of this extension is that for any given θ the scheme (3.15) has time-consistency order equal to two if $\sigma = 1/2$ and one otherwise, independently of F_0 . The parameter θ is typically fixed to be $\theta = 1/2$. Its choice is discussed in [Hun02]. Larger values of θ give stronger damping of implicit terms, whereas lower values typically favour better accuracy. Stability properties of this scheme applied to linear convection-diffusion equations with mixed derivative terms together with technical relations between the parameter θ and σ to guarantee von Neumann stability are investigated in [itHM13, itHW07]. There, setting $\sigma = 1/2$, the preferable value for θ to have stability is $\theta = 1/2 + \sqrt{3}/6$. In [WB03] the authors combine the approach presented above with iterative methods for solving nonlinear systems.

3.2.3 Additive-Multiplicative Operator Splitting (AMOS) schemes

In both schemes (3.14) and (3.15) some explicit terms appear. These may affect stability properties of the methods and, generally, their accuracy. As observed in [BIK01], splitting schemes as directional splitting methods belong to the family of *multiplicative locally one-dimensional* (LOD) schemes. In their general semi-implicit operator form, when a splitting

similar to (3.13) holds, they can be written as:

$$\prod_{i=0}^s (I - \Delta t F_i) U_{n+1} = U_n, \quad (3.16)$$

where, similarly as before, each operator F_i , $1 \leq i \leq s$, is acting just along the i -th direction, whereas F_0 encodes mixed contributions. Dealing with such operator is typically difficult from a numerical point of view since the matrix $(I - \Delta t F_0)$ is, generally, not tridiagonal. For this reason, the scheme (3.15) deals explicitly with such a term. Analogously, explicit components appear also in (3.14) because of the alternating application of forward and backward Euler. As we will point out later, these explicit contributions may create instabilities in the methods considered. Following the strategy presented in [BIK01], our attempt is to modify (3.14) such that no explicit contributions appear. At this stage, the cost of such an operation appears to be the reduction of the accuracy of the method to order one, against the second-order achieved with the classical Peaceman-Rachford method (3.14). In order to preserve such accuracy as well as the symmetry of the method, at each time step two parallel calculations can then be performed:

$$\begin{cases} (I - \Delta t F_1) U_{n^*} = U_n \\ (I - \Delta t F_2) \tilde{U}_{n+1} = U_{n^*} \end{cases} \quad \text{and} \quad \begin{cases} (I - \Delta t F_2) U_{n^*} = U_n \\ (I - \Delta t F_1) \bar{U}_{n+1} = U_{n^*}, \end{cases} \quad (3.17a)$$

which, written in the same form as (3.14) and (3.15), read as:

$$\begin{cases} U_{n^*} = U_n + \Delta t F_1(U_{n^*}) \\ \tilde{U}_{n+1} = U_{n^*} + \Delta t F_2(\tilde{U}_{n+1}) \end{cases} \quad \text{and} \quad \begin{cases} U_{n^*} = U_n + \Delta t F_2(U_{n^*}) \\ \bar{U}_{n+1} = U_n + \Delta t F_1(\bar{U}_{n+1}). \end{cases} \quad (3.17b)$$

To get the numerical solution U_{n+1} we can simply average:

$$U_{n+1} = \frac{\tilde{U}_{n+1} + \bar{U}_{n+1}}{2}, \quad (3.17c)$$

thus ensuring a symmetric splitting. Due to the nature of such a method, we refer to (3.17) as *additive multiplicative operator splitting* (AMOS) ADI method. Note, that this scheme is identical to an earlier version of the well-known *Strang* splitting (see, for instance, [HV03, (1.12), p.329]). For nonlinear problems, such scheme is second-order accurate, in contrast to first-order accuracy of the classical *Strang* splitting scheme (compare [BIK01]). Furthermore, the scheme (3.17) has the advantage of allowing a parallel implementation, as suggested in [LNT92, LNT91, WRV98].

We focus now on the particular case of regular domains $\Omega \subset \mathbb{R}^2$, i.e. we consider in the following $s = 2$. Hence, the F_1 component (F_2 , respectively) will contain operators acting just along the x -direction (y -direction, respectively). When appearing, the term F_0

will deal with the mixed xy -direction. Our aim is to adapt the ADI schemes (3.14), (3.15) and (3.17) to regularised versions of the nonlinear model (TV- H^{-1}).

3.3 Solving (TV- H^{-1}) by ADI splitting

We would like to apply a directional splitting strategy to compute efficiently the numerical solution of the higher-order PDE (TV- H^{-1}) using ADI schemes presented in the previous Section 3.2. As a toy example, we start from considering the easier, linear higher-order biharmonic equation in the following Section 3.3.1, before applying a similar strategy to the nonlinear one in Section 3.3.2.

3.3.1 The linear case: the biharmonic equation

We start by considering the linear, fourth-order parabolic *biharmonic* equation:

$$\frac{\partial u}{\partial t} = -\Delta^2 u \quad \text{in } \Omega \times (0, T). \quad (3.18)$$

Such equation appears in many applied mathematical models such as the Cahn-Hilliard model describing phase transitions and phase separation in binary mixtures (see, for instance, [EF87] and [NCS84]). Some work on ADI schemes applied to the biharmonic equation already exists in literature. It dates back to [CD58] where the authors consider the equation to model vibrational modes for thin plates. The numerical analysis of the resulting ADI schemes has been studied in [WB03, itHM13] where linear stability results are proved as well.

We are looking for the solution U of the following semi-discretised version of (3.18):

$$U'(t) = F(U) := -D_{\Delta^2} U = -D_{xxxx} U - D_{yyyy} U - 2D_{xxyy} U. \quad (3.19)$$

Following (3.13), we decompose the function F into the sum

$$F(U) = F_0(U) + F_1(U) + F_2(U),$$

where the components F_i , $i = 0, 1, 2$, are defined by

$$F_0(u) := -2D_{xxyy} U, \quad F_1(u) := -D_{xxxx} U, \quad F_2(u) := -D_{yyyy} U, \quad (3.20)$$

and the differential operators are discretised as described in Section 3.1.4. Using decomposition (3.20) we can find for every n the approximating solution U_{n+1} of (3.18) using the *Hundsdoerfer* ADI scheme (3.15). Namely, for every Δt and real parameters $\sigma, \theta > 0$,

the scheme reads:

$$\left\{ \begin{array}{l} Y_0 = U_n + \Delta t F(U_n) \\ Y_1 = Y_0 + \theta \Delta t (F_1(Y_1) - F_1(U_n)) \\ Y_2 = Y_1 + \theta \Delta t (F_2(Y_2) - F_2(U_n)) \\ \tilde{Y}_0 = Y_0 + \sigma \Delta t (F(Y_2) - F(U_n)) \\ \tilde{Y}_1 = \tilde{Y}_0 + \theta \Delta t (F_1(\tilde{Y}_1) - F_1(Y_2)) \\ \tilde{Y}_2 = \tilde{Y}_1 + \theta \Delta t (F_2(\tilde{Y}_2) - F_2(Y_2)) \\ U_{n+1} = \tilde{Y}_2. \end{array} \right. \quad (3.21)$$

Now we simplify the problem (3.19) by splitting the fourth-order equation into a mathematically equivalent autonomous system of two partial differential equations of order two. We obtain:

$$\left\{ \begin{array}{l} U'(t) = D_{\Delta} V = D_{xx} V + D_{yy} V = F(U, V), \\ V = -D_{\Delta} U = -D_{xx} U - D_{yy} U = G(U, V). \end{array} \right. \quad (3.22)$$

Then, the *Hundsdofer* scheme (3.21) applied to (3.19) can be equivalently written as a coupled ADI scheme for approximate solutions (U_n, V_n) of (3.22). For positive parameters θ, σ this gives:

$$\left\{ \begin{array}{l} \left(\begin{array}{c} Y_0^2 \\ Y_0^1 \end{array} \right) = \left(\begin{array}{c} G(U_n, V_n) \\ U_n + \Delta t F(U_n, Y_0^2) \end{array} \right), \\ \left(\begin{array}{c} Y_1^1 \\ Y_1^2 \end{array} \right) = \left(\begin{array}{c} Y_0^1 \\ 0 \end{array} \right) + \left(\begin{array}{c} \theta \Delta t F_1(Y_1^1, Y_1^2) \\ G_1(Y_1^1, Y_1^2) - G_1(U_n, V_n) \end{array} \right), \\ \left(\begin{array}{c} Y_2^1 \\ Y_2^2 \end{array} \right) = \left(\begin{array}{c} Y_1^1 \\ 0 \end{array} \right) + \left(\begin{array}{c} \theta \Delta t F_2(Y_2^1, Y_2^2) \\ G_2(Y_2^1, Y_2^2) - G_2(U_n, V_n) \end{array} \right), \\ \left(\begin{array}{c} \tilde{Y}_0^2 \\ \tilde{Y}_0^1 \end{array} \right) = \left(\begin{array}{c} G(Y_2^1, Y_2^2) \\ Y_0^1 + \sigma \Delta t (F(Y_2^1, \tilde{Y}_0^2) - F(U_n, V_n)) \end{array} \right), \\ \left(\begin{array}{c} \tilde{Y}_1^1 \\ \tilde{Y}_1^2 \end{array} \right) = \left(\begin{array}{c} \tilde{Y}_0^1 \\ 0 \end{array} \right) + \left(\begin{array}{c} \theta \Delta t F_1(\tilde{Y}_1^1, \tilde{Y}_1^2) \\ G_1(\tilde{Y}_1^1, \tilde{Y}_1^2) - G_1(Y_2^1, Y_2^2) \end{array} \right), \\ \left(\begin{array}{c} U_{n+1} \\ V_{n+1} \end{array} \right) = \left(\begin{array}{c} \tilde{Y}_1^1 \\ 0 \end{array} \right) + \left(\begin{array}{c} \theta \Delta t F_2(\tilde{Y}_2^1, \tilde{Y}_2^2) \\ G_2(\tilde{Y}_2^1, \tilde{Y}_2^2) - G_2(Y_2^1, Y_2^2) \end{array} \right) \end{array} \right. \quad (3.23)$$

where the functions F, F_1, F_2 and G, G_1, G_2 are given by:

$$\begin{aligned} \begin{pmatrix} F_1(U, V) \\ G_1(U, V) \end{pmatrix} &= \begin{pmatrix} 0 & D_{xx} \\ -D_{xx} & 0 \end{pmatrix} \cdot \begin{pmatrix} U \\ V \end{pmatrix}, \\ \begin{pmatrix} F_2(U, V) \\ F_2(U, V) \end{pmatrix} &= \begin{pmatrix} 0 & D_{yy} \\ -D_{yy} & 0 \end{pmatrix} \cdot \begin{pmatrix} U \\ V \end{pmatrix}, \end{aligned} \quad (3.24)$$

$$F(U, V) = F_1(U, V) + F_2(U, V), \quad G(U, V) = G_1(U, V) + G_2(U, V).$$

Schemes (3.21) and (3.23) are equivalent as one can easily check by substitution. As detailed in the previous Section 3.2.1, ADI schemes are typically applied for evolution equations (as the one for u in (3.22)). Consequently, the scheme for the stationary equation for v is adjusted in order to preserve the equivalence with the corresponding scheme (3.21) solving (3.19). Moreover, in both the explicit steps of the scheme above the order of application of the method is inverted for consistency issues. Namely, in (3.23) we first find consistent approximations for V_{n+1} using them to get consistent approximations of U_{n+1} .

3.3.2 The nonlinear case: the TV- H^{-1} model

We now want to derive in a similar way an ADI method solving the TV- H^{-1} equation (TV- H^{-1}). Expanding the differential operators appearing in the equation generates an intractable number of nonlinear terms of various differential orders. This makes a direct application of the ADI scheme to (TV- H^{-1}) impracticable. Therefore, following the ideas presented in Section 3.3.1, we reduce the original fourth-order equation to an autonomous system of two second-order equations. We first regularise the subgradient of the total variation in (TV- H^{-1}) using the characterisation (3.2). We use the ε -regularisation (see Section 2.3.1) and replace $|\nabla u|$ by $|\nabla u|_\varepsilon := \sqrt{|\nabla u|^2 + \varepsilon}$, $0 < \varepsilon \ll 1$, compare (2.8). This results in the following regularised version of (TV- H^{-1})

$$\frac{\partial u}{\partial t} = -\Delta \nabla \cdot \left(\frac{\nabla u}{|\nabla u|_\varepsilon} \right), \quad \text{or, equivalently,} \quad \begin{cases} \frac{\partial u}{\partial t} = \Delta v, \\ v = -\nabla \cdot \left(\frac{\nabla u}{|\nabla u|_\varepsilon} \right). \end{cases} \quad (3.25)$$

In the following we present two different linearisations of the problem above. Such a choice is important from two different points of view, intrinsically related to each other. The former is the accuracy of the scheme we are considering: rough linearisations (i.e. linearisations which consider most of the nonlinear terms explicitly evaluating them in one or more given approximations of the solution in previous time steps) are likely to present poor accuracy as well as stability issues. This is a general consideration in the

numerical solution of every PDE and has to be taken into account and balanced with the choice of linearisation which might be more accurate and precise, but which could present, on the other hand, difficulties in its implementation and application. The latter point of view is peculiar to our choice of performing a directional splitting scheme. As pointed out above, our purpose is splitting our partial differential operator into the sum of components which are considered both explicitly (as F_0 above) and implicitly (as F_1 and F_2), as in (3.13). The choice of the linearisation affects such a splitting as the F_0 component and the linearised quantities multiplying the differential operators acting in x and y may change accordingly. For instance, the F_0 component might not appear changing the choice of the ADI scheme we want to use.

In the following we proceed by presenting two ADI schemes of the form (3.14) and (3.15) associated to two different linearisations of (3.25). For a given initial condition (U_0, V_0) , our problem consists in finding an approximation (U_{n+1}, V_{n+1}) of the solution $(u(t_{n+1}), v(t_{n+1}))$ to (3.25) for every $n \geq 0$. In both cases, we choose the solution at the previous time step (U_n, V_n) as linearisation point.

The first linearisation

Indicating by \tilde{U} the value of the solution U_n in the previous time step, our first choice of linearisation is the following (compare with [DS12]):

$$\begin{cases} U'(t) = D_\Delta V, \\ V = -\frac{\varepsilon + (D_y^+ \tilde{U})^2}{|D_\nabla^+ \tilde{U}|_\varepsilon^3} D_{xx} U - \frac{\varepsilon + (D_x^+ \tilde{U})^2}{|D_\nabla^+ \tilde{U}|_\varepsilon^3} D_{yy} U + 2 \frac{D_x^+ \tilde{U} D_y^+ \tilde{U}}{|D_\nabla^+ \tilde{U}|_\varepsilon^3} D_{xy} U, \end{cases} \quad (3.26)$$

where (U, V) is the semi-discrete approximation to a solution of (3.25). Here, the linearisation of V constitutes a semi-implicit approximation of the second-order nonlinear diffusion, evaluating all the first-order derivatives of the expansion in the previous time step. As before, we use the following notation for the system (3.26):

$$\begin{pmatrix} U'(t) \\ V \end{pmatrix} = \begin{pmatrix} F(U, V) \\ G(U, V) \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \cdot \begin{pmatrix} U \\ V \end{pmatrix} \quad (3.27)$$

for suitable matrices A, B, C and D in $\mathbb{R}^{NM \times NM}$. We split F and G into the sum of three different terms: F_0 and G_0 containing the mixed derivative term and F_1, G_1 and F_2, G_2 containing the derivatives with respect to x and y only, respectively. This produces the splitting:

$$F(U, V) = F_0(U, V) + F_1(U, V) + F_2(U, V),$$

$$G(U, V) = G_0(U, V) + G_1(U, V) + G_2(U, V), \quad (3.28)$$

with:

$$\begin{aligned} \begin{pmatrix} F_0(U, V) \\ G_0(U, V) \end{pmatrix} &= \begin{pmatrix} A_0 & B_0 \\ C_0 & D_0 \end{pmatrix} \cdot \begin{pmatrix} U \\ V \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 2 \frac{D_x^+ \tilde{U} D_y^+ \tilde{U}}{|D_{\nabla^+}^+ \tilde{U}|_2^3} D_{xy} & 0 \end{pmatrix} \cdot \begin{pmatrix} U \\ V \end{pmatrix}, \\ \begin{pmatrix} F_1(U, V) \\ G_1(U, V) \end{pmatrix} &= \begin{pmatrix} A_1 & B_1 \\ C_1 & D_1 \end{pmatrix} \cdot \begin{pmatrix} U \\ V \end{pmatrix} = \begin{pmatrix} 0 & D_{xx} \\ -\frac{\varepsilon + (D_y^+ \tilde{U})^2}{|D_{\nabla^+}^+ \tilde{U}|_2^3} D_{xx} & 0 \end{pmatrix} \cdot \begin{pmatrix} U \\ V \end{pmatrix}, \end{aligned} \quad (3.29)$$

$$\begin{pmatrix} F_2(U, V) \\ G_2(U, V) \end{pmatrix} = \begin{pmatrix} A_2 & B_2 \\ C_2 & D_2 \end{pmatrix} \cdot \begin{pmatrix} U \\ V \end{pmatrix} = \begin{pmatrix} 0 & D_{yy} \\ -\frac{\varepsilon + (D_x^+ \tilde{U})^2}{|D_{\nabla^+}^+ \tilde{U}|_2^3} D_{yy} & 0 \end{pmatrix} \cdot \begin{pmatrix} U \\ V \end{pmatrix}.$$

Due to the presence of a mixed derivative operator, a simple *Peaceman-Rachford* scheme can not be applied. As discussed before, in these situations the *Hundsdoerfer* scheme (3.15) appears to be more appropriate. Therefore, similarly as for the biharmonic equation in the previous section, we consider the following ADI scheme:

$$\left\{ \begin{aligned} \begin{pmatrix} Y_0^2 \\ Y_0^1 \end{pmatrix} &= \begin{pmatrix} G(U_n, V_n) \\ U_n + \Delta t F(U_n, Y_0^2) \end{pmatrix}, \\ \begin{pmatrix} Y_1^1 \\ Y_1^2 \end{pmatrix} &= \begin{pmatrix} Y_0^1 \\ V_n \end{pmatrix} + \begin{pmatrix} \theta \Delta t (F_1(Y_1^1, Y_1^2) - F_1(U_n, V_n)) \\ G_1(Y_1^1, Y_1^2) - G_1(U_n, V_n) \end{pmatrix}, \\ \begin{pmatrix} Y_2^1 \\ Y_2^2 \end{pmatrix} &= \begin{pmatrix} Y_1^1 \\ V_n \end{pmatrix} + \begin{pmatrix} \theta \Delta t (F_2(Y_2^1, Y_2^2) - F_2(U_n, V_n)) \\ G_2(Y_2^1, Y_2^2) - G_2(U_n, V_n) \end{pmatrix}, \\ \begin{pmatrix} \tilde{Y}_0^2 \\ \tilde{Y}_0^1 \end{pmatrix} &= \begin{pmatrix} G(Y_2^1, Y_2^2) \\ Y_0^1 + \sigma \Delta t (F(Y_2^1, \tilde{Y}_0^2) - F(U_n, V_n)) \end{pmatrix}, \\ \begin{pmatrix} \tilde{Y}_1^1 \\ \tilde{Y}_1^2 \end{pmatrix} &= \begin{pmatrix} \tilde{Y}_0^1 \\ Y_2^2 \end{pmatrix} + \begin{pmatrix} \theta \Delta t (F_1(\tilde{Y}_1^1, \tilde{Y}_1^2) - F_1(Y_2^1, Y_2^2)) \\ G_1(\tilde{Y}_1^1, \tilde{Y}_1^2) - G_1(Y_2^1, Y_2^2) \end{pmatrix}, \\ \begin{pmatrix} U^{n+1} \\ V^{n+1} \end{pmatrix} &= \begin{pmatrix} \tilde{Y}_1^1 \\ Y_2^2 \end{pmatrix} + \begin{pmatrix} \theta \Delta t (F_2(\tilde{Y}_2^1, \tilde{Y}_2^2) - F_2(Y_2^1, Y_2^2)) \\ G_2(\tilde{Y}_2^1, \tilde{Y}_2^2) - G_2(Y_2^1, Y_2^2) \end{pmatrix}. \end{aligned} \right. \quad (3.30)$$

The second linearisation

Another possibility is to linearise the system (3.25) in the following way:

$$\left\{ \begin{array}{l} U'(t) = D_{\Delta} V, \\ V = -D_{\text{div}}^{-} \left(\frac{D_{\nabla}^{+} U}{|D_{\nabla}^{+} \tilde{U}|_{\varepsilon}} \right) = -\frac{1}{|D_{\nabla}^{+} \tilde{U}|_{\varepsilon}} D_{xx} U + \frac{D_x^{+} \tilde{U} D_{xx} \tilde{U} + D_y^{+} \tilde{U} D_x^{-} D_y^{+} \tilde{U}}{|D_{\nabla}^{+} \tilde{U}|_{\varepsilon}^3} D_x^{up} U \\ \quad - \frac{1}{|D_{\nabla}^{+} \tilde{U}|_{\varepsilon}} D_{yy} U + \frac{D_x^{+} \tilde{U} D_x^{+} D_y^{-} \tilde{U} + D_y^{+} \tilde{U} D_{yy} \tilde{U}}{|D_{\nabla}^{+} \tilde{U}|_{\varepsilon}^3} D_y^{up} U, \end{array} \right. \quad (3.31)$$

where again $\tilde{U} = U_n$, the spatial quantities are discretised as above and the discrete operators D_x^{up} and D_y^{up} are defined below. We observe that with this choice no mixed derivative operator acting on U appears. Mixed terms are encoded and considered in the previous time step. On the other hand, we get first derivative operators and not just second-order ones as in (3.26). Writing again the system (3.31) as in (3.27), we now split F and G in the following way:

$$\begin{aligned} F(U, V) &= F_1(U, V) + F_2(U, V), \\ G(U, V) &= G_1(U, V) + G_2(U, V), \end{aligned} \quad (3.32)$$

where

$$\begin{aligned} \begin{pmatrix} F_1(U, V) \\ G_1(U, V) \end{pmatrix} &= \begin{pmatrix} 0 & D_{xx} \\ \frac{1}{|D_{\nabla}^{+} \tilde{U}|_{\varepsilon}} D_{xx} \frac{D_x^{+} \tilde{U} D_{xx} \tilde{U} + D_y^{+} \tilde{U} D_x^{-} D_y^{+} \tilde{U}}{|D_{\nabla}^{+} \tilde{U}|_{\varepsilon}^3} D_x^{up} & 0 \end{pmatrix} \cdot \begin{pmatrix} U \\ V \end{pmatrix}, \\ \begin{pmatrix} F_2(U, V) \\ G_2(U, V) \end{pmatrix} &= \begin{pmatrix} 0 & D_{yy} \\ -\frac{1}{|D_{\nabla}^{+} \tilde{U}|_{\varepsilon}} D_{yy} + \frac{D_x^{+} \tilde{U} D_x^{+} D_y^{-} \tilde{U} + D_y^{+} \tilde{U} D_{yy} \tilde{U}}{|D_{\nabla}^{+} \tilde{U}|_{\varepsilon}^3} D_y^{up} & 0 \end{pmatrix} \cdot \begin{pmatrix} U \\ V \end{pmatrix}. \end{aligned} \quad (3.33)$$

We note that the splitting (3.32) is a two-components splitting. We then design a *Peaceman-Rachford* ADI scheme (3.14) for the system (3.31) as follows:

$$\left\{ \begin{array}{l} \begin{pmatrix} U_{n+1/2} \\ V_{n+1/2} \end{pmatrix} = \begin{pmatrix} U_n + \frac{\Delta t}{2} F_1(U_n, V_n) + \frac{\Delta t}{2} F_2(U_{n+1/2}, V_{n+1/2}) \\ G_1(U_n, V_n) + G_2(U_{n+1/2}, V_{n+1/2}) \end{pmatrix}, \\ \begin{pmatrix} U_{n+1} \\ V_{n+1} \end{pmatrix} = \begin{pmatrix} U_{n+1/2} + \frac{\Delta t}{2} F_1(U_{n+1}, V_{n+1}) + \frac{\Delta t}{2} F_2(U_{n+1/2}, V_{n+1/2}) \\ G_1(U_{n+1}, V_{n+1}) + G_2(U_{n+1/2}, V_{n+1/2}) \end{pmatrix}. \end{array} \right. \quad (3.34)$$

For the discretisation of the first derivative operators in the scheme present in the equation for V in (3.31) we use the standard numerical technique of upwinding, i.e. the sign of the coefficients in front of the first derivatives terms affects in which direction the finite differences are computed. More precisely, we use:

$$\begin{aligned} C_1(\tilde{U})D_x^{up} &= \chi_{\{\text{sign}(C_1>0)\}} D_x^- U + \chi_{\{\text{sign}(C_1<0)\}} D_x^+ U, \\ C_2(\tilde{U})D_y^{up} &= \chi_{\{\text{sign}(C_2>0)\}} D_y^- U + \chi_{\{\text{sign}(C_2<0)\}} D_y^+ U \end{aligned} \quad (3.35)$$

where

$$C_1(\tilde{U}) = \frac{D_x^+ \tilde{U} D_{xx} \tilde{U} + D_y^+ \tilde{U} D_x^- D_y^+ \tilde{U}}{|D_{\nabla}^+ \tilde{U}|_{\varepsilon}^3}, \quad C_2(\tilde{U}) = \frac{D_x^+ \tilde{U} D_x^+ D_y^- \tilde{U} + D_y^+ \tilde{U} D_{yy} \tilde{U}}{|D_{\nabla}^+ \tilde{U}|_{\varepsilon}^3}$$

and χ_S is the characteristic function of the set S .

A numerical discussion pointing out the differences between the ADI schemes (3.30) and (3.34) resulting from the two linearisations (3.26) and (3.31), respectively will follow in Section 3.5.1.

Discussion of stability restrictions for the Hundsdorfer scheme

As we are going to illustrate numerically in Section 3.5.1, a stable application of the ADI schemes to equation (TV- H^{-1}) depends on the choice of the regularising parameter ε . In order to use reasonably large time steps Δt , this parameter has to be taken sufficiently large to get stable results for the numerical solution of (TV- H^{-1}). For the following stability consideration we use the terminology introduced in Definition 3.1.2, where we consider the solution continuous in space and discrete in time.

Fully explicit numerical schemes solving TV gradient flows turn out to show restrictive stability conditions related to the strength of the nonlinearity in the TV subgradient, cf. [CM99, DW05]. On the other hand, schemes solving (3.31) without any operator splitting are unconditionally stable. In particular, we have the following stability theorem:

Theorem 3.3.1. *Let u_0 be a sufficiently regular initial condition and u_n the solution of*

$$u_{n+1} = u_n - \Delta t \Delta \nabla \cdot \left(\frac{\nabla u_{n+1}}{|\nabla u_n|_{\varepsilon}} \right). \quad (3.36)$$

Then, the following stability estimate holds

$$\|\nabla u_{n+1}\|_{\varepsilon} \leq \|\nabla u_0\|_{\varepsilon}, \quad (3.37)$$

where $\|w\|_{\varepsilon} = (\int_{\Omega} (w^2 + \varepsilon))^{1/2}$.

Proof. Multiplying equation (3.36) by $\Delta^{-1}(u_{n+1} - u_n)$ (where Δ^{-1} is the inverse of the

negative Laplacian with zero Neumann boundary conditions) and integrating over Ω we get:

$$\langle u_{n+1} - u_n, \Delta^{-1}(u_{n+1} - u_n) \rangle = \Delta t \langle \nabla \cdot \left(\frac{\nabla u_{n+1}}{|\nabla u_n|_\varepsilon} \right), u_{n+1} - u_n \rangle$$

where $\langle \cdot, \cdot \rangle$ denotes the L^2 inner product. We can rewrite the left hand side of the equation above using the properties of Δ^{-1} and applying the divergence theorem, thus finding:

$$\langle \nabla \Delta^{-1}(u_{n+1} - u_n), \nabla \Delta^{-1}(u_{n+1} - u_n) \rangle + \Delta t \langle \left(\frac{\nabla u_{n+1}}{|\nabla u_n|_\varepsilon} \right), \nabla(u_{n+1} - u_n) \rangle = 0.$$

We can now apply the result provided in [ES07] and summing over all $t_n = n\Delta t$ up to $T = N\Delta t$, finding the following stability estimate:

$$\|\nabla u_{n+1}\|_\varepsilon \leq \Delta t \sum_n \|\partial_t u_{n+1}\| = \frac{u_{n+1} - u_n}{\Delta t} \|_{-1}^2 + \|\nabla u_{N+1}\|_\varepsilon \leq \|\nabla u_0\|_\varepsilon,$$

which gives (3.37). In particular, estimate (3.37) does not depend on the size of ε . \square

Remark 3.3.2. *An alternative proof of the theorem above can be given directly by writing the semi-implicit approximation of the minimising movement scheme in a variational form as follows:*

$$\left\langle \frac{u_{n+1} - u_n}{\Delta t}, v \right\rangle = - \left\langle \Delta \nabla \cdot \left(\frac{\nabla u_{n+1}}{|\nabla u_n|_\varepsilon} \right), v \right\rangle$$

for every $v \in L^2(\Omega)$. In the equation above, the term appearing on the right hand side is then interpreted as the Frechét derivative of the total variation energy in the space $H^{-1}(\Omega)$ and the stability condition (3.37) can be derived similarly as above.

These considerations about explicit and implicit schemes solving without any splitting the problem (TV- H^{-1}) serve as a motivation for the following estimates. We focus on the *Hundsdorfer* scheme (3.23) applied to the TV- H^{-1} equation with the choice (3.28), (3.29). In each iteration the numerical solution is computed from equations consisting of a combination of explicit and implicit quantities. In particular, the explicit quantities might affect the stability properties of the scheme. To motivate this, we focus in the following just on the first three stages of the scheme (3.23) applied to the TV- H^{-1} equation with the choice (3.28), (3.29) and $\theta = 1/2$. Looking at the first three stages of (3.23) only can be justified by the fact that the subsequent three stages of the scheme have a similar structure and are not expected to change the stability properties drastically. Combining the three steps of the scheme (3.23) with the choice (3.28), (3.29) we find the following expression:

$$\begin{aligned} & \frac{u_{n+1} - u_n}{\Delta t} + \frac{1}{2} \partial_{xx}(C_1(u_n) \partial_{xx} u_{n+1}) + \frac{1}{2} \partial_{yy}(C_2(u_n) \partial_{yy} u_{n+1}) \\ & + \frac{\Delta t}{4} \partial_{xx}(C_1(u_n) \partial_{xx} (\partial_{yy}(C_2(u_n) \partial_{yy} u_{n+1}))) \end{aligned} \quad (3.38)$$

$$\begin{aligned}
 &= -\Delta \nabla \cdot \left(\frac{\nabla u_n}{|\nabla u_n|_\varepsilon} \right) + \frac{1}{2} \partial_{xx} (C_1(u_n) \partial_{xx} u_n) + \frac{1}{2} \partial_{yy} (C_2(u_n) \partial_{yy} u_n) \\
 &\quad + \frac{\Delta t}{4} \partial_{xx} (C_1(u_n) \partial_{xx} (\partial_{yy} (C_2(u_n) \partial_{yy} u_n)))
 \end{aligned}$$

where the positive quantities $C_1(u_n)$ and $C_2(u_n)$ come from the linearisation (3.26) and read:

$$C_1(u_n) = \frac{\varepsilon + (\partial_y u_n)^2}{|\nabla u_n|_\varepsilon^3}, \quad C_2(u_n) = \frac{\varepsilon + (\partial_x u_n)^2}{|\nabla u_n|_\varepsilon^3}.$$

We observe that a mixed, eighth-order operator appears, both on the left and on the right hand side of (3.38). In the following stability discussion we neglect these high-order terms which only represent second-order in time contributions. In fact, this simplification may worsen the following considerations on stability which serves us to motivate the unstable behaviour of the *Hundsdorfer* scheme.

We multiply equation (3.38) by u_{n+1} and integrate over the domain Ω . By applying integration by parts twice with respect to the x and the y variables to the second and the third terms of the left hand side of the equation, respectively, we get:

$$\begin{aligned}
 &\frac{1}{2} \left(\int_{\Omega} \partial_{xx} (C_1(u_n) \partial_{xx} u_{n+1}) u_{n+1} \, dx + \int_{\Omega} \partial_{yy} (C_2(u_n) \partial_{yy} u_{n+1}) u_{n+1} \, dx \right) \\
 &= \frac{1}{2} \left(\int_{\Omega} C_1(u_n) (\partial_{xx} u_{n+1})^2 \, dx + \int_{\Omega} C_2(u_n) (\partial_{yy} u_{n+1})^2 \, dx \right) \quad (3.39) \\
 &\geq \frac{1}{2} K(\varepsilon) (\|\partial_{xx} u_{n+1}\|^2 + \|\partial_{yy} u_{n+1}\|^2).
 \end{aligned}$$

Here $K(\varepsilon)$ is a suitable constant that depends on the regularising parameter ε only. A similar strategy is applied to the analogous terms on the right hand side, where we use also Young's inequality with weights δ_1 and δ_2 . We obtain:

$$\begin{aligned}
 &\frac{1}{2} \left(\int_{\Omega} C_1(u_n) \partial_{xx} u_n \partial_{xx} u_{n+1} \, dx + \int_{\Omega} C_2(u_n) \partial_{yy} u_n \partial_{yy} (u_{n+1}) \, dx \right) \\
 &\leq \frac{1}{4\delta_1} \|C_1(u_n) \partial_{xx} u_n\|^2 + \frac{\delta_1}{4} \|\partial_{xx} u_{n+1}\|^2 + \frac{1}{4\delta_2} \|C_2(u_n) \partial_{yy} u_n\|^2 + \frac{\delta_2}{4} \|\partial_{yy} u_{n+1}\|^2.
 \end{aligned}$$

We can use again Young's inequality to deal with the first term on the right hand side of (3.38). We look now at the nonlinear term. By applying the divergence theorem, integration by parts and Young's inequality with weight δ_3 , we get the following estimate for this term:

$$\begin{aligned}
 &-\int_{\Omega} \Delta \nabla \cdot \left(\frac{\nabla u_n}{|\nabla u_n|_\varepsilon} \right) u_{n+1} \, dx = \int_{\Omega} \nabla \nabla \cdot \left(\frac{\nabla u_n}{|\nabla u_n|_\varepsilon} \right) \nabla u_{n+1} \, dx \quad (3.40) \\
 &= -\int_{\Omega} \nabla \cdot \left(\frac{\nabla u_n}{|\nabla u_n|_\varepsilon} \right) \Delta u_{n+1} \, dx \leq \frac{1}{2\delta_3} \left\| \nabla \cdot \left(\frac{\nabla u_n}{|\nabla u_n|_\varepsilon} \right) \right\|^2 + \frac{\delta_3}{2} \|\Delta u_{n+1}\|^2.
 \end{aligned}$$

The second term on the right hand side of the inequality above can be merged with the

corresponding ones in (3.39), choosing δ_1 small enough. Denoting by $C_3(u_n) = \frac{\partial_x u_n \partial_y u_n}{|\nabla u_n|_\varepsilon^3}$, for the curvature term in (3.40) we observe that:

$$\begin{aligned} \left\| \nabla \cdot \left(\frac{\nabla u_n}{|\nabla u_n|_\varepsilon} \right) \right\|^2 &= \int_{\Omega} (C_1(u_n) \partial_{xx} u_n + C_2(u_n) \partial_{yy} u_n + C_3(u_n) \partial_{xy} u_n)^2 dx \\ &\leq 2 \left(2 \int_{\Omega} (C_1(u_n) \partial_{xx} u_n)^2 dx + \int_{\Omega} (C_2(u_n) \partial_{yy} u_n)^2 dx \right) + \int_{\Omega} (C_3(u_n) \partial_{xy} u_n)^2 dx \\ &\leq 4 \left(\frac{1}{\sqrt{\varepsilon}} + \frac{1}{|\partial_y u_n|} \right)^2 \|\partial_{xx} u_n\|^2 + 4 \left(\frac{1}{\sqrt{\varepsilon}} + \frac{1}{|\partial_x u_n|} \right)^2 \|\partial_{yy} u_n\|^2 + \frac{1}{\sqrt{\varepsilon}} \|\partial_{xy} u_n\|^2 \end{aligned}$$

where we used Cauchy's inequality and upper bounds on C_1, C_2 and C_3 . Defining $K_1(\varepsilon)$, $K_2(\varepsilon)$ and $K_3(\varepsilon)$ as

$$K_1(\varepsilon) := 4 \left(\frac{1}{\sqrt{\varepsilon}} + \frac{1}{|\partial_y u_n|} \right)^2, \quad K_2(\varepsilon) := 4 \left(\frac{1}{\sqrt{\varepsilon}} + \frac{1}{|\partial_x u_n|} \right)^2, \quad K_3(\varepsilon) := \frac{1}{\sqrt{\varepsilon}}$$

and using the estimate proved in [CD90] for the mixed derivative term, we get the following bound:

$$\left\| \nabla \cdot \left(\frac{\nabla u_n}{|\nabla u_n|_\varepsilon} \right) \right\|^2 \leq K_1(\varepsilon) \|\partial_{xx} u_n\|^2 + K_2(\varepsilon) \|\partial_{yy} u_n\|^2 + K_3(\varepsilon) \sup_{z \in \{x, y\}} \|\partial_{zz} u_n\|^2.$$

Collecting the previous estimates, choosing δ_1, δ_2 and δ_3 small enough and getting once more upper bounds on C_1 and C_2 we get the following stability estimate

$$\begin{aligned} &\frac{1}{2\Delta t} \|u_{n+1}\|^2 + \|\partial_{xx} u_{n+1}\|^2 + \|\partial_{yy} u_{n+1}\|^2 \\ &\leq \frac{1}{2\Delta t} \|u_n\|^2 + \tilde{K}_1(\varepsilon) \|\partial_{xx} u_n\|^2 + \tilde{K}_2(\varepsilon) \|\partial_{yy} u_n\|^2 + \tilde{K}_3(\varepsilon) \sup_{z \in \{x, y\}} \|\partial_{zz} u_n\|^2 \end{aligned}$$

for scaled constants \tilde{K}_1, \tilde{K}_2 and \tilde{K}_3 which tend to infinity as the regularising parameter $\varepsilon \searrow 0$. For this limit the estimate then blows up, thus indicating possible unstable behaviour when choosing ε small. This will be discussed in more detail in Section 3.5.1.

A stable AMOS ADI scheme

In order to counteract the dependence of the stability properties of the ADI schemes (3.23) and (3.34) on the size of ε , we now consider as an alternative the AMOS operator splitting scheme (3.17) for solving (3.25). Due to the fully implicit character of the scheme, we expect stability properties to improve.

More precisely, both ADI methods (3.23) and (3.34) can be represented in a vectorial, multiplicative form similar to (3.16). For (3.23), the operator F_0 appearing in (3.16) is taken explicitly in time. Thus, when writing the correspondent numerical scheme in a multiplicative form, we have additional explicit terms on the right hand side. Writing

(3.34) in the form (3.16) we again obtain additional explicit terms appearing due to the forward Euler steps in (3.34). This, together with the stability estimates from the previous section, indicates possible stability restrictions for these schemes that will be confirmed by the numerical results in the following Section 3.5.1.

In the following we present an AMOS scheme for solving a slightly modified version of (TV- H^{-1}). In system form, this new equation reads:

$$\begin{cases} \frac{\partial u}{\partial t} = \partial_{xx}v^1 + \partial_{yy}v^2 \\ v^1 = -\partial_x \left(\frac{\partial_x u}{|\nabla u|_\varepsilon} \right), \quad v^2 = -\partial_y \left(\frac{\partial_y u}{|\nabla u|_\varepsilon} \right). \end{cases} \quad (3.41)$$

This equation is ‘more anisotropic’ than the original (TV- H^{-1}) in the sense that the nonlinear diffusion in x and y directions are considered separately, so that the mixed term $-\partial_y \left(\frac{\partial_x u}{|\nabla u|_\varepsilon} \right)$ is not encoded in v_1 and similarly for v_2 . Only the diffusion weighting involves the whole image gradient taking both x and y variations into account. In this way, linearising v_1 and v_2 by considering the diffusion weighting $1/|\nabla \tilde{u}|_\varepsilon$ for a given \tilde{u} , results in an equation with only pure x and y derivatives. Considering such an equation reduces the explicit components appearing in the scheme, allowing a fully implicit treatment of the operators, though still exploiting the advantage of directional splitting by solving along the two directions x and y separately. Applying the AMOS scheme (3.17) to the linearisation of (3.41), we obtain:

$$\begin{cases} \begin{pmatrix} U_* \\ V_*^2 \end{pmatrix} = \begin{pmatrix} U_n + \Delta t F_2(U_*, V_*^2) \\ G_2(U_*, V_*^2) \end{pmatrix}, \\ \begin{pmatrix} \tilde{U}_{n+1} \\ \tilde{V}_{n+1}^1 \end{pmatrix} = \begin{pmatrix} U_* + \Delta t F_1(\tilde{U}_{n+1}, \tilde{V}_{n+1}^1) \\ G_1(\tilde{U}_{n+1}, \tilde{V}_{n+1}^1) \end{pmatrix} \\ \begin{pmatrix} U_* \\ V_*^1 \end{pmatrix} = \begin{pmatrix} U_n + \Delta t F_1(U_*, V_*^1) \\ G_1(U_*, V_*^1) \end{pmatrix}, \\ \begin{pmatrix} \bar{U}_{n+1} \\ \bar{V}_{n+1}^2 \end{pmatrix} = \begin{pmatrix} U_* + \Delta t F_2(\bar{U}_{n+1}, \bar{V}_{n+1}^2) \\ G_2(\bar{U}_{n+1}, \bar{V}_{n+1}^2) \end{pmatrix} \end{cases} \quad (3.42)$$

$$U_{n+1} = \frac{\tilde{U}_{n+1} + \bar{U}_{n+1}}{2}, \quad V_{n+1}^1 = \frac{\tilde{V}_{n+1}^1 + V_*^1}{2}, \quad V_{n+1}^2 = \frac{\bar{V}_{n+1}^2 + V_*^2}{2},$$

where the operators are defined exactly as in (3.33) and upwinding is used for the first derivatives as in (3.35). We recall that the alternating application of the scheme first in the $y - x$ direction and subsequently in the $x - y$ direction allows to achieve order two of accuracy, as detailed in [BIK01]. As we will see in Section 3.5.1, the scheme (3.42) has better stability properties: the time step does not seem to depend on the size of ε , thus

suggesting – at least empirically – unconditional stability.

3.4 Primal-dual formulation of (TV-Wass)

We consider now an alternative numerical approach for the numerical solution of equations of the type (3.1). We do this focusing on the TV-Wasserstein flow (TV-Wass) and characterising the elements of the subdifferential of TV in an alternative way. Namely, instead of considering a characterisation of the type (3.2) for these elements, we use the approach proposed in [Ben11, Mül08] for the classical second-order TV-denoising model and deal with a relaxed primal-dual formulation applied to (TV-Wass) (see Section 2.3.3).

3.4.1 Formulation of the problem

Recalling Definition 2.2.7, we start by writing explicitly what the subdifferential inclusion $q \in \partial|Du|(\Omega)$ exactly means:

$$q \in \partial|Du|(\Omega) \iff |Du|(\Omega) - \int_{\Omega} qu \, dx \leq |Dv|(\Omega) - \int_{\Omega} qv \, dx, \quad \forall v \in L^2(\Omega). \quad (3.43)$$

Equivalently, if $u \in \text{BV}(\Omega) \subset L^2(\Omega)$ achieves the minimum of the following variational problem

$$\min_{u \in \text{BV}(\Omega)} \left\{ |Du|(\Omega) - \int_{\Omega} qu \, dx \right\}, \quad (3.44)$$

then, by minimality, (3.43) is fulfilled and then $q \in \partial|Du|(\Omega)$. Inserting the definition of the total variation seminorm in (3.44) we receive

$$\min_{u \in \text{BV}(\Omega)} \left\{ \sup_{\mathbf{p} \in C_0^\infty(\Omega; \mathbb{R}^2), \|\mathbf{p}\|_\infty \leq 1} \int_{\Omega} u \nabla \cdot \mathbf{p} \, dx - \int_{\Omega} qu \, dx \right\}, \quad (3.45)$$

which is the *primal-dual* formulation of the problem (TV-Wass). The constraint on \mathbf{p} appearing in (3.45) can be relaxed, for instance, by a penalty method. To this end we remove the constraint from the minimisation in (3.45) and add a term to the functional that penalises it if $\|\mathbf{p}\|_\infty > 1$. As in (2.19), a typical example for such a penalty term F is

$$F(s) = \frac{1}{2} \|\max\{s, 0\}\|_{L^2(\Omega)}^2.$$

With these considerations we reformulate (3.45) into the following minimisation problem

$$\min_{u \in \text{BV}(\Omega)} \sup_{\mathbf{p} \in C_0^\infty(\Omega; \mathbb{R}^2)} \left\{ \int_{\Omega} u \nabla \cdot \mathbf{p} \, dx - \frac{1}{\eta} F(|\mathbf{p}| - 1) - \int_{\Omega} qu \, dx \right\}, \quad (3.46)$$

where the parameter $0 < \eta \ll 1$ measures the weight of the penalisation and is decreased throughout the iterations of the numerical scheme we are going to present to ensure conver-

gence to the original, unconstrained problem. We can then find the optimality conditions for both \mathbf{p} and u in (3.46) which, merged with the original equation (TV-Wass), allow us to consider the following, alternative formulation of the TV-Wasserstein model:

$$\begin{cases} \frac{\partial u}{\partial t} = \nabla \cdot (u \nabla q), \\ q = \nabla \cdot \mathbf{p}, \\ 0 = -\nabla u - \frac{1}{\eta} H(\mathbf{p}), \end{cases} \quad (3.47)$$

where the system above H denotes the derivative of the penalty term $F(|\mathbf{p}| - 1)$, i.e.

$$H(\mathbf{p}) = \chi_{\{|\mathbf{p}| \geq 1\}} \operatorname{sgn}(\mathbf{p})(|\mathbf{p}| - 1).$$

3.4.2 A damped Newton system

In the system (3.47) above, we linearise $H(\mathbf{p})$ via its first-order Taylor approximation

$$H(\mathbf{p}) \approx H(\tilde{\mathbf{p}}) + H'(\tilde{\mathbf{p}})(\mathbf{p} - \tilde{\mathbf{p}}),$$

where by H' we indicate the Jacobian of H . In order to guarantee the invertibility of the now linear operator that defines the system, we add an additional damping term in \mathbf{p} , as suggested, for instance, in [Mül08]. Collecting everything, we propose the following numerical scheme for solving (3.47),

$$\begin{cases} \frac{U_{n+1}^{(k)} - U_n}{\Delta t} = D_{\operatorname{div}}^-(U_n D_{\nabla}^+ Q_{n+1}^{(k)}) \\ Q_{n+1}^{(k)} = D_{\operatorname{div}}^- \mathbf{P}_{n+1}^{(k)}, \\ 0 = -D_{\nabla}^+(U_{n+1}^{(k)}) - \frac{1}{\eta} H(\mathbf{P}_{n+1}^{(k-1)}) \\ \quad - \frac{1}{\varepsilon} H'(\mathbf{P}_{n+1}^{(k-1)})(\mathbf{P}_{n+1}^{(k)} - \mathbf{P}_{n+1}^{(k-1)}) - \tau^k (\mathbf{P}_{n+1}^{(k)} - \mathbf{P}_{n+1}^{(k-1)}). \end{cases} \quad (3.48)$$

We observe that the system above corresponds to consider (3.47) where temporal and spatial derivatives have been discretised as usual and the optimality condition involving ∇u has been linearised. We apply Newton's method to solve (3.48). The scheme consists of two nested iterations. The subscripts n are related to the outer time step evolution of the process evolving U . At each time step an implicit approximation of the quantities U_{n+1} , Q_{n+1} and \mathbf{P}_{n+1} is obtained by the application of an inner damped Newton process that runs depending on the superscript k . The sequence of parameters τ^k controls the damping of the Newton iterations: it starts from a large value τ^0 and then decreases, thus ensuring faster convergence. System (3.48) could now be discretised in space as described in Section 3.5.2. For computational simplicity we consider a slightly different penalty term

F :

$$F(\mathbf{p}) = F(p^1, p^2) = \frac{1}{2} \|\max\{|p^1| - 1, 0\}\|_2^2 + \frac{1}{2} \|\max\{|p^2| - 1, 0\}\|_2^2,$$

that results into an anisotropic TV term. With this choice we have:

$$H(p^1, p^2) = \begin{pmatrix} \operatorname{sgn}(p^1)(|p^1| - 1)\chi_{\{|p^1| \geq 1\}} \\ \operatorname{sgn}(p^2)(|p^2| - 1)\chi_{\{|p^2| \geq 1\}} \end{pmatrix}, \quad H'(p^1, p^2) = \begin{pmatrix} \chi_{\{|p^1| \geq 1\}} & 0 \\ 0 & \chi_{\{|p^2| \geq 1\}} \end{pmatrix}.$$

3.5 Numerical results

In this section we report the numerical results for the solution of (TV- H^{-1}) and (TV-Wass) solved by means of ADI splitting and primal-dual quasi-Newton schemes as described in Section 3.3 and Section 3.4.2, respectively. Our discussion starts from illustrative examples and then focuses on some applications of these models to inpainting and denoising problems.

3.5.1 Numerical solution of (TV- H^{-1}) via ADI splitting

Consistently with the description of Sections 3.3.1 and 3.3.2 we start presenting the numerical results obtained applying the *Hundsdoerfer* scheme (3.15) to compute the numerical solution of the biharmonic equation (3.19) and the equivalent system (3.22). Next, we report on numerical experiments for the TV- H^{-1} equation (TV- H^{-1}) using the *Hundsdoerfer* scheme (3.23) with the choice (3.28)-(3.29). The choice of the time step size Δt of this scheme is constrained by very strong stability restrictions related to the size of the regularising parameter ε , as discussed previously. Our numerical tests for the application of the *Peaceman-Rachford* scheme (3.34) to the TV- H^{-1} equation show similar stability behaviour and are therefore not included in the following description. Finally, we show the numerical results for the solution of the modified TV- H^{-1} system (3.41) solved with the AMOS scheme (3.42). This scheme shows stable behaviour independent of the size of Δt and ε .

The biharmonic equation: numerical results

We consider a grid of 100×100 grid points for the discretisation of the spatial domain Ω being the unit square. We analyse the example of the evolution of the biharmonic equation having as initial condition U_0 the Gaussian density $U_{ij}^0 = \exp(-((x_i - 1/2)^2 + (y_j + 1/2)^2)/\gamma^2)$ where the variance γ^2 is equal to 100. The linear system that arises from the application of the *Hundsdoerfer* scheme (3.15) with (3.24) to the biharmonic equation (3.22) is solved by the Schur complement method. Figure 3.1 shows two iterates of the scheme with $\theta = \sigma = 1/2$ for a time-step size $\Delta t = C(\Delta x)^2$, where C now and for the rest of the numerical discussion in this section is equal to 0.1. Even for such a

big choice of Δt the result is stable and the convergence of the solution to the steady state is quick. Considering another initial condition U_0 and taking, for instance, some very oscillatory function does not effect the performance of the method. This confirms the unconditional stability of this scheme when applied to a linear fourth-order equation such as the biharmonic equation (3.19), compare [itHM13].

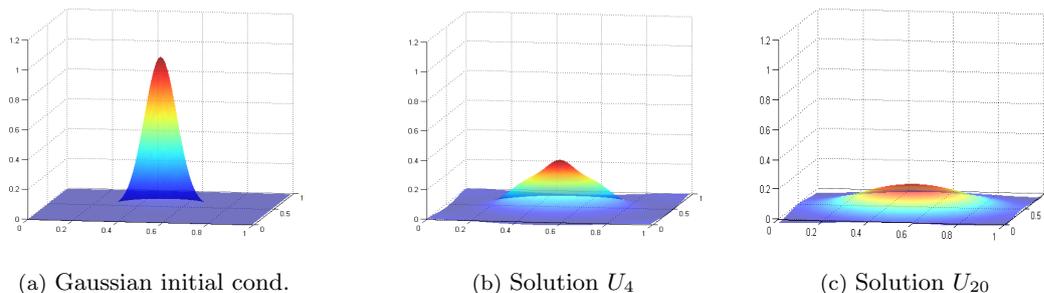


Figure 3.1: Evolution of the biharmonic equation (3.22) solved with the *Hundsdorfer* scheme (3.23) with $\Delta t = C(\Delta x)^2$.

Numerical stability behaviour for the $(TV-H^{-1})$ model

We discuss now the numerical solution of the regularised $TV-H^{-1}$ equation (3.25) by means of an ADI strategy. In particular, we consider *Hundsdorfer* ADI scheme (3.23) with (3.28), (3.29) and the AMOS scheme (3.42). Once again we use the Schur complement technique to solve the linear systems that arise in the numerical solution of these schemes. The expected edge-preserving behaviour of the $TV-H^{-1}$ equation ($TV-H^{-1}$) due to the subgradient of the TV functional is closely related to the size of the regularising parameter ε used in (3.25). This parameter “measures” how close the nonlinear diffusion is to the linear, biharmonic one. Namely, large values of ε result in a smoothing behaviour of (3.25) similar to the one of the biharmonic equation (3.19). In this case the stability properties of the *Hundsdorfer* scheme applied to the nonlinear equation are close to the ones discussed for the biharmonic equation in the previous section. On the other hand, small values of ε keep the regularised version of the subgradient of the total variation close to its exact characterisation and hence solution show edge-preserving features. However, stability issues may occur. The explicit treatment of some of the terms in the *Hundsdorfer* and the *Peaceman-Rachford* schemes (namely, the mixed derivative term in (3.23) and the half-direction forward Euler steps in (3.34)) seems to influence the stability of the schemes in a negative way. In particular, the time step sizes Δt have to be decreased significantly with small values of ε . Moreover, the choice of the initial condition also influences the stability properties. For instance, for smooth Gaussian initial conditions with large support the time steps can be chosen larger than for an oscillatory initial condition, see Figures 3.2-3.3. These issues are resolved by the application of the AMOS scheme (3.42) which did not

show dependence of Δt on the size of ε nor on the type of initial condition in order to get stable results.

We report the numerical results obtained considering the linearisation of the system (3.25) given by (3.26) and solved by the *Hundsdorfer* ADI method (3.23). We consider as initial conditions the Gaussian density $U_{ij}^0 = \exp(-((x_i - 1/2)^2 + (y_j + 1/2)^2)/\gamma^2)$ with $\gamma^2 = 100$ and the oscillating function $U_{ij}^0 = \sin(8\pi x_i) + \cos(8\pi y_j)$. Figure 3.2 shows iterates of the *Hundsdorfer* scheme with $\theta = \sigma = 1/2, \varepsilon = 5$ and $\Delta t = C(\Delta x)^3$ applied to the Gaussian datum. As preliminary experiments showed, larger values of ε and in particular the value $\varepsilon = 5$ is the smallest possible value that can be used in order to get stable solutions of the *Hundsdorfer* scheme with $\Delta t = C(\Delta x)^3$. Figure 3.3 shows the evolution of the process for the oscillatory datum with the same choice of ε as before. In this case to get stable results smaller time steps are needed, highlighting stability dependence on the initial condition.

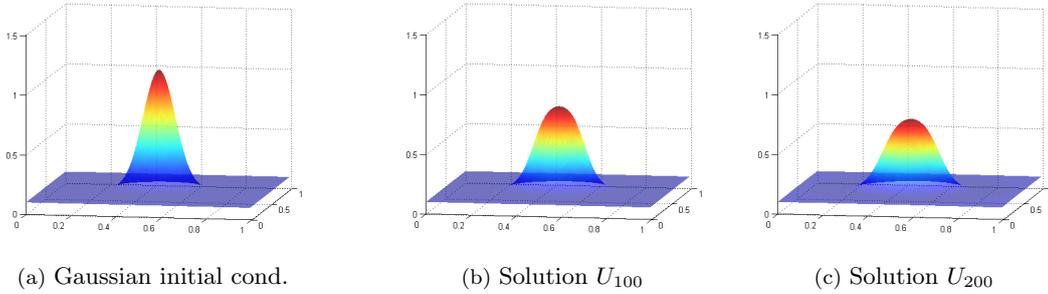


Figure 3.2: Evolution of the TV- H^{-1} equation (TV- H^{-1}) by the *Hundsdorfer* scheme (3.23) with $\Delta t = C(\Delta x)^3$ and $\varepsilon = 5$.

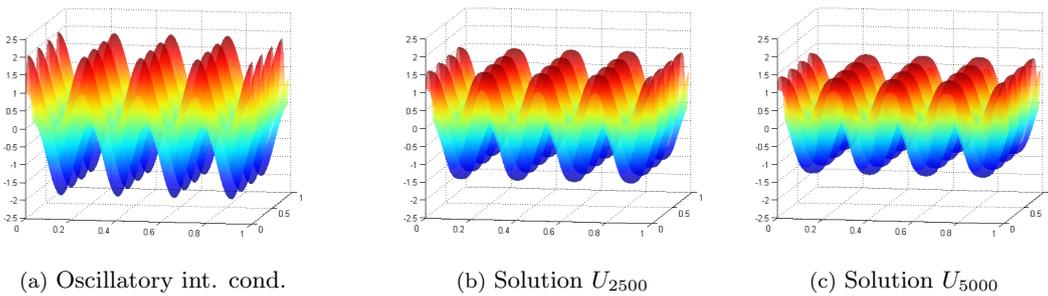


Figure 3.3: Evolution of the TV- H^{-1} equation (TV- H^{-1}) by the *Hundsdorfer* scheme (3.23) with decreased $\Delta t = C(\Delta x)^4$ and $\varepsilon = 5$ to control stability.

In Figures 3.4–3.5 we describe this stability dependence in more detail. For the two different choices of the initial conditions considered above, the smallest values ε needed for stability as the size of Δt increases are plotted. To compare the different setups of the *Hundsdorfer* scheme with respect to θ , these tests were performed for different values of the stabilising parameter $\theta = 0, 1/2, 1/2 + \sqrt{3}/6, 1$ thus resulting into four graphs per

test. Note that different values of θ (compare [HV03, Hun02, itHM13, itHW07] for such choices) result into different weighting of implicit and explicit terms. In other words, this means considering a fully explicit method for $\theta = 0$ and active implicit contributions in the unidirectional steps of (3.23) in the other cases (see Section 3.2.2). For each of these graphs, their epigraph corresponds to the region of stability of the method (according to Definition 3.1.2). To obtain such graphs we have considered different values of the constant C and of the regularising parameter ε for time step sizes of the order $(\Delta x)^k$, $k = 2, 3, 4$. For the choice $\theta = 0$ (explicit methods) stable solutions are only obtained by very restrictive choices of ε . This situation improves for $\theta > 0$. In particular, for values of θ close to 1 the stability constraint on the size of ε is reduced. However, in all cases, these plots show a clear dependence of the stability of the *Hundsdorfer* scheme on the strength of the nonlinearity (encoded by the size of ε). This creates numerical difficulties in the attempt of increasing the time step size to get an efficient numerical scheme for solving the approximated problem (3.25) for sufficiently small values of ε .

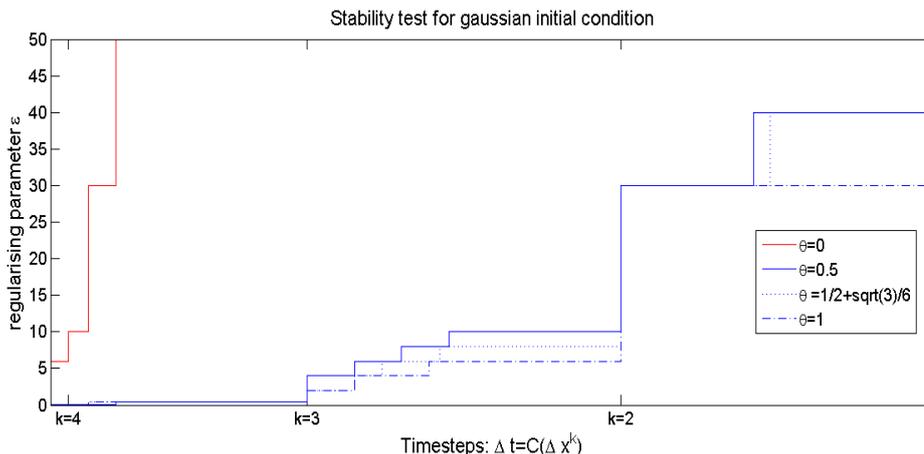


Figure 3.4: Stability test for the numerical solution of (3.26) solved with the *Hundsdorfer* scheme (3.23) with initial condition $U_{ij}^0 = \exp -(((x_i - 1/2)^2 + (y_j + 1/2)^2)/\gamma^2)$ with $\gamma^2 = 100$ for different choices of stabilising parameter θ . For each time step size the minimum value of ε providing stability is plotted.

We do not report here the numerics related to the application of the *Peaceman-Rachford* method (3.34) to the TV- H^{-1} equation (TV- H^{-1}) as the stability issues resemble the ones described above. In order to overcome such problems, we rather present in the following the results related to the application of the ADI AMOS scheme (3.42) solving the slightly modified system (3.41).

A rigorous and formal analysis of the stability of the methods considered in relation to the regularising parameter ε is still missing and represents an interesting topic of future research.

Due to the implicit character of the scheme (3.42), improved stability properties are expected while keeping the advantages of the directional splitting strategy by which each

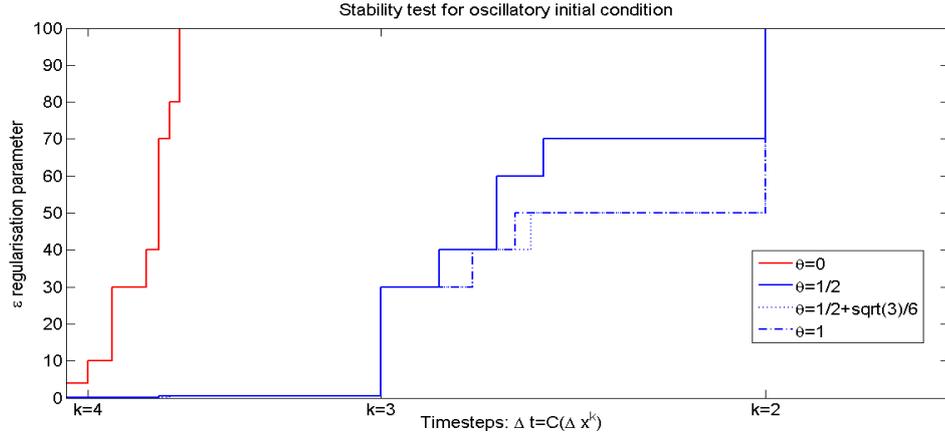


Figure 3.5: Stability test for the numerical solution of (3.26) solved with the *Hundsdorfer* scheme (3.23) with initial condition $U_{ij}^0 = \sin(8\pi x_i) + \cos(8\pi y_j)$. Comparison with Figure 3.4 shows dependence on the initial condition for admissible values of ε providing stability.

substep can be solved very efficiently. In the Figures 3.7-3.8 we show the evolution of the TV- H^{-1} equation (3.41) for the Gaussian initial condition as above for $\Delta t = C(\Delta x)^3$ and $\Delta t = C(\Delta x)^2$ with fixed $\varepsilon = 0.001$. We observe that the time discretisation provides stable results even for large Δt . However, as clearly visible in Figure 3.8, choosing Δt too large badly affects time accuracy.

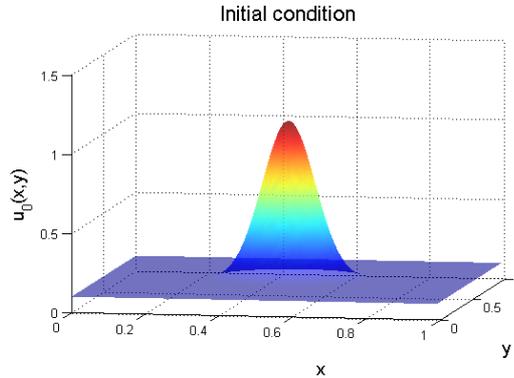


Figure 3.6: Initial condition

The convergence rate of the AMOS scheme (3.42) for the choice of $\Delta t = C(\Delta x)^3$ and regularising parameter $\varepsilon = 0.001$ when applied to the Gaussian initial condition is presented in Figure 3.9. We observe an exponential-type decay for the ℓ^∞ norm of the difference between the iterative solution U_n of (3.42) and the steady state U_∞ which has been computed numerically beforehand by iterating the scheme using as stopping criterion the inequality $\|U_{n+1} - U_n\|_\infty / (MN) \leq 10^{-10}$. For comparisons, the exponential function $(7 \cdot 10^{-6})e^{-0.05x}$, has been plotted. Figure 3.10 shows the decay of the discrete total

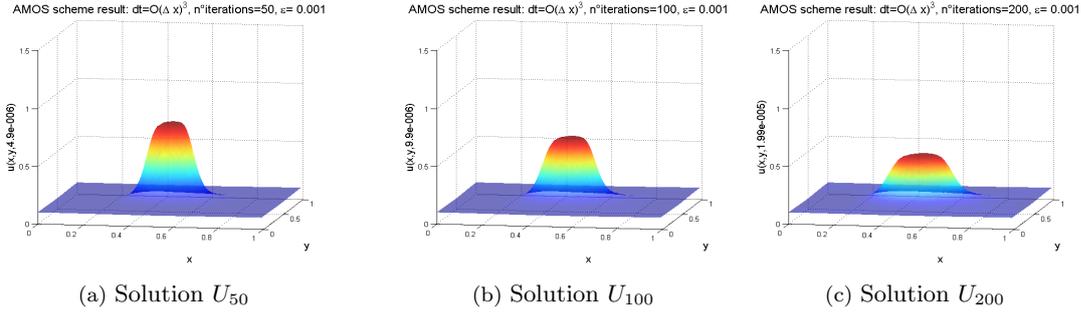


Figure 3.7: Evolution of the modified TV- H^{-1} equation (3.41) with the AMOS scheme (3.42) with $\Delta t = C(\Delta x)^3$ and $\epsilon = 0.001$.

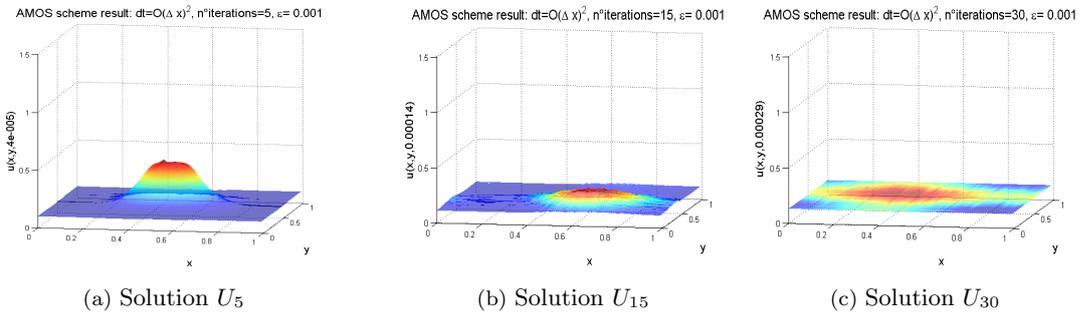


Figure 3.8: Evolution of the modified TV- H^{-1} equation (3.41) with the AMOS scheme (3.42) with $\Delta t = C(\Delta x)^2$ and $\epsilon = 0.001$.

variation energy towards the energy of the steady state U_∞ against the number of time steps. This graph shows that although the modified TV- H^{-1} equation (3.41) is not as the same as the gradient flow (TV- H^{-1}) of the total variation in the space H^{-1} , the total variation energy is still decreasing in every time step.

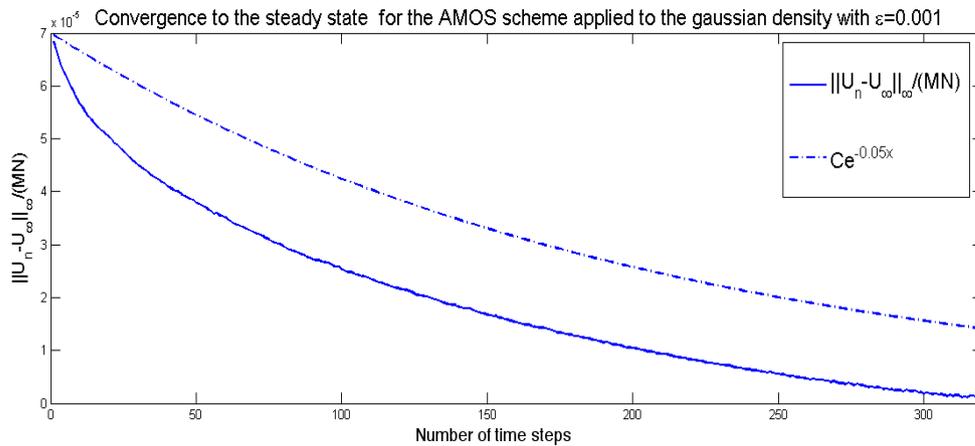


Figure 3.9: Convergence to the steady state for the numerical solution of the modified TV- H^{-1} equation (3.41) computed with the AMOS scheme (3.42).

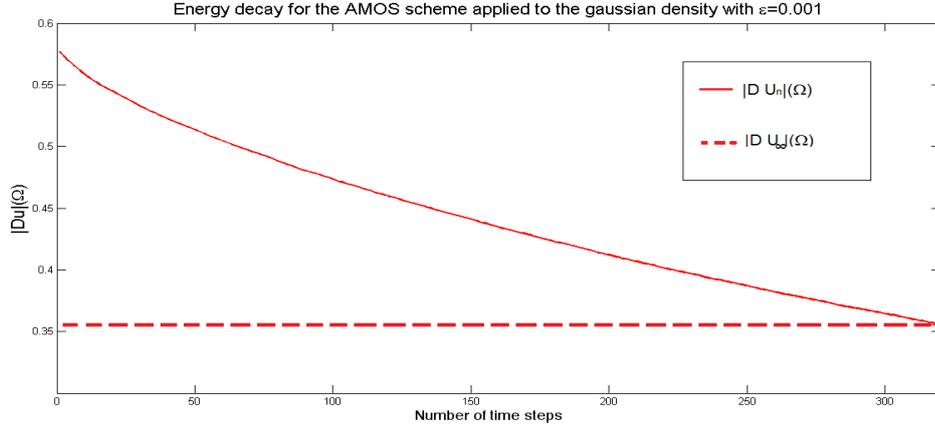


Figure 3.10: Total variation energy decay for the numerical solution of the modified TV- H^{-1} equation (3.41) computed with the AMOS scheme (3.42).

Applications to image inpainting

Motivated by our original purposes of applying the ADI schemes in the imaging framework, we present in Figure 3.11 the scale space properties of system (3.41) solved with the AMOS scheme (3.42) for a 120×120 image. The time step size is $\Delta t = C(\Delta x)^3$ and the regularising parameter $\varepsilon = 0.001$. Due the nonlinear nature of the equation, the diffusion is anisotropic.

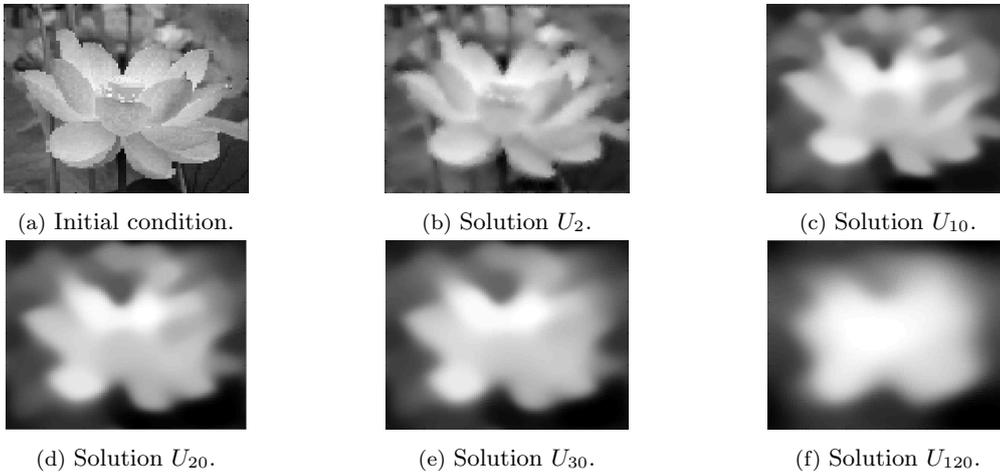


Figure 3.11: Scale space properties for the iterates of the numerical solution of (3.41) computed with the AMOS scheme (3.42), $\Delta t = C(\Delta x)^3$, $\varepsilon = 0.001$.

Finally, we present some numerical results obtained by using the AMOS scheme (3.42) to solve the modified TV- H^{-1} model given by (3.41). The constraint $u = f$ outside of the inpainting domain is approximately enforced by adding the fidelity term $\lambda \cdot \chi_{\Omega \setminus D}(f - u)$ to the equation, where λ measures how close the reconstructed image is to the original one. In Figure 3.12 we show the result for inpainting a 150×150 image of a cross: the initial

condition is inpainted in 1000 iterations using a time step size of $\Delta t = C(\Delta x)^3$ and the regularising parameter ε is chosen to be $\varepsilon = 0.001$, thus allowing the preservation of edges additionally to fulfilling the connectivity principle that is a consequence of the fourth-order of the method. For simplicity, we consider in the following images where the region to inpaint has straight edges, but more general inpainting domains can be considered as well (such as scratches).

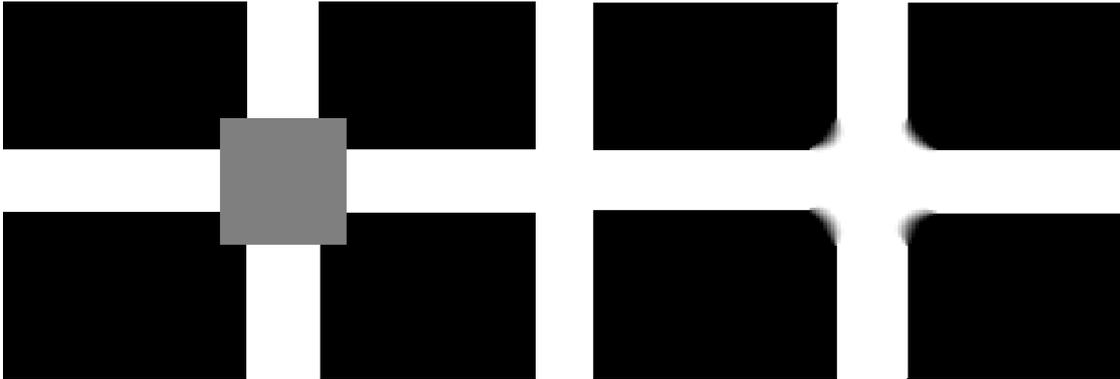


Figure 3.12: Solution of inpainting problem obtained with 1000 iterations of the ADI AMOS scheme (3.42), $\Delta t = C(\Delta x)^3$, $\varepsilon = 0.001$.

As a second example, we consider a greyvalue 300×300 photograph of a toucan in Figure 3.13. Its reconstruction is obtained in only 20 iterations using the AMOS scheme (3.42) with $\Delta t = C(\Delta x)^3$ and $\varepsilon = 0.001$.

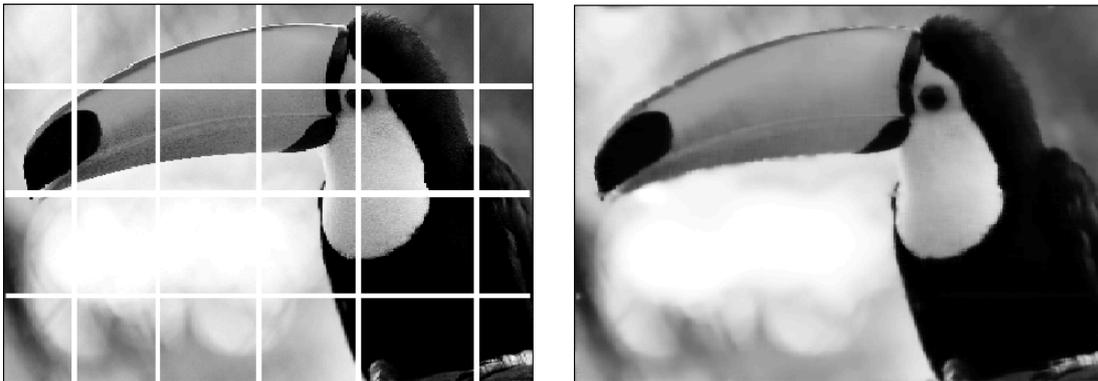


Figure 3.13: Solution of the inpainting problem obtained with ADI AMOS scheme (3.42) after 20 iterations, $\Delta t = C(\Delta x)^3$, $\varepsilon = 0.001$.

3.5.2 Numerical solution of (TV-Wass)

In the following we report some results on the numerical solution of the TV-Wasserstein flow (TV-Wass) using the primal-dual formulation and the time-stepping scheme (3.48)

solving the system (3.48). In each step of Newton's method the block-structure of the Jacobian matrix is exploited by inverting it with a Schur complement strategy. For all the following tests we use the following stopping criterion: $\|U_{n+1}^{(k)} - U_{n+1}^{(k-1)}\|_2 / \|U_{n+1}^{(k)}\|_2 \leq \epsilon_{tol}$, where ϵ_{tol} is the fixed tolerance.

1-D examples: structure of solutions and gradient flow property

We investigate the structure of solutions of the TV-Wasserstein flow (TV-Wass) in its primal-dual formulation (3.48) in the one-dimensional case. That is, we consider in the following the bounded and closed interval $\Omega = [-1, 2]$ and follow the evolution of the numerical solutions of the model for different initial conditions. The outer iterations stop when a maximum number of iterations is achieved. The results are reported in Figure 3.14. We observe that, as expected from TV-type regularisations, discontinuities are preserved and piecewise-constants solutions are favoured. The initial mass of the given densities is preserved, as confirmed from numerical verifications. Similarly as observed in [BFS12], strict positivity of the initial condition is needed for convergence.

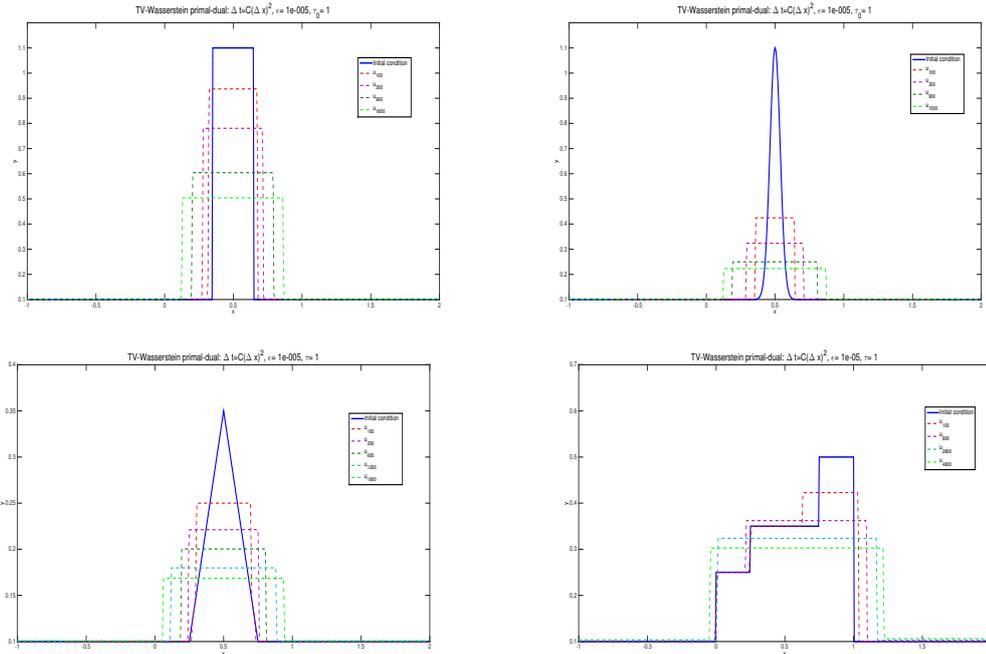


Figure 3.14: Evolution of (TV-Wass) solved with Newton system (3.48) for different initial conditions.

The smoothing effect combined with the self-similarity property of solutions of the model (see [BFS12]) results in the decreasing of the signal intensity as well as in the lengthening of its support. No new edges appear, but the existing ones are squeezed and their magnitude decreases throughout the smoothing process till convergence to a constant

solution, similarly as for the standard TV case, compare Figure 3.15.

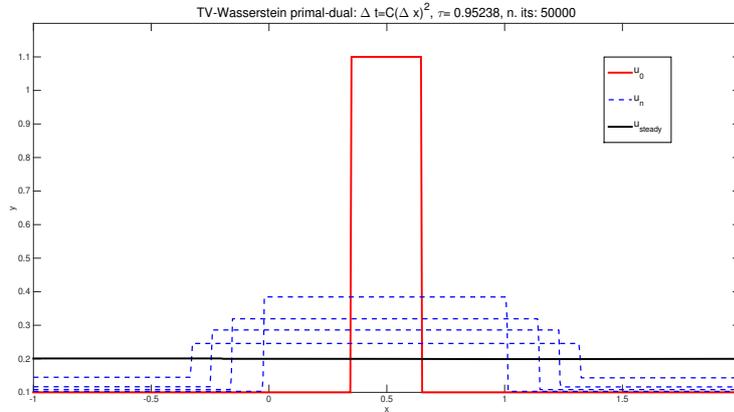


Figure 3.15: Convergence to the steady state.

In Figure 3.16 the decay of the TV energy along the iterations to the steady state is plotted, thus showing the preservation of the gradient flow structure.

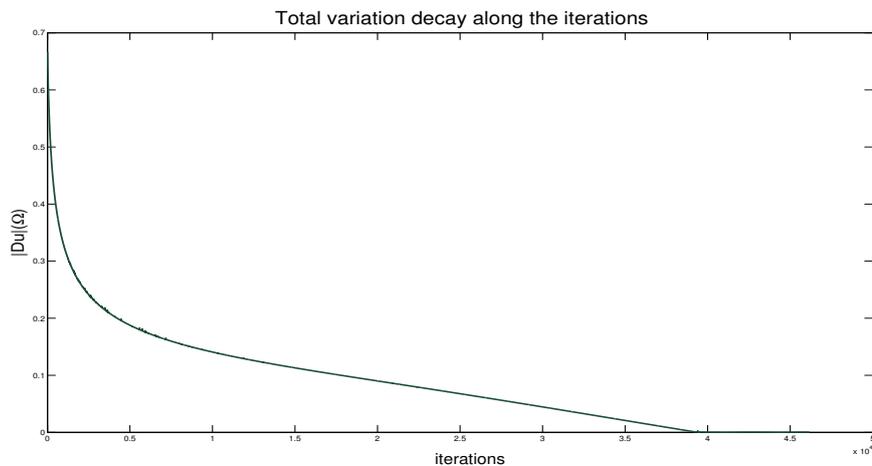


Figure 3.16: TV decay.

In Figure 3.17 we show comparison between numerical solutions of (TV-Wass) and the pure TV flow computed in both cases by using a primal-dual formulation, see [Mül08]. Iterations stop when the maximum number of 5000 iterations is reached. As observed in Figure 3.14, we note that both solutions decrease the intensity of the functions considered as initial condition, but while TV solutions show a rapid increasing of the background with no change in the support of the function, TV-Wasserstein solutions enlarge their support while increasing the intensity of the background with a much slower speed.

To conclude this section, we report in Figure 3.18 a comparison between the numerical solution of (TV-Wass) solved using (3.48) and the explicit self-similar solution explicitly

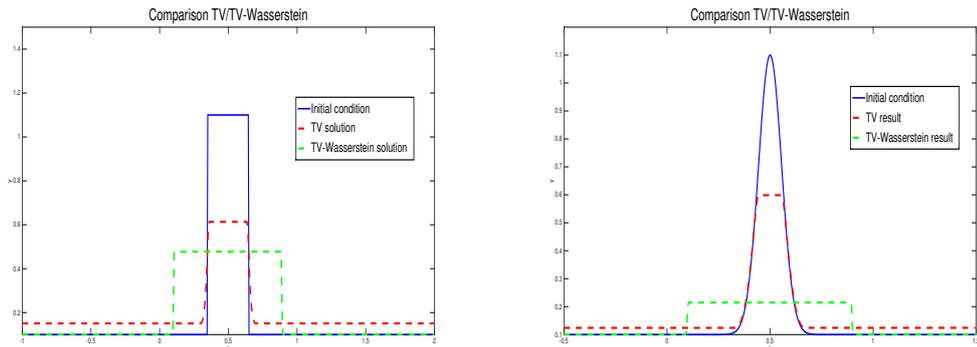


Figure 3.17: Comparison between TV and TV-Wasserstein type solutions.

calculated as in [BFS12, Section 3.4].

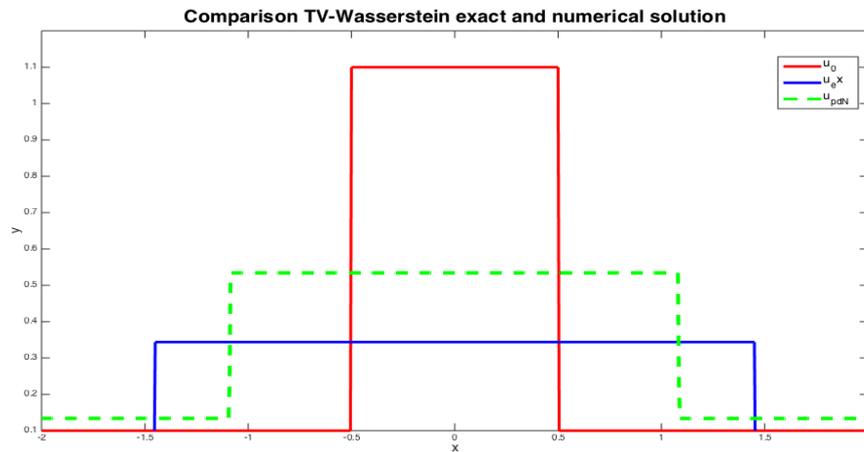


Figure 3.18: Comparison between the exact self-similar solution of (TV-Wass) computed in [BFS12, Section 3.4] and the numerical solution computed using the primal-dual Newton system (3.48).

Applications to image denoising

We now consider (TV-Wass) in the context of image restoration. Figure 3.19 shows the solution of (TV-Wass) computed via (3.48) on a 100×100 pixels initial condition of a square. Analogously as the one-dimensional example in Figure 3.17, we observe that the TV approach decreases the intensity of the square and increases the intensity of the background without changing its support.

Motivated by the use of higher-order PDEs for image smoothing discussed in Section 1.4.1, we now apply our method to a denoising problem. Figure 3.20 shows a pyramidal initial condition and its noisy version obtained by adding Gaussian noise with zero mean

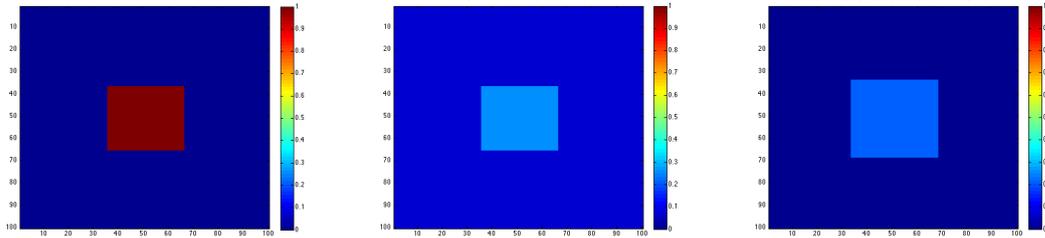
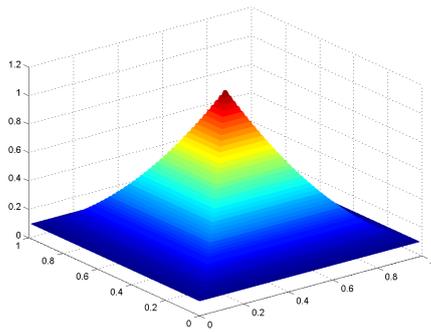
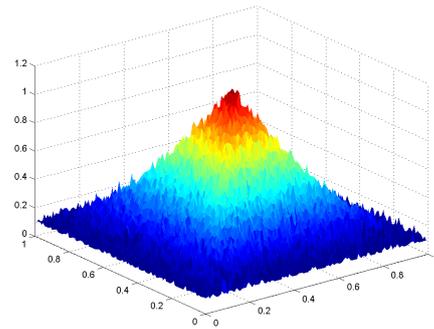


Figure 3.19: Initial condition (l.) and solution of the standard TV (m.) and TV-Wasserstein gradient flow after 5000 iterations (r.). Parameters: $\eta = 10^{-5}$, $\tau^0 = 1$.

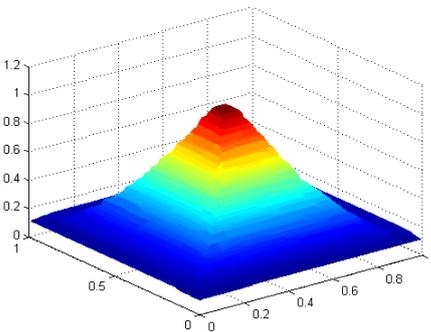
and variance $\sigma^2 = 0.001$. The denoised version is obtained both with the primal-dual TV method with penalty term described in [Mül08] and with our method. We observe that while the simple application of the TV model creates staircaising, the use of higher-order models reduces artefacts and preserves structures.



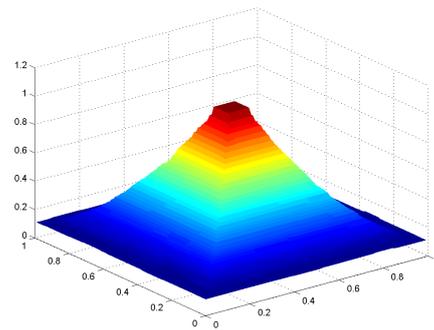
(a) Initial condition.



(b) Noisy pyramid with Gaussian noise of zero mean and variance $\sigma^2 = 0.001$.



(c) Denoised pyramid by TV primal-dual model.



(d) Denoised pyramid by TV-Wasserstein primal-dual system (3.48).

Figure 3.20: TV and TV-Wasserstein primal-dual denoising. $\eta = 10^{-5}$, $\tau^0 = 1$.

Since our numerical experiments are obtained by simulation, i.e. we start from an

original noise-free initial condition U_{ex} , we artificially corrupt it by adding some Gaussian noise and then we compute its reconstructed version U_n by means of (3.48), we report in the following Figure 3.21 the evolution of the error between U_∞ and U_n . We observe that in the early iteration there is a clear decreasing of such quantity due to the noise smoothing properties of (TV-Wass), whereas in the late iterations such error increases due to the properties previously observed in the one-dimensional case related to the lengthening of the support and the squeezing of the existing edges. We can then comment that in denoising applications, in order to compute a reconstruction of the image which is not “too far” from the desired (typically, not accessible) one, iterations should be stopped early enough.

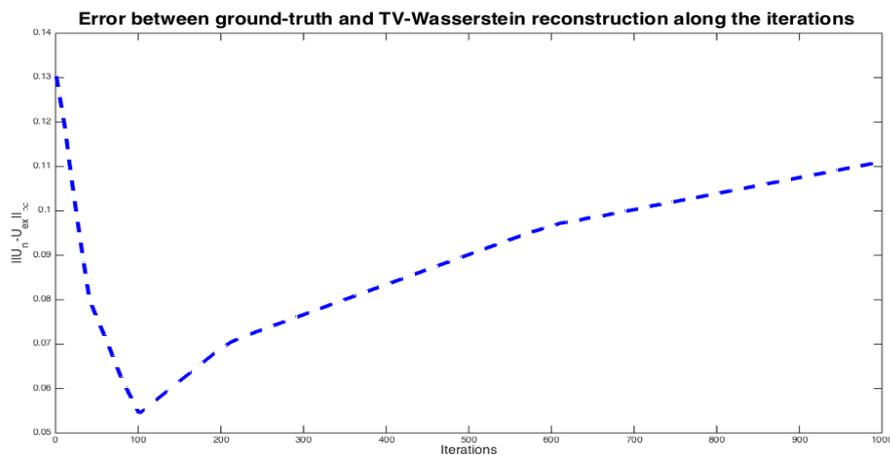
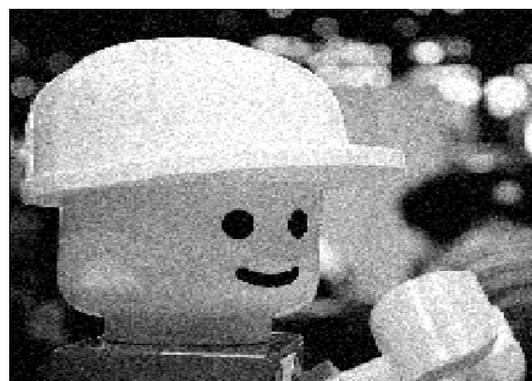


Figure 3.21: Evolution of the error between the original initial condition U_{ex} in Figure 3.20a and the numerical approximation U_n , Figure 3.20d.

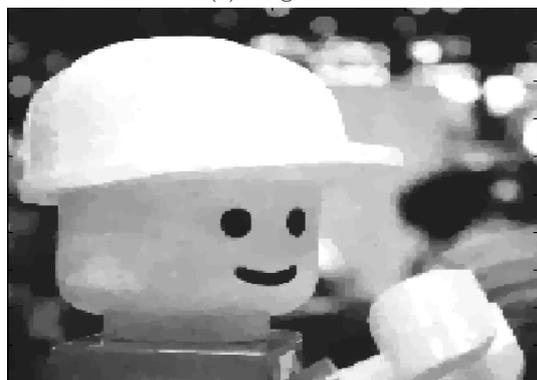
Finally, we consider in Figure 3.22 a real-world image of a LEGO man. The dimension of the image is 200×200 pixels. We add a Gaussian noise with zero mean and variance equal to 0.005 and we show the result of time evolution of the process after some time iterations. A result with the application of the TV primal-dual method is given for comparison as well. Due to the higher differential order an evident reduction of staircasing is observed.



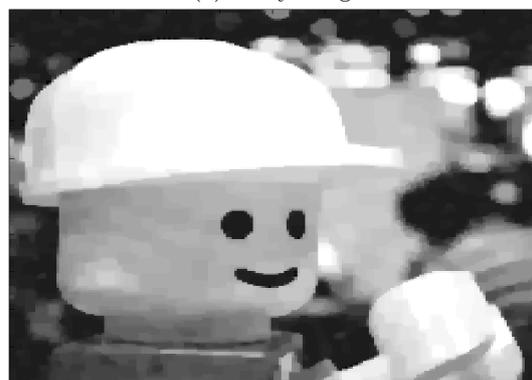
(a) Original.



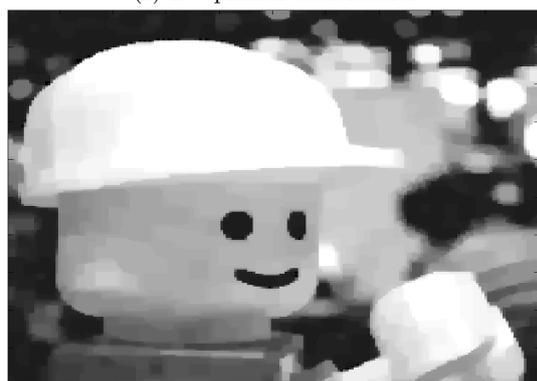
(b) Noisy image.



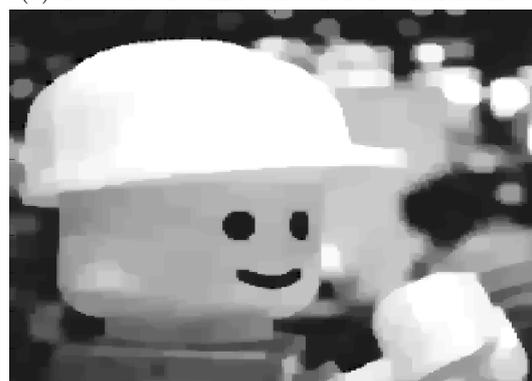
(c) TV primal-dual result.



(d) TV-Wasserstein result after 100 iterations.



(e) TV-Wasserstein result after 200 iterations.



(f) TV-Wasserstein result after 350 iterations.

Figure 3.22: Comparison between TV and TV-Wasserstein primal-dual denoising. $\eta = 10^{-7}$, $\tau^0 = 1$.

Chapter 4

A novel variational model for mixed noise removal

The correct mathematical modelling of the data fidelity term Φ in the general regularisation model (1.4) is crucial for the design of a realistic image reconstruction model which could fit appropriately the given data. Its choice is often driven by physical and statistical considerations on the noise and a Bayesian approach is frequently considered, see [Idi13, Stu10, GB00]. As previously discussed in Section 1.4.2, typically the noise is assumed to be additive, Gaussian-distributed with zero mean and variance σ^2 determining the noise intensity. This assumption is reasonable in most of the applications because of the Central Limit Theorem. One other possibility which is more appropriate for modelling transmission errors affecting only a percentage of the pixels in the image is to consider a different type of noise where the intensity value of only a fraction of pixels in the image is switched to either the maximum/minimum value of its dynamic range or to a random value within it, with positive probability. This type of noise is called *impulse* or “salt & pepper” noise. In some other applications, the additive modelling assumption does not fit the actual physical and statistical properties of the noise considered. For instance, in astronomical imaging applications a *Poisson* distribution of the noise appears more reasonable, since the physical properties of the quantised (discrete) nature of light and the independence property in the detection of photons show dependence on the signal itself, thus making not very accurate the use of a Gaussian modelling.

Recalling Section 1.4.2, from an analytical point of view, in the case when Gaussian noise is assumed an L^2 data fidelity

$$\Phi(u, f) = \int_{\Omega} |u - f|^2 dx, \quad (4.1)$$

is typically considered, see, e.g., [ROF92, CL97]. In the case of impulse noise, variational models enforcing the sparse structure of the noise distribution make use of the L^1 norm

and have been considered, for instance, in [Nik04, DAG09]. Hence, in this case the data fidelity term reads:

$$\Phi(u, f) = \int_{\Omega} |u - f| \, dx. \quad (4.2)$$

Variational models where a Poisson noise distribution is assumed have been approximated by weighted-Gaussian distributions through variance-stabilising techniques, see, e.g., [SMB94, BMPS14]. In [SBMB09, Saw11] a statistically-consistent analytical modelling for Poisson noise distributions has been derived: the resulting data fidelity term is a Kullback-Leibler-type functional Φ which reads

$$\Phi(u, f) = \int_{\Omega} (u - f \log(u)) \, dx. \quad (4.3)$$

As a result of different image acquisition and transmission factors, very often in applications the presence of multiple noise distributions has to be considered. Mixed noise distributions can be observed, for instance, when faults in the acquisition of the image are combined with transmission errors to the receiving sensor. In this case a combination of Gaussian and impulse noise is observed. In other applications, specific tools (such as illumination and/or high-energy beams) are used before the signal is actually acquired. This process is typical, for instance, in microscope and Positron Emission Tomography (PET) imaging applications and may result in a combination of a Poisson-type noise combined to an additive Gaussian noise. From a modelling point of view, the presence of multiple noise distributions has been encoded in the model by combining the data fidelity terms (4.1)-(4.2) and (4.3) in several different ways. In [HL13], for instance, a combined L^1+L^2 TV regularisation model is considered for joint impulse and Gaussian noise removal. A two-phase approach is considered in [CCN08] where two sequential steps with L^1 and L^2 data fidelity are performed to remove the impulse and the Gaussian component in the noise, respectively. Mixtures of Gaussian and Poisson noise have also been considered. In [JCPT15, JCPT12], for instance, the exact log-likelihood estimator of the mixed model is derived and its numerical solution is computed via a primal-dual splitting. A similar model has been considered in [BLCT⁺08] where a scaled gradient semi-convergent algorithm is used to solve the combined model. In [Foi09] the discrete-continuous nature of the model (due to the different support of Poisson-Gaussian distribution, respectively) is approximated to an additive model by using homomorphic variance-stabilising transformations and weighted- L^2 approximations. In [LBU11] a non-Bayesian framework to differentiate Poisson intensities from the Gaussian ones in the Haar wavelet domain is considered. Finally, a general model featuring several noise distributions which combine a linear combination of data fidelities of the type (4.1), (4.2) and (4.3) has been considered in [DLRS13, CDLRS14, CCDLR⁺15] in the context of learning the optimal denoising model from examples.

4.1 Heuristics

In this chapter we present an alternative variational model for the denoising of images corrupted by mixed noise distributions. The derivation of our model is consistent with the statistical assumptions on the data and the resulting, continuous model can be studied and analysed from an analytical point of view using standard tools from calculus of variations and functional analysis. For the numerical realisation of the model we consider a SemiSmooth Newton (SSN) method computing the solution of the model efficiently. By using the classical operation of infimal convolution, our variational model combines data fidelities Φ_i , $i = 1, 2$ associated each to the corresponding noise component in the data appropriately, allowing for a splitting of the noise into its constituting elements. In what follows, we fix the regularisation term to be the total variation. This is just a toy example and extensions to higher-order regularisations such as TV-TV² [PS14] or TGV [BKP10] can be considered as well. We refer to our model as TV-IC to highlight the infimal convolution combination of data fidelities. This has not to be confused with the ICTV model [CL97, HK14] where the same operation is used to combine TV with second-order regularisation terms.

4.1.1 The reference models

We will consider in the following two example problems which extend the inverse problem

$$\text{find } u \text{ s.t. } f = u + n \quad (4.4)$$

to the case of multiple noise distributions.

Sum of multiple noise distributions. In this case we want to reconstruct u from an observed data f which is corrupted by two different noise components combined through an *additive* structure, i.e. (4.4) has the following form

$$\text{find } u \text{ s.t. } f = u + n_1 + n_2, \quad (4.5)$$

where n_1 and n_2 are realisations of two independent random variables with different probability distributions. For instance, we may think of n_1 to be an impulse noise component and n_2 to be a Gaussian noise component.

Sum of signal-dependent and additive noise. In this case we consider the problem of a mixture of noise distributions where a *signal-dependent* component, i.e. depending on the image u we want to reconstruct, is combined to an additional additive noise component.

For simplicity, we focus on a Poisson distributed component z and consider the problem

$$\text{find } u \quad \text{s.t.} \quad f = z + n, \quad z \sim \text{Pois}(u) \quad (4.6)$$

and assume in the following that n is normally distributed.

4.2 Statistical motivation

We start our modelling in the Bayesian framework. In particular, following our presentation in Section 1.2.1, we consider the Maximum A Posteriori estimation (MAP) problem whose solution u_{MAP} is equivalently expressed by considering the negative log-likelihood such that

$$u_{MAP} \in \operatorname{argmin}_u \{-\log P(u|f)\} = \operatorname{argmin}_u \{-\log P(f|u) \log P(u)\}, \quad (4.7)$$

where the second equality follows from the Bayes' theorem and $P(u)$ is a Gibbs' model of the form $P(u) = e^{-R(u)}$, where $R(u)$ is a convex regularising energy functional.

It is well known that, given two independent real-valued random variables V and W with associated probability densities f_V and f_W , the random variable $Z := V + W$ has probability density f_Z given by the convolution of f_V and f_W , i.e.

$$f_Z(z) = \int_{\mathbb{R}} f_V(v) f_W(z-v) dv = (f_V * f_W)(z). \quad (4.8)$$

Following [HUL01, Remark 2.3.2], since f_V and f_W are nonnegative, we can define for $p > 0$ the *convolution of order p* as follows:

$$(f_V *_p f_W)(z) := \left(\int_{\mathbb{R}} (f_V(v) f_W(z-v))^p dv \right)^{1/p}, \quad (4.9)$$

which clearly corresponds to the classical convolution (4.8) if $p = 1$. By letting $p \rightarrow \infty$ it is easy to show that (4.9) converges to the *infinity convolution* of f_V and f_W defined as

$$(f_V *_\infty f_W)(z) := \sup_{v \in \mathbb{R}} f_V(v) f_W(z-v). \quad (4.10)$$

Now, in the special case when the probability densities f_V and f_W have an exponential form of the type $f_V = c_V e^{-g_V}$ and $f_W = c_W e^{-g_W}$, where c_V and c_W are positive normalisation constants and $g_V, g_W : \mathbb{R} \rightarrow \mathbb{R}_+$ are continuous and convex positive functions, (4.10) can be equivalently rewritten as:

$$(f_V *_\infty f_W)(z) = c_V c_W e^{-\inf_{v \in \mathbb{R}} (g_V(v) + g_W(z-v))}. \quad (4.11)$$

Hence, the infinity convolution of f_V and f_W corresponds to the infimal convolution of g_V and g_W after a negative exponentiation and up to multiplication by positive constants.

4.2.1 The impulse and Gaussian case

Let us focus now on the case (4.5) and consider the probability densities classically associated in the literature to impulse and Gaussian noise distributions. They are the Laplace probability density

$$f_{Lapl}(v) = \frac{e^{-g_{Lapl}(v)/\tau}}{2\tau} = \frac{e^{-|v|/\tau}}{2\tau}, \quad v \in \mathbb{R}, \quad \tau > 0, \quad (4.12)$$

for the impulse noise component and the standard Gaussian probability density

$$f_{Gaus}(s) = \frac{e^{-g_{Gaus}(s)/2\sigma^2}}{\sqrt{2\pi\sigma^2}} = \frac{e^{-|s|^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}, \quad s \in \mathbb{R}, \quad \sigma > 0 \quad (4.13)$$

for the Gaussian component. Concerning the choice (4.12), it has been shown in [MWHW78, WF95] that the heavy tails of the Laplace distribution represent good modelling candidates for the impulse noise distribution. Note that f_{Lapl} and f_{Gaus} have the negative exponential structure described above.

We consider now the problem (4.5) in a finite-dimensional statistical setting, that is

$$\text{find } u_i \quad \text{s.t.} \quad f_i = u_i + n_{1,i} + n_{2,i}$$

for each image pixel $i = 1, \dots, M$, where $n_{1,i}$ and $n_{2,i}$ are realisations of two mutually independent random variables. More precisely, the expression above describes the structure of the realisations f_i , $i = 1, \dots, M$ of the random vector $F = (F_1, \dots, F_M)$ in terms of the random vectors $U = (U_1, \dots, U_M)$, $N_1 = (N_{1,1}, \dots, N_{1,M})$ and $N_2 = (N_{2,1}, \dots, N_{2,M})$. Moreover, $n_{1,i}$ and $n_{2,i}$ are, for every $i = 1, \dots, M$, independent realisations of identically distributed Laplace and Gaussian random variables, respectively, i.e. $N_{1,i} \sim \text{Lapl}(0, \tau)$ and $N_{2,i} \sim \mathcal{N}(0, \sigma^2)$ for every i . Thus, at every pixel i the random variable $Z_i := F_i - U_i$ is the sum of the two independent random variables $N_{1,i}$ and $N_{2,i}$ having densities as in (4.12)-(4.13). Therefore, the probability distribution of Z_i is given by the convolution between f_{Lapl} and f_{Gaus} . Let us replace now such classical convolution between these two probability densities with the infinity one defined by (4.10). By (4.11), the probability density of Z_i evaluated in correspondence with one realisation z_i is then expressed as:

$$f_{Z_i}(z_i) = \frac{1}{2\tau\sqrt{2\pi\sigma^2}} e^{-\inf_{v_i \in \mathbb{R}} (g_{Lapl}(v_i)/\tau + g_{Gaus}(z_i - v_i)/2\sigma^2)}, \quad i = 1, \dots, M.$$

By independence of the realisations, we have that the probability density of $Z = (Z_1, \dots, Z_M)$

in correspondence with $z = (z_1, \dots, z_M)$ is expressed by:

$$f_Z(z) = \prod_{i=1}^M f_{Z_i}(z_i).$$

We now compute the negative log-likelihood of $P(z|u)$ in view of applying MAP estimation (4.7). We get:

$$-\log P(z|u) = -\log \prod_{i=1}^M f_{Z_i}(z_i) = -\sum_{i=1}^M \log(f_{Z_i}(z_i)) = -\sum_{i=1}^M \log(f_{Z_i}(f_i - u_i))$$

Hence, we have that the log-likelihood we intend to minimise is proportional to :

$$LL_M(f - u) := \sum_{i=1}^M \inf_{v_i \in \mathbb{R}} \left(\frac{g_{Lapl}(v_i)}{\tau} + \frac{g_{Gaus}(f_i - u_i - v_i)}{2\sigma^2} \right), \quad (4.14)$$

where the terms which do not affect the minimisation over u have been neglected. We want now to pass from a discrete to a continuous representation of the data. Following [Saw11], we assume that the elements in the space \mathbb{R}^M are interpreted as samples of functions defined on the whole image domain Ω . For convenience, we use the same notation to indicate the corresponding, continuous quantities. Considering the indicator function

$$\chi_{D_i}(x) = \begin{cases} 1, & \text{if } x \in D_i, \\ 0, & \text{else,} \end{cases}$$

where D_i is the region in the image occupied by the i -th detector, a discrete data quantity g_i can then be interpreted as mean value as follows

$$g_i = \int_{D_i} g(x) dx = \int_{\Omega} \chi_{D_i}(x) g(x) dx.$$

In this way the negative log-likelihood (4.14) can be rewritten as the following continuous variational data fidelity:

$$\begin{aligned} LL(f - u) &= \int_{\Omega} \inf_{v(x) \in \mathbb{R}} \left(\frac{g_{Lapl}(v(x))}{\tau} + \frac{g_{Gaus}(f(x) - u(x) - v(x))}{2\sigma^2} \right) d\mu(x) \\ &= \inf_{v: \Omega \rightarrow \mathbb{R}} \int_{\Omega} \left(\frac{g_{Lapl}(v(x))}{\tau} + \frac{g_{Gaus}(f(x) - u(x) - v(x))}{2\sigma^2} \right) d\mu(x) \end{aligned} \quad (4.15)$$

where $d\mu(x) = \sum_{i=1}^M \chi_{D_i} dx$ and dx denotes the usual Lebesgue measure in \mathbb{R}^2 . The infimum and the integral operators in (4.15) commute thanks to [Roc76, Theorem 3A] (see also [HM03] for further details) whenever the infimum is taken in the Hilbert space $L^2(\Omega)$, which also ensures that both terms inside the integral are well-defined. Therefore,

the corresponding variational data fidelity representing the combined presence of impulse and Gaussian noise in (4.5) reads:

$$\Phi^{\lambda_1, \lambda_2}(u, f) = \inf_{v \in L^2(\Omega)} \lambda_1 \|v\|_{L^1(\Omega)} + \frac{\lambda_2}{2} \|f - u - v\|_{L^2(\Omega)}^2, \quad (4.16)$$

with $\lambda_1 := 1/\tau$ and $\lambda_2 := 1/\sigma^2$, where the parameters τ and σ^2 quantify the intensity of the noise. Consistently with our statistical assumptions, the impulse and Gaussian noise components in (4.5) are modelled in (4.16) by norms classically used for single noise models and combined in an infimal convolution fashion. The two terms are weighted against each other by the parameters λ_1 and λ_2 , depending on the intensity of the noise distributions considered.

4.2.2 The Poisson and Gaussian case

A similar modelling can be derived for the combination of Gaussian and Poisson noise after some simple algebraic manipulations. Let us recall the model (4.6) and consider in a finite-dimensional setting the problem:

$$f_i = \pi_i + v_i$$

at every pixel $i = 1, \dots, M$. Again, equation above describes the structure of the realisations of the random vector $F = (F_1, \dots, F_M)$ in terms of the mutually independent random vectors $\Pi = (\Pi_1, \dots, \Pi_M)$ and $V = (V_1, \dots, V_M)$, where $\Pi_i \sim \text{Pois}(u_i)$ and $V_i \sim \mathcal{N}(0, \sigma^2)$ for every $i = 1, \dots, M$. In particular, for every $i = 1, \dots, M$, Π_i has the following probability density:

$$f_{\Pi_i}(\pi_i) = \frac{u_i^{\pi_i} e^{-u_i}}{\pi_i!}, \quad \pi_i \in \mathbb{R}, \quad (4.17)$$

where the factorial function is extended to the whole real line by using the classical Gamma or the Pi functions. The negative-exponential structure similar to (4.12)-(4.13) required for the modelling above in terms of the realisations π_i does not seem to be satisfied for such probability density. However, we observe that we can rewrite (4.17) as:

$$f_{\Pi_i}(\pi_i) = \exp\left(-u_i + \log\left(\frac{u_i^{\pi_i}}{\pi_i!}\right)\right) \quad i = 1, \dots, M.$$

Once again, the probability density of F_i is given by the convolution between f_{Π_i} and f_{V_i} . Replacing as before such convolution with the infinity one (4.10) as in (4.11), we have that the probability density of every F_i can be expressed as:

$$f_{F_i}(f_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\inf_{v_i \in \mathbb{R}} \left(\frac{v_i}{2\sigma^2} + u_i - \log\left(\frac{u_i^{f_i - v_i}}{(f_i - v_i)!}\right)\right)\right\}, \quad i = 1, \dots, M.$$

Proceeding similarly as above, i.e. considering the negative log-likelihood of $P(f|u) = f_F(f_1, \dots, f_M)$, neglecting all the constant terms which do not affect the minimisation over u and passing from a discrete to a continuous representation, we obtain in this case the following continuous variational model:

$$\inf_v \frac{\lambda_1}{2} \|v\|_{L^2(\Omega)}^2 + \int_{\Omega} \left(u(x) - \log \left(\frac{u(x)^{f(x)-v(x)}}{(f(x)-v(x))!} \right) \right) d\mu(x), \quad (4.18)$$

where the choice $\lambda_1 := 1/\sigma^2$ has been made and the function space where the supremum is taken has still to be made precise. Let us focus on the second term of the expression above. We have:

$$\begin{aligned} & \int_{\Omega} \left(u(x) - \log \left(\frac{u(x)^{f(x)-v(x)}}{(f(x)-v(x))!} \right) \right) d\mu(x) \\ &= \int_{\Omega} \left(u(x) - \log \left(u(x)^{f(x)-v(x)} \right) + \log \left((f(x)-v(x))! \right) \right) d\mu(x) \\ &= \int_{\Omega} \left(u(x) - (f(x)-v(x)) \log(u(x)) + \log \left((f(x)-v(x))! \right) \right) d\mu(x) \\ &\approx \int_{\Omega} \left(u(x) - (f(x)-v(x)) \log \left(\frac{f(x)-v(x)}{u(x)} \right) - (f(x)-v(x)) \right) d\mu(x) \\ &= D_{KL}(f-v, u), \end{aligned}$$

where we have used the standard Stirling approximation of the logarithm of the factorial function. The functional we end up with is the well-known Kullback-Leibler (KL) functional which has been already used for the design of several imaging models used for Poisson noise denoising, compare [SBMB09, Saw11, LCA07] and see Section 4.3.2 for its definition and properties. After introducing an additional weight λ_2 multiplying the KL term, the variational data fidelity (4.18) reads in this case:

$$\Phi^{\lambda_1, \lambda_2}(u, f) = \inf_{v \in L^2(\Omega) \cap \mathcal{B}} \frac{\lambda_1}{2} \|v\|_{L^2(\Omega)}^2 + \lambda_2 D_{KL}(f-v, u), \quad (4.19)$$

where \mathcal{B} is the subspace of $L^2(\Omega)$ ensuring the positivity of the term $f-v$, i.e. $\mathcal{B} = \{v \in L^2(\Omega) : v \leq f \text{ a.e.}\}$. In order to guarantee that the D_{KL} term is well-defined, an admissible set for u needs to be introduced as well: we give more details on that in the following Section 4.3.

Remark 4.2.1. *Our model heavily relies on the assumption of mutual independence between the random variables associated to the two different noise distributions corrupting the data. This is of course a major simplification of the model, since in real-world applications pixel-wise independence can not be guaranteed. One more realistic assumption could be assuming independence between neighbourhoods of pixels. In this respect, our modelling*

and derivation can be seen as a first move in this direction.

4.2.3 Difference with existing approaches

In the literature, variational models solving the TV denoising problem in the presence of combined impulse and Gaussian noise have been considered in [HL13, CCN08] as well in [DLRS13, CDLRS14, CCDLR⁺15] in the context of learning approaches. The general idea of these models is to combine appropriate data fidelity terms in an simple, typically additive fashion and solve the resulting model with efficient algorithms. Despite producing good results, most of these approaches are not consistent with the statistics in the data, but they rather combine somehow existing models for single noise removal in order to deal with the mixed case.

When a mixture of Gaussian-Poisson noise is assumed, an *exact* log-likelihood model has been considered in [JCPT12, JCPT15]. In the same discrete setting described above in Section 4.2.2, the expression of the negative log-likelihood reads:

$$\sum_{i=1}^M \left(-\log \sum_{\pi=0}^{\infty} \frac{u_i^{\pi} e^{-u_i}}{\pi!} \frac{e^{-\left(\frac{f_i - \pi}{\sqrt{2}\sigma}\right)^2}}{\sqrt{2\pi}\sigma} \right).$$

Here, the complexity of the analytical model derived is due to the presence of an infinite sum which corresponds to the support of the Poisson random variable considered, in contrast with the real-value support of the Gaussian one. In order to design an efficient optimisation strategy, the author first split the expression above into the sum of two different terms, the former being a convex Lipschitz-differentiable function, the latter being proper, convex and l.s.c function. In particular, the authors are able to design a competitive first-order optimisation method based on the use of primal-dual splitting algorithms exploiting the closed-form of the proximal operators associated to the convex Lipschitz-differentiable function. The advantage of such algorithms is that it is only based on the iterative application of the proximal mapping operations, without requiring any matrix inversion. Furthermore, the approximation error accumulating throughout the iterations of the algorithm are shown to be absolutely summable sequences, an essential property which is essential in this framework due to the presence of infinite sums.

Despite being derived from a modification of the classical convolution operation (4.8) and, as such, being not strictly exact from a probabilistic point of view, the variational data fidelity terms (4.16) and (4.19) are related to the mixed statistics of the noise, combining standard data fidelities in an infimal convolution fashion. From a numerical point of view, the simple structure of the models (4.16) and (4.19) allows for the design of efficient (i.e. second-order) numerical schemes. Moreover, our approach is able to decompose the noise in the different statistical components corresponding each to one particular noise

distribution in the data, a feature that, up to knowledge of the author, has not been considered in the corresponding literature.

4.3 The variational model

As usual, let $\Omega \subset \mathbb{R}^2$ a regular bounded image domain and let f be the given, noisy image to reconstruct. Further assumptions on the function spaces where f lies will be specified in the following.

For positive weights $\lambda_1, \lambda_2 > 0$ and admissible sets of functions \mathcal{A} and \mathcal{B} , the proposed Total Variation-Infimal Convolution (TV-IC) denoising model for mixed noise distribution reads:

$$\min_{u \in \text{BV}(\Omega) \cap \mathcal{A}} \left\{ |Du|(\Omega) + \Phi^{\lambda_1, \lambda_2}(u, f) \right\} \quad (\text{TV-ICa})$$

where the data fidelity $\Phi^{\lambda_1, \lambda_2}(u, f)$ has the following infimal convolution structure:

$$\Phi^{\lambda_1, \lambda_2}(u, f) := \inf_{v \in L^2(\Omega) \cap \mathcal{B}} \left\{ \mathcal{F}^{\lambda_1, \lambda_2}(u, v, f) := \lambda_1 \Phi_1(v) + \lambda_2 \Phi_2(u, f - v) \right\}. \quad (\text{TV-ICb})$$

The positive weighting parameters λ_1 and λ_2 in (TV-ICb) balance the trust in the data with the smoothing effect of the regularisation. Moreover, their size weights the fitting of the combined model with respect to the intensity of the single noise statistics in the data.

We now consider the two particular noise combinations considered in Section 4.1.1 and explain how these fit the general model above.

4.3.1 Impulse-Gaussian fidelity

In this case we have $f \in L^2(\Omega)$ and $\Phi_1(v) = \|v\|_{L^1(\Omega)}$ for the impulse noise component and $\Phi_2(u, v, f) = \frac{1}{2} \|f - v - u\|_{L^2(\Omega)}^2$ for Gaussian component. The admissible sets are therefore chosen to be $\mathcal{A} = \mathcal{B} = L^2(\Omega)$. Therefore, the infimal convolution data fidelity in (TV-ICb) reads in this case:

$$\Phi^{\lambda_1, \lambda_2}(u, f) = \inf_{v \in L^2(\Omega)} \left\{ \mathcal{F}^{\lambda_1, \lambda_2}(f, u, v) = \lambda_1 \|v\|_{L^1(\Omega)} + \frac{\lambda_2}{2} \|f - u - v\|_{L^2(\Omega)}^2 \right\}. \quad (4.21)$$

Since the set Ω is bounded, the inclusion $L^2(\Omega) \subset L^1(\Omega)$ holds. The terms in (4.21) are then both well-defined. The following Proposition asserts that the minimisation problem (4.21) is actually well-posed.

Proposition 4.3.1. *Let $f \in L^2(\Omega)$, $u \in \text{BV}(\Omega) \subset L^2(\Omega)$ and $\lambda_1, \lambda_2 > 0$. Then the minimum in the minimisation problem (4.21) is uniquely attained.*

Proof. The function $\Phi^{\lambda_1, \lambda_2}$ can be seen as the *Moreau-envelope* or *proximal map* of the L^1 norm with parameter $1/\lambda_2$ in the Hilbert space $L^2(\Omega)$ evaluated in $f - u$, see [BC11, Section 12.4]. Existence and uniqueness follow directly by standard lower semicontinuity and strict convexity properties of $\mathcal{F}^{\lambda_1, \lambda_2}(f, u, \cdot)$ in $L^2(\Omega)$. \square

Remark 4.3.2. *The unique solution of (4.21) can be explicitly computed using the operation of shrinkage. Nonetheless, in the following we will solve the two minimisation problems in (4.20) jointly using Newton-type numerical solvers.*

4.3.2 Gaussian-Poisson fidelity

In the case (4.19), some additional assumptions need to be specified. The given image is here assumed to be bounded, that is $f \in L^\infty(\Omega)$ (not restrictive in practice for greyscale images) and the data fidelities are taken as $\Phi_1(v) = \frac{1}{2}\|v\|_{L^2(\Omega)}^2$ for the Gaussian component and as the Kullback-Leibler (KL) divergence:

$$\Phi_2(u, f - v) = D_{KL}(f - v, u) = \int_{\Omega} \left(u - (f - v) + (f - v) \log \left(\frac{f - v}{u} \right) \right) d\mu \quad (4.22)$$

for the Poisson component (see below for more properties on the functional). Hence, for $f \in L^\infty(\Omega)$ the variational fidelity in (TV-ICb) in this case reads:

$$\Phi^{\lambda_1, \lambda_2}(u, f) = \inf_{v \in L^2(\Omega) \cap \mathcal{B}} \left\{ \mathcal{F}^{\lambda_1, \lambda_2}(f, u, v) = \frac{\lambda_1}{2} \|v\|_{L^2(\Omega)}^2 + \lambda_2 D_{KL}(f - v, u) \right\} \quad (4.23)$$

with the following choice of the admissible sets in (4.20) for D_{KL} to be well-defined:

$$\mathcal{A} = \{u \in L^1(\Omega), \log u \in L^1(\Omega)\} \quad \text{and} \quad \mathcal{B} = \{v \in L^2(\Omega) : v \leq f \text{ a. e.}\}. \quad (4.24)$$

We note that for $u \in \mathcal{A}$ we have $u \geq 0$ almost everywhere. Therefore, with the choice of the admissible sets as above the D_{KL} functional is well defined.

The Kullback-Leibler functional

We briefly recall here some general definitions and results on the Kullback-Leibler functional which are we going to use for the analysis of the model (4.23). We refer the reader to [Egg93, RA07, BL91, Saw11] for more details on the topic.

Definition 4.3.1 (Kullback-Leibler functional). *Let $\Omega \subset \mathbb{R}^d$ be a regular domain and μ a measure on Ω . The Kullback-Leibler (KL) functional is the function $D_{KL} : L^1(\Omega) \times L^1(\Omega) \rightarrow \mathbb{R}_+ \cup +\infty$ defined by:*

$$D_{KL}(\varphi, \psi) = \int_{\Omega} \left(\varphi \log \left(\frac{\varphi}{\psi} \right) - \varphi + \psi \right) d\mu \quad \text{for every } \varphi, \psi \geq 0 \text{ a. e.} \quad (4.25)$$

Remark 4.3.2. *Here and throughout this chapter, we use the convention $0 \log 0 = 0$ by which we have that the integrand function in (4.25) is nonnegative (by convexity of the function $f(x) = x \log x$) and vanishes if and only if $\varphi = \psi$.*

The following Proposition and Corollary collect some results from [RA07, BL91] on the convexity and weak lower semicontinuity properties of D_{KL} .

Proposition 4.3.3. *Let D_{KL} be defined as in (4.25). The following properties hold:*

- (i) *The function $(\varphi, \psi) \mapsto D_{KL}(\varphi, \psi)$ is convex.*
- (ii) *For any fixed nonnegative $\varphi \in L^1(\Omega)$, the function $D_{KL}(\varphi, \cdot)$ is lower semicontinuous with respect to the weak topology of $L^1(\Omega)$.*
- (iii) *For any fixed nonnegative $\psi \in L^1(\Omega)$, the function $D_{KL}(\cdot, \psi)$ is lower semicontinuous with respect to the weak topology of $L^1(\Omega)$.*
- (iv) *For any nonnegative functions $\varphi, \psi \in L^1(\Omega)$ the following estimate holds:*

$$\|\varphi - \psi\|_{L^1(\Omega)}^2 \leq \left(\frac{2}{3} \|\varphi\|_{L^1(\Omega)} + \frac{4}{3} \|\psi\|_{L^1(\Omega)} \right) D_{KL}(\varphi, \psi). \quad (4.26)$$

Corollary 4.3.4. *Let $\{\varphi_n\}$ and $\{\psi_n\}$ are bounded sequences in $L^1(\Omega)$. Then, estimate (4.26) entails*

$$\lim_{n \rightarrow \infty} D_{KL}(\varphi_n, \psi_n) = 0 \quad \Rightarrow \quad \lim_{n \rightarrow \infty} \|\varphi_n - \psi_n\|_{L^1(\Omega)} = 0.$$

Remark 4.3.5. *Our choice for the Poisson data fidelity (4.22) differs from previous works in Poisson noise modelling [Saw11, DLRS13] where a reduced KL-type data fidelity*

$$\tilde{\Phi}_2(u, f) = \int_{\Omega} (u - f \log u) \, d\mu$$

is used instead. The term above is obtained through Bayesian derivation and MAP estimation similarly as described in Section 4.2 (compare [Saw11, BMPS14, BB11]), but in its expression the terms that do not depend on the function u are neglected. In our combined modelling, however, those quantities have to be taken into account.

Similarly as in Proposition 4.3.1, we now guarantee that the minimisation problem (4.23) is well posed.

Theorem 4.3.6. *Let $f \in L^\infty(\Omega)$, $u \in \text{BV}(\Omega) \cap \mathcal{A}$ and $\lambda_1, \lambda_2 > 0$. Let \mathcal{A} and \mathcal{B} be as in (4.24). Then, the minimum in the minimisation problem (4.21) is uniquely attained.*

Proof. Due to the non-negativity of D_{KL} (see Remark 4.3.2), the functional $\mathcal{F}^{\lambda_1, \lambda_2}$ is bounded from below. Let us consider then a minimising sequence $\{v_n\} \subset L^2(\Omega) \cap \mathcal{B}$ for the functional $\mathcal{F}^{\lambda_1, \lambda_2}(u, \cdot, f)$. Using again the positivity of D_{KL} , we have that:

$$\frac{\lambda_1}{2} \|v_n\|_{L^2(\Omega)}^2 \leq \mathcal{F}^{\lambda_1, \lambda_2}(u, v_n, f) \leq C, \quad \text{for all } n \geq 1.$$

Hence, we can extract a non-relabelled subsequence $\{v_n\}$ weakly converging to v in $L^2(\Omega)$. As Ω is bounded and $L^2(\Omega)$ is continuously embedded $L^1(\Omega)$, $\{v_n\}$ is also converging weakly to v in $L^1(\Omega)$. Due the continuity of the L^2 norm and the weak lower semicontinuity of $D_{KL}(z, \cdot)$ for a fixed nonnegative $z \in L^2(\Omega)$ with respect to the weak topology of $L^1(\Omega)$ guaranteed by Proposition 4.3.3 we have:

$$\mathcal{F}^{\lambda_1, \lambda_2}(u, v, f) = \frac{\lambda_1}{2} \|v\|_{L^2(\Omega)}^2 + \lambda_2 D_{KL}(f - v, u) \leq \liminf_n \mathcal{F}^{\lambda_1, \lambda_2}(u, v_n, f).$$

Hence, $\mathcal{F}^{\lambda_1, \lambda_2}(u, \cdot, f)$ is weakly lower semicontinuous in $L^2(\Omega)$. Now, we need to guarantee that v is an element of \mathcal{B} . This is an immediate consequence of the fact that \mathcal{B} is convex and closed in $L^2(\Omega)$ and hence weakly closed by the Mazur's lemma. Uniqueness of the minimiser follows by the strictly convexity of $\mathcal{F}^{\lambda_1, \lambda_2}(u, \cdot, f)$. \square

Thanks to Proposition 4.3.1 and 4.3.6, we can conclude that in both cases (4.21) and (4.23), the function $\Phi^{\lambda_1, \lambda_2}$ is well-defined and that, for every $u \in \text{BV}(\Omega)$, there exists a unique element $v_u^* \in L^2(\Omega) \cap \mathcal{B}$ minimising the functional $\mathcal{F}^{\lambda_1, \lambda_2}(u, \cdot, f)$. Problem (TV-ICa) can then be rewritten as:

$$\min_{u \in \text{BV}(\Omega) \cap \mathcal{A}} \{|Du|(\Omega) + \lambda_1 \Phi_1(v_u^*) + \lambda_2 \Phi_2(u, f - v_u^*)\}, \quad (4.27)$$

where $v_u^* \in L^2(\Omega) \cap \mathcal{B}$ is the unique solution of (TV-ICb) in one of the two cases (4.21) and (4.23). In particular, there is a positive finite constant C such that:

$$\|v_u^*\|_{L^2(\Omega)} \leq C. \quad (4.28)$$

The question now is if the problem (4.27) is well-posed. Following [CL97, AV94, DLRS13] we show that the answer is positive through relaxation of the TV regularisation functional.

4.3.3 Well-posedness

In this section we prove well-posedness results for the problem (4.27). In view of the numerical realisation of the model based on the use of gradient-based methods (see Section 6.4), our analytical results are presented in the case when the nonsmoothness of the

total variation energy is regularised. This will be needed in the following to avoid the multivaluedness of the subdifferential and deal with unique gradients, a starting point for the design of every desirable optimisation algorithm we intend to use. To do that, we consider in the following a standard Huber-type regularisation of TV depending on a parameter $\gamma \gg 1$, compare with (2.11).

In particular, following [DLRSV15b], the TV term in (4.27) is Huber regularised as follows. Let $Du = \nabla u \mathcal{L}^2 + D_s u$ the Lebesgue decomposition of the two-dimensional distributional gradient Du into its absolutely continuous part $\nabla u \mathcal{L}^2$ and the singular part $D_s u$, similarly as in (2.2). We consider the following Huber-regularisation of the total variation measure $|Du|$:

$$|Du|_\gamma(V) := \int_V |\nabla u|_\gamma \, dx + \int_V |D_s u|, \quad V \in \mathcal{B}(\Omega), \quad (4.29)$$

where, in words, only the absolutely continuous part of Du is regularised using (2.11), while the singular part $D_s u$ is left untouched. In parallel to (4.27), we then consider the following Huber-regularised version of (4.27):

$$\min_{u \in \text{BV}(\Omega) \cap \mathcal{A}} \{J_\gamma(u) := |Du|_\gamma(\Omega) + \lambda_1 \Phi_1(v_u^*) + \lambda_2 \Phi_2(u, f - v^*)\}, \quad (4.30)$$

where $v_u^* \in L^1(\Omega) \cap \mathcal{B}$ is defined as above.

Recalling now the framework for the combination of Gaussian and impulse noise distributions described in Section 4.3.1 and the one for the combination of Gaussian and Poisson noise in Section 4.3.2 with the corresponding assumptions on f , \mathcal{A} and \mathcal{B} in the two different cases, we state and prove here the main well-posedness result.

Theorem 4.3.1. *Let $\lambda_1, \lambda_2 > 0$ and let v^* be the solution of the minimisation problem (TV-ICb) in one of the two cases (4.21) and (4.23) provided by Propositions 4.3.1 and 4.3.6, respectively. Then, there exists a unique solution $u \in \text{BV}(\Omega) \cap \mathcal{A}$ of the minimisation problem (4.30).*

Proof. The functional J_γ in (4.30) is bounded from below by zero. Let then $\{u_n\}$ be a minimising sequence for J_γ . We start from observing that the Huber regularisation of ∇u is coercive and has at most linear growth. Following [DLRS13, Theorem 2.1] we have in fact that if $|\nabla u| \geq \frac{1}{\gamma}$, then trivially

$$|\nabla u|_\gamma = |\nabla u| - \frac{1}{2\gamma} \leq |\nabla u|.$$

On the other hand we have that if $|\nabla u| < \frac{1}{\gamma} < 1$, then

$$|\nabla u| - \frac{1}{2\gamma} \leq |\nabla u|_\gamma = \frac{\gamma}{2} |\nabla u|^2 < \frac{\gamma}{2} |\nabla u|,$$

by convexity of $|\nabla u|^2$. Thus, we have:

$$|Du_n|(\Omega) + \lambda_2 \Phi_2(u_n, f - v_u^*) \leq |Du_n|_\gamma(\Omega) + \lambda_1 \Phi_1(v_u^*) + \lambda_2 \Phi_2(u_n, f - v_u^*) \leq M, \quad \text{for all } n. \quad (4.31)$$

The inequality above can be used to show uniform boundedness of the sequence $\{u_n\}$ in $BV(\Omega)$ thanks to some coercivity conditions.

- If $\Phi_2(u_n, f - v_{u_n}^*) = \frac{1}{2} \|f - v^* - u_n\|_{L^2(\Omega)}^2$ we have that by Young's inequality:

$$M \geq \Phi_2(u_n, f - v_{u_n}^*) = \frac{1}{2} \|f - u_n - v_{u_n}^*\|_{L^2(\Omega)}^2 \geq C_1 \|u_n\|_{L^1(\Omega)}^2 - C_2,$$

where $C_1 := \frac{1}{4|\Omega|}$ and $C_2 := \|f - v_{u_n}^*\|_{L^2}^2$ is finite by (4.28). Hence, $\{u_n\}$ is bounded in $L^1(\Omega)$. Moreover, using the positivity of Φ_2 , we deduce from (4.31) that

$$|Du_n|(\Omega) \leq M, \quad \text{for all } n.$$

- If $\Phi_2(u_n, f - v_{u_n}^*) = D_{KL}(f - v_{u_n}^*, u_n)$, we apply directly the BV-coercivity result for the TV-KL functional proved in [Saw11, Lemma 6.3.2] to (4.31) and get:

$$C \|u_n\|_{BV(\Omega)} \leq |Du_n|(\Omega) + \lambda_2 D_{KL}(f - v_{u_n}^*, u_n) \leq J_\gamma(u_n) \leq M, \quad \text{for every } n.$$

Thus, in both cases, the sequence $\{u_n\}$ is uniformly bounded in $BV(\Omega)$. Since $BV(\Omega)$ is embedded in $L^1(\Omega)$ with continuous embedding (see Theorem 2.1.13), we have that, up to non-relabelled subsequences, there is $u \in BV(\Omega)$ such that:

$$u_n \xrightarrow{*} u \text{ in } BV(\Omega), \quad u_n \rightarrow u \text{ in } L^1(\Omega).$$

Therefore, since the image domain Ω is bounded, a further non-relabelled subsequence $\{u_n\}$ converging pointwise to u a.e. in Ω can be extracted. We claim now that J_γ is weakly lower semicontinuous in $L^1(\Omega)$. To show that, we observe first that the regularisation term $|Du|_\gamma(\Omega)$ is lower-semicontinuous w.r.t. strong convergence in L^1 (compare [DT84]). Additionally, we have that in both cases (4.21) and (4.23), Φ_2 is lower semicontinuous w.r.t. to the weak topology in L^1 .

- If $\Phi_2(u_n, f - v_{u_n}^*) = \frac{1}{2} \|f - v_{u_n}^* - u_n\|_{L^2(\Omega)}^2$, we have:

$$\int_{\Omega} (f - v_{u_n}^* - u)^2 dx \leq \liminf_n \int_{\Omega} (f - v_{u_n}^* - u_n)^2 dx$$

by continuity of the integrand function combined with the pointwise convergence of $\{u_n\}$ and Fatou's lemma;

- If $\Phi_2(u_n, f - v_{u_n}^*) = D_{KL}(f - v_{u_n}^*, u_n)$, this holds thanks to Proposition 4.3.3.

This, combined with the strict convexity of the functional J_γ , ensures existence and uniqueness of the minimiser in $BV(\Omega)$. Furthermore, the minimiser is indeed an element of the admissible set \mathcal{A} as a consequence of the fact that \mathcal{A} (when it is not trivially $L^2(\Omega)$) is a closed and convex subspace of $BV(\Omega)$ and, as such, weakly closed by Mazur's lemma. \square

Corollary 4.3.2. *Let $\lambda_1, \lambda_2 > 0$ and let $v_u^* \in L^2(\Omega) \cap \mathcal{B}$ be the solution of the minimisation problem (TV-ICb). Then, there exists a unique solution $u \in BV(\Omega) \cap \mathcal{A}$ of the nonsmooth minimisation problem (4.27).*

Proof. Starting from (4.31) and using standard lower semicontinuity of $|Du|(\Omega)$ with respect to the strong topology of $L^1(\Omega)$, the result is an immediate consequence of the one proved above. \square

The question is now how to connect the solution of the minimisation problem (4.30) to a solution of the nonsmooth problem (4.27). Recalling Section 2.1.3, we do that via Γ -convergence [DM93].

Proposition 4.3.3. *The sequence of functionals $J_\gamma : BV(\Omega) \times \mathcal{A} \rightarrow \mathbb{R}$ defined in (4.30) Γ -converges to the functional $J : BV(\Omega) \times \mathcal{A} \rightarrow \mathbb{R}$ defined in (4.27) as $\gamma \rightarrow \infty$. Hence, the minimiser of J_γ converges to a minimiser of J as $\gamma \rightarrow \infty$ in $BV(\Omega) \cap \mathcal{A}$.*

Proof. The proof is a standard result based on relaxation techniques for measures, see e.g. [BBB95, BB90]. We observe that as $\gamma \rightarrow \infty$ the functional J_γ converges pointwise, for every $u \in BV(\Omega) \cap \mathcal{A}$ to

$$J(u) = \int_{\Omega} |\nabla u| \, dx + \int_{\Omega} |D_s u| \, dx + \lambda_1 \Phi_1(v_u^*) + \lambda_2 \Phi_2(u, f - v_u^*)$$

since for the absolutely continuous part ∇u there holds $|\nabla u|_\gamma \rightarrow |\nabla u|$ pointwise as $\gamma \rightarrow \infty$. Since the convergence is monotonically increasing, by Proposition 2.1.21, we have that J_γ Γ -converges to J . \square

Therefore, thanks to the results above, the general infimal convolution model (TV-ICa)-(TV-ICb) is well-posed in both cases described in Section 4.3.1 and 4.3.2. For simplicity, we will focus in the following on an equivalent formulation of the nonsmooth problem (TV-ICa)-(TV-ICb) which reads:

$$\min_{\substack{u \in BV(\Omega) \cap \mathcal{A} \\ v \in L^2(\Omega) \cap \mathcal{B}}} \{J(u, v) := |Du|(\Omega) + \lambda_1 \Phi_1(v) + \lambda_2 \Phi_2(u, f - v)\}, \quad (4.32)$$

where the two components of the model, i.e. the reconstructed image u and the noise component v are treated jointly. Analogously, we will consider in the forthcoming Section

6.4 the corresponding Huber-regularised version

$$\min_{\substack{u \in \text{BV}(\Omega) \cap \mathcal{A} \\ v \in L^2(\Omega) \cap \mathcal{B}}} \{J_\gamma(u, v) := |Du|_\gamma(\Omega) + \lambda_1 \Phi_1(v) + \lambda_2 \Phi_2(u, f - v)\},$$

and use it for the design of gradient-based numerical schemes solving (4.32).

4.4 Recovery of the pure noise models

The infimal convolution modelling derived in the previous sections combines variational data fidelities classically used in the literature for single-noise models to deal with the combined case. The reader may ask whether and under which conditions the single noise models can be recovered from our model. In this section we show that this is possible by looking at the behaviour of solutions (u^*, v^*) of (4.32) as the parameters λ_1 and λ_2 become infinitely large. Similarly as our exposition so far, in the following analysis we discuss the Gaussian-impulse (see Section 4.3.1) and the Gaussian-Poisson case (see Section 4.3.2) separately for better clarity.

4.4.1 Recovery of TV- L^1 and TV- L^2 solutions

The following Proposition asserts essentially that TV- L^2 [ROF92] and TV- L^1 [Nik04, DAG09] -type solutions can be recovered ‘asymptotically’ from (4.32) by letting the impulse or the Gaussian fidelity weight go to infinity, respectively. Moreover, when both parameters become infinitely large, a full recovery of the data is obtained. The proof is based on standard energy estimates for the minimisation problem (4.32) in the case when the data fidelity is chosen as in (4.21).

Proposition 4.4.1. *Let $(u^*, v^*) \in \text{BV}(\Omega) \times L^2(\Omega)$ the optimal pair for (4.32) in the Gaussian-impulse case described in Section 4.3.1. If $f \in L^2(\Omega)$ is not identically zero, then the following asymptotical convergences hold:*

- i) *If λ_2 is finite, then $v^* \rightarrow 0$ in $L^1(\Omega)$ as $\lambda_1 \rightarrow +\infty$.*
- ii) *If λ_1 is finite, then $v^* \rightarrow f - u^*$ in $L^2(\Omega)$ as $\lambda_2 \rightarrow +\infty$.*
- iii) *If additionally $f \in \text{BV}(\Omega)$ and is not a constant, then the pair (u^*, v^*) converges to $(f, 0)$ in $L^1(\Omega) \times L^1(\Omega)$ as $\lambda_1, \lambda_2 \rightarrow +\infty$.*

Proof. We start writing the variational inequality satisfied by the optimal pair $(u^*, v^*) \in \text{BV}(\Omega) \times L^2(\Omega)$ in this case. This reads:

$$\begin{aligned} |Du^*|(\Omega) + \lambda_1 \|v^*\|_{L^1(\Omega)} + \frac{\lambda_2}{2} \|f - u^* - v^*\|_{L^2(\Omega)}^2 &\leq \\ |Du|(\Omega) + \lambda_1 \|v\|_{L^1(\Omega)} + \frac{\lambda_2}{2} \|f - u - v\|_{L^2(\Omega)}^2, &\quad \text{for all } u \in \text{BV}(\Omega), v \in L^2(\Omega). \end{aligned} \quad (4.33)$$

For the proof of *i*) we choose in (4.33) $u = v = 0$. We have:

$$\lambda_1 \|v^*\|_{L^1(\Omega)} \leq |Du^*|(\Omega) + \lambda_1 \|v^*\|_{L^1(\Omega)} + \frac{\lambda_2}{2} \|f - u^* - v^*\|_{L^2(\Omega)}^2 \leq \frac{\lambda_2}{2} \|f\|_{L^2(\Omega)}^2,$$

which implies:

$$\|v^*\|_{L^1(\Omega)} \leq \frac{\lambda_2}{2\lambda_1} \|f\|_{L^2(\Omega)}^2.$$

Since λ_2 is positive and fixed, by letting λ_1 to infinity we have that v^* converges to 0 in $L^1(\Omega)$.

Similarly, for *ii*) we choose $u = 0$ and $v = f$ in (4.33) and get:

$$\|f - u^* - v^*\|_{L^2(\Omega)}^2 \leq \frac{2\lambda_1}{\lambda_2} \|f\|_{L^1(\Omega)}.$$

The conclusion follows analogously by letting λ_2 to infinity.

For the proof of *iii*) we choose in (4.33) $u = f \in \text{BV}(\Omega)$ and $v = 0$. Since f is not a constant $|Df|(\Omega) \neq 0$. Thus, we have:

$$\begin{aligned} C_{\lambda_1, \lambda_2} \left(\|v^*\|_{L^1(\Omega)} + \|f - u^* - v^*\|_{L^2(\Omega)}^2 \right) \\ \leq |Du^*|(\Omega) + \lambda_1 \|v^*\|_{L^1(\Omega)} + \frac{\lambda_2}{2} \|f - u^* - v^*\|_{L^2(\Omega)}^2 \leq \frac{\lambda_2}{2} |Df|(\Omega), \end{aligned}$$

where $C_{\lambda_1, \lambda_2} := \min \left\{ \lambda_1, \frac{\lambda_2}{2} \right\}$. Therefore:

$$\|v^*\|_{L^1(\Omega)} + \|f - u^* - v^*\|_{L^2(\Omega)}^2 \leq \frac{1}{C_{\lambda_1, \lambda_2}} |Df|(\Omega).$$

The right hand side of the inequality above goes to zero as $\lambda_1, \lambda_2 \rightarrow +\infty$ since in this case C_{λ_1, λ_2} goes to infinity. Thus, as $\lambda_1, \lambda_2 \rightarrow +\infty$ we have the following convergences:

$$\begin{aligned} v^* &\rightarrow f - u^* && \text{in } L^2(\Omega), \\ v^* &\rightarrow 0 && \text{in } L^1(\Omega). \end{aligned}$$

By uniqueness of the limit in $L^1(\Omega)$ we conclude that $f - u^* = 0$ and, consequently that *iii*) holds. \square

Using the same energy estimates above it is also possible to show that similar convergence results hold whenever one of the two fidelity weights goes to zero.

Corollary 4.4.2. *Let $(u^*, v^*) \in \text{BV}(\Omega) \times L^2(\Omega)$ the optimal pair for (4.32) in the Gaussian-impulse case described in Section 4.3.1. If $f \in L^2(\Omega)$ is not identically zero, then:*

- (i) *If λ_1 is finite, then $|Du^*|(\Omega) \rightarrow 0$ and $v^* \rightarrow 0$ in $L^1(\Omega)$ as $\lambda_2 \rightarrow 0$.*
- (ii) *If λ_2 is finite, then $|Du^*|(\Omega) \rightarrow 0$ and $v^* \rightarrow f - u^*$ in $L^2(\Omega)$ as $\lambda_1 \rightarrow 0$.*

Proof. The proof is similar to the points *i)* and *ii)* of the Proposition 4.4.1. Starting from the variational inequality (4.33) one in fact gets similar estimates as above and the conclusion simply follows by letting the parameters go to zero. \square

4.4.2 Recovery of TV- L^2 and TV-KL solutions

In the Gaussian-Poisson framework described in Section 4.3.2 similar results can be proved. They rely on analogous energy estimates and, essentially, on the estimate (4.26) for the KL fidelity term (4.22). Analogously as before, by letting the Gaussian and Poisson fidelity weights go to infinity, solutions of TV-KL [SBMB09, LCA07] and TV- L^2 [ROF92] model can be recovered as well as the full data.

Proposition 4.4.3. *Let \mathcal{A} and \mathcal{B} be the admissible sets in (4.24). Let $(u^*, v^*) \in (\text{BV}(\Omega) \cap \mathcal{A}) \times (L^2(\Omega) \cap \mathcal{B})$ the optimal pair for (4.32) in the Gaussian-Poisson case described in Section 4.3.2. If $f \in L^\infty(\Omega)$ is not identically zero, then the following asymptotical convergences hold:*

- i) If λ_2 is finite and f is not identically equal to one, then $v^* \rightarrow 0$ in $L^2(\Omega)$ as $\lambda_1 \rightarrow +\infty$.*
- ii) If λ_1 is finite, then $v^* \rightarrow f - u^*$ in $L^1(\Omega)$ as $\lambda_2 \rightarrow +\infty$.*
- iii) If additionally $f \in \text{BV}(\Omega)$ and is not a constant, then the pair (u^*, v^*) converges to $(f, 0)$ in $L^1(\Omega) \times L^2(\Omega)$ as $\lambda_1, \lambda_2 \rightarrow +\infty$.*

Proof. Again, we start writing explicitly the variational inequality satisfied by the optimal pair $(u^*, v^*) \in (\text{BV}(\Omega) \cap \mathcal{A}) \times (L^2(\Omega) \cap \mathcal{B})$ for Gaussian-Poisson combined noise case. This reads:

$$|Du^*|(\Omega) + \frac{\lambda_1}{2} \|v^*\|_{L^2(\Omega)}^2 + \lambda_2 D_{KL}(f - v^*, u^*) \leq |Du|(\Omega) + \frac{\lambda_1}{2} \|v\|_{L^2(\Omega)}^2 + \lambda_2 D_{KL}(f - v, u), \quad \text{for all } u \in (\text{BV}(\Omega) \cap \mathcal{A}) \text{ and } v \in (L^2(\Omega) \cap \mathcal{B}). \quad (4.34)$$

For the proof of *i)* we choose in (4.34) $u = 1_\Omega$, the constant function identically equal to one on Ω and $v = 0$. We deduce:

$$\frac{\lambda_1}{2} \|v^*\|_{L^2(\Omega)}^2 \leq \lambda_2 D_{KL}(f, 1_\Omega).$$

Since by assumption the function f is not identically equal to one, the right hand side of the inequality above is strictly positive and bounded as:

$$0 < D_{KL}(f, 1_\Omega) \leq K := \left(\|f\|_{L^\infty(\Omega)} \|\log f\|_{L^1(\Omega)} + \|f\|_{L^1(\Omega)} \right) < \infty,$$

by Hölder inequality and assumptions on f . Thus, we have:

$$\|v^*\|_{L^2(\Omega)}^2 \leq \frac{2\lambda_2}{\lambda_1} K,$$

which implies the convergence $v^* \rightarrow 0$ in $L^2(\Omega)$ as $\lambda_1 \rightarrow +\infty$.

To prove *ii*), we consider in (4.34) $u = 0$ and $v = f$. For such a choice we have $D_{KL} = 0$. Hence, inequality (4.34) reduces to:

$$\lambda_2 D_{KL}(f - v^*, u^*) \leq \frac{\lambda_1}{2} \|f\|_{L^2(\Omega)}^2,$$

which implies that $D_{KL}(f - v^*, u^*) \rightarrow 0$ as $\lambda_2 \rightarrow +\infty$. We now apply Corollary 4.3.4 and deduce $v^* \rightarrow f - u^*$ in $L^1(\Omega)$.

We can show that *iii*) holds by choosing $v = 0$ and $u = f \in \text{BV}(\Omega)$ with $|Df|(\Omega) \neq 0$ by assumption. Proceeding similarly as in 4.4.1, the convergence result follows immediately after applying once again Corollary 4.3.4. \square

Corollary 4.4.4. *Let \mathcal{A} and \mathcal{B} be the admissible sets in (4.24). Let $(u^*, v^*) \in (\text{BV}(\Omega) \cap \mathcal{A}) \times (L^2(\Omega) \cap \mathcal{B})$ the optimal pair for (4.32) in the Gaussian-Poisson case described in Section 4.3.2. If $f \in L^\infty(\Omega)$ is not identically zero, then:*

(i) *If λ_1 is finite and f is not identically equal to one, then $|Du^*|(\Omega) \rightarrow 0$ and $v^* \rightarrow 0$ in $L^2(\Omega)$ as $\lambda_2 \rightarrow 0$.*

(ii) *If λ_2 is finite, then $|Du^*|(\Omega) \rightarrow 0$ and $v^* \rightarrow f - u^*$ in $L^1(\Omega)$ as $\lambda_1 \rightarrow 0$.*

4.5 Numerical results

In this section we report the numerical results of the solution of general mixed noise model (TV-ICa)-(TV-ICb) (or, equivalently, (4.32)) for the two frameworks described in Sections 4.3.1 and 4.3.2.

We consider a discretised image domain $\Omega = \{(x_i, y_j) : i = 1, \dots, N, j = 1, \dots, M\}$ with cardinality $|\Omega| = NM$. Similarly as in Chapter 3, standard finite difference discretisation is used. In particular, forward and backward finite differences are considered for the discretisation of the divergence and gradient operators, respectively, thus preserving their mutual adjointness property.

For the numerical realisation of the model, we use a **SemiSmooth Newton** (SSN) type algorithm combined with a primal-dual strategy. To do that, we regularise the nonsmooth TV term by Huber-regularised using a parameter $\gamma \gg 1$, thus having unique-gradients. Relations with the original problem are guaranteed by Theorem 4.3.3. A more specific discussion on the numerical aspects for solving similar TV models will be given in Section 5.5. In the following numerical realisation we did not consider any preconditioning method

to improve upon the conditioning of the linear systems appearing after the linearisation in the Newton step. This is an interesting improvement which could be addressed for future research.

Our motivation for using Newton’s algorithms to solve our problem comes from recent works on the use of second-order methods for the numerical solution of similar nonsmooth problems arising in the context of learning approaches, see [DLRS13, CDLRS14, CCDLR⁺15]. Alternatively, other numerical approaches dealing directly with the nonsmoothness of the problems such as dual algorithms [Cha04] and numerical convex solvers such as `cvx` can be used.

For our numerical realisation, we use a combined stopping criterion where the algorithm stops either when the norm of the difference between two different iterates is below a given tolerance or when a maximum number of iterations (typically, 35) is attained.

Our numerical results show the property of our model to deal with the different noise components in the image by means of ‘appropriate’ (i.e. statistically consistent) data fidelity terms. In the following we focus on the formulation (4.32) which allows to decompose the noise into its components.

Parameter choice For our computational tests we consider different images selected either from the Berkeley database¹ in Figure 4.1 and some others we report in Figure 4.2. For each experiment the ground truth image is corrupted with mixed noise distributions of different intensities which are specified for each case in the following numerical tests. For simplicity, we consider square $N \times N$ pixel images (corresponding to a step size $h = 1/N$) and fix the Huber-regularisation parameter for all the following tests to be $\gamma = 1e5$.



Figure 4.1: Some images from Berkeley database.

¹<https://www.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/BSDS300/html/dataset/images.html>

²Moon image: <http://commons.wikimedia.org/wiki/File:FullMoon2010.jpg#/media/File:FullMoon2010.jpg>,
Brain image: <http://www.oasis-brains.org/>, <http://meyerinst.com/confocals/tcs-spe/index.htm>

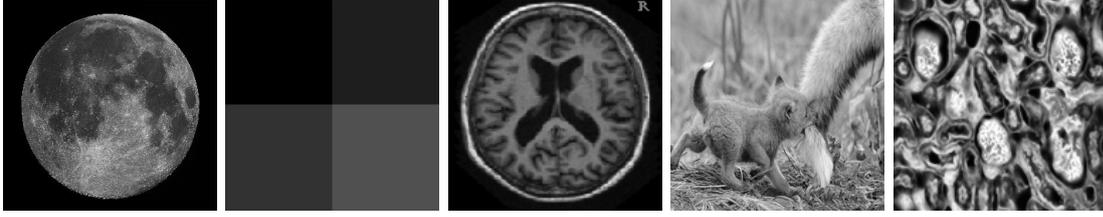


Figure 4.2: Some other images used for our experiments².

4.5.1 Impulse-Gaussian case

We start focusing on the framework considered in Section 4.3.1 for the joint removal of impulse and Gaussian noise. That is, we compute numerically the solution pair of the following Huber-regularised minimisation problem

$$\min_{u,v} \left\{ J_\gamma(u, v) := |Du|_\gamma(\Omega) + \lambda_1 \|v\|_{L^1(\Omega)} + \frac{\lambda_2}{2} \|f - u - v\|_{L^2(\Omega)}^2 \right\}. \quad (4.35)$$

We observe that the L^1 -term in (4.35) dealing with the sparse impulse component of the noise introduces a further non-differentiability obstacle in the design of a numerical gradient-based optimisation method solving (4.35). Therefore, similarly as before, we Huber-regularise this term depending a parameter $\gamma \gg 1$ as we did for the TV term and consider the following optimality conditions:

$$\begin{cases} \frac{\partial J_\gamma}{\partial u} = -\nabla \cdot \left(\frac{\gamma \nabla u}{\max(\gamma |\nabla u|, 1)} \right) - \lambda_2(f - u - v) = 0, \\ \frac{\partial J_\gamma}{\partial v} = \lambda_1 \frac{\gamma v}{\max(\gamma |v|, 1)} - \lambda_2(f - u - v) = 0. \end{cases}$$

The system above can be equivalently written in primal-dual form as:

$$\begin{cases} -\nabla \cdot q - \lambda_2(f - u - v) = 0, \\ \lambda_1 p - \lambda_2(f - u - v) = 0, \\ q = \left(\frac{\gamma \nabla u}{\max(\gamma |\nabla u|, 1)} \right), \\ p = \frac{\gamma v}{\max(\gamma |v|, 1)}. \end{cases}$$

Starting from an appropriate initial guess for u_0 and v_0 (which for our experiments is the

noisy image and a sparse vector, respectively), the SSN iteration reads:

$$\begin{cases} -\nabla \cdot \delta_q + \lambda_2 \delta_u + \lambda_2 \delta_v = -(-\nabla \cdot q + \lambda_2(f - u - v)), \\ \lambda_1 \delta_p + \lambda_2 \delta_u + \lambda_2 \delta_v = -(\lambda_1 p - \lambda_2(f - u - v)), \\ \delta_q - \frac{\gamma \nabla \delta_u}{\max(1, \gamma |\nabla u|)} + \chi_{\mathcal{U}_\gamma} \gamma^2 \frac{\nabla u^T \nabla \delta_u}{\max(1, \gamma |\nabla u|)^2} \frac{q}{\max(1, |q|)} = -q + \frac{\gamma \nabla u}{\max(1, \gamma |\nabla u|)}, \\ \delta_p - \frac{\gamma \delta_v}{\max(1, \gamma |v|)} + \chi_{\mathcal{V}_\gamma} \gamma^2 \frac{v \delta_v}{\max(1, \gamma |v|)^2} \frac{p}{\max(1, |p|)} = -p + \frac{\gamma v}{\max(1, \gamma |v|)}, \end{cases} \quad (4.36)$$

for the increments $\delta_u, \delta_v, \delta_q$ and δ_p and where the active sets \mathcal{U}_γ and \mathcal{V}_γ are defined as: $\mathcal{U}_\gamma := \{x \in \Omega : \gamma |\nabla u(x)| > 1\}$ and $\mathcal{V}_\gamma := \{x \in \Omega : \gamma |v(x)| > 1\}$. As in [DIR12, DLRS13, CDLRS14, CCDLR⁺15], the SSN iteration above has been modified using the properties of the solution on the final active sets \mathcal{U}_γ and \mathcal{V}_γ . Namely, on these sets we have that $q = \frac{\nabla u}{|\nabla u|}$ and $p = \frac{v}{|v|}$ together with $q, p \leq 1$ a.e. in Ω . The standard Newton's iteration can then be modified accordingly, thus obtaining a positive definite Hessian matrix in each iteration which ensures global convergence.

Figure 4.3 shows the numerical denoising results for 312×312 pixel images corrupted with a combination of Gaussian and impulse noise of different intensities. In all experiments we observe that the noise is successfully removed from the original image and it is decomposed in its two impulse and Gaussian components corresponding to the L^1 and L^2 term in (4.35), respectively. Namely, the sparse component of the noise associated to the impulse component is extracted from the noisy image through the use of the L^1 norm of the component v , while a classical L^2 -fidelity is used to denoise the residual Gaussian-noisy image. The solution pair (u^*, v^*) is computed jointly, so the noise removal process can be thought of as an iterative process where in each iteration the impulse component of the noise is extracted and v is updated, while the Gaussian noise component is treated with the L^2 fidelity of the residuum. The parameters λ_1 and λ_2 are optimised in order to obtain the best Peak Signal to Noise Ratio (PSNR). We observe that the reconstructed images may suffer of some loss of contrast corresponding to some loss of structures encoded in the noise components (cf. fifth column of Figure 4.3). This is a well-known drawback of TV regularisation [Mey01] and can be improved by using higher-order imaging models such as TV-TV² [PS14] or TGV regularisation [BKP10, VBK13], or, in alternative, by solving the TV problem using Bregman iteration to reduce such drawback [OBG⁺05]. The application of higher-order regularisation methods for similar mixed noise models is indeed a matter future research.

Figure 4.4 confirms numerically the convergence results of Proposition 4.4.1. The impulse and Gaussian noise component of the model are plotted and their convergence to zero is observed as the corresponding weighting parameter of the model goes to infinity. In particular, this behaviour corresponds to an ‘‘asymptotical’’ convergence of the combined

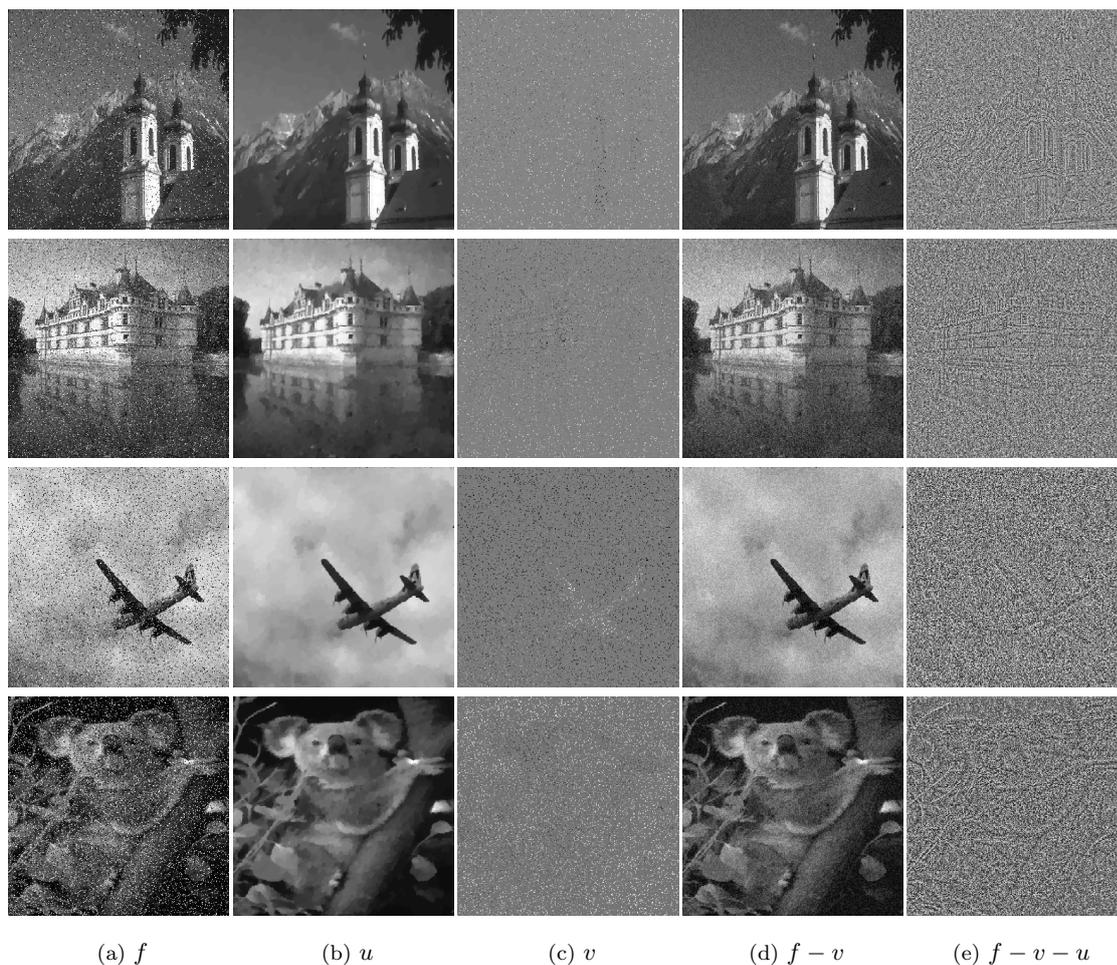


Figure 4.3: **First column:** Images corrupted with zero-mean Gaussian noise and impulse noise. **Second column:** Denoising result. **Third column:** Impulse noise component. **Fourth column:** Gaussian residuum. **Fifth column:** Gaussian noise component. **First row:** impulse noise density $d = 5\%$, Gaussian noise variance $\sigma^2 = 0.001$. Noisy image PSNR=17.91 dB. Denoised version PSNR=29.03 dB. Parameters: $\lambda_1 = 651$, $\lambda_2 = 6779$.

Second row: impulse noise density $d = 5\%$, Gaussian noise variance $\sigma^2 = 0.01$. Noisy image PSNR=16.11 dB. Denoised version PSNR=25.02 dB. Parameters: $\lambda_1 = 622$, $\lambda_2 = 4201$.

Third row: impulse noise density $d = 10\%$, Gaussian noise variance $\sigma^2 = 0.005$. Noisy image PSNR=14.63 dB. Denoised version PSNR=31.43 dB. Parameters: $\lambda_1 = 523$, $\lambda_2 = 5551$.

Fourth row: impulse noise density $d = 15\%$, Gaussian noise variance $\sigma^2 = 0.005$. Noisy image PSNR=12.66 dB. Denoised version PSNR=26.40 dB. Parameters $\lambda_1 = 482$, $\lambda_2 = 5233$.

model (4.35) to the classical TV denoising models for single noise removal (i.e. the classical ROF model [ROF92] for the Gaussian noise case and the TV- L^1 model [DAG09] for the impulse case).

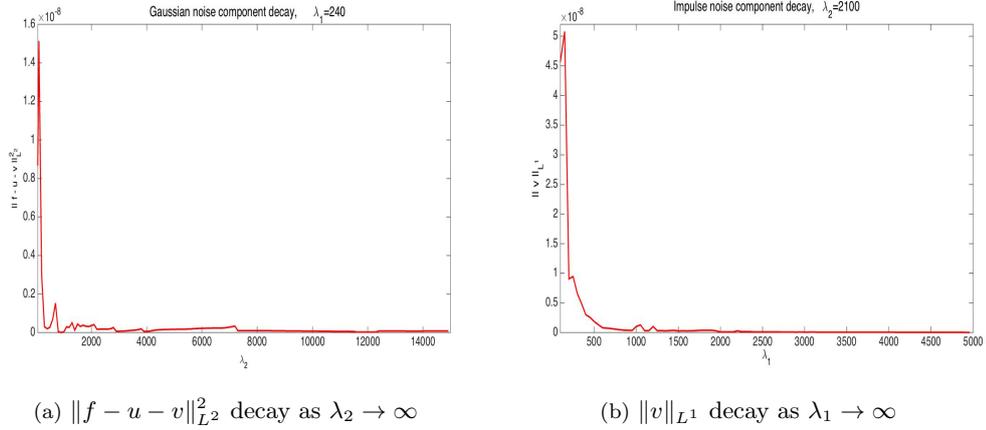


Figure 4.4: Noise components behaviour as parameters λ_1, λ_2 of (4.35) go to infinity.

To have more insight on the type of reconstruction obtained by letting the parameters vary, we compare in Figure 4.5 the solutions computed for the TV-IC model (4.35) for different values of λ_1 and λ_2 . The noisy image considered has been corrupted by a combination of impulse (density of missing pixels $d = 5\%$) and Gaussian noise (with zero mean and $\sigma^2 = 0.005$). For comparisons, we report the denoising results computed with standard denoising models such as TV- L^1 [Nik04, DAG09], TV- L^2 [ROF92] and the TV- L^1 - L^2 combination [HL13]. For reference, we recall the reader the models mentioned:

$$\min_u \{ |Du|_\gamma(\Omega) + \lambda_1 \|f - u\|_{L^1(\Omega)} \} \quad (\text{TV-}L^1)$$

$$\min_u \left\{ |Du|_\gamma(\Omega) + \frac{\lambda_2}{2} \|f - u\|_{L^2(\Omega)}^2 \right\} \quad (\text{TV-}L^2)$$

$$\min_u \left\{ |Du|_\gamma(\Omega) + \lambda_1 \|f - u\|_{L^1(\Omega)} + \frac{\lambda_2}{2} \|f - u\|_{L^2(\Omega)}^2 \right\} \quad (\text{TV-}L^1\text{-}L^2)$$

where, for consistency, we have also Huber-regularised the TV term as in (4.29). As expected from Proposition 4.4.1 and verified numerically in Figure 4.4, we observe that TV- L^1 and TV- L^2 -type solutions can be obtained from (4.35) by considering large weighting parameters λ_2 or λ_1 , respectively. In these situations, we note that only one component of the noise is smoothed, namely the one corresponding to the active (i.e. non-vanishing) fidelity term in the model. Moreover, we observe that the computed solution of the TV-IC model (4.35) is comparable to the one compute using the TV- L^1 - L^2 denoising model, but with the additional property of noise decomposition demonstrated in Figure 4.3 above. Finally, as in *iii*) of Proposition 4.4.1, the noisy image f is recovered by taking large parameters λ_1 and λ_2 .

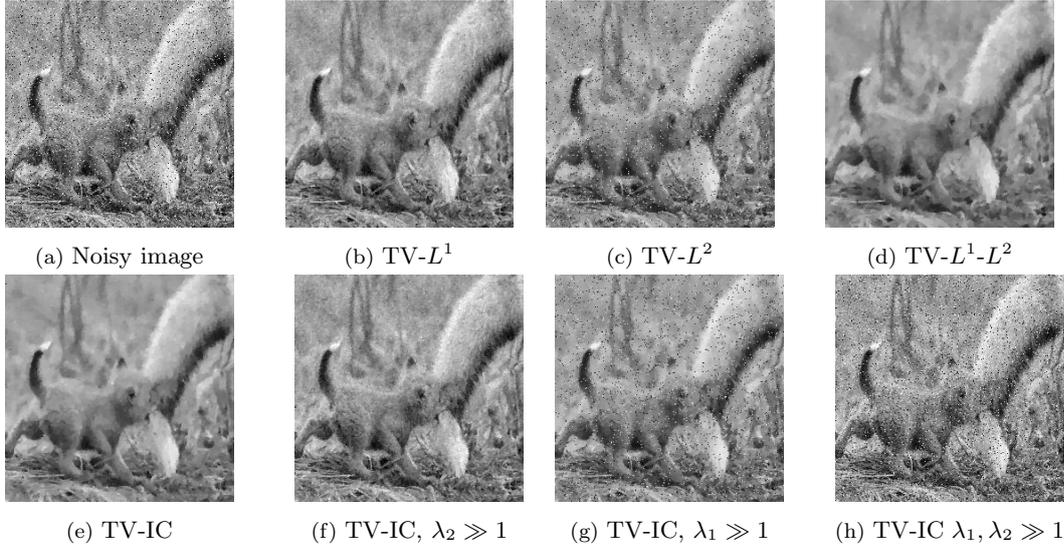


Figure 4.5: Comparison between $(TV-L^1)$, $(TV-L^2)$, $(TV-L^1-L^2)$ and TV-IC (4.35) reconstructions.

First row: (a) noisy image corrupted with impulse noise ($d = 5\%$) and Gaussian noise with zero mean and variance $\sigma^2 = 0.005$, PSNR=17.18 dB. (b) $TV-L^1$ solution with $\lambda_1 = 352$, PSNR=25.53 dB. (c) $TV-L^2$ solution with $\lambda_2 = 2121$, PSNR=20.26 dB. (d) $TV-L^1-L^2$ solution with $\lambda_1 = 351, \lambda_2 = 258$, PSNR=25.62 dB.

Second row: (e) TV-IC solution with $\lambda_1 = 352, \lambda_2 = 2121$, PSNR=26.51 dB (f) TV-IC solution with $\lambda_1 = 352, \lambda_2 = 1e5$, PSNR= 25.61 dB. (g) TV-IC solution with $\lambda_1 = 1e5, \lambda_2 = 2121$, PSNR=20.13 dB. (e) TV-IC solution with $\lambda_1 = \lambda_2 = 1e5$, PSNR=17.17 dB.

4.5.2 Gaussian-Poisson case

For the numerical solution of the Gaussian-Poisson model presented in Section 4.3.2, we relax the unilateral constraints on u and v by adding penalty terms as follows:

$$\min_{u,v} \left\{ J_\gamma(u,v) := |Du|_\gamma(\Omega) + \frac{\lambda_1}{2} \|v\|_{L^2(\Omega)}^2 + \lambda_2 \int_\Omega \left(u - (f-v) + (f-v) \log \left(\frac{f-v}{u} \right) \right) dx + \frac{\eta_1}{2} \|\min(u,0)\|_{L^2(\Omega)}^2 + \frac{\eta_2}{2} \|\min(f-v,0)\|_{L^2(\Omega)}^2 \right\}. \quad (4.37)$$

In the following numerical experiments, we start from initial values $\eta_1^0 = 10$ and $\eta_2^0 = 100$ and increase them throughout the iterations. The optimality conditions for (4.37) read:

$$\begin{aligned} \frac{\partial J_\gamma}{\partial u} &= -\nabla \cdot \left(\frac{\gamma \nabla u}{\max(\gamma |\nabla u|, 1)} \right) + \lambda_2 \left(1 - \frac{f-v}{u} \right) + \eta_1 \chi_{\mathcal{S}_u} u = 0, \\ \frac{\partial J_\gamma}{\partial v} &= \lambda_1 v - \lambda_2 \log \left(\frac{f-v}{u} \right) + \eta_2 \chi_{\mathcal{S}_v} (v-f) = 0 \end{aligned}$$

where $\chi_{\mathcal{J}_u}$ and $\chi_{\mathcal{J}_v}$ are the characteristic functions of the sets $\mathcal{J}_u = \{x \in \Omega : u(x) < 0\}$ and $\mathcal{J}_v = \{x \in \Omega : v(x) < f(x)\}$.

Similarly as before, we express now the system above in primal-dual form and write the modified SSN iteration for the increments $\delta_u, \delta_q, \delta_v$ which reads:

$$\begin{cases} -\nabla \cdot \delta_q + \lambda_2 \left(\frac{f-v}{u^2} \right) \delta_u + \frac{\lambda_2}{u} \delta_v + \eta_1 \chi_{\mathcal{J}_u} \delta_u \\ \qquad \qquad \qquad = - \left(-\nabla \cdot q + \lambda_2 \left(1 - \frac{f-v}{u} \right) + \eta_1 \chi_{\mathcal{J}_u} u \right), \\ \delta_q - \frac{\gamma \nabla \delta_u}{\max(1, \gamma |\nabla u|)} + \chi_{\mathcal{U}_\gamma} \gamma^2 \frac{\nabla u^T \nabla \delta_u}{\max(1, \gamma |\nabla u|)^2} \frac{q}{\max(1, |q|)} = -q + \frac{\gamma \nabla u}{\max(1, \gamma |\nabla u|)}, \\ \lambda_1 \delta_v + \frac{\lambda_2}{u} \delta_u + \frac{\lambda_2}{f-v} \delta_v + \eta_2 \chi_{\mathcal{J}_v} \delta_v \\ \qquad \qquad \qquad = - \left(\lambda_1 v - \lambda_2 \log \left(\frac{f-v}{u} \right) + \eta_2 \chi_{\mathcal{J}_v} (v-f) \right), \end{cases}$$

where, similarly to (4.36), the set \mathcal{U}_γ is defined as $\mathcal{U}_\gamma := \{x \in \Omega : \gamma |\nabla u(x)| > 1\}$.

In Figure 4.6 we report the denoising results for the mixed Gaussian-Poisson model solved via the SSN iteration above. At each pixel (x_i, x_j) of the image domain Ω , the Poisson noise component is distributed with Poisson distribution (4.17) with parameter $u_{ij} = u(x_i, x_j)$, whereas the Gaussian noise component of the examples has zero mean and different intensities (variance) specified in each case. Also in this case we observe that the noise components are decomposed as expected, with the Gaussian one being distributed over the image domain while the Poisson one depending on the intensity of the signal itself in the different points of Ω . In other words, this means that low intensity (darker) areas in the image will be corrupted by a smaller amount of Poisson noise than the high intensity (brighter) regions (compare, for instance the black sky background in the moon image or the synthetic image in the last row), whereas the Gaussian noise component is correctly spread over the whole image domain. As above, we observe some loss of structure in the TV reconstructed image which are encoded in the noise components.

In Figure 4.7 we compare between the reconstructions obtained using single noise TV denoising models with L^2 [ROF92] and KL [Saw11, LCA07] data fidelities, the sum of the two data fidelities above as in [DLRS13, CCDLR⁺15] and the mixed data fidelity terms derived in [JCPT12, JCPT15]. Namely, we compare our method with Huber-regularised versions of the (TV- L^2) model, with

$$\min_u \left\{ |Du|_\gamma(\Omega) + \lambda_2 \int_\Omega (u - f \log u) \, dx \right\}, \quad (\text{TV-KL})$$

$$\min_u \left\{ |Du|_\gamma(\Omega) + \frac{\lambda_1}{2} \|f - u\|_{L^2(\Omega)}^2 + \lambda_2 \int_\Omega (u - f \log u) \, dx \right\}, \quad (\text{TV-}L^2\text{-KL})$$

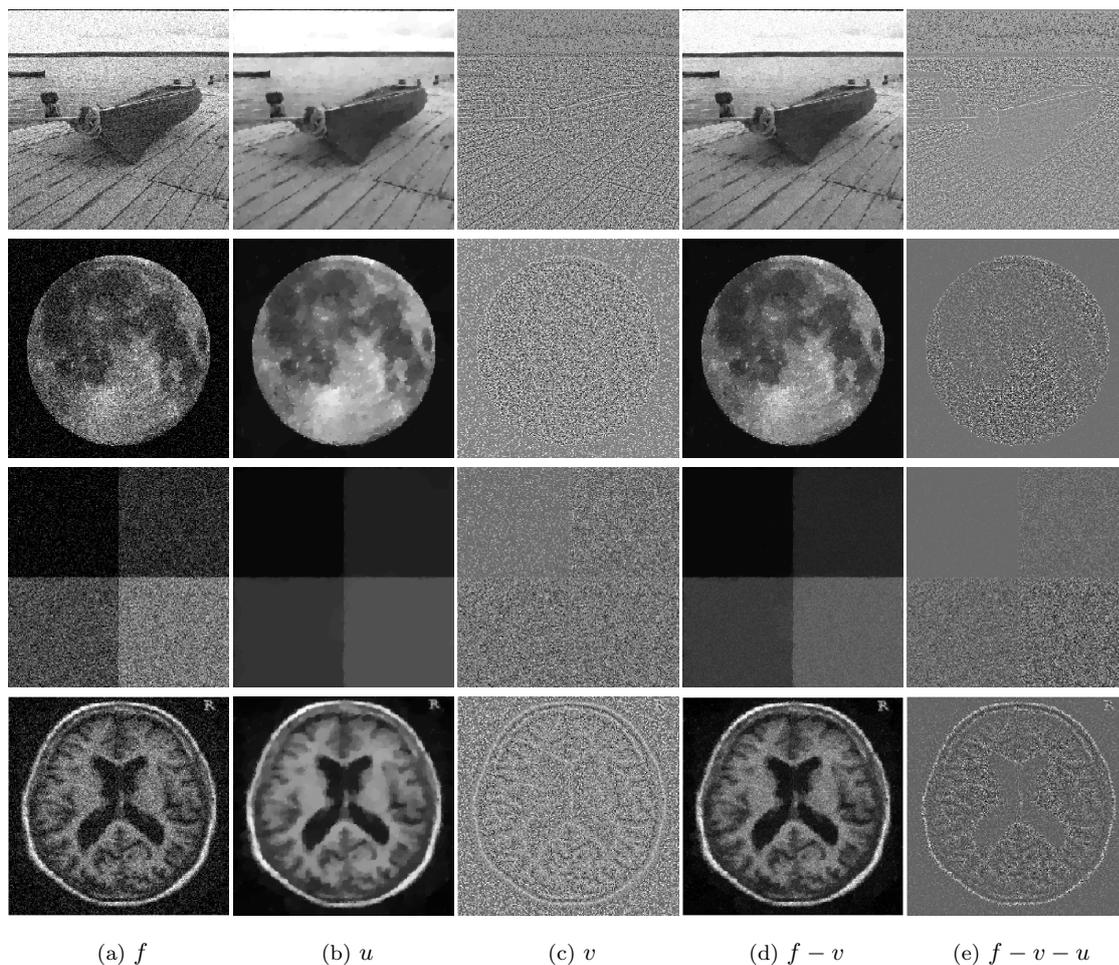


Figure 4.6: **First column:** Images corrupted with Poisson and Gaussian noise. **Second column:** Denoising result. **Third column:** Gaussian noise component. **Fourth column:** Poisson residuum. **Fifth column:** Poisson noise component.

First row: Gaussian noise variance $\sigma^2 = 0.005$. Noisy image PSNR=19.66 dB. Denoised version PSNR=26.03 dB. Parameters: $\lambda_1 = 4289$, $\lambda_2 = 5503$.

Second row: Gaussian noise variance $\sigma^2 = 0.01$. Noisy image PSNR=19.46 dB. Denoised version PSNR=26.19 dB. Parameters: $\lambda_1 = 2903$, $\lambda_2 = 2107$.

Third row: Gaussian noise variance $\sigma^2 = 0.005$. Noisy image PSNR=22.39 dB. Denoised version PSNR=33.04 dB. Parameters: $\lambda_1 = 2105$, $\lambda_2 = 1896$.

Fourth row: Gaussian noise variance $\sigma^2 = 0.05$. Noisy image PSNR=18.62 dB. Denoised version PSNR=23.87 dB. Parameters: $\lambda_1 = 809$, $\lambda_2 = 712$.

$$\min_u \left\{ |Du|_\gamma(\Omega) - \int_\Omega \log \left(\sum_{n=0}^{+\infty} \frac{u^n e^{-u}}{n!} \frac{e^{-\frac{(u-n)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} \right) dx \right\}. \quad (\text{TV-GP})$$

Such comparisons highlight another feature of the TV-IC model solved via SSN primal-dual algorithm (4.37) in comparison with the ones listed above. Numerical algorithms solving TV models for single or mixed Poisson noise removal often suffer of the so-called “checker-

board effect”, i.e. single pixels become visible along the iterative reconstruction process, cf. [BLCT⁺08, Saw11, SBMB09]. In our case, this is clearly visible when using (TV-KL) and (TV- L^2 -KL) models, compare Figure 4.7c and Figure 4.7d. Such unpleasant artefact appears reduced in the TV-IC reconstruction in Figure 4.7f. Note that such reconstruction is very similar to the TV- L^2 one in Figure 4.7b: this is not surprising due to the similarity between Poisson and Gaussian noise for high intensity values in the image (i.e. brighter areas).

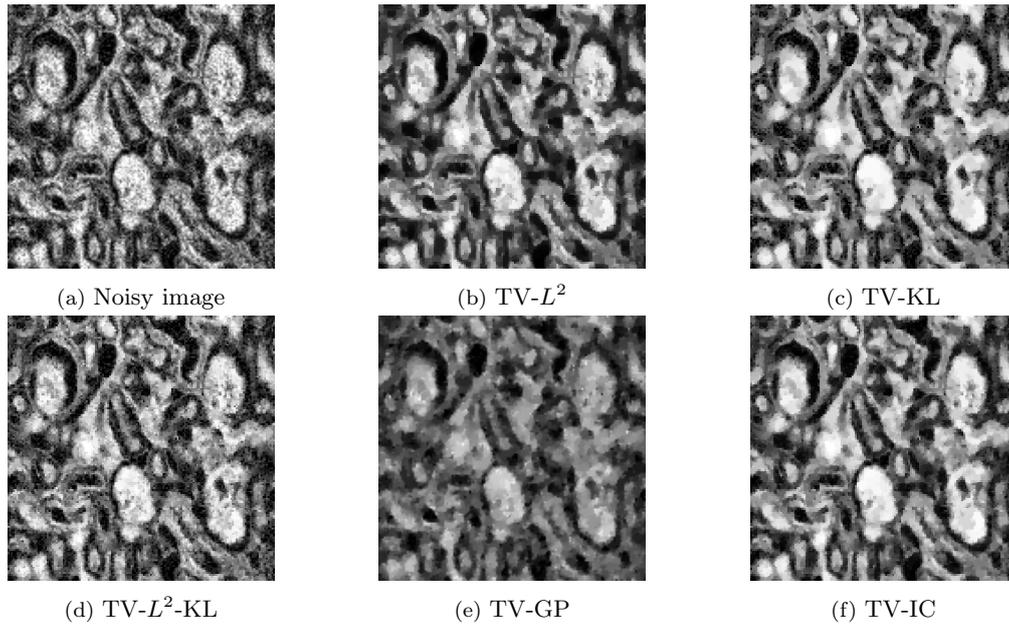


Figure 4.7: Comparison between solutions of TV-IC model (4.6) and solutions of (TV- L^2), (TV-KL), (TV- L^2 -KL) and (TV-GP) models.

First row: (a) noisy image corrupted with Gaussian noise with zero mean and variance $\sigma^2 = 0.005$ and Poisson noise with parameter u , PSNR=18.81 dB. (b) TV- L^2 solution with $\lambda_1 = 1800$, PSNR=22.97 dB. (c) TV-KL solution with $\lambda_2 = 1200$, PSNR=21.19 dB. **Second row:** (d) TV- L^2 -KL solution with $\lambda_1 = 520$, $\lambda_2 = 1100$. PSNR=22.24 dB. (e) Solution of the TV-GP model (TV-GP) model. PSNR=19.65 dB. (f) Solution of TV-IC model (4.37) with $\lambda_1 = 1200$, $\lambda_2 = 1800$, PSNR=22.52 dB.

In Figure 4.8 we report the results of the TV-IC model for large values of the parameters λ_1 and λ_2 to show how the reconstruction of the noisy image changes for different choices of the parameters. As shown in Section 4.4.2 such choices correspond to enforce the (TV-KL) and the (TV- L^2 -KL), respectively.

To conclude, we finally report in the following Figure 4.9 some numerical tests on the asymptotical behaviour of the model (4.37) as the fidelity weights go to infinity. As shown in Proposition 4.4.3, both the TV- L^2 model for Gaussian noise removal (TV- L^2) (whose solution is denoted by u_{TV-L^2}) and the TV-KL one (TV-KL) (whose solution is denoted by u_{TV-KL}) are recovered asymptotically under appropriate norms.

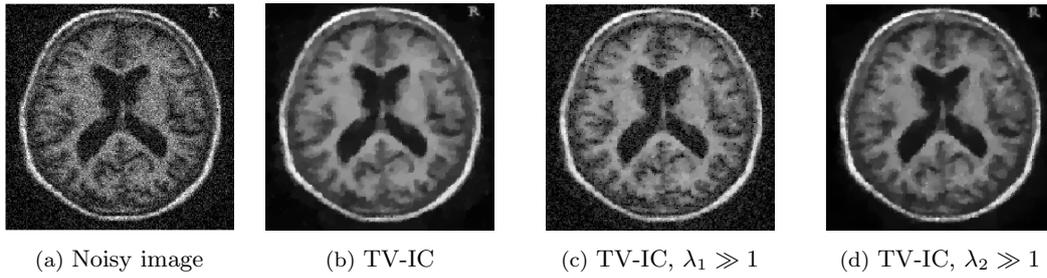


Figure 4.8: Comparison between solutions of TV-IC model (4.6) for large values of the parameters λ_1 and λ_2 .

(a) Noisy image corrupted with Gaussian noise with zero mean and variance $\sigma^2 = 0.05$ and Poisson noise, PSNR=18.62 dB. (b) TV-IC solution with $\lambda_1 = 809$, $\lambda_2 = 712$, PSNR=23.87 dB. (c) TV-IC solution with large λ_1 , $\lambda_2 = 712$, PSNR=19.83 dB. (d) TV-IC solution with large λ_2 , $\lambda_1 = 809$, PSNR=22.71 dB.

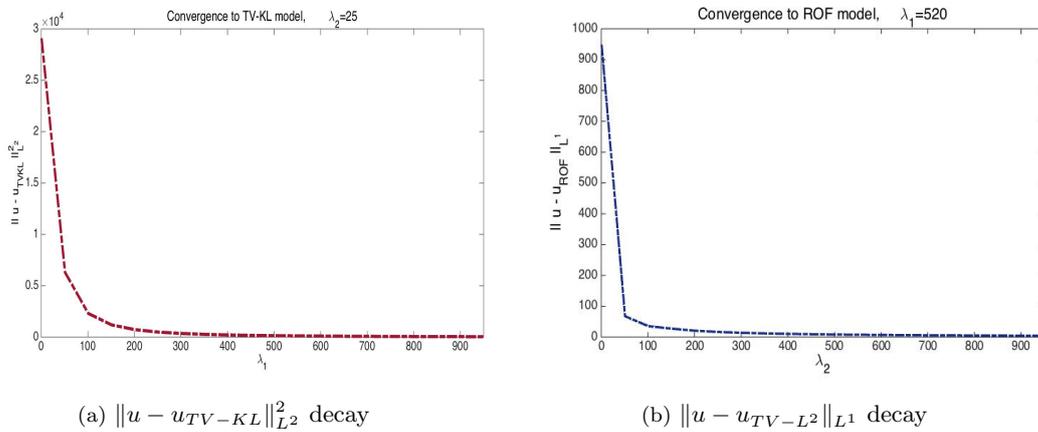


Figure 4.9: Convergence to single noise models as parameters λ_1 , λ_2 of (4.37) go to infinity.

Acknowledgement. The author would like to thank Dr. Anna Jezierska, Department of Modelling and Optimization of Dynamical Systems, Polish Academy of Science, Warsaw, Poland for providing him with the image reported in Figure 4.7e and used for comparison.

Chapter 5

Learning the optimal setup by bilevel optimisation

In the previous chapters of this thesis we focused on two different aspects of the modelling of the minimisation problem

$$\min_u R(u) + \lambda \Phi(Tu, f) \tag{5.1}$$

of the reconstruction of an imperfect image f . Namely, in Chapter 3 we described advantages and numerical challenges in the use of higher-order regularisation terms R encoding *a priori* knowledge on the reconstruction we seek. Conversely, in Chapter 4 we focused on the appropriate choice of the data fidelity term Φ for some specific denoising applications.

Depending on the problem at hand, a tailored choice of the regulariser and of the data fidelity is often very hard. Furthermore, the selection of the optimal balance between the fitting of the data and the strength of the regularisation (encoded in the weighting parameter λ) is a very challenging problem itself and often results in a trial and error-type approach. Finally, an appropriate modelling of the forward operator T is also crucial for the specific application considered.

In this chapter we look at (5.1) for the particular case of image denoising (thus setting T to be the identity operator) from a different point of view. Instead of sticking with *one* particular regularising term R and *one* particular data fidelity term Φ balanced by a *fixed* parameter λ , we review and examine the mathematical contributions in the area of *learning* the ingredients R , Φ and λ for the design of an *optimal* model. Our discussion starts with an overview of the mathematical state-of-the-art methods dealing with this problem and focuses next on the recent developments in the field of nonsmooth bilevel optimisation.

5.1 A Review of learning methods in imaging

In the literature, several approaches have been considered for the optimal design of (5.1). All the strategies vary in their philosophy and mathematics. Nonetheless, we can identify two main attitudes to this problem.

5.1.1 Learning from the model

One approach consists in encoding *a priori* information in the ingredients R, Φ, T and λ of the model we intend to use. This reflects, for instance, in the particular choice of the regularisation term R which, as previously discussed, shall take into account the geometrical structure of the reconstruction we seek, such as the predominance of piecewise constant/linear regions or the presence of objects with a *a priori* known specific geometrical shape, see [CCN07, All08a, PS14, BB13, BKV13, BBBM13], only to name a few. Similarly, a prior knowledge on how and for which type of application the data have been acquired, shall drive the choice of the forward operator T and of the data fidelity term Φ which are typically chosen to be consistent with the statistics of the data, such as the type of noise and blur, and their intensities, see, e.g., [Stu10, VSK85, BB11, HMBS10]. This is the approach we have used in the previous Chapter 3 and Chapter 4, considering suitable tailor-made regularising terms and data fidelities for solving image inpainting and image denoising problems. We refer the reader to these chapters for a more complete list of references.

Despite representing a rigorous and reliable approach in terms of reconstruction guarantees and error and stability estimates, model-driven formulations of (5.1) shows typically little flexibility for real-world applications due to their specificity with respect to the problems considered.

5.1.2 Learning from the data

A different approach consists in learning the ingredients of the model (5.1) *a posteriori*, i.e. from the data itself. This is the case, for instance, of [HT03] where the regularisation functionals are learned from training models. Analogously, in [BTWG08, KTAK11] constrained optimisation problems are considered in large parameter spaces and marginalisation techniques are used for reducing the large amount of data. Dictionary approaches have also been used in the framework of sparse coding in [OF96, MBPS09, GSM10, PF11] and Markov random fields have been used for learning the optimal image prior distributions in the context of Bayesian approaches in [RB05, Tap07].

Data-driven approaches aim to trim the different ingredients in (5.1) in order to best-fit the given data, often at the price of the lack of theoretical justifications and guarantees. In general, they do not offer insights into the structural properties of the solutions of the selected model and stability and error estimates are typically hard to prove.

5.1.3 Bilevel optimisation approaches in function spaces

The idea is then to consider a unified approach combining the pros of *a priori* model-driven and *a posteriori* data-driven modelling to develop a reliable, intelligible and effective model (5.1). One possibility to couple these two approaches is to use a bilevel optimisation strategy where the optimal setup is determined by minimising a suitable cost functional F assessing the quality of the candidate model computed within the setup considered by comparison with a training set of examples serving as ground truths. In other words, bilevel optimisation methods can be seen as semi-supervised learning methods that select and tune adaptively the optimal modelling terms R , Φ and λ by a comparison with a given dataset of measurements and desirable solutions.

In the inverse problem community, optimal model design via bilevel optimisation has been considered, for instance, by Haber and Tenorio in [HT03, HHT10] as well as by Ghattas et al. in [BTWG08, BBG⁺11]. In the context of variational regularisation models like (5.1), recent contributions focused on optimal parameter learning have been considered by Pock et al. in [KP13, CTB12, CYP15], Chung et al. [CEN14] and by Schimdt and Roth in [SR14] for proximal mappings learning. More recent works in the area of bilevel learning are also [RP14, ORBP15] for image segmentation and [KP15] for support vector machines.

5.2 The General learning model

In the following we focus on the approach proposed firstly by Schönlieb and De Los Reyes in [DLRS13] for noise learning via bilevel optimisation. We report the main results and methods used by the author of the thesis in [CDLRS14] for the design of numerical efficient methods in the case of large datasets and in [CCDLR⁺15] where a review on bilevel learning is given by the author in collaboration with Chung, De Los Reyes, Schönlieb and Valkonen. For completeness, we additionally report some recent results regarding the optimal choice of the regularising terms in this framework reported in [DLRSV15b]. As mentioned in Section 1.1.2 of the Introduction, our discussion is set up in infinite dimensional function spaces and the optimisation approach is consequently studied within an *optimise-then-discretise* framework. Our aim is to preserve qualitative properties in images independently of their resolution [VFC12]. Numerically, this reflects numerically the preservation of the desirable properties of resolution and mesh-independence upon discretisation [HS06].

We start following [DLRSV15b] where the bilevel learning problem is presented in a general and abstract functional framework. We state under which conditions of the variational model considered existence of solutions is guaranteed and an optimality system can be derived. Our attention focuses on the problem of image denoising, i.e. $T = Id$ in

(5.1), but generalisations for non-trivial blurring operators T can be considered as well.

Given the noisy data $f \in L^2(\Omega)$ and an abstract function space X , we look for positive parameters $\alpha = (\alpha_1, \dots, \alpha_m)$ and $\lambda = (\lambda_1, \dots, \lambda_d)$ in abstract parameter sets \mathcal{P}_α^+ and \mathcal{P}_λ^+ solving for some convex, proper, weakly* lower semicontinuous cost functional $\mathcal{F} : X \times \mathcal{P}_\alpha^+ \times \mathcal{P}_\lambda^+ \rightarrow \mathbb{R}$ the problem

$$\min_{\alpha \in \mathcal{P}_\alpha^+, \lambda \in \mathcal{P}_\lambda^+} \mathcal{F}(u^{\alpha, \lambda}; \alpha, \lambda) \quad \text{s.t.} \quad u^{\alpha, \lambda} \in \arg \min_{u \in X} J(u; \alpha, \lambda) \quad (5.2a)$$

with

$$J(u; \alpha, \lambda) := \sum_{i=1}^m \int_{\Omega} \alpha_i(x) d|A_i u|(x) + \sum_{j=1}^d \int_{\Omega} \lambda_j(x) \phi_j(u, f) dx. \quad (5.2b)$$

Here, A_i , $i = 1, \dots, m$ are linear operators acting on u which are penalised in the total variation norm and, as such, act as image regularisers, whereas ϕ_j , $j = 1, \dots, d$ are data fitting terms measuring in some sense the distance between the given data model f and the reconstructed image u .

In order to guarantee well-posedness results, the problem (5.2) needs to be regularised. Namely, we consider in the following both a Huber regularisation of the Radon norms similarly as in (2.11) and (4.29) and a quadratic smoothing functional $H : X \rightarrow \mathbb{R}_+$ defined by $H(u) := \frac{1}{2} \|\nabla u\|^2$.

Therefore, for regularising parameters $0 < \varepsilon \ll 1$ and $\gamma \gg 1$, we consider the following regularised problem:

$$\min_{\alpha \in \mathcal{P}_\alpha^+, \lambda \in \mathcal{P}_\lambda^+} F(u^{\alpha, \lambda, \varepsilon, \gamma}; \alpha, \lambda) \quad \text{s.t.} \quad u^{\alpha, \lambda, \varepsilon, \gamma} \in \arg \min_{u \in X} J^{\varepsilon, \gamma}(u; \alpha, \lambda) \quad (5.3a)$$

with

$$J^{\varepsilon, \gamma}(u; \alpha, \lambda) := \frac{\varepsilon}{2} \|\nabla u\|^2 + \sum_{i=1}^m \int_{\Omega} \alpha_i(x) d|A_i u|_\gamma(x) + \sum_{j=1}^d \int_{\Omega} \lambda_j(x) \phi_j(u, f) dx, \quad (5.3b)$$

where we have denoted by $|A_i u|_\gamma$ the Huber-regularisation of the total variation measure of $A_i u$ for every $i = 1, \dots, m$ as defined in (4.29). For what follows, we interpret the choice $\gamma = \infty$ as corresponding to the non-Huber-regularised version of problem (5.3), i.e. to (5.2) with the sole addition of the elliptic term $H(u)$ defined above.

5.2.1 Assessing the quality: the choice of the cost functional

One crucial aspect in this framework is the choice of the cost functional \mathcal{F} appearing in (5.2a) which, heuristically, assesses the optimality of the reconstruction computed in the lower level problem (5.2b) with respect to exemplar training data. A classical choice for

\mathcal{F} is the standard L^2 -squared cost functional

$$\mathcal{F}_{L^2}(u^{\alpha,\lambda}; \alpha, \lambda) := \frac{1}{2} \|f^0 - u^{\alpha,\lambda}\|_{L^2(\Omega)}^2, \quad (5.4)$$

where $f^0 \in L^2(\Omega)$ is the training noise-free version of the given noisy image f . In the following we will mostly make use of such cost functional, which, from an imaging point of view, appears appropriate since minimising (5.4) corresponds to maximising the Signal to Noise Ratio (SNR), a standard quality measure commonly used in imaging applications defined by:

$$SNR(u^{\alpha,\lambda}, f^0) = 20 \log_{10} \left(\frac{\|f^0\|_{L^2(\Omega)}}{\|u^{\alpha,\lambda} - f^0\|_{L^2(\Omega)}} \right). \quad (5.5)$$

Remark 5.2.1. *Alternatively, other quality measures can be used. For instance, in [DLRSV15b, DLRSV15a] the Huberised- L^1 cost functional*

$$\mathcal{F}_{L^1_\gamma \nabla}(u^{\alpha,\lambda}; \alpha, \lambda) := \|D(f^0 - u^{\alpha,\lambda})\|_{L^1(\Omega), \gamma},$$

is used. Here $\|Dw\|_{L^1(\Omega), \gamma} := \int_{\Omega} |Dw|_{\gamma} dx$ and $|Dw|_{\gamma}$ is defined similarly as in (4.29). While this choice introduces some additional difficulties in the analysis of (5.2), such as, for instance, the derivation of ‘classical’ optimality conditions, quality assessments with respect to $\mathcal{F}_{L^1_\gamma \nabla}$ show more visually pleasant results than the ones computed using \mathcal{F}_{L^2} when used for the estimation of the optimal regularisation parameters α_i , $i = 1, \dots, m$, compare [DLRSV15b].

5.3 Learning the image model

One crucial choice in the design of the variational model (5.1) for image reconstruction problems is the regularisation that should be imposed on the image. Typically, a model-driven approach aims to calibrate qualitatively and quantitatively the type and amount of smoothing needed in order to counteract the noise and blur in the data. In the mathematical framework (5.2) this corresponds to choose, for $i = 1, \dots, m$ appropriate image regularisers A_i and the regularisation weights α_i .

In our exposition we have focused in particular on TV-type regularisations. However, we have discussed in the Introduction limits and drawbacks of such models when applied to solve denoising and inpainting tasks and mentioned how higher-order regularisation models can be used to counteract these issues.

5.3.1 ICTV and TGV regularisations

Among the myriad of extensions of TV-based models proposed in the literature, we recall here the Infimal-Convolution Total Variation (ICTV) regularisation model proposed

by Chambolle and Lions in [CL97] and the Total Generalised Variation (TGV) model introduced by Bredies, Kunisch and Pock in [BKP10]. In the following, we give a condensed account of the results presented by Schönlieb, De Los Reyes and Valkonen in [DLRSV15a, DLRSV15b] where these regularisations have been considered as special cases of A_i operators in (5.2) and their strength, encoded in the size of the regularising parameters α_i , has been learned by means of bilevel optimisation techniques.

The ICTV regulariser reads

$$\text{ICTV}_{\alpha,\beta}(u) := \min_{v \in W^{1,1}(\Omega), \nabla v \in \text{BV}(\Omega)} \alpha \|Du - \nabla v\|_{\mathcal{M}(\Omega; \mathbb{R}^2)} + \beta \|D\nabla v\|_{\mathcal{M}(\Omega; \mathbb{R}^{2 \times 2})}. \quad (5.6)$$

The TGV regulariser can be written as the following differential cascade [BV11, VBK13]

$$\text{TGV}_{\alpha,\beta}^2(u) := \min_{w \in \text{BD}(\Omega)} \alpha \|Du - w\|_{\mathcal{M}(\Omega; \mathbb{R}^2)} + \beta \|\mathcal{E}w\|_{\mathcal{M}(\Omega; \text{Sym}^2(\mathbb{R}^2))}. \quad (5.7)$$

Here, the space $\text{BD}(\Omega) := \{w \in L^1(\Omega; \mathbb{R}^2) : \|\mathcal{E}w\|_{\mathcal{M}(\Omega; \mathbb{R}^{2 \times 2})} < \infty\}$ is the space of vector fields of bounded deformation on Ω (see, e.g., [TS80]), \mathcal{E} denotes the *symmetrised gradient* of w (formally defined as $\mathcal{E}w = (Dw + (Dw)^\top)/2$) and $\text{Sym}^2(\mathbb{R}^2)$ is the space of symmetric tensors of order 2 with arguments in \mathbb{R}^2 . In both (5.6) and (5.7) the regularising terms are weighted by the fixed scalar positive parameters α and β . Despite their similar definition, the main difference between (5.6) and (5.7) is the fact that in general there is no guarantee that $w = \nabla v$ for every function v . Consequently, ICTV and TGV regularisations may differ significantly, see for instance [BBBM13]. Nonetheless, for both models, the amount of smoothing heavily depends on the choice of the regularisation parameters (α, β) and their choice can lead to very different results, compare Figure 5.1.

In the following we briefly report some results presented in [DLRSV15b, DLRSV15a] for the estimation of the optimal parameters α and β for TV and TGV type reconstruction models by bilevel optimisation. Since the thorough analysis of the learning model for the estimation of the optimal regulariser is beyond the purpose of this chapter and the research studies of the author of this thesis, we limit ourselves to consider the case where single-Gaussian noise is present, i.e. $\mathcal{P}_\lambda^+ = \{1\}$ and $\lambda = 1$ in the framework above. Furthermore, we consider the regularised problem (5.3) and assess optimality using the L^2 -squared cost (5.4) using a single training pair (f_0, f) of noise-free and noisy image, respectively. In other words, we are interested in finding optimal parameters (α, β) for the variational regularisation model

$$u^{\alpha,\beta} \in \arg \min_{u \in X} R_{\alpha,\beta}(u) + \|f - u\|_{L^2(\Omega)}^2,$$

where f is the noisy image, $R_{(\alpha,\beta)}$ is either TV multiplied by α (without any β), $\text{ICTV}_{\alpha,\beta}$ in (5.6) or $\text{TGV}_{\alpha,\beta}^2$ in (5.7). The function space X is chosen appropriately depending on the regulariser considered.

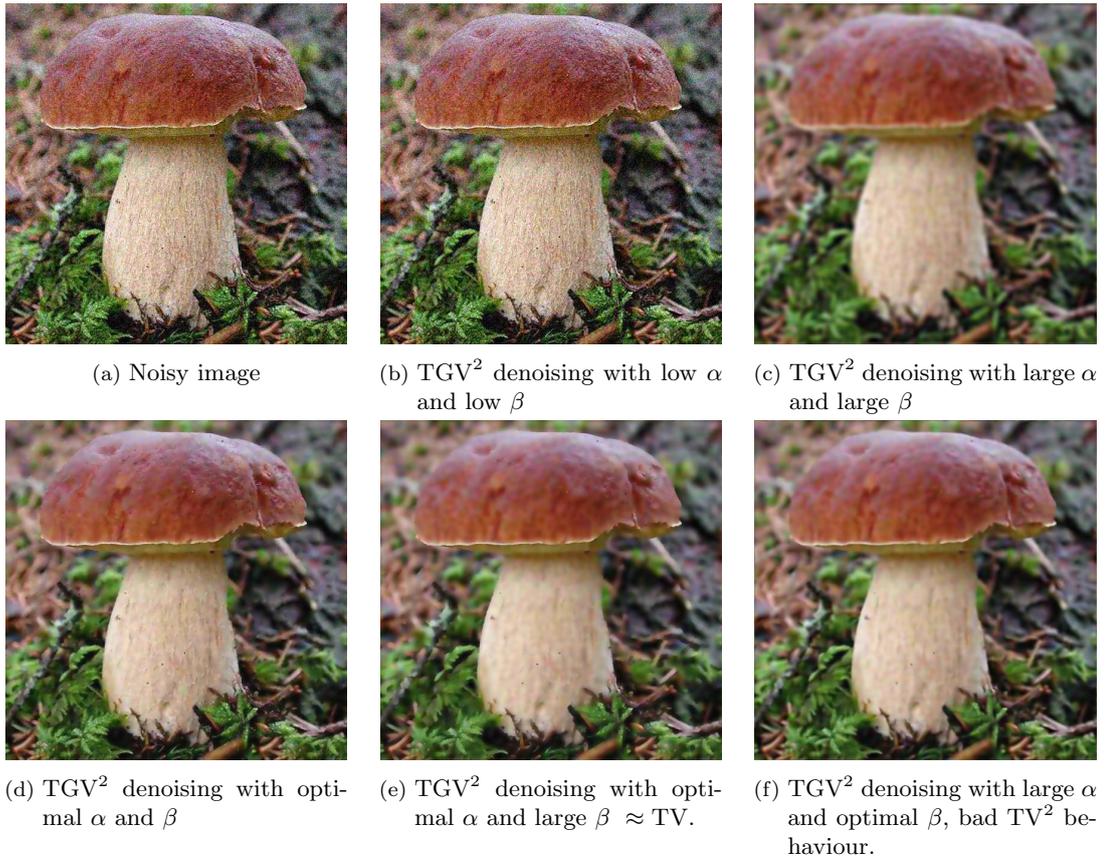


Figure 5.1: The effect of the different choice of the parameters α and β in TGV^2 regularisation (5.7). The image has been corrupted by Gaussian noise with zero mean and variance $\sigma^2 = 0.01$.

We state in the following two results stated and proved in [DLRSV15b] on the existence and the structural properties of optimal regularisation of the TV and TGV^2 models. These results have to be intended as follows. Having fixed one of the two regularisations and under natural assumptions on the given training pair (f_0, f) , there exists a range of regularised scenarios, i.e. regularised versions of the problem (5.2) of the form (5.3) such that in this range an optimal pair of positive parameters (α, β) can be found. Despite being intrinsically interesting, these results are useful also from a computational point of view since they allow to avoid any additional box constraint on the regularising parameters we seek. Furthermore, they do not show existence only for a particular regularised version of (5.2) (i.e. for fixed ε and γ), but for an entire range of scenarios. This property is desirable from a numerical point of view since such property avoids the tedious tuning of the regularising parameters ε and γ .

We start stating the result of existence and structure of optimal regularisation in the case of total variation regularisation. We refer the reader to [DLRSV15b, Corollary 3.1] for its proof and more details.

Theorem 5.3.1 (Total variation Gaussian denoising [DLRSV15b]). *Suppose that the training images f, f_0 are elements in the space $X = \text{BV}(\Omega) \cap L^2(\Omega)$ and that*

$$\text{TV}(f) > \text{TV}(f_0). \quad (5.8)$$

Then, there exist $\bar{\varepsilon}, \bar{\gamma} > 0$ such that any optimal solution $\alpha_{\varepsilon, \gamma} \in [0, \infty]$ to the problem

$$\min_{\alpha \in [0, \infty]} \frac{1}{2} \|f_0 - u^{\alpha, \varepsilon, \gamma}\|_{L^2(\Omega)}^2$$

subject to

$$u^{\alpha, \varepsilon, \gamma} \in \arg \min_{u \in \text{BV}(\Omega)} \left(\frac{\varepsilon}{2} \|\nabla u\|_{L^2(\Omega; \mathbb{R}^2)}^2 + \alpha |Du|_{\gamma}(\Omega) + \frac{1}{2} \|f - u\|_{L^2(\Omega)}^2 \right)$$

satisfies $\alpha_{\varepsilon, \gamma} > 0$ whenever $\varepsilon \in [0, \bar{\varepsilon}]$, $\gamma \in [\bar{\gamma}, \infty]$.

Assumption (5.8) means that, in terms of the total variation, the noisy training image f should present more oscillations than the noise-free image f_0 , as naturally expected.

Under similar assumptions, a similar result can be proved for higher-order regularisations, like, for instance TGV^2 . Again, we refer the reader to [DLRSV15b, Corollary 3.2] for more details.

Theorem 5.3.2 (TGV^2 Gaussian denoising [DLRSV15b]). *Suppose that the training images f, f_0 are elements in the space $L^2(\Omega) \cap \text{BV}(\Omega)$ and that for some $\beta > 0$ the condition*

$$\text{TGV}_{1, \beta}^2(f) > \text{TGV}_{1, \beta}^2(f_0)$$

is satisfied. Then, there exist $\bar{\varepsilon}, \bar{\gamma} > 0$ such that any optimal solution $(\alpha_{\varepsilon, \gamma}, \beta_{\varepsilon, \gamma})$ to the problem

$$\min_{\alpha, \beta \in [0, \infty]} \frac{1}{2} \|f_0 - u^{\alpha, \varepsilon, \gamma}\|_{L^2(\Omega)}^2$$

subject to

$$\begin{aligned} (u^{\alpha, \varepsilon, \gamma}, w^{\alpha, \varepsilon, \gamma}) \in \arg \min_{\substack{u \in \text{BV}(\Omega) \\ w \in \text{BD}(\Omega)}} & \left(\frac{\varepsilon}{2} \|(\nabla u, \nabla w)\|_{L^2(\Omega; \mathbb{R}^2 \times \mathbb{R}^{2 \times 2})}^2 + \alpha |Du - w|_{\gamma}(\Omega) + \beta |\mathcal{E}w|_{\gamma}(\Omega) \right. \\ & \left. + \frac{1}{2} \|f - u\|_{L^2(\Omega)}^2 \right) \end{aligned}$$

satisfies $\alpha_{\varepsilon, \gamma}, \beta_{\varepsilon, \gamma} > 0$ whenever $\varepsilon \in [0, \bar{\varepsilon}]$ and $\gamma \in [\bar{\gamma}, \infty]$.

Furthermore, in [DLRSV15b, Theorem 3.1] the authors prove that the optimal solutions of the regularised models above converge to solutions of the nonsmooth case as the elliptic regularisation vanishes and the Huber regularisation becomes infinitely large, which turns out to be an important property for the numerical realisation of the model.

5.4 Learning the data model

Conversely, we focus now on the choice of the optimal data model. For this sake, we fix the regularisation term to be TV. In (5.2), this corresponds to set $m = 1$, $\mathcal{P}_\alpha^+ = \{1\}$ and to choose the operator Au to be Du , with $d > 1$. In other words, we consider a TV denoising model featuring multiple noise distributions in the data where, for every $i = 1, \dots, d$, the parameter λ_i weights the fitting of the reconstructed image with respect to the data model ϕ_i . Our objective is to compute the optimal λ_i 's. Firstly introduced by Schönlieb and De Los Reyes in [DLRS13], this framework has been considered by the author of the thesis in [CDLRS14].

Motivated by our original purpose of learning by examples, the general model (5.2) is extended here to a training set of pairs of noise-free and noisy images $\{(f_k^0, f_k)\}_{k=1}^N$. Mathematically, this translates in considering (5.2a) subject to a set of N nonsmooth constraints of the type (5.2b). A L^2 -squared cost functional is considered and, similarly as in (5.3), an elliptic-Huber regularised version of the lower-level problem of the form (5.3b) is used.

In the case of scalar parameters λ_i 's the resulting problem reads:

$$\min_{\lambda_i \geq 0, i=1, \dots, d} \frac{1}{2N} \sum_{k=1}^N \|u_k^{\lambda, \varepsilon, \gamma} - f_k^0\|_{L^2(\Omega)}^2 \quad (5.9a)$$

subject to the set of nonsmooth constraints:

$$u_k^{\lambda, \varepsilon, \gamma} = \arg \min_{u \in H_0^1(\Omega) \cap \mathcal{A}} \left(\frac{\varepsilon}{2} \|\nabla u\|_{L^2(\Omega)}^2 + |Du|_\gamma(\Omega) + \sum_{i=1}^d \int_{\Omega} \lambda_i \phi_i(u, f_k) dx \right), \quad k = 1, \dots, N. \quad (5.9b)$$

The space \mathcal{A} is the admissible set for the ϕ_i 's to be well defined. More details on its choice are given later on. We refer the reader to Chapter 4 for classical choices of the data fitting terms ϕ_i 's. We look for solutions in a subspace of the space $H_0^1(\Omega)$, similarly as in [HK04, DLRS13]. Alternatively, it is possible to work in the space $H^1(\Omega)$ and consider alternative boundary conditions. The positive parameters λ_i , $i = 1, \dots, d$ weight the different noise models ϕ 's against each other and balance the fitting of the data against the TV regularisation. In most applications the parameters λ_i 's are chosen to be positive scalar parameters, but in the literature spatially dependent parameters whose structure adapts to the underlying structures and to the presence of inhomogeneous noise have been considered as well, see for instance [BCRS03, ABCH08, DHRC11, FMM12, BDH13, CDLR]. If $d > 1$, the model accommodates multiple noise distributions. In this case, the lower level problem (5.9b) can be considered as an image denoising problem where no prior knowledge of the

noise corrupting the given noisy images f_k is used and the optimal noise model fitting the data is learned directly from the training set. In the case when the noise level is known, classical techniques for choosing the optimal parameters are based on the use of some type of discrepancy principles [Mor66, TSC13] depending on the noise level, cross validation [Wah77] or the L-curve [EHN96, CMRS00]. We refer the reader to the Introduction of the thesis, where more details are given.

Assumptions on data fidelities. For the well-posedness results of the regularised bilevel problem (5.9), we assume that the functions $\phi_i, i = 1, \dots, d$ are convex functions in u , bounded from below and satisfy for every $i = 1, \dots, d$ the following coercivity assumption:

$$\int_{\Omega} \phi_i(u, f) \, dx \geq C_1 \|u\|_{L^p(\Omega)}^p - C_2, \quad \text{for all } u \in L^p(\Omega) \cap \mathcal{A}, \quad (5.10)$$

where $C_1, C_2 \geq 0$ and either $p = 1$ or $p = 2$. Standard noise models fulfil this assumption. In particular we have:

- For the Gaussian noise model (see, e.g., [ROF92]) we have: $\phi(u, f) = |u - f|^2$. Therefore, (5.10) is trivially fulfilled for $\mathcal{A} = L^2(\Omega)$ and $p = 2$.
- For the impulse noise model (see, e.g., [Nik04]), we have $\phi(u, f) = |u - f|$. Condition (5.10) is fulfilled for $\mathcal{A} = L^1(\Omega)$ and $p = 1$.
- For the Poisson impulse model (see [SBMB09]) we have $\phi(u, f) = u - f \log u$. The admissible set is $\mathcal{A} = \{u \in L^1(\Omega), \log u \in L^1(\Omega)\}$. In particular, functions in this space satisfy $u \geq 0$ a.e. in Ω . With this choice, ϕ is convex in u and fulfils (5.10) for $p = 1$, up to an additive constant under the (non-restrictive) assumption that $f \in L^\infty(\Omega)$, see [Bru10]. More precisely, for $u \in \mathcal{A}$ we have:

$$\int_{\Omega} (u - f \log(u + C)) \, dx \geq \|u\|_{L^1(\Omega)} - \|f\|_{L^\infty(\Omega)} \log \|u + C\|_{L^1(\Omega)},$$

for $C > 0$, by Jensen's inequality.

Remark 5.4.1 (Nonsmooth problem in $BV(\Omega)$). *Under the same assumptions on the data fidelity terms ϕ_i 's, in [DLRS13, Theorem 2.2] the authors show the well-posedness of (5.9) in the space $BV(\Omega) \cap \mathcal{A}$ by relaxation. In particular the TV energy is relaxed by Huber-regularisation and no additional H^1 smoothing is introduced. However, the lack of Hilbert space structure does not guarantee the continuity of the solution map $\lambda \mapsto u(\lambda)$ as the Huber regularisation parameter γ goes to infinity. From a numerical point of view this is a serious drawback: derivative-based optimisation techniques solving regularised (i.e. with unique gradient) TV problems require to converge to the solution of the nonsmooth problem as the regularising parameter goes to infinity and this is not a priori guaranteed without any further regularisation.*

5.4.1 Theoretical results

We focus now on the regularised lower level problem (5.9b) and state the following well-posedness result.

Proposition 5.4.2. *For every $k = 1, \dots, N$ there exists a unique solution $u_k^{\lambda, \varepsilon, \gamma}$ of the problem (5.9b). Moreover, the sequence of regularised solutions $(u_k^{\lambda, \varepsilon, \gamma})_\gamma$ converges strongly in $H_0^1(\Omega)$ to $u_k^{\lambda, \varepsilon}$, solution of:*

$$u_k^{\lambda, \varepsilon} = \arg \min_{u \in H_0^1(\Omega) \cap \mathcal{A}} \left(\frac{\varepsilon}{2} \|\nabla u\|_{L^2(\Omega)}^2 + |Du|(\Omega) + \sum_{i=1}^d \int_{\Omega} \lambda_i \phi_i(u, f_k) dx \right), \quad (5.11)$$

as $\gamma \rightarrow \infty$, for every $k = 1, \dots, N$.

Proof. For every $k = 1, \dots, N$ the result can be proved following [DLR11, Theorem 4.2] where maximal monotonicity and coercivity arguments are used together with standard energy estimates. \square

In order to ensure the existence of the bilevel problem (5.9), we follow [DLRS13] and add a Tikhonov-type regularisation to the cost functional in (5.9a). In particular, we allow the parameter vector $\lambda = (\lambda_1, \dots, \lambda_d)$ to be an element of a Hilbert space X^d which coincides with \mathbb{R}_+^d in the case of scalar parameters. Therefore, we consider the following modified version of the cost functional in (5.9):

$$\mathcal{F}(u^{\lambda, \varepsilon, \gamma}; \lambda) := \frac{1}{2N} \sum_{k=1}^N \|u_k^{\lambda, \varepsilon, \gamma} - f_k^0\|_{L^2(\Omega)}^2 + \frac{\eta}{2} \sum_{i=1}^d \|\lambda_i\|_X^2, \quad (5.12)$$

and consider the problem of determining the optimal $\lambda_i \geq 0$, $i = 1, \dots, d$ subject to the constraints (5.9b). Note that under this choice, the parameters λ_i are allowed to be spatially distributed.

Remark 5.4.3. *For the following analysis, we assume $\eta > 0$. For the numerical experiments, the choice $\eta = 0$ will be made in the case of scalar parameters.*

We state now the following Proposition, whose proof is an immediate extension of [DLRS13, Proposition 3.2], which ensures the desired continuity property of the solution map.

Proposition 5.4.4 ([DLRS13, Proposition 3.2]). *Let $(\lambda_n = (\lambda_{n_1}, \dots, \lambda_{n_d}))_n$ a sequence of elements in X^d weakly converging to $\hat{\lambda} \in X^d$ as $n \rightarrow \infty$. Further, let us denote by $u^{\lambda_n, \varepsilon, \gamma}$ the solution vector of (5.9b) associated to λ_n , i.e.:*

$$u^{\lambda_n, \varepsilon, \gamma} := \left(u_1^{\lambda_n, \varepsilon, \gamma}, \dots, u_N^{\lambda_n, \varepsilon, \gamma} \right),$$

and let $u^{\lambda, \varepsilon, \gamma}$ be defined similarly. Then we have:

$$u^{\lambda_n, \varepsilon, \gamma} \rightarrow u^{\lambda, \varepsilon, \gamma} \quad \text{strongly in } (H_0^1(\Omega))^N,$$

as $n \rightarrow \infty$.

Having modified the problem as in (5.12), we now state and prove an extension of [DLRS13, Theorem 3.3] for the multiple constrained case which ensures the existence of the solution of the regularised bilevel problem and its convergence to the corresponding nonsmooth solution as the Huber regularisation parameter γ goes to infinity. We stress that the presence of the H^1 smoothing enforced by the elliptic term is fundamental for the proof of this result. However, L^2 -type gradient smoothings penalise sharp edges and, as such, are not desirable in imaging application (see Section 1.3.1). In our numerical experiments, we circumvent this issue by taking in practice $\varepsilon \ll 1$.

Theorem 5.4.5. *There exists an optimal solution $\bar{\lambda} = (\bar{\lambda}_1, \dots, \bar{\lambda}_d) \in X^d$ of the regularised problem*

$$\min_{\lambda = (\lambda_1, \dots, \lambda_d) \in X^d} \frac{1}{2N} \sum_{k=1}^N \|u_k^{\lambda, \varepsilon, \gamma} - f_k^0\|_{L^2(\Omega)}^2 + \frac{\eta}{2} \sum_{i=1}^d \|\lambda_i\|_X^2, \quad (5.13)$$

subject to (5.9b). Moreover, every weakly convergent sequence of optimal parameters $(\lambda^\gamma = (\lambda_1^\gamma, \dots, \lambda_d^\gamma))_\gamma$ converges, as $\gamma \rightarrow \infty$, to an optimal solution $\lambda^\infty = (\lambda_1^\infty, \dots, \lambda_d^\infty)$ of the nonsmooth problem

$$\min_{\lambda = (\lambda_1, \dots, \lambda_d) \in X^d} \frac{1}{2N} \sum_{k=1}^N \|u_k^{\lambda, \varepsilon} - f_k^0\|_{L^2(\Omega)}^2 + \frac{\eta}{2} \sum_{i=1}^d \|\lambda_i\|_X^2, \quad (5.14)$$

subject to (5.11).

Proof. We follow [DLRS13, Theorem 3.3] and extend the result for the multiple constrained case. Given the positivity of the cost functional (5.12), let us consider a minimising sequence $(\lambda_n)_n$ in X^d . We deduce easily that such sequence is bounded in X^d . Therefore, we can extract a non-relabelled sequence $(\lambda_n)_n$ weakly converging to a limit $\bar{\lambda}$ in X^d . From Proposition 5.4.4 and the weakly lower semicontinuity property of the cost functional in $(L^2(\Omega))^N$, we deduce that $\bar{\lambda}$ is optimal.

Let now $(\lambda^\gamma)_\gamma$ be a sequence of optimal solution of (5.13) subject to (5.9b). Denoting by $u^{\lambda^\gamma, \varepsilon, \gamma} = (u_1^{\lambda^\gamma, \varepsilon, \gamma}, \dots, u_N^{\lambda^\gamma, \varepsilon, \gamma})$ the optimal state vector, we have that optimality of λ^γ combined with feasibility of $(0, 0) \in (H_0^1(\Omega))^N \times X^d$ implies

$$\begin{aligned} \frac{\eta}{2} \sum_{i=1}^d \|\lambda_i^\gamma\|_X^2 &\leq \frac{1}{2N} \sum_{k=1}^N \|u_k^{\lambda^\gamma, \varepsilon, \gamma} - f_k^0\|_{L^2(\Omega)}^2 + \frac{\eta}{2} \sum_{i=1}^d \|\lambda_i^\gamma\|_X^2 = \mathcal{F}(u^{\lambda^\gamma, \varepsilon}; \lambda^\gamma) \\ &\leq \frac{1}{2N} \sum_{k=1}^N \|f_k^0\|_{L^2(\Omega)}^2 = \mathcal{F}(0, 0). \end{aligned}$$

Therefore, $(\lambda^\gamma)_\gamma$ is bounded in X^d . Again, we can extract a non-relabelled subsequence $(\lambda^\gamma)_\gamma$ weakly converging in X^d to a point λ^∞ as $\gamma \rightarrow \infty$. Now, let us denote by $u^{\lambda^\infty, \varepsilon} = (u_1^{\lambda^\infty, \varepsilon}, \dots, u_N^{\lambda^\infty, \varepsilon})$ the optimal state for the problem (5.14) subject to (5.11) associated to λ^∞ whose existence is guaranteed by Proposition 5.4.2. By triangle inequality we have:

$$\sum_{k=1}^N \|u_k^{\lambda^\gamma, \varepsilon, \gamma} - u_k^{\lambda^\infty, \varepsilon, \gamma}\|_{H_0^1(\Omega)} \leq \sum_{k=1}^N \left(\|u_k^{\lambda^\gamma, \varepsilon, \gamma} - u_k^{\lambda^{\gamma^n}, \varepsilon, \gamma}\|_{H_0^1(\Omega)} + \|u_k^{\lambda^{\gamma^n}, \varepsilon, \gamma} - u_k^{\lambda^\infty, \varepsilon, \gamma}\|_{H_0^1(\Omega)} \right),$$

for every sequence $\{\lambda^{\gamma^n}\}$ weakly converging to λ^∞ . The right hand side of the inequality above goes to zero as $\gamma \rightarrow \infty$ by Proposition 5.4.2 and Proposition 5.4.4. Hence, we have

$$u^{\lambda^\gamma, \varepsilon, \gamma} \rightarrow u^{\lambda^\infty, \varepsilon} \quad \text{strongly in } (H_0^1(\Omega))^N.$$

To conclude, we note that by optimality of the pair $(\lambda^\gamma, u^{\lambda^\gamma, \varepsilon, \gamma})$ for the cost functional \mathcal{F} (5.12) we have:

$$\mathcal{F}(u^{\lambda^\gamma, \varepsilon, \gamma}; \lambda^\gamma) \leq \mathcal{F}(u^{\bar{\lambda}, \varepsilon, \gamma}; \bar{\lambda}),$$

where $\bar{\lambda}$ is one optimal solution of (5.13) subject to (5.9b) whose existence is guaranteed by the first part of the proof. By weak lower semicontinuity of \mathcal{F} we can now conclude since:

$$\mathcal{F}(u^{\lambda^\infty, \varepsilon}; \lambda^\infty) \leq \liminf_{\gamma \rightarrow \infty} \mathcal{F}(u^{\lambda^\gamma, \varepsilon, \gamma}; \lambda^\gamma) \leq \liminf_{\gamma \rightarrow \infty} \mathcal{F}(u^{\bar{\lambda}, \varepsilon, \gamma}; \bar{\lambda}) = \mathcal{F}(u^{\bar{\lambda}, \varepsilon, \gamma}; \bar{\lambda}) \leq \mathcal{F}(u^{\lambda^\infty, \varepsilon}; \lambda)$$

for every $u^{\lambda^\infty, \varepsilon} \in (H_0^1(\Omega))^N$ and $\lambda \in X^d$. In other words, $(u^{\lambda^\infty, \varepsilon}, \lambda^\infty)$ is an optimal solution of (5.14) subject to (5.11). \square

Remark 5.4.6. *In the framework presented, some ideas about sparsity regarding the selection of the optimal model as a combination of as few data as possible may be applied. This could be imposed already in the case of scalar parameters λ_i to select the best model fitting the data or in the case of spatially distributed parameters $\lambda = \lambda(x)$ to promote more regularisation only in some specific regions of the image where the noise is stronger, see [DHRC11, BDH13]. The use of an L^1 penalisation on λ to enforce sparsity in space would result in inequality constraints of the second-kind [DLR11] which would make harder the derivation of the optimality system.*

Remark 5.4.7. *As the reader may have noticed, compared to the general models (5.2) and (5.3), we have presented in Section 5.3 and in Section 5.4 two simplifications where either the noise model is assumed to be known or the regularisation one intends to use is fixed. Of course, the combination of the two models which aims to learn both the image and the noise model (in other words, the set of parameters α_i , $i = 1, \dots, m$ and λ_j , $j = 1, \dots, d$) is desirable and is indeed a matter for future research.*

5.5 Numerical optimisation

From a computational point of view, finding the numerical solution of the models above is quite challenging due to the nested bilevel minimisation structure. A standard approach consists in characterising the optimal solutions of the problem in terms of their optimality conditions, see Section 2.4.1. This is used a lot in practice, but appears to be hardly applicable in our case since the nonsmooth problem (5.2) does not have in principle any differentiability property. For this reason, we focus in the following on the regularised version (5.3) depending on the parameters ε and γ , for which additional differentiability properties are expected. In fact, provided that some type of differentiability (Gâteaux or Fréchet) can be proved, solutions of the regularised problems can be well characterised in term of the adjoint states. Once this is done, an asymptotic analysis as $\gamma \nearrow \infty$ and $\varepsilon \searrow 0$ is needed to guarantee the convergence of the solutions of the regularised problem to the ones of the original, nonsmooth problem.

In this section we describe how to characterise solutions of the problem (5.3) in terms of their adjoint states. We start using the Lagrangian formalism to derive the optimality system for the regularised problem which reads:

$$\varepsilon \int_{\Omega} \nabla u \cdot \nabla v \, dx + \sum_{i=1}^m \int_{\Omega} \alpha_i h_{\gamma}(A_i u) \cdot A_i v \, dx + \sum_{j=1}^d \int_{\Omega} \lambda_j \phi_j'(u, f) v \, dx = 0, \quad \forall v \in V, \quad (5.15a)$$

$$\varepsilon \int_{\Omega} \nabla p \cdot \nabla v \, dx + \sum_{i=1}^m \int_{\Omega} \alpha_i h_{\gamma}'(A_i u)^* A_i p \cdot A_i v \, dx + \sum_{j=1}^d \int_{\Omega} \lambda_j \phi_j''(u, f) p \cdot v \, dx = -F'(u)v, \quad \forall v \in V, \quad (5.15b)$$

$$\int_{\Omega} h_{\gamma}(A_i u) A_i p (\eta - \alpha_i) \, dx \geq 0, \quad \forall \eta \geq 0, \quad i = 1, \dots, m, \quad (5.15c)$$

$$\int_{\Omega} \phi_i(u, f) p (\zeta - \lambda_j) \, dx \geq 0, \quad \forall \zeta \geq 0, \quad j = 1, \dots, d. \quad (5.15d)$$

Here, V is a subspace of $H^1(\Omega)$ endowed with appropriate boundary conditions, $p \in V$ is the associated adjoint state and h_{γ} is a smoother version of the derivative of the Huber

regularisation of TV 2.11. In particular, we consider the following C^1 function:

$$h_\gamma(z) := \begin{cases} \frac{z}{|z|} & \text{if } \gamma|z| - 1 \geq \frac{1}{2\gamma} \\ \frac{z}{|z|} \left(1 - \frac{\gamma}{2} \left(1 - \gamma|z| + \frac{1}{2\gamma}\right)^2\right) & \text{if } \gamma|z| - 1 \in \left(-\frac{1}{2\gamma}, \frac{1}{2\gamma}\right) \\ \gamma z & \text{if } \gamma|z| - 1 \leq -\frac{1}{2\gamma}. \end{cases} \quad (5.16)$$

which enjoys the additional differentiability property required in the adjoint equation (5.15b). We remark that the optimality system above is derived formally for the sake of having some insight on its general structure. Depending on the specific function space, regulariser and model considered, every step in the derivation above has to be rigorously justified.

Remark 5.5.1. *Gradient information is fundamental for the computation of the numerical solution of the learning models considered. Typically, an easier expression of the gradient of the reduced cost functional $F(\alpha, \lambda) := \mathcal{F}(u^{\alpha, \lambda}; \alpha, \lambda)$ is given in terms of the adjoint equation (5.15b). In our case, we get from (5.15b):*

$$(\nabla_\alpha \tilde{F})_i = \int_\Omega h_\gamma(A_i u) A_i p \, dx, \quad (\nabla_\lambda \tilde{F})_j = \int_\Omega \phi'_j(u, f) p \, dx, \quad (5.17)$$

for $i = 1, \dots, m$ and $j = 1, \dots, d$, respectively. Furthermore, in the case of scalar (i.e. non-spatially distributed) parameters the variational inequalities (5.15c)-(5.15d) become actually equalities. This is very important from a numerical point of view since the gradient formulas (5.17) above can be used without any additional projection step.

5.5.1 Adjoint-based methods

In many situations the numerical solution of the optimality system (5.15) may be very challenging to compute, due mainly to the large size of both the images considered and of the training database (see Section (5.6) for more details). To overcome this issue, we follow [DLRS13, CCDLR⁺15, DLRSV15b] and focus in the following on the application of quasi-Newton optimisation methods based on the characterisation of the gradient of the cost functional in terms of the adjoint states as described in the previous section. We also refer the reader to the relevant parts in Section 2.4.2 of this thesis, where a review on adjoint-based methods for the numerical solution of PDE constrained optimisation problems is given.

In particular, we use the characterisation of the gradient of the reduced cost functional given in (5.17) in terms of the adjoint state to solve the upper level problem by means of BFGS updates (2.38) combined with an Armijo line search strategy (2.42). In the simple case of TV regularisation, such numerical strategy guarantees the positivity of the parameters estimated without any further projection step, whereas for more involved

regularisation, such as TGV or ICTV extra conditions need to be added to guarantee their positivity along the iterations, see [DLRSV15b].

An efficient solver is also required for the lower-level denoising problem at each BFGS iteration. In our numerical experiments we use primal-dual SemiSmooth Newton (SSN) algorithms (see Section 2.4.2). In their general form, these algorithms make use of the dual variables q_i , $i = 1, \dots, m$ and translate the sufficient condition (5.15a) into the system:

$$\begin{cases} \varepsilon \int_{\Omega} \nabla u \cdot \nabla v \, dx + \sum_{i=1}^m \int_{\Omega} q_i \cdot A_i v \cdot \, dx + \sum_{j=1}^d \int_{\Omega} \lambda_j \phi'_j(u, f) v \, dx = 0, & \forall v \in X, \\ q_i = \alpha_i h_{\gamma}(A_i u), & \text{a.e. in } \Omega, \quad i = 1, \dots, m, \end{cases}$$

where again a Huber regularisation term h_{γ} is used. In the following numerical experiments we consider $h_{\gamma}(z) = \frac{z}{\max(1, \gamma|z|)}$, but smoother versions can be chosen, similarly as in (5.16).

The SSN step has the following Jacobi matrix

$$\begin{pmatrix} -\varepsilon \Delta \cdot + \sum_{j=1}^d \phi''_j(u, f) & A_1^* & \dots & A_m^* \\ -\alpha_1 \left[\mathcal{R}(A_1 u) - \chi_1 \frac{A_1 u \otimes A_1 u}{|A_1 u|^3} \right] A_1 & I & 0 & 0 \\ \vdots & 0 & \ddots & 0 \\ -\alpha_m \left[\mathcal{R}(A_m u) - \chi_m \frac{A_m u \otimes A_m u}{|A_m u|^3} \right] A_m & 0 & 0 & I \end{pmatrix}$$

where $\chi_i(x)$ is the indicator function of the set $\{x : \gamma|A_i u| > 1\}$ and $\mathcal{R}(A_i u) := \frac{\min(1, \gamma|A_i u|)}{|A_i u|}$, for $i = 1, \dots, m$. In our numerical implementations we will follow [HS06] and modify systematically the matrix above replacing the terms $\frac{A_i u \otimes A_i u}{|A_i u|^3}$ by $\frac{q_i}{\max(|q_i|, \alpha_i)} \otimes \frac{A_i u}{|A_i u|^2}$ for every $i = 1, \dots, m$. Such modification improves the convergence properties of the standard SSN method and the resulting algorithm exhibits both a global and a local superlinear convergent behaviour.

We will consider a warm start of the modified SSN solver initialising the algorithm with the the result computed in the previous BFGS iteration. The computation of the adjoint states needed for the evaluation of the gradient of the cost functional through (5.17) is performed using sparse linear solvers.

Remark 5.5.2. *The full optimality system of equations (5.15) may be alternatively solved at once using efficient generalised Newton solvers. This approach has been considered, for instance, in [KP13] and [CDLR] in a finite- and infinite-dimensional framework, respectively. However, in such approaches the nonsmoothness of the regularisation terms and the positivity constraints may require a modification of the full Newton matrix to get good convergence properties through the introduction of additional feasibility steps. Furthermore, domain decomposition methods can also be used for improving upon the efficiency of the*

method in the case of large images.

As previously specified, these approaches are considered in a *optimise-then-discretise* framework. Numerically, we expect to obtain resolution independent algorithms which, thanks to their second-order convergence properties, can compute the optimal solutions of the model in a feasible time.

5.5.2 Optimality system

Let us consider the framework described in Section 5.4 and in particular let us focus on the numerical optimisation of the bilevel optimisation problem (5.9) to learn the optimal noise model. In order to derive rigorously an optimality system of the form (5.15a), the Gâteaux or Fréchet differentiability property of the solution operator $\mathcal{S} : X^d \rightarrow (H_0^1(\Omega))^N$ assigning to each parameter λ the corresponding solution $u^{\lambda,\varepsilon,\gamma}$ is required. Then, using standard Lax-Milgram theory the following Theorem ([DLRS13, Theorem 3.3]) can be proved.

Theorem 5.5.3 ([DLRS13, Proposition 2 and Theorem 3.2]). *The solution operator $\mathcal{S} : X^d \rightarrow (H_0^1(\Omega))^N$ is Gâteaux differentiable. Moreover, if $(\lambda, u^{\lambda,\varepsilon,\gamma}) \in X^d \times (H_0^1(\Omega))^N$ is an optimal solution of the problem (5.13) subject to (5.9b), then there exist Lagrange multipliers $(p, \mu) \in (H_0^1(\Omega))^N \times X^d$ such that the following optimality system holds:*

$$\varepsilon \int_{\Omega} \nabla u_k^{\lambda,\varepsilon,\gamma} \cdot \nabla v \, dx + \int_{\Omega} h_{\gamma}(\nabla u_k^{\lambda,\varepsilon,\gamma}) \cdot \nabla v \, dx + \sum_{i=1}^d \int_{\Omega} \lambda_i \phi'_i(u_k^{\lambda,\varepsilon,\gamma}, f_k) v \, dx = 0, \quad \forall v \in H_0^1(\Omega) \quad (5.18a)$$

$$\varepsilon \int_{\Omega} \nabla p_k \cdot \nabla v \, dx + \int_{\Omega} h'_{\gamma}(\nabla u_k^{\lambda,\varepsilon,\gamma})^* \nabla p_k \cdot \nabla v \, dx + \sum_{i=1}^d \int_{\Omega} \lambda_i \phi''_i(u_k^{\lambda,\varepsilon,\gamma}, f_k) p_k v \, dx = -\frac{1}{N} \int_{\Omega} (u_k^{\lambda,\varepsilon,\gamma} - f_k^0) v \, dx, \quad \forall v \in H_0^1(\Omega), \quad (5.18b)$$

$$\mu_i = \eta(\lambda_i, \xi - \lambda_i)_X + \int_{\Omega} \nabla \phi_i(u^{\lambda,\varepsilon,\gamma}, f) \cdot p(\xi - \lambda_i) \, dx \geq 0, \quad i = 1, \dots, d, \quad \forall \xi \in X : \xi \geq 0 \text{ a.e. in } \Omega, \quad (5.18c)$$

for every $k = 1, \dots, N$.

Remark 5.5.4. *In the case of scalar parameters ($X^d = \mathbb{R}_+^d$), the inequality condition*

(5.18c) becomes an equality that reads:

$$\begin{aligned} \mu_i &= \eta\lambda_i + \int_{\Omega} \nabla\phi_i(u^{\lambda,\varepsilon,\gamma}, f) \cdot p \, dx, & i &= 1, \dots, d, \\ \mu_i &\geq 0, \quad \lambda_i \geq 0, \quad \mu_i\lambda_i = 0, & i &= 1, \dots, d. \end{aligned}$$

Similarly as discussed in Remark 5.5.1, the gradient of the reduced cost functional $\tilde{F}(\lambda) := \mathcal{F}(u^{\lambda,\varepsilon,\gamma}; \lambda)$ can then be expressed as:

$$(\nabla_{\lambda}\tilde{F})_i = \eta\lambda_i + \int_{\Omega} \nabla\phi_i(u^{\lambda,\varepsilon,\gamma}, f) \cdot p \, dx, \quad i = 1, \dots, d.$$

The expression above plays a fundamental role in the design of the optimisation algorithms we are going to use to solve the problem (5.9), as discussed in Section 5.5.1.

5.5.3 Lagrangian derivation of the optimality system for TV-IC model

We now come back to the TV-IC denoising model (4.20) presented in Chapter 4 for mixed noise statistics. Here, the choice of the optimal weighting parameters λ_1 and λ_2 is essential both for an optimal decomposition of the noise into its constituting components and for the balancing between TV regularisation and the fitting with the given data models. In the following we formally derive the optimality system for (4.20) through Lagrangian formalism for the case of impulse-Gaussian infimal convolution fidelity as discussed in 4.3.1. A similar derivation accommodating also the Gaussian-Poisson case is not straightforward since the definition of the admissible sets \mathcal{A} and \mathcal{B} in (4.24) in this case poses additional pointwise bounds on the state variables that need to be taken into account. Similarly as above, we introduce both a Huber-regularisation of the TV term and an elliptic regularisation weighted by a parameter $\varepsilon \ll 1$. Furthermore, for simplicity, we fix the training database to be composed of $N = 1$ image only. In summary, we are considering the following bilevel TV-IC problem:

$$\min_{\lambda=(\lambda_1,\lambda_2)\geq 0} \frac{1}{2} \|u^{\lambda,\varepsilon,\gamma} - f^0\|_{L^2(\Omega)}^2, \quad (5.19a)$$

subject to:

$$(u^{\lambda,\varepsilon,\gamma}, v^{\lambda,\varepsilon,\gamma}) = \arg \min_{\substack{u \in H_0^1(\Omega), \\ v \in L^2(\Omega)}} \left\{ \frac{\varepsilon}{2} \|\nabla u\|_{L^2(\Omega)}^2 + |Du|_{\gamma}(\Omega) + \lambda_1 \Phi_1(v) + \lambda_2 \Phi_2(f, v, u) \right\}, \quad (5.19b)$$

with

$$\Phi_1 = \|v\|_{L^1(\Omega),\gamma} \quad \Phi_2(f, v, u) = \frac{1}{2} \|f - u - v\|_{L^2(\Omega)}^2, \quad (5.19c)$$

where $\|\cdot\|_{L^1(\Omega),\gamma}$ is a further Huber-type regularisation (5.16) of the L^1 term associated to the impulse noise component Φ_1 used to guarantee its differentiability. Let us focus now

on (5.19b). In correspondence with an optimal pair (\tilde{u}, \tilde{v}) we have that:

$$\begin{aligned} \varepsilon \int_{\Omega} \nabla \tilde{u} \cdot \nabla w \, dx + \int_{\Omega} h_{\gamma}(\nabla \tilde{u}) \cdot \nabla w \, dx \\ + \lambda_2 \int_{\Omega} \frac{\delta \Phi_2}{\delta u}(f, \tilde{v}, \tilde{u}) w \, dx = 0 \quad \text{for all } w \in H_0^1, \end{aligned} \quad (5.20a)$$

$$\lambda_1 \int_{\Omega} \Phi_1'(\tilde{v}) z \, dx + \lambda_2 \int_{\Omega} \frac{\delta \Phi_2}{\delta v}(f, \tilde{v}, \tilde{u}) z \, dx = 0 \quad \text{for all } z \in L^2(\Omega). \quad (5.20b)$$

Equivalently, the optimality conditions above can be expressed by the following PDEs:

$$-\varepsilon \Delta \tilde{u} - \nabla \cdot (h_{\gamma}(\nabla \tilde{u})) + \lambda_2 \frac{\delta \Phi_2}{\delta u}(f, \tilde{v}, \tilde{u}) = 0, \quad (5.21a)$$

$$\lambda_1 \Phi_1'(\tilde{v}) + \lambda_2 \frac{\delta \Phi_2}{\delta v}(f, \tilde{v}, \tilde{u}) = 0 \quad (5.21b)$$

where equalities hold, for instance, in the sense of distributions. We now make use of the Lagrangian formalism presented in Section 2.4.1 and define the Lagrangian functional associated to the problem, which reads:

$$\begin{aligned} \mathcal{L}(\lambda_1, \lambda_2, u, v, p_1, p_2) := \frac{1}{2} \int_{\Omega} (u - f^0)^2 \, dx + \int_{\Omega} \left(\varepsilon \Delta u + \nabla \cdot (h_{\gamma}(\nabla u)) - \lambda_2 \frac{\delta \Phi_2}{\delta u}(f, v, u) \right) p_1 \, dx \\ - \int_{\Omega} \left(\lambda_1 \Phi_1'(v) + \lambda_2 \frac{\delta \Phi_2}{\delta v}(f, v, u) \right) p_2 \, dx. \end{aligned}$$

where the regularity of the adjoint states p_1 and p_2 still needs to be defined. Now, denoting by $\Lambda = (\Lambda_1, \Lambda_2)$, $S = (S_1, S_2)$ and P the first, the second and the third pair of arguments for \mathcal{L} corresponding to the pair of control, state and adjoint variables, respectively, we write the optimality conditions of the type (2.26) holding for the optimal solution $(\tilde{\lambda}_1, \tilde{\lambda}_2, \tilde{u}, \tilde{v})$ of (5.19) in correspondence with adjoint states p_1 and p_2 that need to be specified. Such conditions read:

$$\begin{aligned} \mathcal{L}_S(\tilde{\lambda}_1, \tilde{\lambda}_2, \tilde{u}, \tilde{v}, p_1, p_2) = 0, \\ \mathcal{L}_{\Lambda}(\tilde{\lambda}_1, \tilde{\lambda}_2, \tilde{u}, \tilde{v}, p_1, p_2)((\alpha, \beta)^T - (\tilde{\lambda}_1, \tilde{\lambda}_2)^T) \geq 0, \quad \text{for every } \alpha, \beta \geq 0. \end{aligned} \quad (5.22)$$

To make more explicit the structure of the optimality system, let us now compute formally such derivatives in correspondence with the optimal solution:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial S_1}(\tilde{\lambda}_1, \tilde{\lambda}_2, \tilde{u}, \tilde{v}, p_1, p_2)[w_1] = \int_{\Omega} (\tilde{u} - f^0) w_1 \, dx + \varepsilon \int_{\Omega} \Delta w_1 p_1 \, dx + \int_{\Omega} h_{\gamma}'(D\tilde{u})^* \nabla p_1 \cdot \nabla w_1 \, dx \\ - \tilde{\lambda}_2 \int_{\Omega} \frac{\delta^2 \Phi_2}{\delta^2 u}(f, \tilde{v}, \tilde{u}) p_1 w_1 \, dx - \tilde{\lambda}_2 \int_{\Omega} \frac{\delta^2 \Phi_2}{\delta u \delta v}(f, \tilde{v}, \tilde{u}) p_2 w_1 \, dx = 0, \quad \forall w_1 \in H_0^1(\Omega), \end{aligned}$$

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial S_2}(\tilde{\lambda}_1, \tilde{\lambda}_2, \tilde{u}, \tilde{v}, p_1, p_2)[w_2] &= -\tilde{\lambda}_2 \int_{\Omega} \frac{\delta^2 \Phi_2}{\delta u \delta v}(f, \tilde{v}, \tilde{u}) p_1 w_2 dx \\ &- \tilde{\lambda}_1 \int_{\Omega} \Phi_1''(\tilde{v}) p_2 w_2 dx - \tilde{\lambda}_2 \int_{\Omega} \frac{\delta^2 \Phi_2}{\delta^2 v}(f, \tilde{v}, \tilde{u}) p_2 w_2 dx = 0, \quad \forall w_2 \in L^2(\Omega). \end{aligned}$$

Or, equivalently:

$$\begin{aligned} \varepsilon \int_{\Omega} \nabla p_1 \cdot \nabla w_1 dx + \int_{\Omega} h'_{\gamma}(D\tilde{u})^* \nabla p_1 \cdot \nabla w_1 dx - \tilde{\lambda}_2 \int_{\Omega} \frac{\delta^2 \Phi_2}{\delta^2 u}(f, \tilde{v}, \tilde{u}) p_1 w_1 dx & \quad (5.24a) \\ - \tilde{\lambda}_2 \int_{\Omega} \frac{\delta^2 \Phi_2}{\delta u \delta v}(f, \tilde{v}, \tilde{u}) p_2 w_1 dx = - \int_{\Omega} (\tilde{u} - f^0) w_1 dx, & \quad \forall w_1 \in H_0^1(\Omega), \end{aligned}$$

$$\begin{aligned} \tilde{\lambda}_2 \int_{\Omega} \frac{\delta^2 \Phi_2}{\delta u \delta v}(f, \tilde{v}, \tilde{u}) p_1 w_2 dx + \tilde{\lambda}_2 \int_{\Omega} \frac{\delta^2 \Phi_2}{\delta^2 v}(f, \tilde{v}, \tilde{u}) p_2 w_2 dx & \quad (5.24b) \\ = -\tilde{\lambda}_1 \int_{\Omega} \Phi_1''(\tilde{v}) p_2 w_2 dx & \quad \forall w_2 \in L^2(\Omega). \end{aligned}$$

Proceeding in an analogous way for optimality condition (5.22) we have:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \Lambda_1}(\tilde{\lambda}_1, \tilde{\lambda}_2, \tilde{u}, \tilde{v}, p_1, p_2)(\alpha - \lambda_1^{\gamma}) &= \left(\int_{\Omega} \Phi_1'(\tilde{v}) p_2 dx \right) (\tilde{\lambda}_1 - \alpha) \geq 0, & (5.25a) \\ \frac{\partial \mathcal{L}}{\partial \Lambda_2}(\tilde{\lambda}_1, \tilde{\lambda}_2, \tilde{u}, \tilde{v}, p_1, p_2)(\beta - \tilde{\lambda}_2) &= \left(\int_{\Omega} \left(\frac{\delta \Phi_2}{\delta u}(f, \tilde{v}, \tilde{u}) p_1 + \frac{\delta \Phi_2}{\delta v}(f, \tilde{v}, \tilde{u}) p_2 \right) dx \right) (\tilde{\lambda}_2 - \beta) \geq 0 \end{aligned}$$

for every $\alpha, \beta \geq 0$. Therefore, formally, in correspondence with an optimal quadruplet $(\tilde{\lambda}_1, \tilde{\lambda}_1, \tilde{u}, \tilde{v}) \in \mathbb{R}_+ \times \mathbb{R}_+ \times H_0^1(\Omega) \times L^2(\Omega)$ there exist $(p_1, p_2) \in H_0^1(\Omega) \times L^2(\Omega)$ such that the optimality system (5.20)-(5.24)-(5.25) is satisfied. A rigorous justification of all the steps above such as, for instance, the differentiability property of the solution operator is a matter of future research.

Preliminary results for noise model selection

In the following we report some preliminary numerical experiments on the optimal noise model selection for the TV-IC impulse+Gaussian denoising model (5.19). In particular, we consider the case of an image corrupted only by impulse noise with a percentage of missing pixels $d = 5\%$, see Figure 5.2, and report in Figure 5.3 the behaviour of the cost functional (5.19a), highlighting with a red cross the optimal selected pair of parameters $(\tilde{\lambda}_1, \tilde{\lambda}_2)$. In our numerical simulations, we are not using the optimality system derived above. Here, we are rather interested in having a first insight on how the bilevel approach performs in the selection of the optimal noise model in the case when no *a priori* assumptions on the noise are available. Namely, in the particular case considered, we expect our model to select optimal parameters $\tilde{\lambda}_1$ and $\tilde{\lambda}_2$ enforcing a TV- L^1 type model (TV- L^1) which, as previously discussed, has been shown to be suitable for this type of

problems [Nik04, DAG09]. In Figure 5.3 we observe that the optimum is indeed achieved in correspondence of an optimal pair $(\tilde{\lambda}_1, \tilde{\lambda}_2)$ which does correspond to a TV- L^1 type model, as discussed in Section 4.4.1. In particular, such plot shows that the optimal choice of parameters corresponds to $\tilde{\lambda}_1 = 281$ and a very large $\tilde{\lambda}_2$, which actually enforces a TV- L^1 model, as the Gaussian noise residuum decays to zero as $\lambda_2 \rightarrow \infty$, see Proposition 4.4.1. Therefore, from this preliminary test we are inclined to think the combined model (5.19) could be used in the framework for a blind selection of the optimal noise model best fitting the given data.

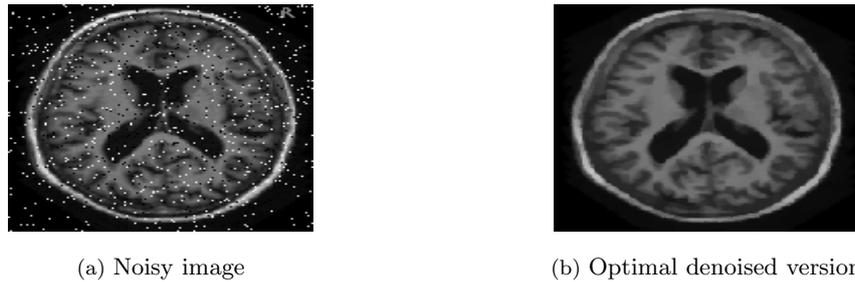


Figure 5.2: Impulse noise TV-IC denoising with optimal parameters $(\tilde{\lambda}_1, \tilde{\lambda}_2)$.

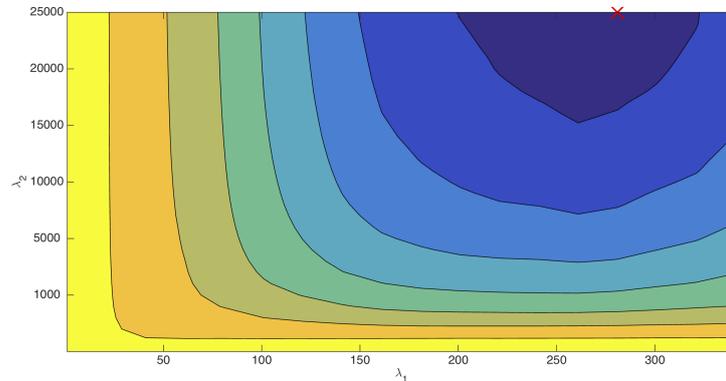


Figure 5.3: Contour plots of the functional value versus λ_1 and λ_2 for the problem (5.19a). The value of the optimal parameters $\tilde{\lambda}_1$ and $\tilde{\lambda}_2$ for TV-IC denoising model is indicated with a red cross. Such choice enforces the Gaussian noise component of the model to be zero. The estimated optimal impulse parameter is $\tilde{\lambda}_1 = 281$.

5.6 Robust estimation for large training sets

In many imaging applications such as Magnetic Resonance Imaging (MRI) or Positron Emission Tomography (PET) the accuracy, i.e. the noise level of the measurements, can be tuned by means of a database of real data. Such accuracy can depend, for instance, on the acquisition time of the machinery used. Thus, ground truth images or reasonably

accurate approximations f_k^0 can be found by using the maximal accuracy possible, while their noisy versions f_k can be measured within a usual clinical setup. Such a strategy is used already in the medical imaging community where the good quality measurements or template shapes are used as priors in image reconstruction and image segmentation problems. The dataset can also be simulated: even in real world applications such as MRI, simulated databases are used to tune image retrieval systems, see for instance [MSG01]. Alternatively, we can imagine the retrieval of such a test set for a specific application using phantoms and their noisy acquisitions. Ideally, the use of a very rich database (i.e. $N \gg 1$) would correspond to a more robust estimation of the parameters. Mathematically, this corresponds to consider a very large set of constraints of the type (5.9b) to be solved in each iteration of the optimisation loop. The computational solution of such an optimisation problem renders expensive and therefore challenging due to the large-scale nature of the problem (5.9) and due to the nonsmooth nature of each constraint.

In order to deal with such large-scale problems various approaches have been presented in literature. They are based on the common idea of solving not *all* the nonlinear PDE constraints, but just a sample of them, whose size varies according to the approach one intends to use. In *stochastic approximation* (SA) methods ([RM51, Nes09]) typically a single data point is sampled per iteration, thus producing a generally inexpensive but noisy step. In *sample* or *batch average approximation* methods (see e.g. [BCNN11, BCNW12]) larger samples are used to compute an approximating (batch) gradient: the computation of such steps is normally more reliable, but generally more expensive. The development of parallel computing, however, has improved upon the computational costs of batch methods: independent computation of functions and gradients can now be performed in parallel processors, so that the reliability of the approximation can be supported by more efficient methods. Some other efficient algorithms have been proposed in [BB08, HCH12, PJ92] under the name of Stochastic Average Approximation methods (SAA) and have shown good performance.

In this section we focus on a stochastic approximation method proposed by Byrd et al. in [BCNW12] called *dynamic sample size* method. The main idea of this method is to consider an initial, small, training sample of the dictionary to start the algorithm with and *dynamically* increasing its size, if needed, throughout the different steps of the optimisation process. The criterion to decide whether or not the sample size has to be increased is a check on the sample variance estimates on the batch gradient. The desired trade-off between efficiency and accuracy is then obtained by starting with a small sample and gradually increasing its size till reaching the requested level of accuracy.

The following discussion extends the work of [BCNW12] in two directions: firstly, in [BCNW12] the linearity of the solution map is required which is not fulfilled for our problem (5.9). We are going to show that the strategy of Byrd et al. can be modified for nonlinear solution maps as the one we are considering. Secondly, in [BCNW12] the

optimisation algorithm is of gradient-descent type. Using a BFGS method to solve (5.9) we extend their approach incorporating also *second order* information in form of an approximation of the Hessian by evaluations of the sample gradient in the iterations of the optimisation algorithm.

Notation. In the following, we denote the vector of scalar positive parameters we aim to estimate by $\lambda = (\lambda_1, \dots, \lambda_d) \in \mathbb{R}_+^d$. We also define by \mathcal{S} the solution map that, for each constraint $k = 1, \dots, N$ of (5.9b), associates to λ and to the noisy image f_k the corresponding solution $u_k := u_k^{\lambda, \varepsilon, \gamma}$ so that $\mathcal{S}(\lambda, f_k) = u_k$. Since we are dealing with scalar parameters, recalling Remark 5.4.3, hereafter we set $\eta = 0$ in (5.12) and consider the reduced cost functional $F(\lambda) := \mathcal{F}(u^{\varepsilon, \lambda, \gamma}; \lambda)$ defined as

$$F(\lambda) = \frac{1}{2N} \sum_{k=1}^N \|\mathcal{S}(\lambda, f_k) - f_k^0\|_{L^2(\Omega)}^2. \quad (5.26)$$

We also define:

$$l(\lambda, f_k) := \|\mathcal{S}(\lambda, f_k) - f_k^0\|_{L^2(\Omega)}^2, \quad k = 1, \dots, N \quad (5.27)$$

as the *loss functions* of the functional F defined in (5.26) for each $k = 1, \dots, N$. In our case, these functions are simply the mean square error between f_k^0 and u_k , but the following considerations may be adapted for more general (differentiable) loss functions as well. Finally, for every sample $S \subset \{1, \dots, N\}$ of the database, we introduce the batch objective function defined as

$$F_S(\lambda) := \frac{1}{2|S|} \sum_{k \in S} l(\lambda, f_k). \quad (5.28)$$

5.6.1 Dynamic sampling

We want to solve (5.9) by using a quasi-Newton method (namely, the Broyden-Fletcher-Goldfarb-Shanno algorithm BFGS described in Section 2.4.2) together with an Armijo backtracking line search rule. We combine such algorithm with a modified version of the dynamic sampling algorithm presented in [BCNW12, Section 3]. In order to compare our algorithm with the Newton-Conjugate Gradient method presented in [BCNW12, Section 5], we highlight that in our optimisation algorithm the Hessian matrix is never computed, but approximated efficiently by the BFGS matrix.

Heuristically, our algorithm starts by selecting from the whole dataset a sample S whose size $|S|$ is small compared to the original size N . In the following iterations, if the approximation computed produces an improvement in the cost functional F , then the sample size is kept unchanged and the optimisation process continues selecting in the next iteration a new sample of the same size. Otherwise, if the approximation computed is not a good one, a new, larger, sample size is selected and a new sample S of this new size is

used to compute the new step. By starting with small sample sizes it is hoped that in the early stages of the algorithm the solution can be computed efficiently in each iteration. The key point in this procedure is clearly the rule that checks throughout the progression of the algorithm, whether the approximation we are performing is good enough, i.e. the sample size is big enough, or has to be increased. Because of this systematic check on the quality of approximation in each step of the algorithm, such sampling strategy is called *dynamic*.

We consider a condition on the batch gradient ∇F_S which imposes in every stage of the optimisation that the vector $-\nabla F_S$ is a descent direction for F at λ if the following condition holds:

$$\|\nabla F_S(\lambda) - \nabla F(\lambda)\|_2 \leq \theta \|\nabla F_S(\lambda)\|_2 \quad \theta \in [0, 1). \quad (5.29)$$

Theorem 5.6.1. *Let $\theta \in [0, 1)$. If condition (5.29) holds, then $-\nabla F_S(\lambda)$ is a descent direction for J .*

Proof. We prove the result by contradiction. Therefore, suppose that there holds

$$\nabla F_S(\lambda)^\top \nabla F(\lambda) \leq 0. \quad (5.30)$$

By squaring both sides of inequality (5.29) we receive:

$$\|\nabla F_S(\lambda)\|_2^2 - 2 \nabla F_S(\lambda)^\top \nabla F(\lambda) \leq \theta^2 \|\nabla F_S(\lambda)\|_2^2,$$

which implies:

$$0 < (1 - \theta^2) \|\nabla F_S(\lambda)\|_2^2 \leq 2 \nabla F_S(\lambda)^\top \nabla F(\lambda) \leq 0,$$

since $\theta \in [0, 1)$ and by (5.30). Hence the converse of (5.30) holds, that is $-\nabla F_S(\lambda)$ is a descent direction. \square

The computation of ∇F may be very expensive for applications involving large databases and nonlinear constraints, so we rewrite (5.29) as an estimate of the variance of the random vector $\nabla F_S(\lambda)$. In order to do that, recalling definitions (5.27) and (5.28) we first observe that

$$\nabla F_S(\lambda) = \frac{1}{2|S|} \sum_{k \in S} \nabla l(\lambda, f_k) \quad (5.31)$$

We can compute (5.31) in correspondence to an optimal solution $\hat{\lambda}$ by using Remark 5.5.1 and the characterisation of ∇F given in terms of the adjoint states p_k . By linearity, the extension to the multiple-constrained case reads:

$$\nabla F_S(\hat{\lambda}) = \sum_{k \in S} \sum_{i=1}^d \int_{\Omega} \phi'_i(u, f_k) p_k \, dx \quad (5.32)$$

Thanks to this characterisation, we can now extend the dynamic sampling algorithm in [BCNW12] to the case where the solution map \mathcal{S} is nonlinear: by taking (5.32) into account and following [BCNW12, Section 3] we can rewrite (5.29) as a condition on the variance of the batch gradient that reads as follows. The expected value of the left hand side of (5.29) we want to estimate over all possible samples $S \subset \{1, \dots, N\}$ satisfies:

$$\mathbb{E}_S \left(\|\nabla F_S(\lambda) - \nabla F(\lambda)\|_2^2 \right) = \|\text{Var}_S(\nabla F_S(\lambda))\|_1,$$

since $\mathbb{E}_S(\nabla F_S(\lambda)) = \nabla F(\lambda)$. Hence, the problem boils down to estimate the d -dimensional vector $\text{Var}_S(\nabla F_S(\lambda))$, where S is a sample chosen from N without replacement. In our applications this means that we do not want to consider the same PDE, i.e. the same image pair in the database more than once at each iteration of the optimisation loop. Hence, classical statistical theory estimates $\text{Var}_S(\nabla F_S(\lambda))$ in this case as:

$$\text{Var}_S(\nabla F_S(\lambda)) = \frac{\text{Var}(\nabla l(\lambda, f_k))}{|S|} \frac{N - |S|}{N - 1},$$

where $k \in \{1, \dots, N\}$. Since $\text{Var}(\nabla l(\lambda, f_k))$ is too expensive to compute, we approximate its value with the sample variance:

$$\text{Var}_{k \in S}(\nabla l(\lambda, f_k)) = \frac{1}{|S| - 1} \sum_{k \in S} \left(\nabla l(\lambda, f_k) - \nabla F_S(\lambda) \right)^2$$

where the square is taken component-wise. In other words, we are considering the following modified version of (5.29):

$$\frac{\|\text{Var}_{k \in S}(\nabla l(\lambda, f_k))\|_1}{|S|} \frac{N - |S|}{N - 1} \leq \theta^2 \|\nabla F_S(\lambda)\|_2^2. \quad (5.33)$$

Condition (5.33) is the responsible for possible changes of the sample size in the optimisation algorithm and has to be checked in every iteration. If inequality (5.33) is not satisfied, a larger sample \hat{S} whose size satisfies the descent condition (5.33) needs to be considered. By assuming that the change in the sample size is gradual enough such that, for any given λ the following approximations hold:

$$\begin{aligned} \|\text{Var}_{k \in \hat{S}}(\nabla l(\lambda, f_k))\|_1 &\approx \|\text{Var}_{k \in S}(\nabla l(\lambda, f_k))\|_1, \\ \|\nabla F_{\hat{S}}(\lambda)\|_2 &\approx \|\nabla F_S(\lambda)\|_2, \end{aligned}$$

we see that condition (5.33) is satisfied whenever we choose $|\hat{S}|$ such that

$$|\hat{S}| \geq \left\lceil \frac{N - \|\text{Var}_{k \in S}(\nabla l(\lambda, f_k))\|_1}{\|\text{Var}_{k \in S}(\nabla l(\lambda, f_k))\|_1 + \theta^2(N - 1)\|\nabla F_S(\lambda)\|_2^2} \right\rceil. \quad (5.34)$$

Conditions (5.33) and (5.34) are the key points in the optimisation algorithm we are going to present: by checking the former, one can control whether the sampling approximation is accurate enough and if this is not the case at any stage of the algorithm, by imposing the latter a new larger sample size is determined.

We remark that these two conditions force the direction $-\nabla F_S$ to be a descent direction. Steepest descent methods are known to be slowly convergent. Algorithms incorporating information coming from the Hessian are generally more efficient. However, normally the computation of the Hessian is very expensive, so Hessian-approximating methods are commonly used. In [BCNW12] a Newton-CG method is employed. There, an approximation of the Hessian matrix $\nabla^2 F_S$ is computed only on a subsample H of S such that $|H| \ll |S|$. As the sample S is dynamically changing, the subsample H will change as well (with a fixed, constant ratio) and the computation of the new conjugate gradient direction can be performed efficiently. Here, we consider the BFGS method to compute an approximation of the Hessian because of its efficiency and low computational costs.

Further, we use the Armijo line search (2.42). The positivity of the parameters is always preserved along the iterations so no additional projection steps are required.

We present now the BFGS optimisation with dynamic sampling for solving (5.9): compared to [BCNW12, Algorithm 5.2] we stress once more that the gain in efficiency is obtained thanks to the use of BFGS instead of the Newton-CG sampling method.

Algorithm 1 Dynamic Sampling BFGS for solving (5.9)

- 1: Initialize: λ_0 , sample \mathcal{S}_0 with $|\mathcal{S}_0| \ll N$ and model parameter θ , $k = 0$.
 - 2: **while** BFGS not converging, $k \geq 0$
 - 3: sample $|\mathcal{S}_k|$ PDE constraints to solve;
 - 4: update of the BFGS matrix;
 - 5: compute direction d_k by BFGS and steplength α_k by Armijo cond. (2.42);
 - 6: define new iterate: $\lambda_{k+1} = \lambda_k + \alpha_k d_k$;
 - 7: **if** condition (5.33) **then**
 - 8: maintain the sample size: $|\mathcal{S}_{k+1}| = |\mathcal{S}_k|$;
 - 9: **else** augment \mathcal{S}_k such that condition (5.34) is verified.
 - 10: **end**
-

Convergence and sensitivity. In Figure 5.4 we report two graphs regarding the convergence behaviour and the accuracy level of Algorithm 1. Namely, in Figure 5.4a we represent the decreasing of the cost functional along the BFGS iterations. Because of the sampling strategy, in the early iterations of BFGS the problem considered varies quite a lot, thus showing oscillations. However, evolving the process the convergence appears superlinear, as expected. To reduce the initial oscillations, one strategy may be, for instance, a different initialisation of the BFGS matrix B_0 , which is typically chosen as the identity

matrix. One alternative could be setting B_0 to be the corresponding Hessian matrix of a linearised version of (5.9) where the nonlinear terms have been neglected.

In Figure 5.4b we represent the sensitivity with respect to the accuracy parameter θ in (5.33): smaller values of θ penalise larger variances on ∇F_S , thus favouring larger samples. Larger values of θ allow larger variances on ∇F_S and, consequently, smaller sample sizes. In this case, efficiency improves, but accuracy suffers, see Table 5.1.

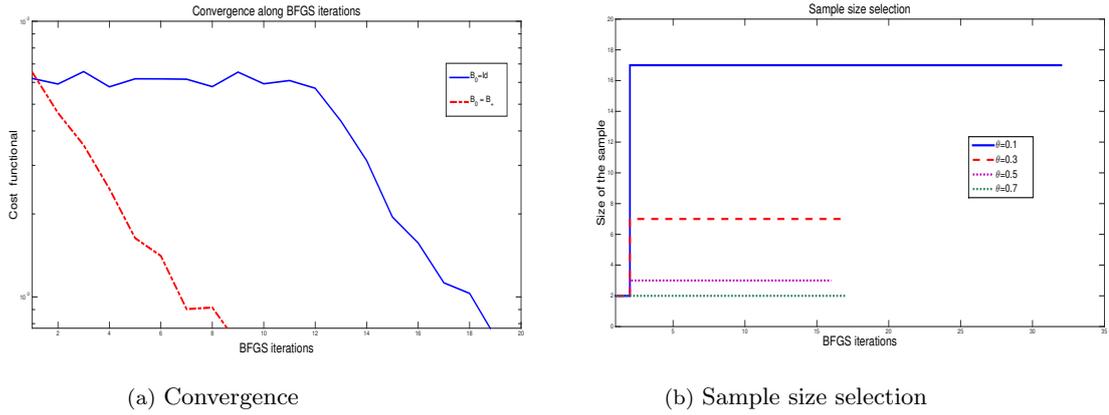


Figure 5.4: Left: evolution of BFGS with dynamic sampling Algorithm 1 along the iterations. The blue line corresponds to the initialisation $B_0 = Id$, while the red line to $B_0 = B_+$, where B_+ is the Hessian matrix of the linearised problem computed in correspondence of the initial value λ_0 .

Right: samples size changes in Algorithm 1 for different values of θ . For each value of θ , the result is plotted till convergence. For this example $N = 20$, $|S_0| = 2$.

θ	Efficiency	Difference
0.1	516	0.07%
0.3	246	4.3%
0.5	92	5.9%
0.7	68	15%

Table 5.1: As θ increases, efficiency improves as smaller samples are allowed. However, the relative difference with the value estimated without sampling shows that accuracy suffers.

A more thorough analysis of the Algorithm proposed and further studies on its improvements regarding different initialisations of the B_0 matrix and on an ‘adaptive’ choice of θ for the selection of an optimal sample size balancing accuracy and efficiency are interesting topics of future research.

5.7 Numerical results

In this section we focus on the numerical solution of (5.9) via quasi-Newton methods (compare relevant parts in Section 2.4.2) combined with the Armijo backtracking line search rule (2.42). In particular, we consider BFGS updates of the quasi-Newton matrix as in (2.38) only in the case when the curvature condition (2.41) is satisfied. The characterisation of the gradient of the cost functional in terms of the adjoint states is given as in Theorem 5.5.2 (see also Remark 5.5.4). The nonsmooth lower level constraints (5.9b) are solved by SemiSmooth Newton (SSN) algorithms whose form depends on the choice and combination of the fidelity terms ϕ_i 's in (5.9b) and which are described in detail for every numerical experiment.

We fix the following setting:

- We consider for simplicity square images with equal number of pixels in the x and in the y direction and approximate the differential operators by finite difference discretisation schemes with mesh step size $h = 1/(\text{number of pixels in the } x\text{-direction})$. We use forward difference for the discretisation of the divergence operator and backward differences for the gradient to guarantee the adjointness property in the discrete case as well to guarantee the property (3.11). The Laplace operator is discretised by using the usual five point formula (3.10).
- At each iteration of the outer BFGS loop, a warm start of the SSN denoising algorithm is used by taking as initial iterate the denoised image computed in the previous outer iteration of BFGS. The algorithm terminates if either the difference between two consecutive iterates is smaller than $\epsilon_{tol} = 10^{-5}$ or if the maximum number of iterations $\text{maxiter} = 35$ is reached.
- The value of the Armijo condition parameter in (2.42) is chosen to be 10^{-4} .
- The Tikhonov regularisation parameter η is set to zero for the scalar case, whereas is chosen to be strictly positive in the case of spatially dependent λ_i 's.
- In order to simplify the notation of this section, we set $u := u^{\lambda, \varepsilon, \gamma}$ to be the solution of the lower-level Huber-regularised elliptic variational problem depending on the parameter λ to optimise.

As repeatedly mention in this section, in the case of scalar parameters we observe in our simulations that positivity is preserved along the iterations. Therefore, no additional projection step is needed in practice in this case.

The following exposition combines the numerical results reported in De Los Reyes-Schönlieb [DLRS13] with the ones computed by the author of the thesis in [CDLRS14, CCDLR⁺15]. For the sake of clarity, we did not choose to divide the following discussion

into subsections according to original/non-original contributions. An appropriate credit to the relevant references and contributions is given for each example reported.

5.7.1 Single noise estimation

In this section we consider the single noise distribution case, i.e. $d = 1$ in (5.9). As a start, we consider the case of a single training pair (f^0, f) of noise-free and noisy version of the same image, i.e. set $N = 1$ in (5.9a). In the following examples the fidelity term ϕ will change according to the different assumptions on the statistics in the data, i.e. on the noise distribution considered. In our first examples, the problem reduces to find the optimal balance between TV regularisation and data fidelity provided the original image is known in advance and, as such, is a bit unrealistic for applications. As motivated in Section 5.6, in real-world applications large training sets of images are used to learn the optimal parameters which are then used for the denoising of new images not contained in the training set.

Gaussian noise

We start considering the problem of computing the optimal fidelity weight λ^* for an image corrupted by Gaussian distributed noise. In this case the data term reads $\phi(u, f) = (u - f)^2$. The optimisation problem assumes in this case the following form:

$$\min_{\lambda \geq 0} \frac{1}{2} \|u - f^0\|_{L^2(\Omega)}^2 \quad (5.35a)$$

subject to the following variational inequality expressing the optimality of u :

$$\varepsilon \int_{\Omega} \nabla u \cdot \nabla v \, dx + \int_{\Omega} h_{\gamma}(\nabla u) \cdot \nabla v \, dx + \int_{\Omega} \lambda(u - f)v \, dx = 0 \quad \text{for all } v \in H_0^1(\Omega), \quad (5.35b)$$

where $h_{\gamma}(\nabla u)$ is the Huber-type regularisation of ∇u . As discussed in Section 5.5.1, the numerical solution of the lower-level problem is computed using a SSN algorithm. Expressed in primal-dual form, the optimality condition (5.35b) reads:

$$\begin{cases} -\varepsilon \Delta u - \nabla \cdot q + \lambda(u - f) = 0, \\ q = \frac{\gamma \nabla u}{\max(\gamma |\nabla u|, 1)}. \end{cases}$$

Starting from an appropriate initial guess for u_0 and q_0 the SSN iteration reads

$$\begin{cases} -\varepsilon \Delta \delta u - \nabla \cdot \delta q + \lambda \delta u = -(-\varepsilon \Delta u - \nabla \cdot q + \lambda(u - f)) \\ \delta q - \frac{\gamma \nabla \delta u}{\max(1, \gamma |\nabla u|)} + \chi \mathcal{U}_{\gamma} \gamma^2 \frac{\nabla u^{\top} \nabla \delta u}{\max(1, \gamma |\nabla u|)^2} \frac{q}{\max(1, |q|)} = -(q - \frac{\gamma \nabla u}{\max(\gamma |\nabla u|, 1)}), \end{cases}$$

for the increments δ_u and δ_q . Similarly as in the SSN system (4.36), the active set \mathcal{U}_γ is defined as $\mathcal{U}_\gamma := \{x \in \Omega : \gamma|\nabla u(x)| > 1\}$. Furthermore, as discussed in Section 5.5.1, the SSN iteration is modified using the property of the solution on the final active set where $q = \frac{\nabla u}{|\nabla u|}$ and $q \leq 1$ a.e. in Ω .

As example, we report the results computed in [DLRS13] regarding the estimation of the optimal parameter λ^* in (5.36) for a noisy image distorted by Gaussian noise with zero mean and variance 0.02. The optimisation result is reported in Figure 5.5 and has been obtained with the following choice of the parameter values $\varepsilon = 1e - 12$, $\gamma = 100$ and $h = 1/177$.

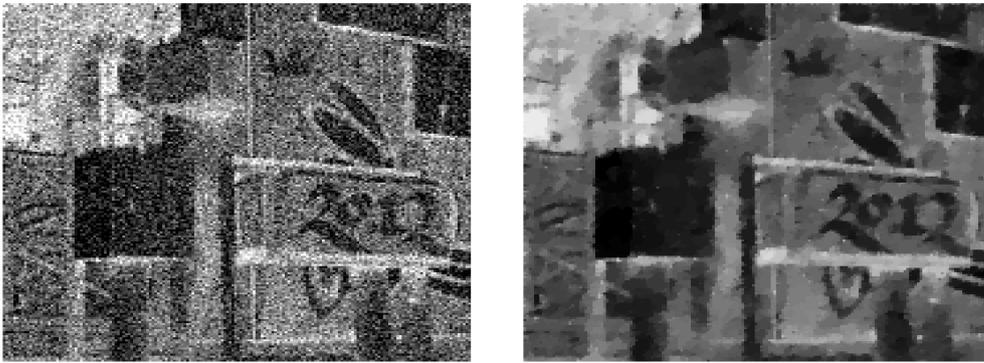


Figure 5.5: Noisy (left) and optimal denoised (right) image. Noise variance: 0.02. Optimal parameter $\lambda^* = 1770.9$. Image from [DLRS13].

As observed by the authors in other experiments, increasing the variance of the noise changes significantly the value of the optimal parameter λ^* (see [DLRS13, Section 4.1]). This can be easily explained: as the image becomes noisier less information that can be directly obtained. In that case, the TV-regularisation term plays an increasingly important role.

In order to assess the optimality of the λ^* computed with respect to standard quality measure used in imaging, such as SNR (5.5), let us consider the 80×80 pixel bottom left corner of the noisy image in Figure 5.5. In Figure 5.6 the values of the cost functional and of the SNR for parameter values between 150 and 1200 are plotted. In both graphs, the values corresponding to the computed optimal parameter $\lambda^* = 885.5$ are shown with a cross. As expected, the two optima agree. Here $h = 1/80$.

Impulse noise

Following [DLRS13, Section 4.4], we now report the results for the estimation of the optimal fidelity weight λ^* in the case when impulse noise is assumed in the data. The

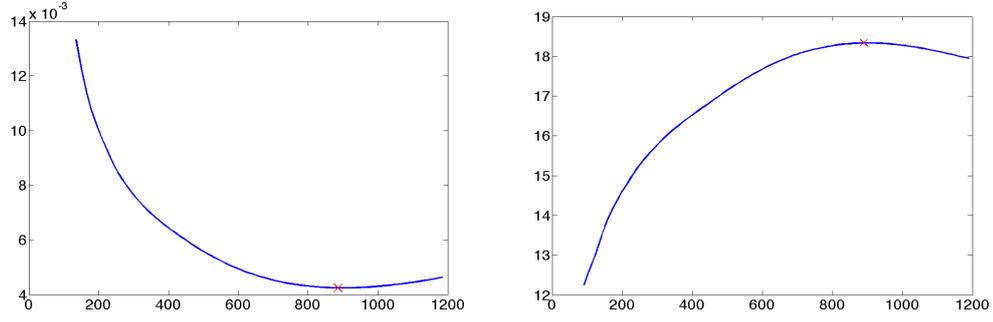


Figure 5.6: Plot of the cost functional value (left) and the SNR (right) vs. the parameter λ . Parameters: the input is the 80×80 pixel crop of the bottom left corner of the noisy image in Figure 5.5, $h = 1/80$, $\gamma = 100$, $\varepsilon = 1e - 12$. The red cross in the plot corresponds to the optimal $\lambda^* = 885.5$.

appropriate data term in this case is $\phi(u, f) = |u - f|$ and the bilevel problem reads:

$$\min_{\lambda \geq 0} \frac{1}{2} \|u - f^0\|_{L^2(\Omega)}^2 \quad (5.36a)$$

subject to:

$$u = \arg \min_v \left\{ \frac{\varepsilon}{2} \|\nabla v\|_{L^2}^2 + |Dv|_{\gamma}(\Omega) + \lambda \int_{\Omega} |v - f| \, dx \right\}. \quad (5.36b)$$

The presence of the L^1 fidelity term introduces a further non-differentiability obstacle in the lower level problem (5.36b). Therefore, similarly as in Section 4.5.1 we introduce a further Huber regularisation for such term depending also on a parameter $\gamma \gg 1$. Expressed in primal-dual form, the optimality conditions for (5.36b) read:

$$\begin{cases} -\varepsilon \Delta u - \nabla \cdot q + \lambda p = 0, \\ q = \left(\frac{\gamma \nabla u}{\max(\gamma |\nabla u|, 1)} \right), \\ p = \frac{\gamma (u-f)}{\max(\gamma |u-f|, 1)}. \end{cases}$$

Analogously as before, the modified SSN iteration for the increments δ_u, δ_q and δ_p reads in this case:

$$\begin{cases} -\varepsilon \Delta \delta_u - \nabla \cdot \delta_q + \lambda \delta_p = -(-\varepsilon \Delta u - \nabla \cdot q + \lambda p), \\ \delta q - \frac{\gamma \nabla \delta u}{\max(1, \gamma |\nabla u|)} + \chi_{\mathcal{U}_{\gamma}} \gamma^2 \frac{\nabla u^{\top} \nabla \delta u}{\max(1, \gamma |\nabla u|)^2} \frac{q}{\max(1, |q|)} = -(q - \frac{\gamma \nabla u}{\max(\gamma |\nabla u|, 1)}), \\ \delta p - \frac{\gamma \delta u}{\max(1, \gamma |u-f|)} + \chi_{\mathcal{S}_{\gamma}} \gamma^2 \frac{(u-f) \delta u}{\max(1, \gamma |u-f|)^2} \frac{p}{\max(1, |p|)} = -(p - \frac{\gamma (u-f)}{\max(1, \gamma |u-f|)}), \end{cases}$$

where \mathcal{U}_{γ} is defined as above and, similarly, \mathcal{S}_{γ} is the set $\mathcal{S}_{\gamma} = \{x \in \Omega : \gamma |u(x) - f(x)| > 1\}$.

In Figure 5.7 we report the denoising result for a noisy image with a percentage of missing pixels $d = 10\%$. For this numerical test $\varepsilon = 1e - 10$, $\gamma = 1e4$ and $h = 1/201$.



Figure 5.7: Noisy image (left) and optimal denoised (right) image. Density of missing pixels $d = 10\%$. Optimal parameter $\lambda^* = 89.32$.

Poisson noise

We now consider the case of images corrupted by Poisson noise. In this case an appropriate data fidelity is the Kullback-Leibler-type fidelity $\phi(u, f) = u - f \log u$, which imposes the additional constraint for u to be nonnegative. Numerically, we enforce this constraint by using a standard penalty method and solve:

$$\min_{\lambda \geq 0} \frac{1}{2} \|u - f^0\|_{L^2(\Omega)}^2,$$

subject to:

$$u = \arg \min_v \left\{ \frac{\varepsilon}{2} \|\nabla v\|_{L^2}^2 + |Dv|_\gamma(\Omega) + \lambda \int_\Omega (v - f \log v) dx + \frac{\rho}{2} \|\min(v, 0)\|_{L^2}^2 \right\}, \quad (5.37)$$

where $\rho \gg 1$ is a parameter weighting the relaxation of the positivity constraint. Again, we write the primal-dual form of the optimality condition for (5.37) which reads :

$$\begin{cases} -\varepsilon \Delta u - \nabla \cdot q + \lambda \left(1 - \frac{f}{u}\right) + \rho \chi_{\mathcal{T}} u = 0, \\ q = \frac{\gamma \nabla u}{\max(\gamma |\nabla u|, 1)}. \end{cases} \quad (5.38)$$

Here, \mathcal{T} is the active set $\mathcal{T} := \{x \in \Omega : u(x) < 0\}$. Similarly as before we now use the system above to design the following modified SSN iteration for solving (5.38):

$$\begin{cases} -\varepsilon \Delta \delta u - \nabla \cdot \delta q + \lambda \frac{f}{u^2} \delta u + \rho \chi_{\mathcal{T}} \delta u = -(-\varepsilon \Delta u - \nabla \cdot q + \lambda \left(1 - \frac{f}{u}\right) + \rho \chi_{\mathcal{T}} u), \\ \delta q - \frac{\gamma \nabla \delta u}{\max(1, \gamma |\nabla u|)} + \chi_{\mathcal{U}_\gamma} \gamma^2 \frac{\nabla u^\top \nabla \delta u}{\max(1, \gamma |\nabla u|)^2} \frac{q}{\max(1, |q|)} = -(q - \frac{\gamma \nabla u}{\max(\gamma |\nabla u|, 1)}), \end{cases}$$

for the increments δu and δq . As above, here \mathcal{U}_γ is the set $\mathcal{U}_\gamma = \{x \in \Omega : \gamma |\nabla u(x)| > 1\}$.

Figure 5.8 shows the optimal denoising result for the exemplary image of the planet Saturn¹ with $\lambda^* = 1013.76$.

¹Source: <http://nssdc.gsfc.nasa.gov/image/planetary/saturn/saturn.jpg>



Figure 5.8: Optimal Poisson denoising: original (left), noisy (center) and optimal denoised (right) image. Parameters: $\gamma = 1e3$, $\varepsilon = 1e - 10$, $h = 1/128$, $\rho = 1e4$. Optimal weight: $\lambda^* = 1013.76$.

Multiple constraints

In this section we present the numerical results obtained using the dynamic sampling Algorithm 1 applied to compute the numerical solution of (5.9) for a database of large size $N \gg 1$.

As a toy example, we consider the case when the noise in the images is normally distributed. In (5.9), this reflects in the estimation of just one parameter λ that weights the fidelity term $\phi(u, f_k) = (u - f_k)^2$ in each constraint. Considering the training database $\{(f_k^0, f_k)\}_{k=1, \dots, N}$ of clean and noisy images, the problem reduces to:

$$\min_{\lambda \geq 0} \frac{1}{2N} \sum_{k=1}^N \|u_k - f_k^0\|_{L^2(\Omega)}^2 \quad (5.39)$$

where, for each k , u_k is the solution of the Huber-regularised version of the PDE (5.9b):

$$-\varepsilon \Delta u_k - \nabla \cdot \left(\frac{\gamma \nabla u_k}{\max(\gamma |\nabla u_k|, 1)} \right) + \lambda (u_k - f_k) = 0, \quad k = 1, \dots, N. \quad (5.40)$$

Again, we use the characterisation of the gradient of the cost functional provided by Theorem 5.5.2 to compute the adjoint states p_k . Recalling also equations (5.31)-(5.32), we then apply Algorithm 1 to solve the multiple constrained problem.

In the following example the noise in the images has normal distribution $\mathcal{N}(0, 0.005)$. The parameter θ of the Algorithm 1, is chosen to be $\theta = 0.5$. See Figure 5.4 and Table 5.1 for more details. Figure 5.9 shows an example of database of brain images² together with the optimal denoised version obtained by Algorithm 1 for single Gaussian noise estimation.

²OASIS online database: <http://www.oasis-brains.org/>.

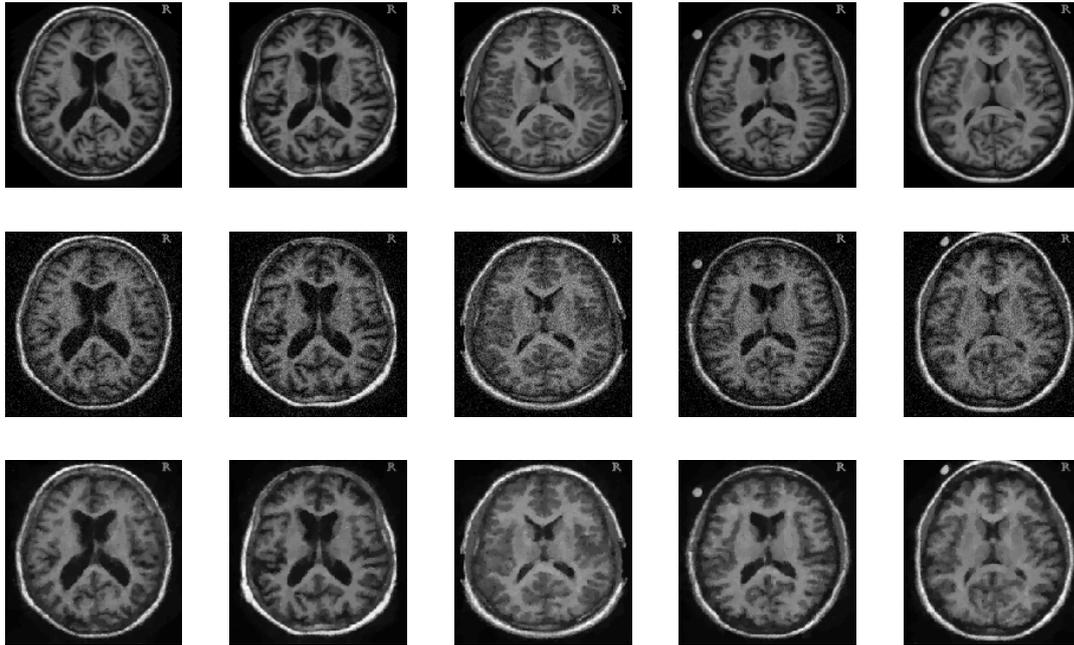


Figure 5.9: Sample of 5 images of the OASIS MRI brain database: original images (upper row), noisy images (middle row) and denoised images with optimal parameter $\lambda_S^* = 3306.4$ computed with dynamic sampling Algorithm 1 (bottom row).

Dynamic sampling: discussion. Table 5.2 shows the numerical value of the optimal parameter λ^* when varying the size of the dictionary. We measure the efficiency of the algorithms used in terms of the number of nonlinear PDEs solved during the BFGS optimisation and we compare the efficiency of solving the bilevel problem without and with the dynamic sampling strategy. We observe a clear improvement in efficiency when using dynamic sampling: the number of PDEs solved in the optimisation process is very much reduced. We note that this corresponds to an increasing number of BFGS iterations which does not appear to be an issue as BFGS iterations are themselves very fast. For the sake of computational efficiency, what really matters is the number of PDEs that need to be solved in *each* iteration of BFGS. Moreover, thanks to modern parallel computing methods and to the decoupled nature of the constraints in each BFGS iteration, solving such a reduced amount of PDEs makes the computational efforts very reasonable. In fact, we note that the size of the sample is generally maintained very small in comparison to N or just slightly increased. Computing also the relative error between the optimal parameter computed by solving all the PDEs and the one computed with dynamic sampling method, we note a good level of accuracy: the difference between the two values remains below 5%.

N	λ^*	λ_S^*	$ S_0 $	$ S_{end} $	eff.	eff. DS	BFGS its.	BFGS-DS its.	diff.
10	3334.5	3427.7	2	3	140	84	7	21	2.7%
20	3437.0	3475.1	4	4	240	120	7	15	1.1%
30	3436.5	3478.2	6	6	420	180	7	15	1.2%
40	3431.5	3358.3	8	9	560	272	7	16	2.1%
50	3425.8	3306.4	10	10	700	220	7	11	3.5%
60	3426.0	3543.4	12	12	840	264	7	11	3.3%
70	3419.7	3457.7	14	14	980	336	7	12	1.1%
80	3418.1	3379.3	16	16	1120	480	7	15	< 1%
90	3416.6	3353.5	18	18	1260	648	7	18	2.3%
100	3413.6	3479.0	20	20	1400	520	7	13	1.9%

Table 5.2: N is the size of the database, λ^* is the optimal parameter for (5.39)-(5.40) computed by solving all the N constraints, whereas λ_S^* is computed with Algorithm 1. The initial size S_0 is chosen to be $|S_0| = 20\%N$. $|S_{end}|$ of the sample at the end of the optimisation algorithm. Efficiency of the algorithms is measured in terms of the number of PDEs solved. Accuracy of the approximation is computed in terms of the difference $\|\lambda_S^* - \lambda^*\|_1 / \|\lambda_S^*\|_1$.

Robustness. To test the generalising power of the learning model, we report here the results presented in [CCDLR⁺15] and check the robustness of our learning approach applied to the Berkeley segmentation data set BSDS300³, see Fig. 5.10. We apply artificial single Gaussian noise of variance $\sigma^2 = 0.01$ to the images in the data set and then split it into two halves, each of 100 images. We learned the regularisation parameter α for each half individually, and then denoised the images of the other half with that parameter. The results for TV regularisation with L^2 cost and fidelity are in Table 5.3. As can be seen, the parameters learned using each half, hardly differ. The average PSNR, when denoising using the optimal parameter, learned using the same half, and the parameter learned from the other half, also differs insignificantly. We can therefore conclude that the bilevel learning model has good generalising power.

³Berkeley database, available online at: <http://www.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/BSDS300/html/dataset/images.html>

Validation	Learning	α	Average PSNR
1	1	0.0190	31.3679
1	2	0.0190	31.3672
2	1	0.0190	31.2619
2	2	0.0190	31.2612

Table 5.3: Cross-validated computations on the BSDS300 data set split into two halves, both of 100 images. Total variation regularisation with L^2 cost and fidelity. ‘Learning’ and ‘Validation’ indicate the halves used for learning α , and for computing the average PSNR and SSIM, respectively. Noise variance $\sigma = 0.01$.

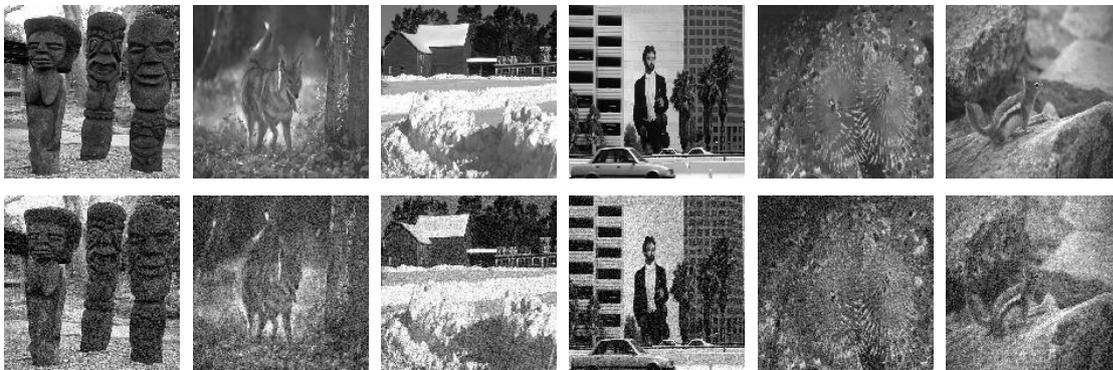


Figure 5.10: Noise-free and noisy versions of images from the Berkeley database. The Gaussian noise distribution is 0 mean and variance $\sigma^2 = 0.01$.

Distributed parameters

We conclude this section on single noise estimation with an example of spatially-dependent λ reported in [CDLR] and previously considered in [DHRC11, BDH13]. There, the parameter space is chosen to be the space $V = \{v \in H^1(\Omega) : \partial_n v|_\Gamma = 0 \text{ on } \partial\Omega\}$. Again, we consider a single Gaussian noise model. For this example the noisy image is distorted non-uniformly, compare Figure 5.11: a Gaussian noise with zero mean and variance 0.04 is present on the whole image and an additional noise with variance 0.06 is added on the area enclosed with a red line.

Since the spatially dependent parameter does not allow to get rid of the positivity constraints in an automatic way, the optimality system is solved by means of a SSN method combined with a Schwarz domain decomposition method. Specifically, the image domain is decomposed first and then Newton algorithm is applied in each subdomain. The detailed numerical performance of this approach is reported in [CDLR].

The results shown in Figure 5.11 have been obtained with the following choice of parameters $\varepsilon = 1e - 16$, $\gamma = 25$ and $h = 1/500$. The values of λ^* on whole image domain

are between 100 to 400. From the right image in Figure 5.11 we can see the dependence of the optimal parameter λ^* on the distribution of noise. As expected, at the high-level noise area in the input image, values of λ^* are lower (darker area) than in the rest of the image.



Figure 5.11: Noisy image (left), denoised image (center) and intensity of λ^* (right). Image from [CDLR].

5.7.2 Multiple noise estimation

In this section we focus on the case when the bilevel optimisation model is used to estimate the optimal modelling and data fitting for images corrupted by multiple noise statistics. Similarly as in (5.9b), we will consider a linear combination of data fidelities ϕ 's modelling each one specific fitting with particular statistics on the data (i.e. Gaussian, impulse or Poisson). The use of alternative combinations of data fidelities discussed in Chapter 4 and applied in this framework is a matter of future research. A preliminary result in this direction is the optimality system formally derived in Section 5.5.3, but further research needs to be done for its rigorous justification and numerical realisation.

Gaussian and impulse noise

We consider the case where a combination of Gaussian and impulse noise is present and a training database $\{(f_k^0, f_k)\}$ of images, with $k = 1, \dots, N$, $N \gg 1$ is considered. The fidelity term for the impulse distributed component is $\phi_1(u, f_k) = |u - f_k|$, whereas, as above, for the Gaussian noise we consider the fidelity $\phi_2(u, f_k) = (u - f_k)^2$, for every k . Each fidelity term is weighted by a parameter $\lambda_i, i = 1, 2$. Thus, we aim to solve:

$$\min_{(\lambda_1, \lambda_2), \lambda_i \geq 0} \frac{1}{2N} \sum_{k=1}^N \|u - f_k^0\|_{L^2(\Omega)}^2 \quad (5.41)$$

where, for each k , u_k is now the solution of the regularised PDE:

$$\begin{cases} -\varepsilon\Delta u - \nabla \cdot q + \lambda_1 h_\gamma^1(u - f_k) + \lambda_2(u - f_k) = 0, \\ q = \frac{\gamma \nabla u}{\max(\gamma |\nabla u|, 1)}, \\ p = \frac{\gamma (u - f)}{\max(\gamma |u - f|, 1)}, \quad k = 1, \dots, N, \end{cases} \quad (5.42)$$

where, similarly as before, we consider a Huber-regularisation of both ∇u of the $|u - f_k|$ term (denoted by h_γ^1), for every k . The adjoint states are computed through the optimality system (5.18) and then used as before for the computation of the gradient of the cost functional. We compute our results using again the dynamic sampling Algorithm 1 to solve (5.41)-(5.42) with the choice $\varepsilon = 1e - 12$, $\gamma = 100$ and $\theta = 0.5$.

We take as example slices of the brain database shown in Figure 5.9 corrupted with both Gaussian noise distributed as $\mathcal{N}(0, 0.005)$ and impulse noise with fraction of missing pixels $d = 5\%$, and again solve (5.9a)-(5.9b) by solving the PDE constraints all at once and by using dynamic sampling for different N . In Table 5.4 we report the results for the estimation of λ_1 and λ_2 for different values of N , thus confirming the effectiveness of the dynamic sampling strategy also in this case. In Figure 5.12 we show the denoising result for one image in the database of brain slices computed in correspondence with the optimal parameters.

N	λ_{1_S}	λ_{2_S}	$ S_0 $	$ S_{end} $	eff.	eff. Dyn.S.	diff.
10	86.31	28.43	2	7	180	70	5.2%
20	90.61	26.96	4	6	920	180	5.3%
30	94.36	29.04	6	7	2100	314	5.6%
40	88.88	31.56	8	8	880	496	1.2%
50	88.92	29.81	10	10	2200	560	< 1%
60	89.64	28.36	12	12	1920	336	1.9%
70	86.09	28.09	14	14	2940	532	3.3%
80	87.68	29.97	16	16	3520	448	< 1%

Table 5.4: λ_{1_S} and λ_{2_S} are the optimal weights for (5.41)-(5.42) estimated with dynamic sampling Algorithm 1. We observe again a clear improvement in efficiency (i.e. number of PDEs solved). As above, $|S_0| = 20\%N$ and $\theta = 0.5$.

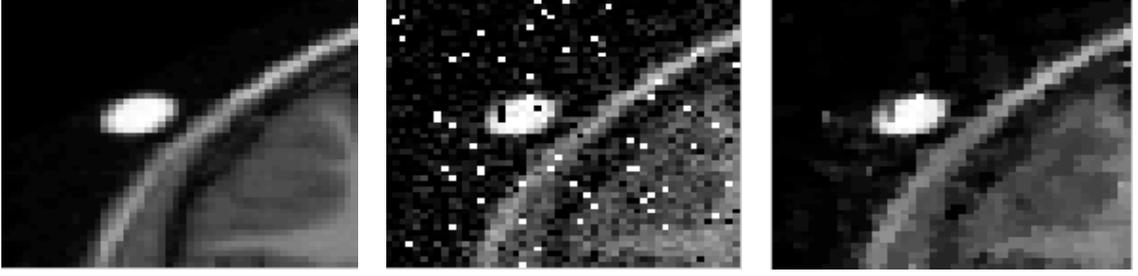


Figure 5.12: Noise-free and noisy version of a slice of the database in Figure 5.9 and optimal denoising result with $\lambda_1^* = 88.92$ and $\lambda_2^* = 29.81$.

Gaussian and Poisson noise

We consider now the case where a combination of Gaussian and Poisson noise is corrupting the data for the easier case of a single image in order not to make this section too heavy. As before, the Gaussian fidelity reads $\phi_1(u, f) = (u - f)^2$, whereas a KL-type term $\phi_2(u, f) = (u - f \log u)$ is used for the Poisson noise modelling. This combined model has been considered in [DLRS13] and, although producing good results, does not seem to show statistical consistency with the noise in the data, compare Chapter 4. We believe that a modelling of the form 4.3.2 may be preferable, but its use in a learning context presents some difficulties as briefly discussed in Section 5.5.3 and it is indeed a matter of future research. The problem then consists to determine the optimal weighting (λ_1, λ_2) solving the model:

$$\min_{(\lambda_1, \lambda_2), \lambda_i \geq 0} \|u - f^0\|_{L^2(\Omega)}^2$$

subject to the Huber-regularised TV constraint:

$$u = \arg \min_v \left(\frac{\varepsilon}{2} \|\nabla v\|_{L^2(\Omega)}^2 + |Dv|_{\gamma}(\Omega) + \frac{\lambda_1}{2} \|v - f\|_{L^2(\Omega)}^2 + \lambda_2 \int_{\Omega} (v - f \log v) dx + \frac{\rho}{2} \|\min(v, 0)\|_{L^2(\Omega)}^2 \right),$$

for the training pair (f_0, f) corrupted by Gaussian and Poisson noise and where, again, $\rho \gg 1$ is a penalty parameter enforcing the positivity constraint on u . We express the corresponding optimality condition in a primal-dual form as before:

$$\begin{cases} -\varepsilon \Delta u - \nabla \cdot q + \lambda_1(u - f) + \lambda_2 \left(1 - \frac{f}{u}\right) + \rho \chi_{\mathcal{T}} u = 0, \\ q = \frac{\gamma \nabla u}{\max(\gamma |\nabla u|, 1)}, \end{cases} \quad (5.43)$$

where, as before, \mathcal{T} is the active set $\mathcal{T} := \{x \in \Omega : u(x) < 0\}$. We design again a modified SSN iteration for (5.43) and compute the optimal parameters using BFGS update based on the computation of the gradient using adjoint information. Results are reported in Figure 5.13. For this numerical test, the image has been corrupted with a combination

of Poisson noise and Gaussian noise with variance 0.005 parameters have been chosen as: $\varepsilon = 1e - 10, \gamma = 1e4$ and $\rho = 1e4$.

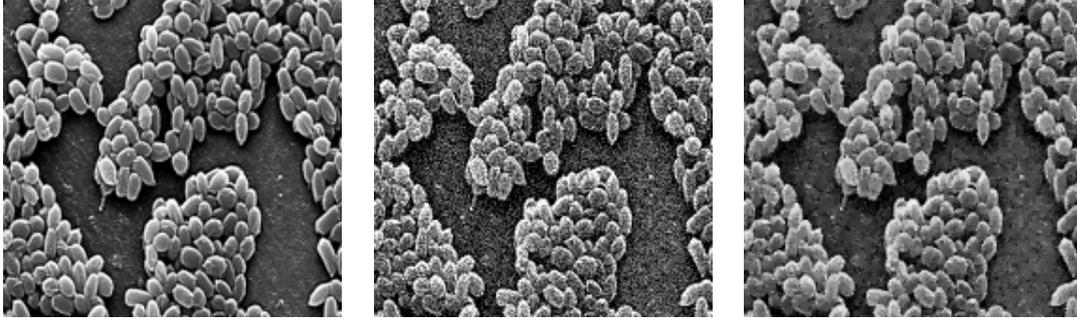


Figure 5.13: Optimal Gaussian-Poisson denoising: original (left), noisy (center) and optimal denoised (right) image. Parameters: $\gamma = 1e4, \varepsilon = 1e - 10, h = 1/170, \rho = 1e4$. Optimal weights: $\lambda_1^* = 761.3, \lambda_2^* = 165.1$.

Chapter 6

Graph clustering and object measurement in images

In this last chapter, we look at TV regularisation PDE models from a completely different point of view. Namely, we present a TV-based model in the discrete framework of graphs and exploit similarities between the vertices of the image graph in terms of their distinctive characteristics (features) for the solution of the problem of image segmentation.

Classical mathematical methods for image segmentation are mainly formalised in terms of variational problems in which the segmented image is a minimiser of an appropriate energy. The most common image feature encoded in such energies is the magnitude of the image gradient, detecting regions (or contours) where sharp variations of the intensity values occur. Examples include the Mumford-Shah segmentation approach [MS89], the snakes and geodesic active contour models [KWT88, CKS97]. Moreover, in [CV01] Chan and Vese proposed an instance of the Mumford-Shah model for piecewise constant images whose energy is based on the mean grey values of the image inside and outside of the segmented region rather than the image gradient and hence does not require strong edges for segmentation. In [CSL00] the extension to vector-valued images is considered.

The nonsmoothness of most of the segmentation energies renders their numerical minimisation usually difficult. In the case of the Mumford-Shah segmentation model the numerical realisation is additionally complicated by its dependency on the image function as well as the object contour. To overcome this, several regularisation methods and approximations have been proposed in the literature, e.g. [AT90, Bra98, BDM97, VC97]. In the context of TV-based segmentation models, the Ginzburg-Landau functional has a particularly important role. Originally considered for the modelling of physical phenomena such as phase transition and phase separation (see [BS96] for a survey on the topics) such functional is often used in imaging for approximating the TV energy, compare [ET06, Ese04, ES02], which relate to previous works by Ambrosio and Tortorelli on diffuse interface approximation models [AT90, AT92].

6.1 A review on classical image segmentation models

We review in this section three different approaches classically used for image segmentation problems: the Canny method for edge detection [Can86], the Mumford-Shah energy minimisation [MS89] and the Chan-Vese model [CV01].

6.1.1 Canny edge detection

One of the first approaches for image segmentation considered by Canny in [Can86] is an edge detection algorithm which, by reducing the amount of information in an image, can still preserve its main structural and geometrical properties (edges) for further partitioning of the image into regions. This algorithm is still used a lot in practice. It consists of five steps.

- (i) **Denoising:** pre-processing step where possible blur/noise in the image is removed via Gaussian filtering. In the following, we use the Rudin, Osher and Fatemi TV denoising model [ROF92].
- (ii) **Large magnitude gradient detection:** the regions in the image where the gradient has large magnitude are detected, thus potential edges are identified.
- (iii) **Suppression:** only local maxima in the gradient image are marked as edges, while all the other points are discarded. In this step the algorithm thin the blurred edges detected in Step 2 in order to get sharper edges.
- (iv) **Double thresholding:** this step refines the suppression Step 3 by using two thresholding values to identify *strong* edges (i.e. with gradient magnitude larger than the higher threshold), *weak* edges (gradient magnitude between the two thresholds) and to discard false edges (gradient magnitude lower than the lower threshold).
- (v) **Edge tracking by hysteresis:** false edges not connected to strong edges from the previous step are discarded. Only the final, potential edges are maintained and the final edge set is determined.

Despite being very easy and widely applicable, the Canny algorithm described above is not robust with respect to noise or to the presence of texture.

6.1.2 The Mumford-Shah functional

In order to identify edges of an image, a simplified approach consists in approximating the original image by a piecewise-regular function which presents changes in its regularity in correspondence with its edges. By constructing such an approximation, one would in fact construct the desired segmentation. Mumford and Shah proposed in 1989 a variational method that relies on this simplification. Denoting by $f : \Omega \subset \mathbb{R}^2 \rightarrow \mathbb{R}$ the given greyscale

image to segment, the Mumford-Shah model [MS89] aims to approximate the image f with a piecewise-smooth function u which solves the minimisation problem

$$\min_{u,C} \mathcal{J}(u,C) := \mu \text{Length}(C) + \lambda \int_{\Omega} (f(x) - u(x))^2 dx + \int_{\Omega \setminus C} |\nabla u(x)|^2 dx \quad (6.1)$$

for positive parameters $\mu, \lambda > 0$. In (6.1) the piecewise-smooth function u is allowed to have discontinuities in correspondence with the edge set curve C . The first term of the functional prohibits C to degenerate, the second one is a classical L^2 fidelity term and the third one ensures the differentiability of u outside C . Minimising \mathcal{J} corresponds to select the edge set C as the segmentation boundary. Due to the lack of differentiability of \mathcal{J} with respect to appropriate norms, the classical Euler-Lagrange equations approach turns out not to be applicable. Several approximation methods have been considered in the literature in order to deal with smoother functionals $\mathcal{J}_{\varepsilon}, \varepsilon > 0$ defined on more regular Sobolev spaces and not depending on the edge set C , compare [AT90, Bra98, BDM97] where also Γ -limits as $\varepsilon \rightarrow 0$ are studied. Such smoother versions of \mathcal{J} are used for the numerical solution of (6.1) (see [VC97]), but very often turn out to be quite complicated and computationally expensive. For this reason, simplified models have been considered. The given image f may be approximated, for instance, by a function u which is piecewise *constant* in each connected component of $\Omega \setminus C$. In this case, the minimisation problem reads

$$\min_{u,C} \tilde{\mathcal{J}}(u,C) := \mu \text{Length}(C) + \lambda \int_{\Omega} (f(x) - u(x))^2 dx. \quad (6.2)$$

In this situation, the edge set curve C needs to be necessarily the boundary of a closed set. Namely, C is going to be composed of closed curves. From an analytical point of view, the model (6.2) is nonconvex. Nonetheless, existence of its solution is proved in [MS89].

6.1.3 The Chan-Vese model

Another well-known model is the Chan-Vese model presented in [CV01] which can be considered as a simplification of the the Mumford-Shah model where the image u we seek is assumed *a priori* to have only two values:

$$u(x) = \begin{cases} c_1, & \text{if } x \text{ is inside } C, \\ c_2, & \text{if } x \text{ is outside } C. \end{cases} \quad (6.3)$$

In the definition above, C is the boundary of a closed set. The variational Chan-Vese model seeks the optimal c_1 and c_2 (i.e. the optimal u of the form (6.3)) and contour C to solve the following problem:

$$\min_{c_1, c_2, C} \mathcal{F}(c_1, c_2, C) := \mu \text{Length}(C) + \nu \text{Area}(\text{int}(C))$$

$$+ \lambda_1 \int_{\text{int}(C)} (f(x) - c_1)^2 dx + \lambda_2 \int_{\text{ext}(C)} (f(x) - c_2)^2 dx \quad (6.4)$$

where $\mu, \nu, \lambda_1, \lambda_2 > 0$. In (6.4) the first term penalises the length of C , ensuring its regularity. The size of C is controlled by the second term which is penalising the area in the interior of C , while the two other terms penalise the discrepancy between the fitting of the piecewise-constant model u and the given image f in the interior/exterior of C , respectively. Computing the local minima of (6.4) one retrieves the optimal binary approximations u of f . Existence and regularity of solutions of (6.4) are proved using similar results as the ones reported in [MS12] for the solution of a simplified Mumford-Shah model. An equivalent formulation of (6.4) in terms of level sets representing C can also be considered. Namely, one writes that C is the zero level set of a Lipschitz function $\psi(x)$. Then, the energy function \mathcal{F} in (6.4) is expressed as function of ψ as

$$\begin{aligned} \mathcal{F}(c_1, c_2, \psi) = & \mu \int_{\Omega} \delta(\psi(x)) |\nabla \psi(x)| dx + \nu \int_{\Omega} H(\psi(x)) dx \\ & + \lambda_1 \int_{\Omega} (f(x) - c_1)^2 H(\psi(x)) dx + \lambda_2 \int_{\Omega} (f(x) - c_2)^2 (1 - H(\psi(x))) dx, \end{aligned} \quad (6.5)$$

where H denotes the Heaviside function and δ its distributional derivative, i.e. the Dirac mass. For numerical implementations, H and δ are generally regularised. In order to find the minima of (6.5) an alternating strategy is used, updating c_1, c_2 and ϕ successively in each iteration. The level set representation of C is initially constructed as

$$\psi(x) = \begin{cases} +1, & \text{if } x \text{ is inside } C, \\ -1, & \text{if } x \text{ is outside } C. \end{cases}$$

Thus, the inside and the outside of C are distinguished by the sign of ψ . Several extensions of the Chan-Vese model have been considered in literature. In particular, in [CSL00] the model is extended for vector-valued data, the only difference with (6.4) being the use of Euclidean vector norms in the integrands.

Despite being very popular and widely used in applications, the Chan-Vese model and its extensions present intrinsic limitations. Firstly, due to the lack of uniqueness of the model, the segmentation result is strongly dependent on the initial condition. Different initialisations can lead to different minima, so the segmentation result may vary significantly. This is a significant drawback in terms of automation of the method, which would require a careful initialisation of the contour. Secondly, due to the modelling assumption (6.3), the Chan-Vese model works well for images characterised by intensity homogeneity since, by assumption, the intensity in each region is assumed to be constant. If this is not the case, the evolving contour curve C may be misleadingly guided by inappropriate image information.

6.2 Image segmentation as graph clustering

Variational image segmentation still faces many problems in the presence of low contrast and the absence of clear boundaries separating regions. Their main drawback is that these methods are limited to consider image features which can be mathematically formalised in terms, e.g., of an image gradient and encoded within a segmentation energy. In recent years dictionary based methods have become more and more popular in the image processing community, complementing more classical variational segmentation methods. By learning the distinctive features of the region to be segmented from examples provided by the user, these methods are able to segment the desired regions in the image correctly.

In the following, we consider the method proposed in [BF12, GMB⁺14b, GMB⁺14a] for image segmentation and labelling. This approach goes beyond the standard variational approach in two respects. Firstly, the model is set up in the purely discrete framework of graphs. This is rather unusual for variational models where one normally considers functionals and function spaces defined on subdomains of \mathbb{R}^2 in order to exploit properties and tools from convex and functional analysis and calculus of variations. Secondly, the new framework allows for more flexibility in terms of the features considered. Additional features like texture, light intensity or others, can be considered as well without encoding them directly in the function space or in the regularity of the functions considered. However, due to the possibly very large size of the image (nowadays of the order of megapixels for professional cameras) and the large number of features considered, the construction of the problem may be computationally expensive and often requires reduction techniques [Nys29, NDRM07, FBCM04].

In previous papers (see, e.g., [SM00, SW97, GM98]) graph segmentation problems were approached by means of the graph cut objective function. We rather follow [BF12] and present in the following the mathematical background for the design a graph segmentation approach based on the use of the Ginzburg-Landau (GL) functional. The image segmentation problem is then rephrased as a minimisation approach on a graph defined by appropriate features computed from the image.

6.2.1 The Ginzburg-Landau functional as approximation of TV

Several physical problems modelling phase transition and phase separation phenomena are built around the well-known GL functional:

$$GL(u) := \frac{\varepsilon}{2} \int_{\Omega} |\nabla u(x)|^2 dx + \frac{1}{\varepsilon} \int_{\Omega} W(u(x)) dx. \quad (6.6)$$

The functional above is defined in the continuous setting. Here, Ω represents an open subset of \mathbb{R}^d , $d = 2, 3$, $u : \Omega \rightarrow \mathbb{R}$ is the density of a two-phase material and $W(u)$ is a double-well potential, e.g. $W(u) = \frac{1}{4}(u^2 - 1)^2$. The two wells ± 1 of W correspond to the

two phases of the material. The parameter $\varepsilon > 0$ is the spatial scale. Variational models built around this functional are also referred to as *diffuse interface* models because of the interface appearing between the two regions containing the phases (i.e. the two wells of W) due to the competition between the two terms of the functional (6.6). Nonetheless, some smoothness preventing u from having jumps between the two regions is ensured by the first regularisation term.

The use of the GL functional has become very popular in image processing due to its connections with the TV seminorm. In [MM77a, MM77b], for instance, Γ -convergence properties of (6.6) to the TV functional are shown. Thus, the GL functional is very often used as a quadratic approximation of TV. Fast numerical schemes relying on these connections have been designed for many imaging problems, thus overcoming the issues related to nonsmooth TV minimisation [AT92, ET06, CV01]. The functional considered often is of the form

$$E(u) := GL(u) + \lambda \phi(u, u_0), \quad (6.7)$$

where $\phi(u, u_0)$ is the fidelity term measuring the distance of the reconstructed image u to the given image u_0 . Its choice depends on the applications considered, see Section 5.4 for more details. The parameter $\lambda > 0$ determines the influence of the data fit, compared to the regularisation. Taking the L^2 gradient descent of (6.7) we get the following evolutionary PDE, known in the literature as the Allen-Cahn equation [AC79] with an additional forcing term due to the fidelity ϕ :

$$\frac{\partial u}{\partial t} = -GL'(u) - \lambda \phi'(u) = \varepsilon \Delta u - \frac{1}{\varepsilon} W'(u) - \lambda \phi'(u, u_0). \quad (6.8)$$

Steady states of equation (6.8) are the stationary points of the energy E in (6.7). Note that E is not convex so uniqueness is not guaranteed and, consequently, the long time behaviour for solutions of (6.8) will depend on the initial condition. The linear diffusion term weighted by ε appearing in (6.8) allows for fast solvers using for instance the Fast Fourier Transform (FFT) which translates the Laplace operator into a multiplication operator on the Fourier modes.

Connections with binary segmentation In the following, we rely on the method presented in [BF12] for high-dimensional data classification on graphs relying on the use of the GL functional (6.7) and the corresponding gradient descent (6.8) which has been applied to several imaging problems [GMB⁺14b, GMB⁺14a], showing good performance and robustness. We consider the problem of binary image segmentation where we want to partition a given image into two components where each component is a set of pixels (also called a cluster, or a class) and represents a certain object or group of objects. Typically, some *a priori* information describing the object(s) we want to extract is given and serves as initial input for the segmentation algorithm. For image labelling, in [BF12] two images

are taken as input: the first one has been manually segmented in two classes and the objective is to automatically segment the second image using the information provided by the segmentation of the first one.

6.2.2 Towards the modelling: the graph framework

We revise in the following the main ingredients of the model considered and start from a quick review of concepts in graph theory.

We consider an image I as a collection of pixels arranged in a rectangle of size $S := N \times M$, i.e. $I := \{x \in \mathbb{Z}^2 : 0 \leq x_1 \leq N-1 \text{ and } 0 \leq x_2 \leq M-1\}$. For each $x \in I$, we define the image neighbourhood of x as the set $\mathcal{N}(x) := \{y \in I : |x_1 - y_1| \leq \tau \text{ and } |x_2 - y_2| \leq \tau\}$, with $\tau \in \mathbb{N}$ fixed, i.e. $\mathcal{N}(x)$ contains the pixels in a $(2\tau + 1) \times (2\tau + 1)$ sized square centred at x . For some appropriate $K \in \mathbb{N}$, we associate to every vertex $x \in I$ a vector $z \in \mathbb{R}^K$ encoding selected characteristics of the neighbourhood $\mathcal{N}(x)$. These characteristics are related to the grey or RGB intensity values as well as the texture features of the neighbourhood. In Section 6.2.5, we will explain in more detail our feature vector construction. The map $\psi : I \rightarrow \mathbb{R}^K, x \mapsto z$ is called the feature function. With an eye to constructing the feature vectors, it will be useful to associate a neighbourhood vector $\nu(x) := (x_j)_{j \in \mathcal{N}(x)} \in I^{(2\tau+1) \times (2\tau+1)}$ to each neighbourhood, such that the ordering of the x_j in $\nu(x)$ is consistent between pixels x , e.g., order the pixels from each square $\mathcal{N}(x)$ from left to right and top to bottom. The specific choice of ordering is not important, as long as it is consistent for each pixel neighbourhood.

Next, we construct a simple weighted undirected graph $G = (V, E, w)$ whose vertices correspond to the pixels in I and with edges whose weights depend on the feature function ψ . Let V be a vertex set of cardinality S . To emphasize that each vertex in V corresponds to exactly one pixel in I , we will label the vertex corresponding to $x \in I$ by x as well. Let $w : V \times V \rightarrow \mathbb{R}$ be a symmetric and nonnegative function, i.e.

$$w(x_i, x_j) = w(x_j, x_i), \quad \text{and} \quad w(x_i, x_j) \geq 0, \quad \text{for each } x_i, x_j \in V. \quad (6.9)$$

We define the edge set E as the collection of all undirected edges connecting nodes x_i and x_j for which $w(x_i, x_j) > 0$ [Chu97]. The function w restricted to $E \subset V \times V$ is then a positive edge weight function.

In our applications we define w as

$$w(x_i, x_j) := \hat{w}(\psi(x_i), \psi(x_j)) = \hat{w}(z_i, z_j),$$

where $\hat{w} : \mathbb{R}^K \times \mathbb{R}^K \rightarrow \mathbb{R}$ is a given function and ψ is the feature function.

In operator form, the **weight matrix** $W \in \mathbb{R}^{S \times S}$ is the nonnegative symmetric matrix whose elements are $w_{i,j} = w(x_i, x_j)$. In the following, we will not distinguish between the

two functions w and \hat{w} and, with a little abuse of notation, we will write $w(z_i, z_j)$ for $\hat{w}(z_i, z_j)$.

Remark 6.2.1. *Weight functions express the similarities between vertices and will be used in the following to partition V into clusters such that the sum of the edge weights between the clusters is small. There are many different mathematical approaches to attempt this partitioning. When formulated as a balanced cut minimisation, the problem is NP-complete [vL07], which inspired relaxations which are more amenable to computational approaches, many of which are closely related to spectral graph theory [SM00]. We refer the reader to [Chu97] for a monograph on the topic. The method we use in this paper can be understood (at least in spirit, if not technically, [vGB12, vGGOB14]) as a nonlinear extension of the linear relaxed problems.*

To solve the segmentation problem, we minimise a *discrete GL functional* (which is formulated in the graph setting, instead of the continuum setting), via a gradient descent method. In this setting the Laplacian in (6.8) will be a (negative) normalised graph Laplacian. We will use the spectral decomposition of u with respect to the eigenfunctions of this Laplacian and use the Nyström method to quickly compute this decomposition. Let us start introducing the graph Laplacian and graph GL functional.

The discrete operators

We now define the differential operators in the graph framework. For each vertex x , we consider now the **degree** of x ,

$$d : V \rightarrow \mathbb{R}, \quad d(x) := \sum_{y \in V} w(x, y).$$

In operator form, the **degree matrix** $D \in \mathbb{R}^{S \times S}$ is defined as the diagonal matrix having as elements $d_{i,i} = d(x_i)$.

A subset A of the vertex set V is **connected** if any two vertices in A can be connected by a path such that all the points of the path lie in A . A finite family of sets A_1, \dots, A_t is called a **partition** of the graph if $A_i \cap A_j = \emptyset$ for $i \neq j$ and $\bigcup_i A_i = V$.

We now have all the ingredients to define the **graph Laplacian**. Denoting by \mathcal{V} the space of all the functions $V \rightarrow \mathbb{R}$, the graph Laplacian is the operator $L : \mathcal{V} \rightarrow \mathcal{V}$ such that:

$$Lu(x) = \sum_{y \in V} w(x, y)(u(x) - u(y)), \quad x \in V. \quad (6.10)$$

We are considering a finite graph of size S , so real valued functions can be identified as vectors in \mathbb{R}^S . We can then write the graph Laplacian in matrix form as $L = D - W$. It is worth mentioning (see Remark 6.2.2 below) that this graph Laplacian is a

positive semidefinite operator. Note that by convention the sign of the discrete Laplacian is opposite to that of the (negative semidefinite) continuum Laplacian.

The associated quadratic form of L is

$$Q(u, Lu) := \frac{1}{2} \sum_{x,y \in V} w(x,y) (u(x) - u(y))^2. \quad (6.11)$$

The quadratic form Q can be interpreted as the energy whose optimality condition corresponds to the vanishing of the graph Laplacian (6.10).

Remark 6.2.2. *The operator L has S non-negative real-valued eigenvalues $\{\lambda_i\}_{i=1}^S$ which satisfy: $0 = \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_S$. The eigenvector corresponding to λ_1 is the constant S -dimensional vector $\mathbf{1}_S$, see [vL07].*

The operator in (6.10) is not the only graph Laplacian appearing in the literature. To set it apart from others, it is also referred to as the unnormalised or combinatorial graph Laplacian. Such operator can be related to the standard continuous differential one through nonlocal calculus [GO09a]. More precisely, the eigenvectors of L converge to the eigenvectors of the standard Laplacian, but in the large sample size limit a proper scaling of L is needed to guarantee stability of convergence to the continuum operator [BF12, GMB⁺14a]. Hence, we consider in the following the normalisation of L given by the symmetric graph Laplacian

$$L_s := D^{-1/2} L D^{-1/2} = I - D^{-1/2} W D^{-1/2}. \quad (6.12)$$

Clearly, the matrix L_s is symmetric. Other normalisations of L are possible, like, for instance the random walk graph Laplacian (see [Chu97, vL07, vGGOB14]).

In [SM00, Section 5] a quick review on the connections between the use of the symmetric graph Laplacian (6.12) and spectral graph theory is given. Computing the eigenvalues of the normalised symmetric Laplacian corresponds to the computation of the generalised eigenvalues used to compute normalised graph cuts in a way that the standard graph Laplacian may fail to do, compare [Chu97]. Normally, spectral clustering algorithms for binary segmentation base the partition of a connected graph on the eigenvector corresponding to the second eigenvalue of the normalised Laplacian, using, for example, k -means. For further details and a comparison with other methods we refer the reader to [SM00] and to [BF12, Section 2.3] where a detailed explanation on the importance of the normalisation of the Laplacian is given.

We want to formulate a SemiSupervised Learning (SSL) technique to label unknown elements of the graph considered by using a known and labelled set of vertices. A common approach is to learn a real valued function $u \in [-1, 1]$ and then threshold this output to determine the desired class of labels. In the following we will make use of the discrete

Ginzburg-Landau functional to do this. In particular, we will use it to determine a function u which will be approximately equal to one on a subset and equal to minus one on the other and which will present a transition region between the two. By thresholding such result, we will obtain the desired binary segmentation of the graph into the two desired classes.

Recalling (6.6) -(6.7) and (6.11), we define the discrete GL functional¹ as

$$GL_d(u) := \frac{\varepsilon}{2} Q(u, L_s u) + \frac{1}{\varepsilon} \sum_{x \in V} W(u(x)) + \sum_{x \in V} \frac{\chi(x)}{2} (u(x) - u_0(x))^2. \quad (6.13)$$

Here u_0 represents known data provided by the user and where the two classes have been labelled with the values one and minus one. As before, $W(u(x)) = \frac{1}{4}(u^2(x) - 1)^2$ is the double-well potential. The function $\chi : V \rightarrow \{0, 1\}$ is the characteristic function of the subset of labelled vertices $V_{lab} \subset V$, i.e. $\chi = 1$ on V_{lab} and $\chi = 0$ on $V_{unlab} := V_{lab}^c$. Hence, the corresponding fidelity term enforces the fitting between u and u_0 in correspondence of the the known labels on the set V_{lab} , while the labelling for the pixels in V_{unlab} is driven by the minimisation of the first two regularising terms in (6.13).

The corresponding ℓ^2 gradient flow for (6.13) reads

$$\frac{\partial u}{\partial t} = -\varepsilon L_s u - \frac{1}{\varepsilon} \sum_{x \in V} (u^3(x) - u(x)) - \sum_{x \in V} \chi(x)(u(x) - u_0(x)). \quad (6.14)$$

The idea is to design a semi-supervised learning (SSL) approach where *a priori* information for the set V_{lab} (i.e. cluster labels) is used to label the points in the set V_{unlab} . The comparison uses the weight function defined in (6.9) to build the graph by comparing the feature vectors at each point.

Remark 6.2.3 (The choice of the weight function). *As pointed out in [BF12, Section 2.5], the main criteria driving the choice of the weight function are the desired outcome (in the case of segmentation, this reflects on the choice of an appropriate metric on a vector space) and the computational efforts required to diagonalise the corresponding matrix W . A common weight function is the Gaussian function*

$$w(x, y) = \exp(-\|x - y\|^2/\sigma^2), \quad \sigma > 0, \quad x, y \in V, \quad (6.15)$$

where in our applications we choose the standard Euclidean norm. An alternative proposed by Zelnik-Manor and Perona in [LP05] introduces local saling weights, thus allowing spatial dependence of the parameter σ in (6.15) and is particularly useful in the case of segmentation when multiple scales need to be segmented simultaneously.

¹Discrete GL functional with a data fidelity term' would be a more accurate name, but we opt for brevity here.

Several approaches to SSL using graph theory have been considered in literature. In [CLL⁺05] Coifman proposes a geometric diffusion approach to learn classes based on some spectral properties. Gilboa and Osher propose in [GO07] a similar technique with a stiffness constraint due to an explicit forward time stepping modelled on the use of the graph Laplacian. Explicit methods tackling the problem using TV have been considered in [GO09a]. The approach presented here adapts fast algorithms available for the efficient minimisation of the continuous GL functional to the minimisation of the discrete one in (6.13). In particular, to overcome the high computational costs, we present in the following a splitting operator scheme and a matrix completion technique applied to our problem.

Towards the numerical realisation. The numerical strategy we intend to use is based on the following steps (see Section 6.2.5 for more details) :

- At each time step $n\Delta t$, $n \geq 0$, consider at every point the spectral decomposition of the numerical approximation $U_n = u(n\Delta t)$ with respect to the eigenvectors v_k of the operator L_s :

$$U_n(x) = \sum_k \alpha_n^k(x) v_k(x), \quad x \in V, \quad (6.16)$$

with coefficients α_n . Similarly, use spectral decomposition in the $\{v_k\}$ basis for the numerical approximation of the other nonlinear quantities appearing in (6.14).

- Having fixed the basis of eigenfunctions, the numerical approximation in the next time step U_{n+1} is computed by determining the new coefficients α_{n+1}^k in (6.16) for every k .

The numerical bottlenecks of this strategy appear to be the computation of the eigenvectors v_k of the operator L_s , which, in practice, can be computationally costly for large and non-sparse matrices W , and the efficient computation of the numerical approximation U_{n+1} given U_n . To mitigate the former problem, we use the Nyström extension (Section 6.2.3), whereas we consider a convex splitting strategy for the latter (Section 6.2.4). The numerical details and the pseudocode of the whole algorithm are presented in Section 6.2.5.

6.2.3 The Nyström extension for matrix completion

Following the detailed discussion in [BF12, Section 3.2], we present here the Nyström technique for matrix completion [Nys29] used in previous works by the graph theory community [FBCM04, FBCM02] and applied later to several imaging problems [NDRM07]. In our problem, the Nyström extension is used to find an approximation of the eigenvectors

v_k of the operator L_s . We will freely switch between the representation of eigenvectors (or eigenfunctions) as a real-valued functions on the vertex set V and as a vectors in \mathbb{R}^S .

Consider a fully connected graph with vertices V and the set of corresponding feature vectors $\psi(V) = \{z_i\}_{i=1}^S$. A vector v is an eigenvector of the operator L_s in (6.12) with eigenvalue λ if and only if v is an eigenvector of the operator $D^{-1/2}WD^{-1/2}$ with eigenvalue $1 - \lambda$, since

$$L_s v = v - D^{-1/2}WD^{-1/2}v = \lambda v \iff D^{-1/2}WD^{-1/2}v = (1 - \lambda)v. \quad (6.17)$$

Thus, finding the spectral decomposition of L_s boils down to diagonalise the operator $D^{-1/2}WD^{-1/2}$. This is not obviously easier, as the matrix W , despite being nonnegative and symmetric, may be large and non-sparse, so the computation of its spectrum may be computationally hard. Here, however, we take advantage of the Nyström extension. Given the eigenvalue problem

$$\text{find } \theta \in \mathbb{R} \text{ and } v : V \rightarrow \mathbb{R}, v \neq 0 \quad \text{s. t.} \quad \sum_{x \in V} w(x, y) v(x) = \theta v(y), \quad (6.18)$$

for every point $y \in V$, we approximate the sum on the left hand side of the equation above using a standard quadrature rule where the interpolation points are chosen by randomly selecting a subset of L points from the set V and the interpolation weights are chosen correspondingly. The Nyström extension for (6.18) then

$$\text{approximates (6.18) by } \sum_{i=1}^L w(y, x_i) v(x_i) \approx \sum_{x \in V} w(y, x) v(x) = \theta v(y), \quad (6.19)$$

where $X := \{x_i\}_{i=1}^L \subset V$ is a set of randomly chosen vertices.

The set X defines a partition of V into X and $Y := X^c$. In (6.19) we approximate the value $v(y)$, for an eigenvector v of W and $y \in Y$, only knowing the values $v(x_i)$, $i = 1, \dots, L$, by solving the linear problem

$$\sum_{i=1}^L w(y, x_i) v(x_i) = \theta v(y). \quad (6.20)$$

With this method we can approximate the values of an eigenvector v of W , corresponding to the eigenvalue θ , in the *whole* set of points V using its values in the subset X and solving the interpolated eigenvalue equation above. Generally, this is not as immediate as it sounds since the eigenvectors of W are not known in advance. However, by choosing $y = x_j$, $j = 1, \dots, L$, in (6.20), we find an eigenvalue problem for the known matrix with

entries $w(x_j, x_i)$, which is a much smaller matrix than the full matrix W :

$$\sum_{i=1}^L w(x_j, x_i)v(x_i) = \theta v(x_j). \quad (6.21)$$

If L is small enough such that this eigenvalue problem can be solved, then θ and $v(x_i)$, $i = 1, \dots, L$, can be computed, which in turn can be substituted back into (6.20) to find an approximation to $v(y)$, for any $y \in V$. In short, we approximate the eigenvectors in (6.18) by extensions of the eigenvectors in (6.21), using the extension equation (6.20), and we approximate the eigenvalues in (6.18) by the eigenvalues from (6.21). The main Nyström assumption is that these approximated eigenvectors and eigenvalues approximately diagonalise W . For numerical implementation and further clarification of the method, it is useful to write it in matrix language.

Let us define first the sub matrices $W_{XX} \in \mathbb{R}^L \times \mathbb{R}^L$ and $W_{XY} \in \mathbb{R}^L \times \mathbb{R}^{S-L}$ as

$$W_{XX} = \begin{pmatrix} w(x_1, x_1) & \cdots & w(x_1, x_L) \\ \vdots & \ddots & \vdots \\ w(x_L, x_1) & \cdots & w(x_L, x_L) \end{pmatrix}, \quad (6.22)$$

$$W_{XY} = \begin{pmatrix} w(x_1, y_1) & \cdots & w(x_1, y_{S-L}) \\ \vdots & \ddots & \vdots \\ w(x_L, y_1) & \cdots & w(x_L, y_{S-L}) \end{pmatrix}.$$

Analogous definitions hold for W_{YY} and W_{YX} . Each of these matrices represents the sub matrix having as elements the weights between the points in X , Y or between the two sets. With this notation, the whole matrix $W \in \mathbb{R}^S \times \mathbb{R}^S$ can be written in block-form as

$$W = \begin{pmatrix} W_{XX} & W_{XY} \\ W_{YX} & W_{YY} \end{pmatrix}, \quad W_{YX} = W_{XY}^T.$$

Similarly, vectors $v \in \mathbb{R}^S$ can be written as $v = (v_X^T \ v_Y^T)^T$. We focus on the spectral decomposition of the first block of W , that is W_{XX} . Since this matrix is symmetric, calling Θ_X the matrix $\Theta_X = \text{diag}(\theta_1, \dots, \theta_L)$ containing the eigenvalues of W_{XX} , then by the spectral theorem $W_{XX} = V_X \Theta_X V_X^T$ (compare with (6.21)), with V_X be the orthogonal matrix having as columns the eigenvectors of W_{XX} . Writing (6.20) for $y \in Y$, in operator form, we obtain V_Y as

$$V_Y = W_{YX} V_X \Theta_X^{-1}.$$

The approximated eigenvectors of W can then be written in matrix form as

$$V = \begin{pmatrix} V_X \\ W_{YX}V_X\Theta_X^{-1} \end{pmatrix}. \quad (6.23)$$

Now, let us observe that

$$\begin{aligned} V\Theta_XV^T &= \begin{pmatrix} V_X \\ W_{YX}V_X\Theta_X^{-1} \end{pmatrix} \Theta_X [V_X^T \quad (W_{YX}V_X\Theta_X^{-1})^T] = \\ &= \begin{pmatrix} V_X\Theta_XV_X^T & W_{XY} \\ W_{YX} & W_{YX}W_{XX}^{-1}W_{XY} \end{pmatrix} = \begin{pmatrix} W_{XX} & W_{XY} \\ W_{YX} & W_{YX}W_{XX}^{-1}W_{XY} \end{pmatrix} \approx W. \end{aligned} \quad (6.24)$$

Therefore, Nyström extension can be interpreted as the approximation $W \approx V\Theta_XV^T$, under the approximation of W_{YY} given by $W_{YY} \approx W_{YX}W_{XX}^{-1}W_{XY}$. The quality of the approximation of the full W is quantified by the norm of the Schur complement $\|W_{YY} - W_{YX}W_{XX}^{-1}W_{XY}\|$, see [FBCM04] for more details.

Recalling the definition of the symmetric graph Laplacian L_s given by (6.12) and the relation between the spectral decomposition of W and the one of W in (6.17), we observe that a normalisation step now needs to be computed for obtaining the spectral decomposition of L_s . Defining $\mathbf{1}_L$ as the L -dimensional vector consisting of ones and $\mathbf{1}_{S-L}$ analogously, we use (6.24) and start computing the nonnegative vector $d = (d_X^T d_Y^T)^T$ by

$$\begin{pmatrix} d_X \\ d_Y \end{pmatrix} = \begin{pmatrix} W_{XX} & W_{XY} \\ W_{YX} & W_{YX}W_{XX}^{-1}W_{XY} \end{pmatrix} \begin{pmatrix} \mathbf{1}_L \\ \mathbf{1}_{S-L} \end{pmatrix} \quad (6.25)$$

$$= \begin{pmatrix} W_{XX}\mathbf{1}_L + W_{XY}\mathbf{1}_{S-L} \\ W_{YX}\mathbf{1}_L + W_{YX}W_{XX}^{-1}W_{XY}\mathbf{1}_{S-L} \end{pmatrix}. \quad (6.26)$$

Therefore, the matrices W_{XX} and W_{XY} can be normalised simply by considering:

$$\hat{W}_{XX} = W_{XX} ./ (\sqrt{d_X} \otimes \sqrt{d_X}^T) \quad \text{and} \quad \hat{W}_{XY} = W_{XY} ./ (\sqrt{d_Y} \otimes \sqrt{d_Y}^T), \quad (6.27)$$

where the division is intended element-wise and \otimes is the standard vector tensor product.

A further step of normalisation is now needed since the approximated eigenvectors of W , i.e. the columns of the matrix V in (6.23) may not be orthogonal. Such normalisation may be obtained by using auxiliary unitary matrices. We refer the reader to [BF12, Section 3.2] for more details on this.

Once these additional normalisation steps are completed, a spectral decomposition of W in terms of its eigenvalues $\hat{\lambda}_i$ and the corresponding normalised eigenvectors v_i , $i = 1, \dots, S$ is obtained. Therefore, recalling (6.17), the spectral decomposition of L_s is given in terms of the eigenvalue $1 - \hat{\lambda}_i$ and the same eigenvectors v_i .

Remark 6.2.4. *One may think of the idea of partitioning the set of vertices into several subsets such that $X = X_1 \cup \dots \cup X_k$, $X_i \cap X_j = \emptyset$, $i \neq j$ and using the Nyström idea to approximate the weight function starting to one of them, say X_1 , to compute its value in the all others, repeatedly. Nonetheless, as observed above, the quality of the Nyström approximation is checked by computing the norm of the Schur complement for each pair of subsets whose computation can be quite costly in principle. Moreover, in case of an inaccurate estimation errors may propagate. Consequently, we believe that the current one-step strategy may be preferable since it relies on one single partitioning step and, therefore, does not propagate errors in the further approximations.*

6.2.4 Convex splitting

Recalling the general introduction on splitting methods for the study of PDEs given in Section 3.2, we focus here on convex splitting, which is used to numerically solve problems with a general gradient flow structure of the form

$$\frac{\partial u}{\partial t} = -\nabla_V \mathcal{H}(u) \quad \text{in } \Omega, \quad u(x, 0) = u_0(x) \quad \text{in } \Omega. \quad (6.28)$$

Here, $\Omega \subset \mathbb{R}^d$ is a regular domain, \mathcal{H} is the energy functional to minimise and ∇_V indicates formally the Fréchet derivative with respect to the metric in a Banach space V . The main idea of convex splitting is to split the functional \mathcal{H} into a convex and a concave part:

$$\mathcal{H}(u) = \mathcal{H}_1(u) - \mathcal{H}_2(u), \quad (6.29)$$

where both \mathcal{H}_1 and \mathcal{H}_2 are convex for all u . Denoting by U_n the numerical approximation of $u(\cdot, n\Delta t)$, $\Delta t > 0$, $n \geq 0$, a semi-implicit discretisation of (6.28)-(6.29) reads

$$U_{n+1} - U_n = -\Delta t(\nabla_V \mathcal{H}_1(U_{n+1}) - \nabla_V \mathcal{H}_2(U_n)). \quad (6.30)$$

The advantage of this method consists in treating the convex part implicitly in time and the concave part explicitly. Typically, nonlinearities are considered in the explicit part of the splitting and their instability is balanced by the effect of the implicit terms.

Such methods have been applied in finite element and finite difference approximations (see, e.g., [Eyr98]). In [SB11] such approach is applied to the problem of image inpainting. There, convex splitting is generalised to the case where the gradient flow considered does not follow directly a variational principle, but it can be interpreted as the sum of two dif-

ferent gradient flows with respect to two different norms. Nonetheless, an unconditionally stable numerical method is derived.

Following [BF12, Section 3.1], the splitting choice for the discrete GL functional (6.13) is based on the convex splitting of its continuous version with fidelity term, cf. [BEG08, SB11], and reads

$$GL_{d,1}(u) := \frac{\varepsilon}{2} Q(u, L_s u) + \frac{C}{2} \sum_{x \in V} u^2(x), \quad (6.31a)$$

$$GL_{d,2}(u) := -\frac{1}{4\varepsilon} \sum_{x \in V} (u^2(x) - 1)^2 + \frac{C}{2} \sum_{x \in V} u^2(x) - \sum_{x \in V} \frac{\chi(x)}{2} (u(x) - u_0(x))^2, \quad (6.31b)$$

where the constant $C > 0$ has to be chosen large enough such that $GL_{d,2}$ is convex for u around the wells of W . The differential operator contained in the implicit component of the splitting, $GL_{d,1}$, is the symmetric graph Laplacian, which can be diagonalised quickly and inverted using Fourier transform methods. In [BF12, Section 3.1], more details of the splitting are presented. The time-discretised scheme (6.30) now reads

$$\begin{aligned} U_{n+1}(x) - U_n(x) = & -\Delta t (\varepsilon L_s(U_{n+1}(x)) + C U_{n+1}(x)) \\ & - \Delta t \left(-\frac{1}{\varepsilon} (U_n^3(x) - U_n(x)) + C U_n(x) - \chi(x) (U_n(x) - U_0) \right), \quad x \in V. \end{aligned} \quad (6.32)$$

Recalling the decomposition (6.16), the idea now is to use a spectral decomposition in terms of the eigenvectors v_k of L_s computed through Nyström completion for all the nonlinear quantities appearing in the scheme (6.32).

6.2.5 Numerical realisation and algorithm

We present here the pseudocode combining all the different steps described above for the realisation of the GL minimisation. We recall that ε is the scale parameter of the GL functional (6.13), σ is the variance used in the Gaussian similarity function (6.15), C is the convex splitting parameter in (6.31a)-(6.31b) and L is the number of sample points in (6.19). Regarding the choice of the parameter σ in (6.15), we observed in our numerical experiments that larger values of the parameter allowing more diversity between pixel features resulted in only almost negligible changes in the segmentation results. In all our numerical experiments, we fixed the parameter $\sigma = 15$.

We will now give further details. First we randomly select L pixels from I . As described in Section 6.2.2 we now create a vertex set $V \cong I$, which we partition into a set X , consisting of the vertices corresponding to the L randomly chosen pixels, and a set $Y := V \setminus X$. We now compute the feature vectors of each vertex in V . If I is a grey scale image, we can represent features by an intensity map $f : V \rightarrow \mathbb{R}$. If I is an RGB colour image instead, we use a vector-valued (red, green, and blue) intensity map $f : V \rightarrow \mathbb{R}^3$ of the

Algorithm 2 GL-minimisation with Nyström extension for image segmentation

-
- 1: PARAMETERS: $L \ll S$, σ , ε , C .
 - 2: select L random points and build the set $X \subset V$
 - 3: get a partition $V = X \cup Y$, $Y := X^c$
 - 4: determine features and edge weights of X and Y using (6.15) and build W_{XX} , W_{XY}
 - 5: use **Nyström extension** to compute W_{XY} (6.24) and matrix V of eigenvectors of W (6.23)
 - 6: get normalised \hat{W}_{XX} and \hat{W}_{XY} using (6.25)-(6.27)
 - 7: normalise V and get eigenvalues-eigenvectors of W ($\hat{\lambda}_i, v_i$)
 - 8: output \leftarrow eigenvalues-eigenvectors ($1 - \hat{\lambda}_i, v_i$) of L_s used as GL minimisation input
 - 9: **convex splitting** for GL minimisation through Fourier transform methods, as described in Section 6.2.4
 - 10: output \leftarrow the binary segmentation.
-

form $f(x) = (f_R(x), f_G(x), f_B(x))$. We mirror the boundary to define neighbourhoods also on the image edges. The feature function $\psi : V \rightarrow \mathbb{R}^K$ concatenates the intensity values in the neighbourhood of a pixel into a vector: $\psi(x) := (f(\nu_1(x)), \dots, f(\nu_{\tilde{\tau}}(x)))^T$, where $\nu(x) = (\nu_1(x), \dots, \nu_{\tilde{\tau}}(x)) \in \mathbb{R}^{\tilde{\tau}}$ is the neighbourhood vector of $x \in V$ and $\tilde{\tau} = (2\tau + 1)^2$, the size of the neighbourhood of x . Note that $K = \tilde{\tau}$ if I is a grey scale image and $K = 3\tilde{\tau}$ if I is an RGB colour image.

Additional features can be considered, such as texture, for instance. In particular, we consider the eight MR8 filter responses [VZ05] as texture features on a grey scale image and choose the function $t : V \rightarrow \mathbb{R}^8$ as $t(x) = (\text{MR8}_1(x), \dots, \text{MR8}_8(x))$.

MR8 filters for texture classification We use the MR8 filter bank described in [VZ05]. The set of filters is actually composed by 38 filters, but only the maximum 8 filter responses across all the orientations are computed. Namely, the filter bank consists of two isotropic Gaussian and Laplacian filters, one anisotropic edge filter considered at three different scales in six different orientations and an anisotropic bar filter considered at the same scales and orientations. Only the maximum response across orientations is measured, thus reducing the number of filter responses from 38 to 8 (3 scales for 2 filters plus 2 isotropic filters), see Figure 6.1.

The resulting filters are rotationally invariant and contain both isotropic and anisotropic filters at multiple orientations. Hence, good features for all type of textures are expected to be described by this type of texture features. Furthermore, due to the low dimensionality of the filters, the computational costs are reduced and the clustering result is simplified, still remaining accurate.

The feature function ψ can be defined in this case as $\psi(x) := (t(\nu_1(x)), \dots, t(\nu_{\tilde{\tau}}(x)))^T$ where $\nu(x)$ and $\tilde{\tau}$ are defined as above. Here, $K = 8\tilde{\tau}$. Of course, a combination of colour and texture features can be considered as well by considering ψ defined as $\psi(x) :=$

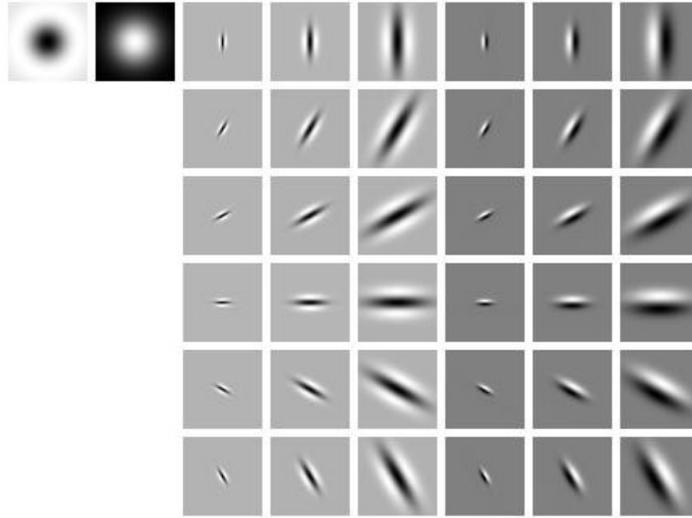


Figure 6.1: The MR8 filter bank. Image from [VZ05]

$(f(\nu_1(x)), t(\nu_1(x)), \dots, f(\nu_{\tilde{\tau}}(x)), t(\nu_{\tilde{\tau}}(x)))$ for every x in V . In this case, when dealing with RGB colour images, the dimension of the feature vector is therefore $K = 11\tilde{\tau}$.

Using (6.15), the matrices W_{XX} and W_{XY} from (6.22) can now be constructed and the Nyström extension can be performed for approximating the eigenvectors and eigenvalues of W as described in Section 6.2.3, which are then used to compute the eigenvectors $\{v_k\}$ of L_s and corresponding eigenvalues $\{\lambda_k\}$, compare (6.17). Recalling (6.16), those eigenvectors are used as basis functions for U_n , the numerical approximation of u in the n -th iteration of the GL minimisation. Considering (6.32) and writing the nonlinear quantities appearing in terms of $\{v_k\}$ similarly as in (6.16), we have

$$(U_n(x))^3 = \sum_k \beta_n^k(x) v_k(x), \quad \chi(x)(U^n(x) - u_0(x)) = \sum_k \gamma_n^k(x) v_k(x), \quad x \in V.$$

The computation of U in the next iteration reduces to finding the coefficients α_{n+1}^k in the expression

$$U_{n+1}(x) = \sum_k \alpha_{n+1}^k(x) v_k(x), \quad x \in V,$$

in terms of β_n^k, γ_n^k and the other parameters involved, that is the scale parameter ε in (6.13), the parameter $C > 0$ appearing in the splitting (6.31) and the time step Δt . Using (6.32), we compute α_{n+1}^k simply as

$$\alpha_{n+1}^k = \mathcal{D}_k^{-1} \left(\left(1 + \frac{\Delta t}{\varepsilon} + C\Delta t \right) \alpha_n^k - \frac{\Delta t}{\varepsilon} \beta_n^k - \Delta t \gamma_n^k \right),$$

where the diagonal matrix \mathcal{D}_k is defined as $\mathcal{D}_k := 1 + \Delta t(\varepsilon\lambda_k + C)$.

6.3 The Hough transform

We address now the issue of detecting objects in an image with geometrical properties that are *a priori* known. An example is the detection of lines and circles. These objects can be identified by mapping them onto an auxiliary space where relevant geometrical properties (such as linear alignment and roundness) are represented as peaks of specific auxiliary functions. In this chapter, we use the Hough transform [Hou62] to detect measurement tools (rulers, concentric circles of fixed radii) with the intent of providing quantitative, scale-independent measurements of the region segmented by one of the techniques described in the sections above. In this way, an absolute measurement of the region of interest in the image is possible, independently of the scale of the image itself which could depend, for instance, on the distance to the objective of the camera.

For our purposes, we focus on straight lines (for which the Hough transform was originally introduced and considered [Hou62]) and circles, [DH72]. Other applications of this transformation for more general curves exist as well. In [BMP13, MPCB15] the Hough transform is used in the context of astronomical and medical images for specific class of curves (*Lamet* and *Watt* curves). In [Gra14] applications to cellular mitosis are presented. There, the Hough transform recognises the cells (as circular/elliptical objects) and tracks them in the process of cellular division.

6.3.1 Line detection

We start from the typical *slope-intercept form* of a line:

$$y = mx + b, \quad m, b, x, y \in \mathbb{R}. \quad (6.33)$$

Traditionally, the equation above is considered as a function of points with coordinates (x, y) satisfying equation (6.33) for *fixed* values of m and b . In other words, these values identify a specific straight line in the x - y plane, cf. Figure 6.2a. Rewriting (6.33) as $b = y - mx$ and keeping fixed the coordinates (x, y) we obtain a new equation of a straight line in the m - b plane, cf. Figure 6.2b, depicting the parameter space. Intersection points of lines in the parameter space are the pair of values m' and b' identifying the lines in the x - y plane where the points considered lie. In other words, straight lines intersecting in the point (m', b') in the parameter space m - b correspond to points lie on the straight line $y = m'x + b'$ in the x - y plane.

Hence, if we are given a black and white image in the x - y plane, and for all coordinates (x, y) of black locations in the image, we draw the corresponding lines in the m - b plane, intersection points of those lines will tell us which (x, y) locations in the image lie on the same line. Of course any two points lie on a line, thus we are specifically interested in intersection points in the m - b plane in which many different lines intersect, indicating the

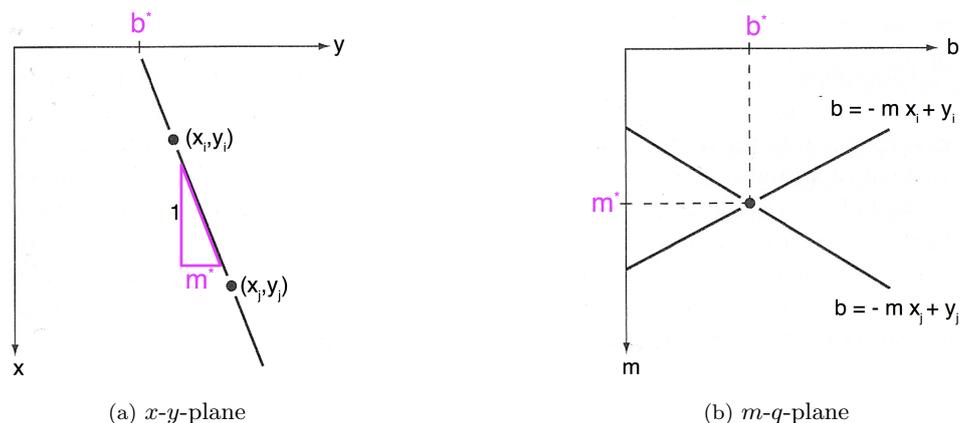


Figure 6.2: Slope-intercept form, (6.33). Images edited from [GW06].

presence of an actual black line segment in the original image.

Drawbacks of this parametrisation are the need for an unbounded parameter space to describe near vertical lines and the impossibility to describe a vertical line. One alternative is the *normal parametrisation* which views a straight line in the x - y space as the tangent line to a circle with radius ρ , touching the circle at angular coordinate θ , as illustrated as in Figure 6.3a, [DH72]. In ρ - θ parameter spaces this leads to

$$\rho = x \cos \theta + y \sin \theta, \quad \theta \in [0, \pi] \quad (6.34)$$

The objects in the parameter space are now sinusoidal curves, but again intersection points identify parameters for the points lying on the same straight line in the x - y plane. Figures 6.3b and 6.3c show an exemplary binary image with two black straight lines and the corresponding parameter space. The bright spots in the parameter space indicate a large number of intersections, thus identifying the two lines in the original image. With parametrisation (6.34) the Hough transform can be interpreted in terms of the Radon transform, as detailed in [PIK92]. The Radon transform describes mathematically several medical imaging acquisition problems (PET and CT, for example) and, as such, is widely used in their mathematical modelling. We refer the reader to [PIK92, BMPS14] and the references therein for a detailed variational analysis of such problems in appropriate measurement and image spaces.

6.3.2 Circle detection

Analogously to what we did as above, when looking for circular structures in a given image, we consider, for $(x, y) \in \mathbb{R}^2$ the parametric representation of a circle:

$$r^2 = (x - c_1)^2 + (y - c_2)^2 \quad (6.35)$$

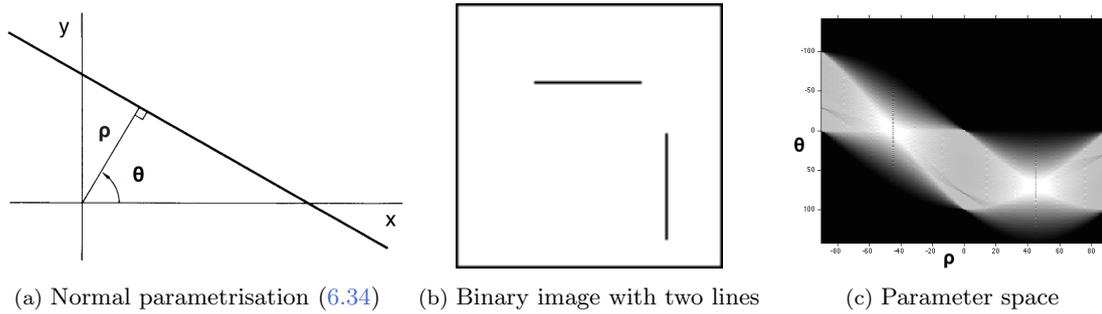


Figure 6.3: Normal form: image and parameter spaces.

where $r > 0$ is the radius of the circle and $(c_1, c_2) \in \mathbb{R}^2$ are the coordinates of its centre. Every point (x, y) lying on the circle, satisfies equation (6.35) for fixed r , c_1 and c_2 . As before, we now consider equation (6.35) in the three-dimensional parameter space $c_1 - c_2 - r$ for fixed x and y . Here, the objects of interest turn out to be cone-shaped surfaces, as shown in Figure 6.4a. Their intersection identifies the desired values of r, c_1 and c_2 in equation (6.35), see Figure 6.4b. As example, we look at a slice of the parameter space along the direction r at level r_i and consider the point $p_1 = (x_1, y_1)$ in the plane $c_1 - c_2$ on the boundary of the circle to detect. This point is surrounded by a circle C_1 with radius r_i whose interior contains all the points (c_1, c_2) corresponding to the coordinates of all the centres of the circles of radius r_i on which p_1 can lie. The same holds for other points p_2 and p_3 surrounded by the circles C_2 and C_3 , respectively. The intersection point of the circles C_1, C_2 and C_3 identifies the true centre coordinates of the circle C in the original image. The accumulator array has a value of 3 in that region.

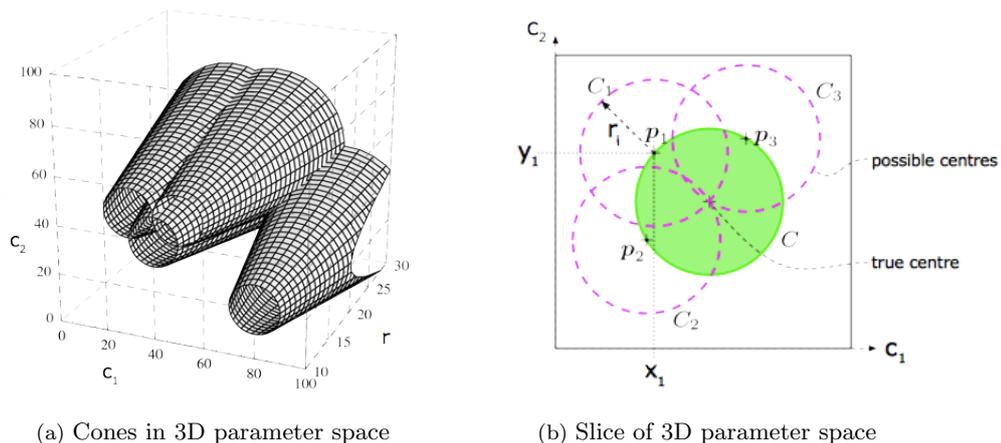


Figure 6.4: Circular Hough transform. Images edited from [BB09].

Numerical strategy We start using the Canny method (see 6.1.1) to identify the edges in the image. Having decided which geometrical shape we are interested in (and, as such, its general parametric representation), the corresponding parameter space is subdivided

into accumulator arrays (cells) whose dimension depends on the dimension of the parameter space itself (2D in the case of straight lines, 3D in the case of circles). Each accumulator array groups a range of parameter values. The accumulator array is initialised to 0 and incremented every time an object in the parameter space passes through the cell. In this way, one looks for the peaks over the set of accumulator arrays as they indicate a high value of intersecting objects for a specific cell. In other words, they are indicators of potential objects having the specific geometrical shape we are interested in.

6.3.3 Pseudocode

Numerically, dealing with the Hough transform consists in looking for peaks of the accumulator arrays in the parameter space which the original image is mapped onto. For straight lines detection we use the MATLAB routines `hough`, `houghpeaks` and `houghlines`: the `hough` function applied to the binary edge image gives back the values of θ and ρ and the Hough transform matrix obtained in correspondence of those values; `houghpeaks` look for peaks of the accumulator arrays in the parameter space and `houghlines` selects from the Hough matrix the values of θ and ρ determined by `houghpeaks` and identifying lines. For circle detection, we use the MATLAB routine `imfindcircles` which finds the centre of the circles in the image by considering at first candidates on the high intensity gradient regions (i.e. on the possible boundaries of the circles) which are allowed to ‘cast votes’ in the accumulator array by counting the number of intersections of the circles with fixed radii with the ones drawn around other points on the circumference, see Figure 6.5.

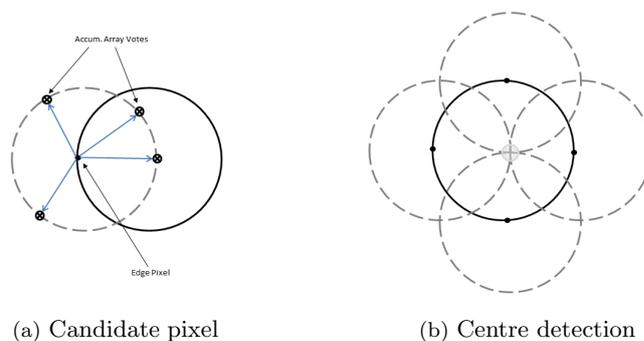


Figure 6.5: The classical circular Hough transform voting pattern.

The accuracy and the number of detections for such routines can be tuned by some parameters, such as, for instance, the maximum number of peaks one wants to consider, obj_{max} , or the array peak threshold, $thresh$, i.e. the minimum number of elements for an accumulator array to be considered a peak. The user also has to specify an initial range of pixel values $[s_{min}, s_{max}]$ as a very rough approximation of the measurement scale. Namely, in the case of line detection this determines a minimum/maximum *spacing* between lines, whereas for circle detection this serves as a rough approximation of the range of values

Algorithm 3 Hough transform for line and circle detection

-
- 1: PARAMETERS: $[s_{min}, s_{max}]$, α , obj_{max} , $thresh$
 - 2: preprocessing: Canny edge detection
 - 3: compute the Hough transform of the edge image
 - 4: set up detection accuracy depending on α and use $[s_{min}, s_{max}]$ as rough initial guess
 - 5: determine at most obj_{max} peaks in the parameter space, thresholding using $thresh$
 - 6: output \leftarrow peaks in the parameter space, corresponding to objects of interest in the original image
-

for the circles' radii. This rough approximation may be given, for example, from average data which the user knows *a priori*. We explain this with some examples in Section 6.4. Accuracy of the detection algorithm is tuned by a parameter α . In case of linear objects this corresponds to choose the maximum number of pixels between two line segments to consider them as one single line, whereas for circle detection this corresponds to the circularity of an object to be considered a circle.

6.4 Applications to object measurement in images

We report in this section the numerical results obtained by the combination of the methods presented above for the detection and quantitative measurement of objects in an image.

To avoid confusion, we will distinguish in the following between two different meanings of scale. Namely, by *image scale* we intend the proportion between the real dimensions (length, width) of objects in the image and their corresponding dimensions quantified in pixel count. Dealing with measurement tools, we talk about *measurement scale* to intend the ratio between a fixed unit of measure (*mm* or *cm*) marked the measurement tool considered and the correspondent number of pixels on the image.

6.4.1 Male pied flycatcher's blaze segmentation

Our first application consists in the segmentation and measurement of the white forehead patch (blaze) for a population of male pied flycatchers (*Ficedula hypoleuca*). We start from giving some biological and behavioural motivations for it, describing in detail how the methods described in Sections 6.2 and 6.3 can be used to solve this problem.

Biological motivations

Measurement of the white forehead patch (blaze) of male pied flycatchers has been studied with regard to sexual selection in [SMS⁺97]. The forehead patch is known to vary between individuals [LA92] and can be subject to both intra- [JLC13] and intersexual [PM99] selection with female pied flycatchers from Spain preferring males with large patches. Forehead patch size has been shown to signal male phenotypic quality through plasma ox-

idative stress and antioxidant capacity [MVRDCn+11]. However, in all studies to date the measurements of patches have been inconsistent and generally inaccurate. For example, some studies have simply measured patch height [DSLS99], whereas Potti and Montalvo [PM99] assumed the shape to be a trapezium with area equal to $0.5(B + b)H$, B being the white patch width, b the bill width and H the height of the white patch. Morales et al. [MMM+07] measured the length and breadth of the forehead patch with callipers to the nearest 0.01 mm and its size (mm^2) was calculated as the area of a rectangle. Other studies have measured the patches from photographs, e.g., Järvistö et al [JLC13] Ruuskanen et al [RLN+13] and Sirkiä et al. [SL09] who photographed the forehead with a scale bar included in each picture, and measured the patch as the white area in mm^2 using IMAGEJ software [AMaR04]. But none of these three papers provide methods of how the measurement was actually achieved, e.g, whether patches were delineated or roughly estimated with a simple shape. Most recently Moreno et al. [MVRDCn+11] analysed digital photos of forehead patches with Adobe PhotoShop CS version 11.0. relating the distance of 1 mm on the ruler to number of pixels, and used this to estimate length. Zooming to 400% and using the paintbrush tool with 100% hardness and 25% spacing the authors delineate the patch and measure the area of the white areas on forehead. While this is the best measurement method to date, it still is subject to human measurement error and subjective assessment of patch boundaries.

Outline of the method

Here, we consider a mathematically robust approach to segment the blaze and to provide an accurate measurement of forehead patch area. We present the numerical results obtained by applying sequentially Algorithms 2 and 3.

Our image database consists of 132 images of individuals from a particular population of male pied flycatchers. Images are 3648×2736 pixels and have been taken by a Canon 350D camera with Canon zoom lens EFD 18 – 55 mm , see Figure 6.6. In each image one of two types of measurement tool is present: a standard ruler or a surface on which two concentric circles of fixed diameter (1 cm the inner one, 3 cm the outer one) are drawn. In the following we will refer to these tools as *linear* and *circular* ruler, respectively. Here, the measurement scale corresponds to the distance between ruler marks for linear rulers and to the radius of the inner circles for circular rulers.

Figure 6.6 shows clearly that the scale of the images in the database may vary significantly because of the different positioning of the camera in front of the flycatcher. In order to study possible correlations between the dimensions (i.e. perimeter, area) of the blaze and significant behavioural factors, the task then is to segment the blaze and detect automatically the scale of the image considered to provide scale-independent measurements.

Due to the very irregular contours and the fine texture on the flycatcher’s forehead, standard variational segmentation methods such as Canny edge detection (Section 6.1.1)

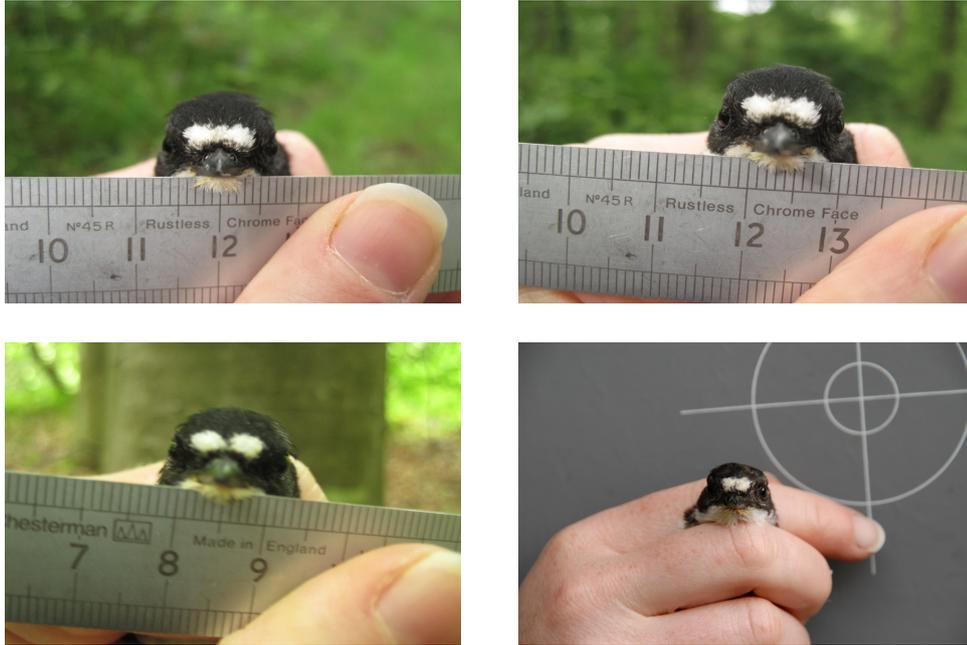


Figure 6.6: The blaze segmentation and measurement problem: the white forehead spot needs to be segmented and measured accurately by detecting the scale on the measurement tool.

or Mumford-Shah models, [MS89, AT92, Bra98, BDM97], are not suitable for our task, as preliminary tests showed. Chan-Vese (Section 6.1.3) is not suitable either, mainly because of its dependence on the initial condition and the small scale detection limits which in practice prevents us from an automatic and accurate segmentation of the tiny, yet characteristic feathers composing the blaze.

On the other hand, the graph GL method exploits similarities and differences between pixels in term of their RGB intensities and texture. In our images these similarities and differences are very distinctive and will guide the segmentation step, with less dependence on the image scale and the initial condition.

We divide our task into different steps:

1. For a given, unsegmented image, we detect the head of the pied flycatcher through a comparison with a user prepared dictionary (see Figure 6.7) using GL segmentation Algorithm 2. Further computations are restricted to the head only.
2. Starting from the reduced image, a second step similar to Step 1 is now performed for the segmentation of the blaze, using again Algorithm 2. A dictionary of blazes is used an extended set of features is considered.
3. A refinement step is now performed in order to reduce the outliers detected in the process of segmentation.

4. We use the Hough transform based Algorithm 3 to detect in the image objects with *a priori* known geometrical shape (lines for linear rulers, circles for circular rulers) for the computation of the measurement scale.
5. The final step is the measurement of the values we are interested in (i.e. the perimeter of the blaze, its area and the width and height of the different blaze components). In the case of linear rulers our results are given up to some error (due to the uncertainty in the detection of the measurement scale computed as average between ruler marks distances).

In order to establish relations with behavioural and biological data confirming or contradicting the initial assumption of correlation between blaze size and higher attractiveness conjectured in [PM99], we have implemented a user ready program for the quantitative analysis and measurements of the size of the bird blazes which is currently used by the Department of Zoology of the University of Cambridge. The results of this study will be the topic of a forthcoming paper [CvGS⁺].

Numerical realisation

We now give an exhaustive description of all the steps described above.

Step 1: Head detection. We consider unlabelled images in the database and compare each of them with a dictionary of previously labelled images, see Figure 6.7. The training regions (i.e. the heads) are labelled with a value 1, the background with value -1 . Unlabelled regions are initialised with value 0.

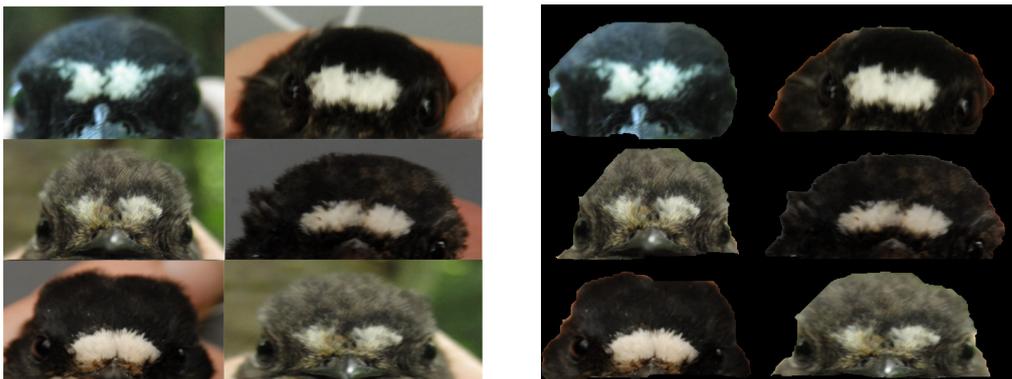


Figure 6.7: Training dictionary for head detection: the heads are manually selected by the user and separated from the background. Then, the corresponding regions are labelled with 1 while the background is labelled by -1 .

The main computational difficulties in this step are due to size of the images considered. This may affect the performance of the algorithm as in order to apply the Nyström

completion technique described in Section 6.2.3 one has to choose an adequate number of points whose features will approximate well the whole matrix. The larger and more heterogeneous the image is, the larger will be the number of points needed to produce a sensible approximation. We circumvent this issue noticing that at this stage of the algorithm, we only need a rough detection of the head which will be used in the following for the accurate segmentation step. Thus, downscaling the image to a lower resolution (in our practice, reducing the resolution by ten times the original one) allows us to use a small number of Nyström sample points (typically 150–200) to produce an accurate result.

For this first step we use as features simply the RGB intensities and proceed as described in Section 6.2.5. Once the head is detected, the resulting image is upscaled again to its original resolution. The solutions computed for the images in Figure 6.6 are presented in Figure 6.8.



Figure 6.8: Head detection from images in Figure 6.6 using dictionaries in Figure 6.7

Step 2: Blaze segmentation. We consider now the reduced image from which we want to extract the flycatcher’s blaze. Again, a dictionary of different blazes is manually created by the user (see Figure 6.9). Similarly as before, training regions (the blazes) are labelled with value 1 and the black feathers in the background with value -1 . As before, unlabelled regions are initialised with value 0. At this stage, RGB intensities alone are not enough to differentiate the blazes from the background consistently in a large number of bird images, due to the colour difference between different blazes. For this step, an additional feature to be considered is the texture of the blaze. For this purpose, we use the MR8 texture features presented in [VZ05] and proceed as detailed in Section 6.2.5. For 3×3 neighbourhoods, the feature vector for each pixel will be an element in \mathbb{R}^{99} , see Section 6.2.5. The Ginzburg-Landau minimisation provides the segmentation results shown in Figure 6.10.

Step 3: Segmentation refinement. This step uses very simple morphological operations in order to remove false detections obtained after Step 2. These can occur due to the choice of colour-texture based features used to compute the feature vectors in Step 2.

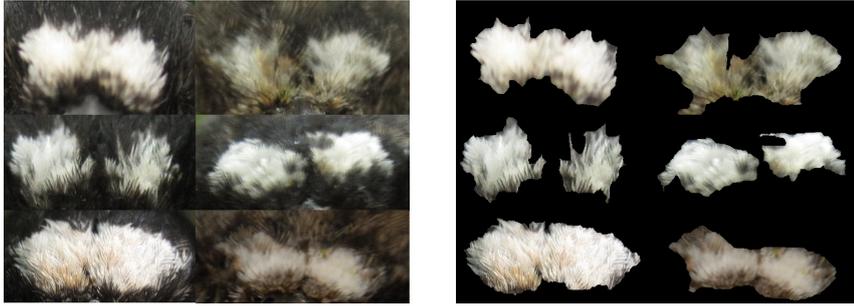


Figure 6.9: Training dictionary for blaze segmentation. As in Figure 6.8 blazes are manually selected by the user and labelled with 1, while black feathers on the background are labelled with -1 .

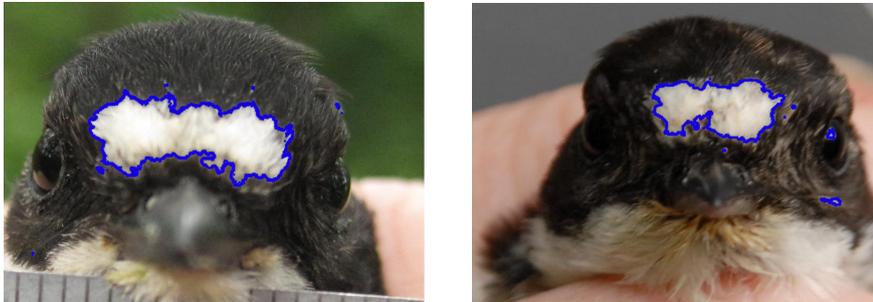


Figure 6.10: Blaze segmentation

For instance, when looking at Figure 6.10 (right) we observe that some bits on the left pied flycatcher's cheek have been detected as they exhibit similar texture properties as the ones on the blaze. In order to prevent this, our software asks the user to confirm whether the segmentation result provided is the expected one or if there are additional unwanted regions detected. If that is the case, using the MATLAB routine `bwconncomp` we label all the connected components segmented in the previous step, discarding among them all the ones whose area is smaller than a fixed percentage (typically 10%) of the largest detected component (presumably, the blaze). This works well in practice, see Figure 6.11. If the user is not satisfied he or she can remove manually the unwanted regions. Figure 6.12 shows some blaze segmentation results after the refinement step.

Step 4: Measurement scale detection. The images in our database divide into two groups: the first is characterised by the presence of linear rulers, whereas the second contains circular rulers (Figure 6.6). We thus need to use the Hough transform based Algorithm 3 to detect lines or circles, respectively. The user is then required to tell the software which objects he or she wants to detect. In both cases, in order to avoid false detections (such as “aligned” objects erroneously detected as lines, or circle-like objects

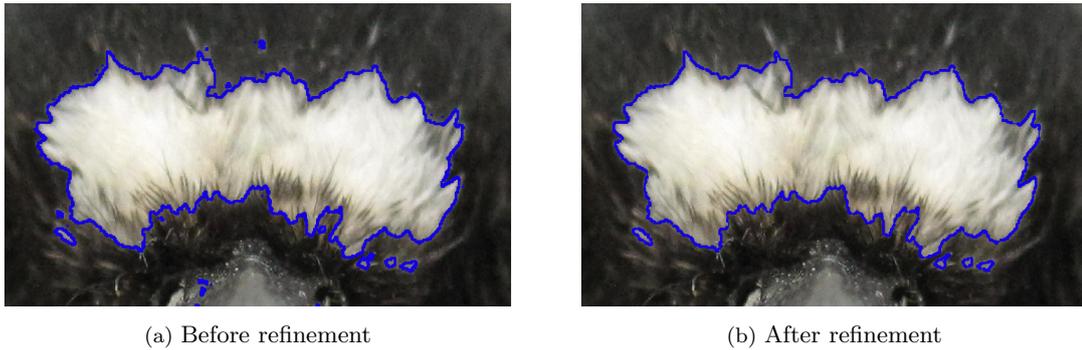


Figure 6.11: Example of segmentation refinement

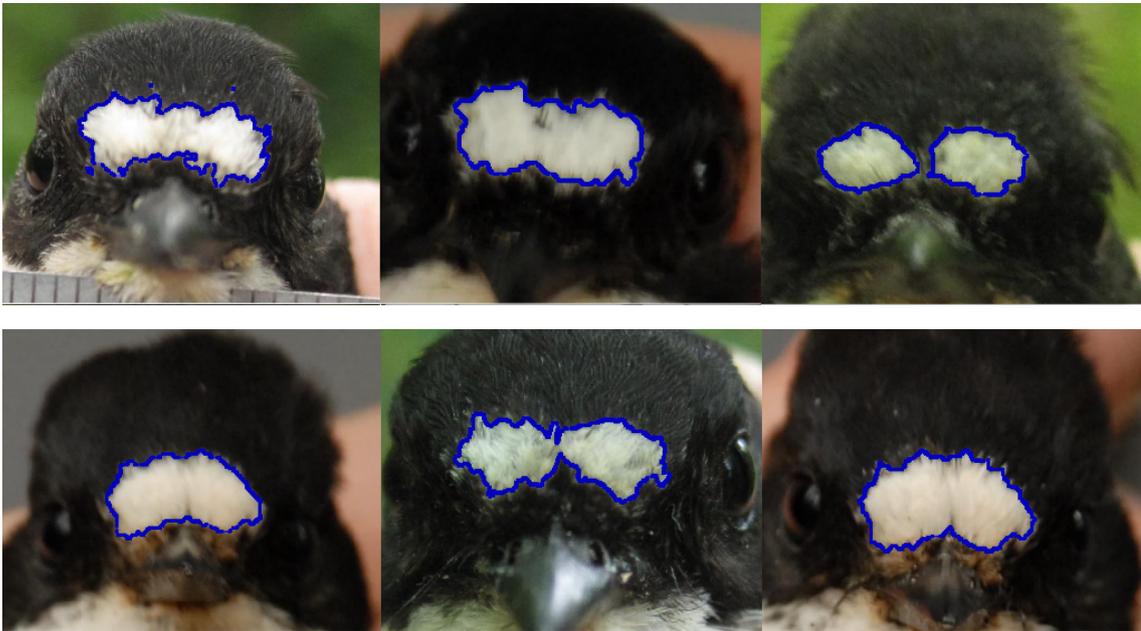


Figure 6.12: Segmentation results after refinement step

wrongly considered as circles, see Figure 6.13), a good candidate for a rough, sensible approximation of the measurement scale is needed as described in Section 6.3.3. In order to get this, we proceed as follows: after detecting the head as in Step 1, we use the option `EquivDiam` of the MATLAB routine `regionprops` to detect the diameter of the head region (in pixels). We then compare such measurement with pre-collected average measurements of head diameters of male pied flycatchers of a similar population (in *cm*), thus obtaining an initial approximation of the measurement scale. In the case of images containing linear rulers, this will serve as a spacing parameter s for the algorithm. In other words, only lines distant at least s pixels from each other will be considered. In the case of circular rulers, the same rough approximation will serve similarly as an indication of the range of values in which the Hough transform based MATLAB function `imfindcircles` will look

for circles' radii. For linear ruler images, the algorithm will look only for parallel lines aligned with a prescribed direction. We set this direction as the one perpendicular to the longest line in the image (since the expectation is, that this longest line is the edge of the ruler). Results of this step are shown in Figure 6.14.

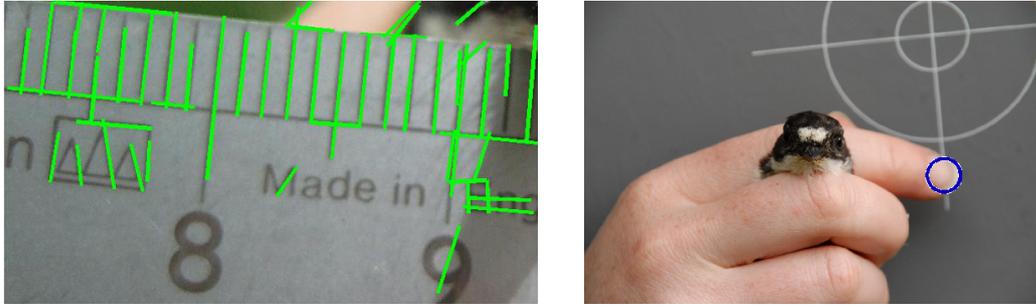


Figure 6.13: Shadows, blur, noise or other objects in the image may disturb the detection.

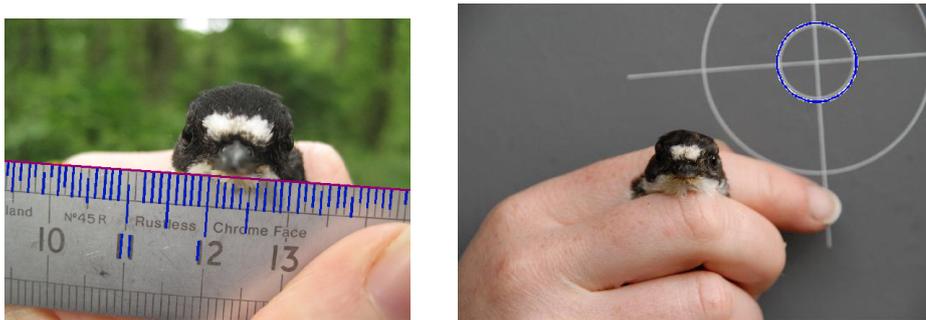


Figure 6.14: Hough transform used for detecting geometrical objects. Left: lines detection using MATLAB routines `hough`, `houghlines`, `houghpeaks`. Right: circle detection using MATLAB routines `imfindcircles`, `viscircles`.

Outliers removal for linear rulers. The scale detection step described above may miss some lines on the ruler. This can be due to an oversmoothing in the denoising step, to high threshold values for edge detection or also to the choice of a large spacing parameter. Furthermore, as we can see from Figures 6.6 and 6.14, we can reasonably assume that the ruler lies on a plane, but its bending can distort some distances between lines. Moreover, few other false line detections can occur (like the number 11 marked on the ruler main body in Figure 6.14). To exclude these cases, we compute the distance (in pixels) between all the consecutive lines detected and eliminate possible outliers using the standard interquartile range (*IQR*) formula [UC97] for outliers' removal. Indicating by Q_1 and Q_3 the lower quartile and the third quartile, an outlier is every element not contained in the interval $[Q_1 - 1.5 * (Q_3 - Q_1), Q_3 + 1.5 * (Q_3 - Q_1)]$. Finally, we compute

the empirical mean, variance and standard deviation (SD) of the values within this range, thus getting a final indication of the scale of the ruler together with an indicator of the precision of the method.

Step 5: Measurement. Once the measurement scale has been detected, it is easy to get all the required measurements. We are interested in the perimeter, the area of the blaze and also in the height and width of the whole blaze component. For linear rulers, due to the error committed in the scale detection step, these values present some uncertainty and variability (see above). In Table 6.1 we show the results of numerical tests on a sample of 30 images with linear rulers. For every image in the sample we compute the SD error and report in the table the minimum, maximum, and average standard (SD) deviation error over the single ones compute, together with the relative standard deviation (RSD) which gives a percentage indication of the error committed.

$$RSD := \frac{\sigma}{\bar{X}} \cdot 100,$$

where σ is the sample SD and \bar{X} is the sample mean of measurements. We observe a minimum and maximum SD of 4.00 and 10.67 pixels, respectively, which, compared to the dimension of the original image (3648×2736 pixels) suggests a reasonable precision. This is confirmed by the average SD value over the sample which is found to be 6.81 pixels. In percentage, the average error over the sample is 11.99%. For circular rulers, we observed in all our experiments that an initial approximation of the range of values for the circle radius (see Step 4 above) results in a robust and typically outlier-free detection of the circular ruler and consequently in an accurate measurement of its radius; the only possible cause of variability and error is its bending.

Uncertainty in the measurements of lengths and areas is calculated with standard formulas in propagation of errors.

SD min	SD max	mean SD	RSD min	RSD max	mean RSD
4.01 pixels	10.67 pixels	6.81 pixels	6.59 %	17.36 %	11.99 %

Table 6.1: Precision of the measurement scale detection for linear rulers on a sample of 30 images.

Despite these variabilities, our method is a flexible and semi-supervised approach for this type of problem. Further tests on the whole set of images and improvements on its accuracy are a matter of future research. The analysis of the resulting data measurements for the particular problem of flycatchers' blaze segmentation will be the topic of the forthcoming paper [CvGS⁺].

6.4.2 Moles monitoring for melanoma staging

In this section we focus on another application of the scale detection Algorithm 3 in the context of melanoma (skin cancer) monitoring, see Figure 6.15. Early signs of melanoma are sudden changes in existing moles and are encoded in the mnemonic ABCD rule. They are **A**symmetry, irregular **B**orders, variegated **C**olour and **D**iameter². In the following we focus on the D sign.

Due to their dimensions and their irregular shapes, moles are often very hard to measure. Typically, a common dermatological practice consists in positioning a ruler under the mole and then taking a picture with a professional camera, see Figure 6.15³. Sudden changes in the evolution of the mole are then observed by comparison between different pictures taken over time. Hence, their quantitative measurement may be an indication of a malignant evolution.

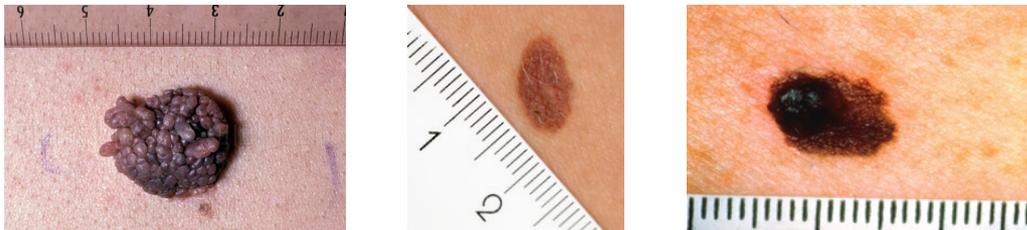


Figure 6.15: The monitoring and measuring of moles is essential for the early diagnosis of melanoma. Normally, due to their small size, they can be measured by juxtaposing a small ruler with them.

In the following examples reported in Figure 6.16, we use the graph segmentation approach described in Algorithm 2 where texture-characteristic regions are present (see Figure 6.16a) and the Chan-Vese model (6.4) for images characterised by homogeneity of the mole and skin regions and the regularity of mole boundaries (Figures (6.16b)-(6.16c)). For the numerical implementation, we use the freely available online IPOL Chan-Vese segmentation code [Get12]. Let us point out here that previous works using variational models for accurate melanoma segmentation already exist in literature, see [CS13, AFSEC14], but in those no measurement technique is considered.

² *Prevention: ABCD's of Melanoma.* American Melanoma Foundation, <http://www.melanomafoundation.org/prevention/abcd.htm>.

³Mole images from <http://www.diomedia.com/>, ©Phototake RM/ ISM, <http://www.medicalprotection.org/uk/practice-matters-issue-3/skin-lesion-photography>, <http://www.medicalprotection.org/uk/practice-matters-issue-3/skin-lesion-photography>, ©Chassenet/Science Photo Library <http://en.wikipedia.org/wiki/Melanoma> (public domain).

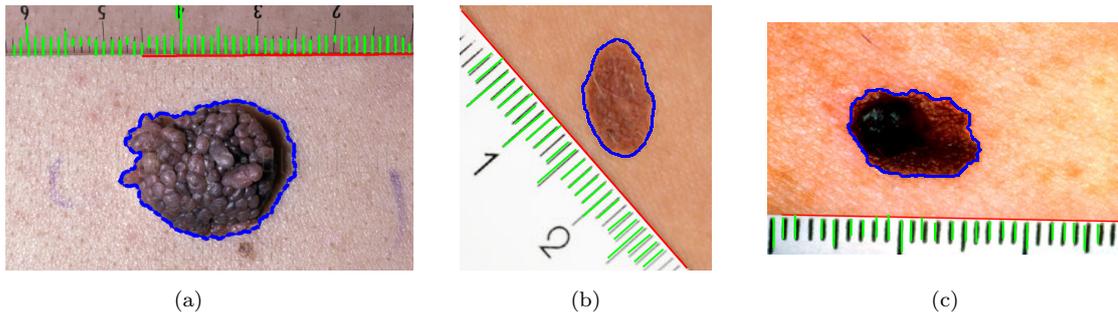


Figure 6.16: Moles' detection using GL Algorithm 2 (a), the Chan-Vese model (6.4) ((b),(c)), and measurement scale detection by Hough transform (Algorithm 3).

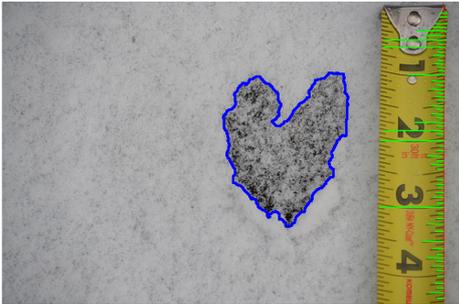
6.4.3 Other applications: animal tracks and archaeological finds

We conclude the chapter presenting some possible other applications for the combined segmentation and scale detection models presented above.

The first application is the identification and classification of animals living in a given area through their soil, snow and mud footprints. Their quantitative measurement is also interesting in the study of the age and size of a particular animal species. As in the problems above, such measurement very often reduces to a very inaccurate measurement performed with a ruler placed next to the footprint image. In Figure 6.17a⁴ our combined method is applied for the measurement of a white-tailed deer footprint.

As a final application, we focus on archaeology. In many archaeological finds, objects need to be measured for comparisons and historical studies [HZK10]. Figure 6.17b shows the application of our method to coin measurements. Due to its circular shape, for this image a combined Hough transform method for circle and line detection has been used. The example image is taken from [HZK10] where the authors propose a gradient threshold based method combined with a Fourier transform approach. Despite being quite efficient for the particular applications considered, such approach relies in practice on the good experimental setting in which the image is taken: almost noise-free images and very regular objects with sharp boundaries (mainly coins) and homogeneous backgrounds are considered. Furthermore, results are reported only for rulers with vertical orientation and no bending.

⁴Image from <http://mamajoules.blogspot.co.uk/2015/01/a-naturalists-thoughts-on-animal-tracks.html>.



(a) White-tailed deer tracks measurement



(b) Coin measurement, image taken from [HZK10]

Figure 6.17: The measurement scale has been detected only in a portion of the figure for the sake of reading clarity.

Chapter 7

Conclusions

My PhD thesis is concerned with the design and the efficient numerical solution of Total Variation (TV) PDE models and with their application to solve several imaging tasks. In each of its chapters this thesis focuses on one specific ingredient of the general variational regularisation model

$$\min_u \text{TV}(u) + \lambda \Phi(u, f) \quad (7.1)$$

and investigates new approaches to refine its choice in order to provide a better reconstruction of the imperfect, observed data f . Namely, in Chapter 3 I considered two non-standard gradient flows of the TV energy and used the resulting higher-order PDE models derived to counteract well-known TV drawbacks evident especially in denoising and inpainting applications. For the efficient numerical realisation of these models I used dimensional splitting (ADI) and quasi-Newton schemes. More focus on the choice of the appropriate data fitting term Φ is given in Chapter 4 where I discussed an infimal convolution combination of classical data fidelities for the particular case of image denoising in the presence of multiple noise statistics. Again, Newton-type methods are used there for an efficient numerical realisation of the model. In Chapter 5 I reviewed and examined some recent contributions in the field of optimal design of imaging models by means of nonsmooth bilevel optimisation. In particular, I focused on the use of training databases to learn from examples the optimal noise model and the ratio λ between the fitting of the data and the image regularisation. Again, the numerical solution of the resulting bilevel models has been obtained by means of Newton's methods combined with sampling strategies for large databases. Finally, in Chapter 6 I described how TV models of the form (7.1) can be effectively adapted and used also in the framework of graphs for some image segmentation problems arising in real-world applications.

In the following I will list some interesting open problems which could be addressed for a better understanding of the topics considered in this thesis. For the sake of clarity, they will be presented following the order of the chapters described above.

- The results of unconditional stability observed in Chapter 3 for the AMOS scheme

(3.42) are only empirical. We are inclined to think that this property should in fact be guaranteed by the implicit character of the scheme, but a rigorous proof is still missing. Due to the nonlinear nature of the model considered, linear stability methods based on Fourier representations are not appropriate. Here, stability needs rather to be proved by using energy estimates, similarly as in [SB11]. Additionally, our efficient numerical solvers could be used for having more insights on the structure of the solution of the (TV- H^{-1}) and (TV-Wass) models and for a deeper understanding of the theory of fourth-order nonlinear parabolic PDEs of this type.

- The analysis and the results reported in Chapter 4 for the TV-IC model (4.20) focus on the particular case where TV regularisation is used for the smoothing of noise in the data which can be expressed as a combination of two noise statistics. The use of higher-order regularisations, such as ICTV [CL97] or TGV [BKP10] would probably result in a more accurate noise decomposition where no structures are encoded in the noise residuals, compare Figure 4.3. Additionally, the extension of the TV-IC model to the case when three or more noise distributions are assumed could also be interesting as well as the applicability of the model to data fidelities modelling signal-dependent noise distributions [AA08].
- The bilevel optimisation approaches for learning the optimal model setup presented in Chapter 5 naturally give room to several open questions. For instance, the case of spatially distributed parameters $\lambda = \lambda(x)$ would be interesting to consider in situations where the noise is present only in some specific regions of the image, see [DHRC11, BDH13]. Preliminary results on this are presented in [CDLR] where a H^1 penalisation is considered. One alternative could be the use of an L^1 penalisation on λ to enforce sparsity in space, which would introduce several complications in the derivation of the optimality system and would result in inequality constraints of the second-kind [DLR11]. The use of bilevel optimisation approaches for learning the optimal parameters for the TV-IC model (4.20) appears also to be promising for obtaining better denoising results in the case of mixed noise statistics, as preliminary showed in Section 5.5.3. From a numerical point of view, the dynamic sampling Algorithm 1 could further be improved by introducing, for instance, an adaptive selection of the accuracy parameter θ to balance the sample size selection and the accuracy throughout the BFGS iterations.
- Finally, the graph modelling considered in 6 is a flexible and new interesting framework for imaging applications. Interesting topics of research could explore new feature functions different from colour and texture, such as illumination or HSV intensities, for instance. Furthermore, it would also be interesting to see how the clustering algorithm presented performs with respect to state-of-the-art algorithms used for data predictions in the increasingly popular field of data science.

Bibliography

- [AA08] G. Aubert and J.-F. Aujol, *A variational approach to removing multiplicative noise*, SIAM Journal on Applied Mathematics **68** (2008), no. 4, 925–946, <http://dx.doi.org/10.1137/060671814>. (Cited on pages 39 and 220).
- [ABCH08] A. Almansa, C. Ballester, V. Caselles, and G. Haro, *A TV based restoration model with local constraints*, Journal of Scientific Computing **34** (2008), no. 3, 209–236, <http://dx.doi.org/10.1007/s10915-007-9160-x>. (Cited on page 153).
- [AC79] S. M. Allen and J. W. Cahn, *A microscopic theory for the antiphase boundary motion and its application to antiphase domain coarsening*, Acta Metallurgica **27** (1979), 1085–1095. (Cited on page 190).
- [AC05] J.-F. Aujol and A. Chambolle, *Dual norms and image decomposition models*, International Journal of Computer Vision **63** (2005), no. 1, 85–104, <http://dx.doi.org/10.1007/s11263-005-4948-3>. (Cited on page 81).
- [ACB⁺03] L. Ambrosio, L. A. Caffarelli, Y. Brenier, G. Buttazzo, and C. Villani, *Optimal transportation and applications*, Lecture Notes in Mathematics, vol. 1813, Springer-Verlag, Berlin; Centro Internazionale Matematico Estivo (C.I.M.E.), Florence, 2003, <http://dx.doi.org/10.1007/b12016>. (Cited on page 79).
- [AFP00] L. Ambrosio, N. Fusco, and D. Pallara, *Functions of bounded variation and free discontinuity problems*, Oxford University Press, USA, 2000. (Cited on pages 33, 47, and 50).
- [AFSEC14] Q. Abbas, I. Fondon, A. Sarmiento, and M. Emre Celebi, *An improved segmentation method for non-melanoma skin lesions using active contour model*, Image Analysis and Recognition, Lecture Notes in Computer Science, no. 8815, Springer, 2014, pp. 193–200. (Cited on page 216).

- [AGS08] L. Ambrosio, N. Gigli, and G. Savaré, *Gradient flows in metric spaces and in the space of probability measures*, second ed., Lectures in Mathematics ETH Zürich, Birkhäuser Verlag, Basel, 2008. (Cited on pages 79 and 81).
- [AK06] G. Aubert and P. Kornprobst, *Mathematical problems in image processing: Partial Differential Equations and the Calculus of Variations*, vol. 147, Springer-Verlag, New York, 2006. (Cited on pages 22 and 31).
- [All08a] W. Allard, *Total variation regularization for image denoising, I. Geometric theory*, SIAM Journal on Mathematical Analysis **39** (2008), no. 4, 1150–1190, <http://dx.doi.org/10.1137/060662617>. (Cited on pages 35 and 146).
- [All08b] ———, *Total variation regularization for image denoising, II. Examples*, SIAM Journal on Imaging Sciences **1** (2008), no. 4, 400–417, <http://dx.doi.org/10.1137/070698749>. (Cited on page 35).
- [AMaR04] M. D. Abramoff, P. G. Magalhães, and S. J. Ram, *Image processing with ImageJ*, Biophotonics international **11** (2004), no. 7, 36–42. (Cited on page 208).
- [Amb89] L. Ambrosio, *A compactness theorem for a new class of functions of bounded variation*, Bollettino della Unione Matematica Italiana. Sezione B. Articoli di Ricerca Matematica **3** (1989), 857–881. (Cited on pages 33, 34, and 55).
- [AT90] L. Ambrosio and V. M. Tortorelli, *Approximation of functionals depending on jumps by elliptic functionals via Γ -convergence*, Communications on Pure and Applied Mathematics **43** (1990), no. 8, 999–1036, <http://dx.doi.org/10.1002/cpa.3160430805>. (Cited on pages 33, 185, and 187).
- [AT92] ———, *On the approximation of free discontinuity problems*, Bollettino della Unione Matematica Italiana. Sezione B. Serie VII **6** (1992), no. 1, 105–123. (Cited on pages 185, 190, and 209).
- [AV94] R. Acar and C. R. Vogel, *Analysis of bounded variation penalty methods for ill-posed problems*, Inverse Problems **10** (1994), no. 6, 1217–1229, <http://dx.doi.org/10.1088/0266-5611/10/6/003>. (Cited on pages 35, 60, 61, and 127).
- [AV97] G. Aubert and L. Vese, *A variational method in image recovery*, SIAM Journal on Numerical Analysis **34** (1997), 1948–1979, <http://dx.doi.org/10.1137/S003614299529230X>. (Cited on page 35).

- [AZSG08] S. P. Awate, H. Zhang, T. J. Simon, and J. C. Gee, *Multivariate segmentation of brain tissues by fusion of MRI and DTI data*, 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro., 2008, <http://dx.doi.org/10.1109/ISBI.2008.4540970>, pp. 213–216. (Cited on page 28).
- [BB90] G. Bouchitté and G. Buttazzo, *New lower semicontinuity results for non-convex functionals defined on measures*, *Nonlinear Analysis* **15** (1990), 679–692. (Cited on page 130).
- [BB00] J.-D. Benamou and Y. Brenier, *A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem*, *Numerische Mathematik* **84** (2000), no. 3, 375–393, <http://dx.doi.org/10.1007/s002110050002>. (Cited on page 82).
- [BB08] L. Bottou and O. Bousquet, *The tradeoffs of large scale learning*, *Advances in Neural Information Processing Systems*, vol. 20, NIPS Foundation, 2008, <http://leon.bottou.org/papers/bottou-bousquet-2008>, pp. 161–168. (Cited on page 166).
- [BB09] W. Burger and M. J. Burge, *Digital image processing: an algorithmic introduction using Java*, Springer Science & Business Media, 2009. (Cited on page 205).
- [BB11] M. Benning and M. Burger, *Error estimates for general fidelities*, *Electronic Transactions on Numerical Analysis* **38** (2011), 44–68. (Cited on pages 126 and 146).
- [BB13] ———, *Ground states and singular vectors of convex variational regularization methods*, *Methods and Applications of Analysis* **20** (2013), no. 4, 295–334, <http://dx.doi.org/10.4310/MAA.2013.v20.n4.a1>. (Cited on page 146).
- [BBB95] G. Bouchitté, G. Buttazzo, and A. Braides, *Relaxation results for some free discontinuity problems.*, *Journal für die reine und angewandte Mathematik (Crelles Journal)* **458** (1995), 1–18. (Cited on page 130).
- [BBBM13] M. Benning, C. Brune, M. Burger, and J. Müller, *Higher-order TV methods – Enhancement via Bregman iteration*, *Journal of Scientific Computing* **54** (2013), no. 2, 269–310, <http://dx.doi.org/10.1007/s10915-012-9650-3>. (Cited on pages 146 and 150).
- [BBFAC04] J. Bect, L. Blanc-Fraud, G. Aubert, and A. Chambolle, *A ℓ^1 -unified variational framework for image restoration*, *Computer Vision - ECCV*

- 2004, Lecture Notes in Computer Science, vol. 3024, Springer Berlin Heidelberg, 2004, http://dx.doi.org/10.1007/978-3-540-24673-2_1, pp. 1–13. (Cited on page 36).
- [BBG⁺11] L. Biegler, G. Biros, O. Ghattas, M. Heinkenschloss, D. Keyes, B. Mallick, Y. Marzouk, L. Tenorio, B. van Bloemen Waanders, and K. Willcox, *Large-scale inverse problems and quantification of uncertainty*, Wiley Series in Computational Statistics, John Wiley & Sons, Ltd., Chichester, 2011. (Cited on page 147).
- [BC11] H. H. Bauschke and P. L. Combettes, *Convex analysis and monotone operator theory in Hilbert spaces*, CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC, Springer, New York, 2011. (Cited on page 125).
- [BCC08] A. Blanchet, V. Calvez, and J. A. Carrillo, *Convergence of the mass-transport steepest descent scheme for the subcritical Patlak-Keller-Segel model*, SIAM Journal on Numerical Analysis **46** (2008), no. 2, 691–721, <http://dx.doi.org/10.1137/070683337>. (Cited on page 82).
- [BCDS13] M. Benning, L. Calatroni, B. Düring, and C.-B. Schönlieb, *A primal-dual approach for a total variation Wasserstein flow*, Geometric science of information, Lecture Notes in Comput. Sci., vol. 8085, Springer, Heidelberg, 2013, http://dx.doi.org/10.1007/978-3-642-40020-9_45, pp. 413–421. (Cited on pages 42 and 47).
- [BCM⁺00] P. Blomgren, T. F. Chan, P. Mulet, Vese L., and W. L. Wan, *Variational PDE models and methods for image processing*, Research Notes in Mathematics, vol. 420, Chapman & Hall/CRC, 2000, pp. 43–67. (Cited on page 61).
- [BCNN11] R. H. Byrd, G. M. Chin, W. Neveitt, and J. Nocedal, *On the use of stochastic Hessian information in optimization methods for machine learning*, SIAM Journal on Optimization **21** (2011), no. 3, 977–995, <http://dx.doi.org/10.1137/10079923X>. (Cited on page 166).
- [BCNW12] R. H. Byrd, G. M. Chin, J. Nocedal, and Y. Wu, *Sample size selection in optimization methods for machine learning*, Mathematical Programming **134** (2012), no. 1, Ser. B, 127–155, <http://dx.doi.org/10.1007/s10107-012-0572-5>. (Cited on pages 166, 167, 169, and 170).
- [BCRS03] M. Bertalmio, V. Caselles, B. Rougé, and A. Solé, *TV based image restoration with local constraints*, Journal of Scientific Com-

- puting **19** (2003), no. 1-3, 95–122, <http://dx.doi.org/10.1023/A:1025391506181>. (Cited on page 153).
- [BCW10] M. Burger, J. A. Carrillo, and M.-T. Wolfram, *A mixed finite element method for nonlinear diffusion equations*, *Kinetic and Related Models* **3** (2010), no. 1, 59–83, <http://dx.doi.org/10.3934/krm.2010.3.59>. (Cited on page 82).
- [BDH13] K. Bredies, Y. Dong, and M. Hintermüller, *Spatially dependent regularization parameter selection in total generalized variation models for image restoration*, *International Journal of Computer Mathematics* **90** (2013), no. 1, 109–123, <http://dx.doi.org/10.1080/00207160.2012.700400>. (Cited on pages 153, 157, 180, and 220).
- [BDM97] A. Braides and G. Dal Maso, *Non-local approximation of the Mumford-Shah functional*, *Calculus of Variations and Partial Differential Equations* **5** (1997), no. 4, 293–322, <http://dx.doi.org/10.1007/s005260050068>. (Cited on pages 185, 187, and 209).
- [BEG08] A. Bertozzi, S. Esedoglu, and A. Gillette, *Analysis of a two-scale Cahn-Hilliard model for binary image inpainting*, *Multiscale Modeling and Simulation* **6** (2008), no. 3, 913–936, <http://dx.doi.org/10.1137/060660631>. (Cited on pages 38 and 200).
- [Ben11] M. Benning, *Singular regularization of inverse problems. Bregman distances and applications to variational frameworks with singular regularization energies*, Ph.D. thesis, University of Münster, 2011. (Cited on page 99).
- [BF12] A. L. Bertozzi and A. Flenner, *Diffuse interface models on graphs for classification of high dimensional data*, *Multiscale Modeling & Simulation* **10** (2012), no. 3, 1090–1118, <http://dx.doi.org/10.1137/11083109X>. (Cited on pages 31, 35, 46, 47, 189, 190, 193, 194, 195, 198, and 200).
- [BF02] S. Belongie, C. Fowlkes, F. Chung, and J. Malik, *Spectral partitioning with indefinite kernels using the Nyström extension*, *Computer Vision - ECCV 2002, Lecture Notes in Computer Science*, vol. 2352, Springer Berlin Heidelberg, 2002, http://dx.doi.org/10.1007/3-540-47977-5_35, pp. 531–542. (Cited on page 195).
- [BFMS08] W. Baatz, M. Fornasier, P. A. Markowich, and C.-B. Schönlieb, *Inpainting of ancient Austrian frescoes*, *Conference Proceedings of Bridges 2008*, Leeuwarden, 2008, pp. 150–156. (Cited on page 28).

- [BFS12] M. Burger, M. Franek, and C.-B. Schönlieb, *Regularised regression and density estimation based on optimal transport*, Applied Mathematics Research Express (AMRX) **2012** (2012), no. 2, 209–253. (Cited on pages 43, 81, 82, 109, and 111).
- [BGH⁺14] M. Benning, L. Gladden, D. Holland, C.-B. Schönlieb, and T. Valkonen, *Phase reconstruction from velocity-encoded MRI measurements – A survey of sparsity-promoting variational approaches*, Journal of Magnetic Resonance **238** (2014), 26–43, <http://dx.doi.org/10.1016/j.jmr.2013.10.003>. (Cited on pages 36 and 37).
- [BHS09] M. Burger, L. He, and C.-B. Schönlieb, *Cahn-Hilliard inpainting and a generalization for grayvalue images*, SIAM Journal on Imaging Sciences **2** (2009), no. 4, 1129–1167, <http://dx.doi.org/10.1137/080728548>. (Cited on pages 38, 42, 78, and 79).
- [BIK01] D. Barash, M. Israeli, and R. Kimmel, *An accurate operator splitting scheme for nonlinear diffusion filtering*, 2001. (Cited on pages 84, 86, 87, and 98).
- [BKP10] K. Bredies, K. Kunisch, and T. Pock, *Total generalized variation*, SIAM Journal on Imaging Sciences **3** (2010), no. 3, 492–526, <http://dx.doi.org/10.1137/090769521>. (Cited on pages 37, 38, 117, 137, 150, and 220).
- [BKV13] K. Bredies, K. Kunisch, and T. Valkonen, *Properties of L^1 -TGV² : The one-dimensional case*, Journal of Mathematical Analysis and Applications **398** (2013), no. 1, 438 – 454, <http://dx.doi.org/10.1016/j.jmaa.2012.08.053>. (Cited on pages 37 and 146).
- [BL91] J. M. Borwein and A. S. Lewis, *Convergence of best entropy estimates*, SIAM Journal on Optimization **1** (1991), no. 2, 191–205, <http://dx.doi.org/10.1137/0801014>. (Cited on pages 125 and 126).
- [BLCT⁺08] F. Benvenuto, A. La Camera, C. Theys, A. Ferrari, H. Lantéri, and M. Bertero, *The study of an iterative method for the reconstruction of images corrupted by Poisson and Gaussian noise*, Inverse Problems **24** (2008), no. 3, <http://dx.doi.org/10.1088/0266-5611/24/3/035016>. (Cited on pages 39, 116, and 143).
- [BMK08] S. D. Babacan, R. Molina, and A. K. Katsaggelos, *Parameter estimation in TV image restoration using variational distribution approximation*, IEEE transactions on Image Processing **17** (2008), no. 3, 326–339, <http://dx.doi.org/10.1109/TIP.2007.916051>. (Cited on page 41).

- [BMP13] M. C. Beltrametti, A. M. Massone, and M. Piana, *Hough transform of special classes of curves*, SIAM Journal on Imaging Sciences **6** (2013), no. 1, 391–412, <http://dx.doi.org/10.1137/120863794>. (Cited on page 203).
- [BMPS14] M. Burger, J. Müller, E. Papoutsellis, and C.-B. Schönlieb, *Total variation regularization in measurement and image space for PET reconstruction*, Inverse Problems **30** (2014), no. 10, 105003, 42. (Cited on pages 36, 116, 126, and 204).
- [BNS14] F. Baus, M. Nikolova, and G. Steidl, *Fully smoothed ℓ_1 -TV models: bounds for the minimizers and parameter choice*, Journal of Mathematical Imaging and Vision **48** (2014), no. 2, 295–307, <http://dx.doi.org/10.1007/s10851-013-0420-0>. (Cited on page 45).
- [BP10] M. Bergounioux and L. Piffet, *A second-order model for image denoising*, Set-Valued and Variational Analysis **18** (2010), no. 3-4, 277–306, <http://dx.doi.org/10.1007/s11228-010-0156-6>. (Cited on page 37).
- [BPPS15] M. Burger, K. Papafitsoros, E. Papoutsellis, and C.-B. Schönlieb, *Infimal convolution regularisation functionals of BV and L^p spaces. Part i: The finite p case*, submitted (2015), <http://arxiv.org/abs/1504.01956>. (Cited on page 62).
- [Bra98] A. Braides, *Approximation of free-discontinuity problems*, Lecture Notes in Mathematics, vol. 1694, Springer-Verlag, Berlin, 1998. (Cited on pages 185, 187, and 209).
- [Bra02] ———, *Γ -convergence for beginners*, Oxford lecture series in mathematics and its applications, 2002. (Cited on page 55).
- [Bre83] H. Brezis, *Analyse fonctionnelle*, vol. 5, Masson, 1983. (Cited on page 49).
- [Bre14] K. Bredies, *Recovering piecewise smooth multichannel images by minimization of convex functionals with total generalized variation penalty*, Efficient Algorithms for Global Optimization Methods in Computer Vision, Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2014, http://dx.doi.org/10.1007/978-3-642-54774-4_3, pp. 44–77. (Cited on page 37).
- [Bru10] C. Brune, *4D imaging in tomography and optical nanoscopy*, Ph.D. thesis, University of Muenster, Germany, july 2010. (Cited on pages 82 and 154).

- [BS96] M. Brokate and J. Sprekels, *Hysteresis and phase transitions*, Applied Mathematical Sciences, vol. 121, Springer-Verlag, New York, 1996. (Cited on page 185).
- [BSW⁺13] C. Brune, A. Sawatzky, F. Wübbeling, T. Kösters, and M. Burger, *EM-TV methods for inverse problems with poisson noise*, Level Set and PDE Based Reconstruction Methods in Imaging, Lecture Notes in Mathematics, Springer International Publishing, 2013, http://dx.doi.org/10.1007/978-3-319-01712-9_2, pp. 71–142. (Cited on page 39).
- [BTWG08] T. Bui-Thanh, K. Willcox, and O. Ghattas, *Model reduction for large-scale systems with high-dimensional parametric input space*, SIAM Journal on Scientific Computing **30** (2008), no. 6, 3270–3288, <http://dx.doi.org/10.1137/070694855>. (Cited on pages 146 and 147).
- [BV04] S. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge University Press, Cambridge, 2004, <http://dx.doi.org/10.1017/CB09780511804441>. (Cited on page 63).
- [BV11] K. Bredies and T. Valkonen, *Inverse problems with second-order total generalized variation constraints*, Proceedings of SampTA 2011 - 9th International Conference on Sampling Theory and Applications, Singapore, 2011. (Cited on pages 37 and 150).
- [Can86] J. Canny, *A computational approach to edge detection*, IEEE Transactions on Pattern Analysis and Machine Intelligence **8** (1986), no. 6, 679–698, <http://dx.doi.org/10.1109/TPAMI.1986.4767851>. (Cited on page 186).
- [CCDLR⁺15] L. Calatroni, C. Cao, J.C. De Los Reyes, C.-B. Schönlieb, and T. Valkonen, *Bilevel approaches for learning of variational imaging model*, submitted (2015). (Cited on pages 42, 43, 44, 47, 116, 123, 135, 137, 141, 147, 159, 172, and 179).
- [CCN07] V. Caselles, A. Chambolle, and M. Novaga, *The discontinuity set of solutions of the TV denoising problem and some extensions*, Multiscale Modeling & Simulation **6** (2007), no. 3, 879–894, <http://dx.doi.org/10.1137/070683003>. (Cited on pages 35, 37, and 146).
- [CCN08] J.-F. Cai, R. H. Chan, and M. Nikolova, *Two-phase approach for deblurring images corrupted by impulse plus Gaussian noise*, Inverse Problems and Imaging **2** (2008), no. 2, 187–204, <http://dx.doi.org/10.3934/ipi.2008.2.187>. (Cited on pages 39, 43, 116, and 123).

-
- [CD58] S. D. Conte and R. T. Dames, *An alternating direction method for solving the biharmonic equation.*, Mathematical Tables and Other Aids to Computation **12** (1958), 198–205. (Cited on page 88).
- [CD90] W. Chen and Z. Ditzian, *Mixed and directional derivatives*, Proceedings of the American Mathematical Society **108** (1990), no. 1, 177–185, <http://dx.doi.org/10.2307/2047711>. (Cited on page 97).
- [CD09] A. Chambolle and J. Darbon, *On total variation minimization and surface evolution using parametric maximum flows*, International Journal of Computer Vision **84** (2009), no. 3, 288–307, <http://dx.doi.org/10.1007/s11263-009-0238-9>. (Cited on page 36).
- [CDLR] C. Chung and J.C. De Los Reyes, *Learning optimal spatially-dependent regularization parameters in total variation image restoration*, in preparation. (Cited on pages 153, 160, 180, 181, and 220).
- [CDLRS14] L. Calatroni, J. C. De Los Reyes, and C.-B. Schönlieb, *Dynamic sampling schemes for optimal noise learning under multiple nonsmooth constraints*, System Modeling and Optimization, Springer Berlin Heidelberg, 2014, pp. 85–95. (Cited on pages 42, 43, 45, 47, 116, 123, 135, 137, 147, 153, and 172).
- [CDLRS15] L. Calatroni, J. C. De Los Reyes, and C.-B. Schönlieb, *A variational model for total variation mixed noise removal*, in preparation (2015). (Cited on pages 42, 43, and 47).
- [CDS14] L. Calatroni, B. Düring, and C.-B. Schönlieb, *ADI splitting schemes for a fourth-order nonlinear partial differential equation from image processing*, Discrete and Continuous Dynamical Systems. Series A **34** (2014), no. 3, 931–957, <http://dx.doi.org/10.3934/dcds.2014.34.931>. (Cited on pages 42 and 47).
- [CE05] T. F. Chan and S. Esedoglu, *Aspects of total variation regularized L^1 function approximation*, SIAM Journal on Applied Mathematics **5** (2005), no. 65, 1817–1837, <http://dx.doi.org/10.1137/040604297>. (Cited on page 39).
- [CEN07] T. Chan, S. Esedoglu, and K. Ni, *Histogram based segmentation using Wasserstein distances*, Scale Space and Variational Methods in Computer Vision, Lecture Notes in Computer Science, vol. 4485, Springer Berlin Heidelberg, 2007, http://dx.doi.org/10.1007/978-3-540-72823-8_60, pp. 697–708. (Cited on page 82).

- [CEN14] J. Chung, M. I. Español, and T. Nguyen, *Optimal regularization parameters for general-form Tikhonov regularization*, ArXiv e-prints (2014), <http://arxiv.org/abs/1407.1911>. (Cited on page 147).
- [CGM99] T. F. Chan, G. H. Golub, and P. Mulet, *A nonlinear primal-dual method for total variation-based image restoration*, SIAM Journal on Scientific Computing **20** (1999), no. 6, 1964–1977, <http://dx.doi.org/10.1137/S1064827596299767>. (Cited on page 60).
- [Cha04] A. Chambolle, *An algorithm for total variation minimization and applications*, Journal of Mathematical Imaging and Vision **20** (2004), no. 1, 89–97, <http://dx.doi.org/10.1023/B:JMIV.0000011325.36760.1e>. (Cited on pages 36, 60, 63, 81, 84, and 135).
- [Chu97] F. R. K. Chung, *Spectral graph theory*, CBMS Regional Conference Series in Mathematics, vol. 92, Published for the Conference Board of the Mathematical Sciences, Washington, DC, 1997. (Cited on pages 31, 191, 192, and 193).
- [CKS97] V. Caselles, R. Kimmel, and G. Sapiro, *Geodesic active contours*, International Journal of Computer Vision **22** (1997), no. 1, 61–79, <http://dx.doi.org/10.1023/A:3A1007979827043>. (Cited on page 185).
- [CKS01] T.F. Chan, S.H. Kang, and J. Shen, *Total variation denoising and enhancement of color images based on the CB and HSV color models*, Journal of Visual Communication and Image Representation **12** (2001), no. 4, 422 – 435, <http://dx.doi.org/10.1006/jvci.2001.0491>. (Cited on page 35).
- [CKS02] ———, *Euler’s elastica and curvature-based inpainting*, SIAM Journal on Applied Mathematics **63** (2002), no. 2, 564–592, <http://dx.doi.org/10.1137/S0036139901390088>. (Cited on pages 32, 34, 37, and 79).
- [CL97] A. Chambolle and P. L. Lions, *Image recovery via total variation minimization and related problems*, Numerische Mathematik **76** (1997), 167–188, <http://dx.doi.org/10.1007/s002110050258>. (Cited on pages 35, 37, 39, 60, 115, 117, 127, 150, and 220).
- [Cla12] C. Clason, *L^∞ fitting for inverse problems with uniform noise*, Inverse Problems **28** (2012), no. 10, 104007, <http://dx.doi.org/10.1088/0266-5611/28/10/104007>. (Cited on page 39).

-
- [CLL⁺05] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, F. Warner, and S. Zucker, *Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps*, Proceedings of the National Academy of Sciences, 2005, pp. 7426–7431. (Cited on page 195).
- [CM99] T. F. Chan and P. Mulet, *On the convergence of the lagged diffusivity fixed point method in total variation image restoration*, SIAM Journal on Numerical Analysis **36** (1999), no. 2, 354–367, <http://dx.doi.org/10.1137/S0036142997327075>. (Cited on pages 35, 60, 61, 77, and 94).
- [CM09] J. A. Carrillo and J. S. Moll, *Numerical simulation of diffusive and aggregation phenomena in nonlinear continuity equations by evolving diffeomorphisms*, SIAM Journal on Scientific Computing **31** (2009), no. 6, 4305–4329, <http://dx.doi.org/10.1137/080739574>. (Cited on page 82).
- [CMRS00] D. Calvetti, S. Morigi, L. Reichel, and F. Sgallari, *Tikhonov regularization and the L-curve for large discrete ill-posed problems*, Journal of Computational and Applied Mathematics **123** (2000), no. 1-2, 423 – 446, [http://dx.doi.org/10.1016/S0377-0427\(00\)00414-3](http://dx.doi.org/10.1016/S0377-0427(00)00414-3). (Cited on pages 41 and 154).
- [CP11] A. Chambolle and T. Pock, *A first-order primal-dual algorithm for convex problems with applications to imaging*, Journal of Mathematical Imaging and Vision **40** (2011), no. 1, 120–145, <http://dx.doi.org/10.1007/s10851-010-0251-1>. (Cited on pages 36, 60, and 64).
- [CS01] T. F. Chan and J. Shen, *Nontexture inpainting by curvature-driven diffusions*, Journal of Visual Communication and Image Representation **12** (2001), no. 4, 436–449, <http://dx.doi.org/10.1006/jvci.2001.0487>. (Cited on pages 32 and 79).
- [CS03] ———, *On the role of the BV image model in image restoration*, Recent Advances in Scientific Computing and Partial Differential Equations (Providence AMS, ed.), Contemp. Math., vol. 330, 2003, pp. 25–41. (Cited on page 35).
- [CS05a] ———, *Image processing and analysis*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2005, <http://dx.doi.org/10.1137/1.9780898717877>. (Cited on pages 22, 47, and 60).
- [CS05b] ———, *Variational image inpainting*, Communications on Pure and Applied Mathematics **58** (2005), no. 5, 579–619, <http://dx.doi.org/10.1002/cpa.20075>. (Cited on pages 34, 36, and 38).

- [CS13] P. G. Cavalcanti and J. Scharcanski, *Macroscopic pigmented skin lesion segmentation and its influence on lesion classification and diagnosis*, Color Medical Image Analysis, Lecture Notes in Computational Vision and Biomechanics (Springer, ed.), vol. 6, 2013. (Cited on page 216).
- [CSL00] T. F. Chan, B. Y. Sandberg, and Vese L., *Active contours without edges for vector-valued images*, Journal of Visual Communication and Image Representation **11** (2000), 130–141. (Cited on pages 185 and 188).
- [CSZ06] T. F. Chan, J. Shen, and H.-M. Zhou, *Total variation wavelet inpainting*, Journal of Mathematical Imaging and Vision **25** (2006), no. 1, 107–125, <http://dx.doi.org/10.1007/s10851-006-5257-3>. (Cited on page 36).
- [CTB12] Y. Chen, Pock. T., and H. Bischof, *Learning ℓ_1 -based analysis and synthesis sparsity priors using bilevel optimization*, Workshop on Analysis Operator Learning vs. Dictionary Learning, NIPS 2012, 2012. (Cited on page 147).
- [CV01] T. F. Chan and L. A. Vese, *Active contours without edges*, IEEE Transactions in Image Processing **10** (2001), no. 2, 266–277, <http://dx.doi.org/10.1109/83.902291>. (Cited on pages 185, 186, 187, and 190).
- [CvGS⁺] L. Calatroni, Y. van Gennip, C.-B. Schönlieb, H. Rowland, et al., *Quantification and robust measurement of male pied flycatcher’s blaze via mathematical imaging*. (Cited on pages 47, 210, and 215).
- [CvGS⁺15] L. Calatroni, Y. van Gennip, C.-B. Schönlieb, H. Rowland, and A. Flenner, *Graph clustering, variational image segmentation methods and hough transform scale detection for object measurement in images*, in preparation (2015). (Cited on pages 42, 46, and 47).
- [CYP15] Y. Chen, W. Yu, and T. Pock, *On learning optimized reaction diffusion processes for effective image restoration*, IEEE Conference on Computer Vision and Pattern Recognition, (to appear) 2015, <http://arxiv.org/abs/1503.05768>. (Cited on pages 45 and 147).
- [DAG09] V. Duval, J.F. Aujol, and Y. Gousseau, *The TVL1 model: a geometric point of view*, Multiscale Modeling & Simulation **8** (2009), no. 1, 154–189, <http://dx.doi.org/10.1137/090757083>. (Cited on pages 35, 39, 43, 116, 131, 138, 139, and 165).
- [DH72] R. O. Duda and P. E. Hart, *Use of the Hough transformation to detect lines and curves in pictures*, Communications ACM **15** (1972), no. 1, 11–

- 15, <http://doi.acm.org/10.1145/361237.361242>. (Cited on pages 203 and 204).
- [DHRC11] Y. Dong, M. Hintermüller, and M. M. Rincon-Camacho, *Automated regularization parameter selection in multi-scale total variation models for image restoration*, *Journal of Mathematical Imaging and Vision* **40** (2011), no. 1, 82–104, <http://dx.doi.org/10.1007/s10851-010-0248-9>. (Cited on pages 153, 157, 180, and 220).
- [DLR11] J. C. De Los Reyes, *Optimal control of a class of variational inequalities of the second kind*, *SIAM Journal on Control and Optimization* **49** (2011), no. 4, 1629–1658, <http://dx.doi.org/10.1137/090764438>. (Cited on pages 155, 157, and 220).
- [DIR12] J. C. De los Reyes, *Optimization of mixed variational inequalities arising in flow of viscoplastic materials*, *Computational Optimization and Applications* **52** (2012), no. 3, 757–784, <http://dx.doi.org/10.1007/s10589-011-9435-x>. (Cited on page 137).
- [DLR15] J. C. De Los Reyes, *Numerical PDE-constrained optimization*, Springer Briefs in Optimization, Springer, Cham, 2015, <http://dx.doi.org/10.1007/978-3-319-13395-9>. (Cited on pages 47, 49, 65, 72, and 75).
- [DLRLM14] J. C. De Los Reyes, E. Loyaza, and P. Merino, *Second-order orthant-based methods with enriched hessian information for sparse- ℓ^1 minimization*, *ModeMat Report 14-06* (2014), <http://arxiv.org/abs/1407.1096>. (Cited on page 62).
- [DLRS13] J. C. De Los Reyes and C.-B. Schönlieb, *Image denoising: learning the noise model via nonsmooth PDE-constrained optimization*, *Inverse Problems and Imaging* **7** (2013), no. 4, 1183–1214. (Cited on pages 39, 43, 45, 47, 62, 116, 123, 126, 127, 128, 135, 137, 141, 147, 153, 154, 155, 156, 159, 161, 172, 174, and 183).
- [DLRSV15a] J. C. De Los Reyes, C.-B. Schönlieb, and T. Valkonen, *Bilevel parameter learning for higher-order total variation regularisation models*, submitted (2015). (Cited on pages 45, 149, and 150).
- [DLRSV15b] J.C. De Los Reyes, C.-B. Schönlieb, and T. Valkonen, *The structure of optimal parameters for image restoration problems*, to appear in *Journal of Mathematical Analysis and Applications* (2015). (Cited on pages 45, 47, 62, 128, 147, 149, 150, 151, 152, 159, and 160).

- [DM93] G. Dal Maso, *Introduction to Γ -convergence*, Birkhauser, 1993. (Cited on pages 47, 55, and 130).
- [DMM10] B. Düring, D. Matthes, and J. P. Milišić, *A gradient flow scheme for nonlinear fourth order equations*, Discrete and Continuous Dynamical Systems. Series B. A Journal Bridging Mathematics and Sciences **14** (2010), no. 3, 935–959, <http://dx.doi.org/10.3934/dcdsb.2010.14.935>. (Cited on page 82).
- [DS96] D. Dobson and F. Santosa, *Recovery of blocky images from noisy and blurred data*, SIAM Journal on Applied Mathematics **56** (1996), no. 4, 1181–1198, <http://dx.doi.org/10.1137/S003613999427560X>. (Cited on page 37).
- [DS12] B. Düring and C.-B. Schönlieb, *A high-contrast fourth-order PDE from imaging: numerical solution by ADI splitting*, Multi-scale and high-contrast PDE: from modelling, to mathematical analysis, to inversion, Contemp. Math., vol. 577, Amer. Math. Soc., Providence, RI, 2012, <http://dx.doi.org/10.1090/comm/577/11465>, pp. 93–103. (Cited on pages 43, 81, and 91).
- [DSLS99] S. Dale, T. Slagsvold, H. M. Lampe, and G. P. Sætre, *Population divergence in sexual ornaments: the white forehead patch of norwegian pied flycatchers is small and unsexy*, Evolution (1999), 1235–1246. (Cited on page 208).
- [DT84] F. Demengel and R. Temam, *Convex functions of a measure and applications*, Indiana University Mathematics Journal **33** (1984), 673–709. (Cited on page 129).
- [DW05] S. Didas and B. Weickert, J. and Burgeth, *Stability and local feature enhancement of higher order nonlinear diffusion filtering*, Pattern Recognition, Lecture Notes in Computer Science, vol. 3663, Springer Berlin Heidelberg, 2005, http://dx.doi.org/10.1007/11550518_56, pp. 451–458. (Cited on pages 78 and 94).
- [DWB09] S. Didas, J. Weickert, and B. Burgeth, *Properties of higher order nonlinear diffusion filtering*, Journal of Mathematical Imaging and Vision. **35** (2009), 208–226, <http://dx.doi.org/10.1007/s10851-009-0166-x>. (Cited on page 77).
- [EF87] C. M. Elliott and D. A. French, *Numerical studies of the Cahn-Hilliard equation for phase separation*, IMA Journal of Applied Mathematics **38**

- (1987), no. 2, 97–128, <http://dx.doi.org/10.1093/imamat/38.2.97>. (Cited on page 88).
- [Egg93] P. P. B. Eggermont, *Maximum entropy regularization for Fredholm integral equations of the first kind*, SIAM Journal on Mathematical Analysis **24** (1993), no. 6, 1557–1576. (Cited on page 125).
- [EHN96] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of inverse problems*, Mathematics and its Applications, vol. 375, Kluwer Academic Publishers Group, Dordrecht, 1996, <http://dx.doi.org/10.1007/978-94-009-1740-8>. (Cited on pages 41 and 154).
- [ELB08] A. Elmoataz, O. Lezoray, and S. Boughleux, *Nonlocal discrete regularization on weighted graphs: A framework for image and manifold processing*, Image Processing, IEEE Transactions on **17** (2008), no. 7, 1047–1060. (Cited on page 31).
- [ES02] S. Esedoglu and J. Shen, *Digital inpainting based on the Mumford-Shah-Euler image model*, European Journal of Applied Mathematics **13** (2002), no. 4, 353–370, <http://dx.doi.org/10.1017/S0956792502004904>. (Cited on pages 35 and 185).
- [ES07] C. M. Elliott and S. A. Smitheman, *Analysis of the TV regularization and H^{-1} fidelity model for decomposing an image into cartoon plus texture*, Communications on Pure and Applied Analysis **6** (2007), no. 4, 917–936, <http://dx.doi.org/10.3934/cpaa.2007.6.917>. (Cited on pages 81 and 95).
- [ES09] ———, *Numerical analysis of the TV regularization and H^{-1} fidelity model for decomposing an image into cartoon plus texture*, IMA Journal of Numerical Analysis **29** (2009), no. 3, 651–689, <http://dx.doi.org/10.1093/imanum/drn025>. (Cited on page 81).
- [Ese04] S. Esedoglu, *Blind deconvolution of bar code signals*, Inverse Problems **20** (2004), no. 1, 121–135, <http://dx.doi.org/10.1088/0266-5611/20/1/007>. (Cited on pages 35 and 185).
- [ET76] I. Ekeland and R. Temam, *Convex analysis and variational problems*, vol. 1, North Holland, 1976. (Cited on page 63).
- [ET06] S. Esedoglu and Y.-H. R. Tsai, *Threshold dynamics for the piecewise constant Mumford-Shah functional*, Journal of Computational Physics **211** (2006), no. 1, 367–384, <http://dx.doi.org/10.1016/j.jcp.2005.05.027>. (Cited on pages 35, 185, and 190).

BIBLIOGRAPHY

- [Eva10] L. C. Evans, *Partial differential equations*, second ed., Graduate Studies in Mathematics, vol. 19, American Mathematical Society, 2010. (Cited on pages 31, 47, 49, and 58).
- [Eyr98] D. Eyre, *An unconditionally stable one-step scheme for gradient systems*, unpublished, 1998. (Cited on page 199).
- [FBCM04] C. Fowlkes, S. Belongie, Fan Chung, and J. Malik, *Spectral grouping using the Nyström method*, IEEE Transactions on Pattern Analysis and Machine Intelligence **26** (2004), no. 2, 214–225, <http://dx.doi.org/10.1109/TPAMI.2004.1262185>. (Cited on pages 189, 195, and 198).
- [FMM12] K. Frick, P. Marnitz, and A. Munk, *Shape-constrained regularization by statistical multiresolution for inverse problems: asymptotic analysis*, Inverse Problems **28** (2012), no. 6, 065006, 31, <http://dx.doi.org/10.1088/0266-5611/28/6/065006>. (Cited on page 153).
- [Foi09] A. Foi, *Clipped noisy images: Heteroskedastic modeling and practical denoising*, Signal Processing **89** (2009), no. 12, 2609 – 2629, Special Section: Visual Information Analysis for Security. (Cited on page 116).
- [Fol95] G. B. Folland, *Introduction to Partial Differential Equations*, Princeton University Press, 1995. (Cited on page 31).
- [Fol99] ———, *Real analysis*, second ed., Pure and Applied Mathematics (New York), John Wiley & Sons, Inc., New York, 1999, Modern techniques and their applications, A Wiley-Interscience Publication. (Cited on page 54).
- [GB00] J. D. Gibson and A. Bovik, *Handbook of image and video processing*, 1st ed., Academic Press, Inc., Orlando, FL, USA, 2000. (Cited on page 115).
- [Get12] P. Getreuer, *Chan-Vese segmentation*, Image Processing On Line **2** (2012), 214–224, <http://dx.doi.org/10.5201/ipo1.2012.g-cv>. (Cited on page 216).
- [GG84] S. Geman and D. Geman, *Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images*, IEEE Transactions on Pattern Analysis and Machine Intelligence (1984), no. 6, 721–741. (Cited on page 30).
- [Giu84] E. Giusti, *Minimal surfaces and functions of bounded variation*, Birkhäuser, 1984. (Cited on pages 33, 47, and 50).
- [GM98] S. Guattery and G. L. Miller, *On the quality of spectral separators*, SIAM Journal on Matrix Analysis and Applications **19** (1998), no. 3,

- 701–719, <http://dx.doi.org/10.1137/S0895479896312262>. (Cited on page 189).
- [GMB⁺14a] C. Garcia-Cardona, E. Merkurjev, A. L. Bertozzi, A. Flenner, and A. G. Percus, *Diffuse interface methods for multiclass segmentation of high-dimensional data*, Applied Mathematical Letters **33** (2014), 29–34, <http://dx.doi.org/10.1016/j.aml.2014.02.008>. (Cited on pages 46, 189, 190, and 193).
- [GMB⁺14b] ———, *Multiclass data segmentation using diffuse interface methods on graphs*, IEEE Transactions on Pattern Analysis and Machine Intelligence **36** (2014), no. 8, 1600–1613, <http://doi.ieeecomputersociety.org/10.1109/TPAMI.2014.2300478>. (Cited on pages 46, 189, and 190).
- [GMmM01] F. Guichard, L. Moisan, and J. m. Morel, *A review of PDE models in image processing and image analysis*, Journal de Physique IV **2002** (2001), 137–154. (Cited on page 31).
- [GO07] G. Gilboa and S. Osher, *Image regularization and supervised segmentation*, Multiscale Modeling & Simulation **6** (2007), no. 2, 595–630, <http://dx.doi.org/10.1137/060669358>. (Cited on page 195).
- [GO09a] ———, *Nonlocal operators with applications to image processing*, Multiscale Modeling & Simulation **7** (2009), no. 3, 1005–1028, <http://dx.doi.org/10.1137/070698592>. (Cited on pages 193 and 195).
- [GO09b] T. Goldstein and S. Osher, *The split Bregman method for L^1 regularized problems*, SIAM Journal on Imaging Sciences **2** (2009), no. 2, 323–343, <http://dx.doi.org/10.1137/080725891>. (Cited on pages 36 and 82).
- [Gra06] L. Grady, *Random walks for image segmentation*, Pattern Analysis and Machine Intelligence, IEEE Transactions on **28** (2006), no. 11, 1768–1783. (Cited on page 31).
- [Gra14] J. Grah, *Methods for automatic mitosis detection and tracking in phase contrast images*, Master’s thesis, University of Münster, 2014. (Cited on page 203).
- [GSM10] Y. Guoshen, G. Sapiro, and S. Mallat, *Image modeling and enhancement via structured sparse model selection*, Image Processing (ICIP), 2010 17th IEEE International Conference on, Sept 2010, <http://dx/doi.org/10.1109/ICIP.2010.5653853>, pp. 1641–1644. (Cited on page 146).

- [GW06] R. C. Gonzalez and R. E. Woods, *Digital image processing*, 3rd ed., Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2006. (Cited on pages 22, 24, and 204).
- [HCH12] E. Haber, M. Chung, and F. Herrmann, *An effective method for parameter estimation with PDE constraints with multiple right-hand sides*, SIAM Journal on Optimization **22** (2012), no. 3, 739–757, <http://dx.doi.org/10.1137/11081126X>. (Cited on page 166).
- [HHH12] H. Huang, E. Haber, and L. Horesh, *Optimal estimation of ℓ_1 -regularization prior from a regularized empirical Bayesian risk standpoint*, Inverse Problems and Imaging **6** (2012), no. 3, 447–464, <http://dx.doi.org/10.3934/ipi.2012.6.447>. (Cited on page 45).
- [HHT10] E. Haber, L. Horesh, and L. Tenorio, *Numerical methods for the design of large-scale nonlinear discrete ill-posed inverse problems*, Inverse Problems **26** (2010), no. 2, 025002, 14, <http://dx.doi.org/10.1088/0266-5611/26/2/025002>. (Cited on page 147).
- [HK04] M. Hintermüller and K. Kunisch, *Total bounded variation regularization as a bilaterally constrained optimization problem*, SIAM Journal on Applied Mathematics **64** (2004), no. 4, 1311–1333, <http://dx.doi.org/10.1137/S0036139903422784>. (Cited on page 153).
- [HK14] M. Holler and K. Kunisch, *On infimal convolution of TV-type functionals and applications to video and image reconstruction*, SIAM Journal on Imaging Sciences **7** (2014), no. 4, 2258–2300, <http://dx.doi.org/10.1137/130948793>. (Cited on page 117).
- [HL13] M. Hintermüller and A. Langer, *Subspace correction methods for a class of nonsmooth and nonadditive convex variational problems with mixed L^1/L^2 data-fidelity in image processing*, SIAM Journal on Imaging Sciences **6** (2013), no. 4, 2134–2173, <http://dx.doi.org/10.1137/120894130>. (Cited on pages 39, 43, 116, 123, and 139).
- [HM03] O.-A. Hafsa and J.-P. Mandallena, *Interchange of infimum and integral*, Calculus of Variations and Partial Differential Equations **18** (2003), no. 4, 433–449, <http://dx.doi.org/10.1007/s00526-003-0211-3>. (Cited on page 120).
- [HMBS10] D. J. Holland, D. M. Malioutov, A. Blake, and Gladden L. F. Sederman, A. J., *Reducing data acquisition times in phase-encoded velocity imaging*

- using compressed sensing*, Journal of Magnetic Resonance **203** (2010), 236–246. (Cited on page 146).
- [Hol13] M. Holler, *Higher order regularization for model based data decompression*, Ph.D. thesis, University of Graz, 2013. (Cited on page 37).
- [Hou62] P. V. C. Hough, *A method and means for recognizing complex patterns*, US Patent: 3,069,654, December 1962. (Cited on pages 48 and 203).
- [HS06] M. Hintermüller and G. Stadler, *An infeasible primal-dual algorithm for total bounded variation-based inf-convolution-type image restoration*, SIAM Journal on Scientific Computing **28** (2006), no. 1, 1–23, <http://dx.doi.org/10.1137/040613263>. (Cited on pages 147 and 160).
- [HT03] E. Haber and L. Tenorio, *Learning regularization functionals—a supervised training approach*, Inverse Problems **19** (2003), no. 3, 611–626, <http://dx.doi.org/10.1088/0266-5611/19/3/309>. (Cited on pages 146 and 147).
- [HUL93] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex analysis and minimization algorithms. II*, Grundlehren der Mathematischen Wissenschaften, vol. 306, Springer-Verlag, Berlin, 1993, Advanced theory and bundle methods. (Cited on page 63).
- [HUL01] J.-B. Hiriart-Urruty and C. Lemaréchal, *Fundamentals of convex analysis*, Grundlehren Text Editions, Springer-Verlag, Berlin, 2001. (Cited on page 118).
- [Hun02] W. Hundsdorfer, *Accuracy and stability of splitting with stabilizing corrections*, Applied Numerical Mathematics **42** (2002), no. 1-3, 213–233, [http://dx.doi.org/10.1016/S0168-9274\(01\)00152-0](http://dx.doi.org/10.1016/S0168-9274(01)00152-0). (Cited on pages 84, 85, 86, and 104).
- [HV03] W. Hundsdorfer and J. Verwer, *Numerical solution of time-dependent advection-diffusion-reaction equations*, Springer Series in Computational Mathematics, vol. 33, Springer-Verlag, Berlin, 2003, <http://dx.doi.org/10.1007/978-3-662-09017-6>. (Cited on pages 47, 84, 85, 87, and 104).
- [HZK10] M. Herrmann, S. Zambanin, and M. Kampel, *Image-based measurement of ancient coins*, Making History Interactive. Computer Applications and Quantitative Methods in Archaeology (CAA), Proceedings of the 37th International Conference, Archaeopress, Oxford, 2010, pp. 117–121. (Cited on pages 217 and 218).

- [Idi13] J. Idier, *Bayesian approach to inverse problems*, John Wiley & Sons, 2013. (Cited on pages 29 and 115).
- [itHM13] K. J. in 't Hout and C. Mishra, *Stability of ADI schemes for multidimensional diffusion equations with mixed derivative terms*, Applied Numerical Mathematics **74** (2013), 83–94, <http://dx.doi.org/10.1016/j.apnum.2013.07.003>. (Cited on pages 86, 88, 102, and 104).
- [itHW07] K. J. in 't Hout and B. D. Welfert, *Stability of ADI schemes applied to convection-diffusion equations with mixed derivative terms*, Applied Numerical Mathematics **57** (2007), no. 1, 19–35, <http://dx.doi.org/10.1016/j.apnum.2005.11.011>. (Cited on pages 83, 86, and 104).
- [JCPT12] A. Jezierska, E. Chouzenoux, J.-C. Pesquet, and H. Talbot, *A primal-dual proximal splitting approach for restoring data corrupted with Poisson-Gaussian noise*, IEEE International Conference on Acoustics, Speech and Signal Processing (Kyoto), 2012. (Cited on pages 39, 43, 116, 123, and 141).
- [JCPT15] ———, *A Convex Approach for Image Restoration with Exact Poisson-Gaussian Likelihood*, SIAM Journal on Imaging Science **8** (2015), no. 4. (Cited on pages 43, 116, 123, and 141).
- [JKO98] R. Jordan, D. Kinderlehrer, and F. Otto, *The variational formulation of the Fokker-Planck equation*, SIAM Journal on Mathematical Analysis **29** (1998), no. 1, 1–17, <http://dx.doi.org/10.1137/S0036141096303359>. (Cited on page 81).
- [JLC13] P. E. Järvisvö, T. Laaksonen, and S. Calhim, *Forehead patch size predicts the outcome of male-male competition in the pied flycatcher*, Ethology **119** (2013), no. 8, 662–670. (Cited on pages 207 and 208).
- [Kel76] J. B. Keller, *Inverse problems*, The American Mathematical Monthly **83** (1976), no. 2, pp. 107–118, <http://www.jstor.org/stable/2976988>. (Cited on page 29).
- [Kel99] C. T. Kelley, *Iterative methods for optimization*, Frontiers in Applied Mathematics, vol. 18, SIAM, Philadelphia, PA, 1999, <http://dx.doi.org/10.1137/1.9781611970920>. (Cited on pages 49 and 69).
- [KP13] K. Kunisch and T. Pock, *A bilevel optimization approach for parameter learning in variational models*, SIAM Journal on Imaging Sciences **6** (2013), no. 2, 938–983, <http://dx.doi.org/10.1137/120882706>. (Cited on pages 45, 147, and 160).

-
- [KP15] T. Klatzer and T. Pock, *Continuous hyper-parameter learning for support vector machines*, Computer Vision Winter Workshop (CVWW), 2015. (Cited on page 147).
- [KS89] R. V. Kohn and P. Sternberg, *Local minimisers and singular perturbations*, Proceedings of the Royal Society of Edinburgh. Section A. Mathematics **111** (1989), no. 1-2, 69–84, <http://dx.doi.org/10.1017/S0308210500025026>. (Cited on page 35).
- [KTAK11] V. Kolehmainen, T. Tarvainen, S. R. Arridge, and J. P. Kaipio, *Marginalization of uninteresting distributed parameters in inverse problems—application to diffuse optical tomography*, International Journal for Uncertainty Quantification **1** (2011), no. 1, 1–17, <http://dx.doi.org/10.1615/Int.J.UncertaintyQuantification.v1.i1.10>. (Cited on page 146).
- [KWT88] M. Kass, A. Witkin, and D. Terzopoulos, *Snakes: Active contour models*, International Journal of Computer Vision **1** (1988), no. 4, 321–331, <http://dx.doi.org/10.1007/BF00133570>. (Cited on page 185).
- [LA92] A. Lundberg and R. V. Alatalo, *The pied flycatcher*, London: Poyser, 1992. (Cited on page 207).
- [LBLZ] F. Li, Z. Bao, R. Liu, and G. Zhang, Journal of Visual Communication and Image Representation. (Cited on page 36).
- [LBU11] F. Luisier, T. Blu, and M. Unser, *Image denoising in mixed Poisson-Gaussian noise*, IEEE Transactions on Image Processing **20** (2011), 696–708. (Cited on pages 39 and 116).
- [LCA07] T. Le, R. Chartrand, and T.J. Asaki, *A variational approach to reconstructing images corrupted by Poisson noise*, Journal of Mathematical Imaging and Vision **27** (2007), no. 3, 257–263, <http://dx.doi.org/10.1007/s10851-007-0652-y>. (Cited on pages 36, 39, 43, 122, 133, and 141).
- [LH95] C. L. Lawson and R. J. Hanson, *Solving least squares problems*, Classics in Applied Mathematics, SIAM, 1995. (Cited on page 41).
- [LL01] E. H. Lieb and M. Loss, *Analysis*, vol. 14 of Graduate studies in Mathematics series, American Mathematical Society, Providence, RI., 2001. (Cited on page 49).
- [LLSV14] J. Lellmann, D. Lorenz, C.-B. Schönlieb, and T. Valkonen, *Imaging with Kantorovich–Rubinstein discrepancy*, SIAM Journal on Imaging Sciences

- 7** (2014), no. 4, 2833–2859, <http://dx.doi.org/10.1137/140975528>. (Cited on page 82).
- [LNT91] T. Lu, P. Neittaanmäki, and X.-C. Tai, *A parallel splitting up method and its application to Navier-Stokes equations*, Applied Mathematics Letters **4** (1991), no. 2, 25–29, [http://dx.doi.org/10.1016/0893-9659\(91\)90161-N](http://dx.doi.org/10.1016/0893-9659(91)90161-N). (Cited on page 87).
- [LNT92] ———, *A parallel splitting-up method for partial differential equations and its applications to Navier-Stokes equations*, RAIRO Modélisation Mathématique et Analyse Numérique **26** (1992), no. 6, 673–708. (Cited on page 87).
- [LP05] Zelnik-Manor L. and Perona P., *Self-tuning spectral clustering*, Advances in Neural Information Processing Systems 17, MIT Press, 2005, pp. 1601–1608. (Cited on page 194).
- [LT06] M. Lysaker and X. C. Tai, *Iterative image restoration combining total variation minimization and a second-order functional*, International Journal of Computer Vision **66** (2006), no. 1, 5–18, <http://dx.doi.org/10.1007/s11263-005-3219-7>. (Cited on page 79).
- [LV08] L. H. Lieu and L. A. Vese, *Image restoration and decomposition via bounded total variation and negative Hilbert-Sobolev spaces*, Applied Mathematics and Optimization **58** (2008), no. 2, 167–193, <http://dx.doi.org/10.1007/s00245-008-9047-8>. (Cited on pages 78, 79, and 81).
- [MBBS14] M. Möller, E. M. Brinkmann, M. Burger, and T. Seybold, *Color Bregman TV*, SIAM Journal on Imaging Sciences **7** (2014), no. 4, 2771–2806, <http://dx.doi.org/10.1137/130943388>. (Cited on page 36).
- [MBBT15] E. Merkurjev, E. Bae, A. Bertozzi, and X.-C. Tai, *Global binary optimization on graphs for classification of high-dimensional data*, Journal of Mathematical Imaging and Vision **52** (2015), no. 3, 414–435, <http://dx.doi.org/10.1007/s10851-015-0567-y>. (Cited on page 46).
- [MBPS09] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, *Online dictionary learning for sparse coding*, Proceedings of the 26th Annual International Conference on Machine Learning (New York, NY, USA), ICML '09, ACM, 2009, <http://doi.acm.org/10.1145/1553374.1553463>, pp. 689–696. (Cited on page 146).
- [Mem11] F. Memoli, *Gromov-Wasserstein distances and the metric approach to object matching*, Foundations of Computational Mathematics **11** (2011),

- no. 4, 417–487, <http://dx.doi.org/10.1007/s10208-011-9093-5>. (Cited on page 82).
- [Mey01] Y. Meyer, *Oscillating patterns in image processing and nonlinear evolution equations*, vol. 22, American Mathematical Society, 2001. (Cited on pages 35, 78, and 137).
- [MM77a] L. Modica and S. Mortola, *Il limite nella Γ -convergenza di una famiglia di funzionali ellittici*, Bollettino della Unione Matematica Italiana, Serie A (5) **14** (1977), no. 3, 526–529. (Cited on pages 35 and 190).
- [MM77b] ———, *Il limite nella Γ -convergenza di una famiglia di funzionali ellittici*, Bollettino della Unione Matematica Italiana, Serie B (5) **14** (1977), no. 3, 285–299. (Cited on pages 35 and 190).
- [MMM⁺07] J. Morales, J. Moreno, S. Merino, J. J. Sanz, G. Tomas, and E. et al. Arriero, *Female ornaments in the pied flycatcher *ficedula hypoleuca*: associations with age, health and reproductive success*, Ibis **149** (2007), no. 2, 245–254. (Cited on page 208).
- [Mor66] V. A. Morozov, *On the solution of functional equations by the method of regularization*, Soviet Math. Dokl. **7** (1966), 414–417. (Cited on pages 41 and 154).
- [MPCB15] A. M. Massone, A. Perasso, C. Campi, and M. C. Beltrametti, *Profile detection in medical and astronomical images by means of the Hough transform of special classes of curves*, Journal of Mathematical Imaging and Vision **51** (2015), no. 2, 296–310, <http://dx.doi.org/10.1007/s10851-014-0521-4>. (Cited on page 203).
- [MS89] D. Mumford and J. Shah, *Optimal approximations by piecewise smooth functions and associated variational problems*, Communications on Pure and Applied Mathematics **42** (1989), no. 5, 577–685, <http://dx.doi.org/10.1002/cpa.3160420503>. (Cited on pages 31, 185, 186, 187, and 209).
- [MS12] J.-M. Morel and S. Solimini, *Variational methods in image segmentation: with seven image processing experiments*, vol. 12, Springer Science & Business Media, 2012. (Cited on page 188).
- [MSG01] M. D. Mantle, A. J. Sederman, and L. F. Gladden, *Single- and two-phase flow in fixed-bed reactors: MRI flow visualisation and lattice-Boltzmann simulations*, Chemical Engineering Science **56** (2001), no. 2, 523 – 529. (Cited on page 166).

- [Mül08] J. Müller, *Parallel total variation minimization*, Diploma thesis, University of Muenster, WWU, november 2008. (Cited on pages 47, 64, 99, 100, 110, and 112).
- [Mül13] ———, *Advanced image reconstruction and denoising – Bregmanized (higher order) total variation and application in PET*, Ph.D. thesis, University of Münster, 2013. (Cited on pages 36 and 37).
- [MVRDCn⁺11] J. Moreno, A. Velando, R. Ruiz-De-Castañeda, A. Cantarero, S. Gonzalez-Braojos, and A. Redondo, *Plasma antioxidant capacity and oxidative damage in relation to male plumage ornamental traits in a montane iberian pied flycatcher *ficedula hypoleuca* population*, *Acta Ornithologica* **46** (2011), no. 1, 65–70. (Cited on page 208).
- [MWHW78] R. J. Marks, G. L. Wise, D. G. Haldeman, and J. L. Whited, *Detection in Laplace noise*, *IEEE Transactions on Aerospace and Electronic Systems* **14** (1978), no. 6, 866–872. (Cited on page 119).
- [Mye98] T. G. Myers, *Thin films with high surface tension*, *SIAM Review* **40** (1998), no. 3, 441–462, <http://dx.doi.org/10.1137/S003614459529284X>. (Cited on page 78).
- [NCS84] A. Novick-Cohen and L. A. Segel, *Nonlinear aspects of the Cahn-Hilliard equation*, *Physica D: Nonlinear Phenomena* **10** (1984), no. 3, 277–298, [http://dx.doi.org/10.1016/0167-2789\(84\)90180-5](http://dx.doi.org/10.1016/0167-2789(84)90180-5). (Cited on page 88).
- [NDRM07] M. M. Naeini, G. Dutton, K. Rothley, and G. Mori, *Action recognition of insects using spectral clustering*, MVA 2007 IAPR Conference on Machine Vision Applications, 2007. (Cited on pages 189 and 195).
- [Nes09] Y. Nesterov, *Primal-dual subgradient methods for convex problems*, *Mathematical Programming* **120** (2009), no. 1, Ser. B, 221–259, <http://dx.doi.org/10.1007/s10107-007-0149-x>. (Cited on page 166).
- [Nik02] M. Nikolova, *Minimizers of cost-functions involving nonsmooth data-fidelity terms. Application to the processing of outliers*, *SIAM Journal on Numerical Analysis* **40** (2002), no. 3, 965–994, <http://dx.doi.org/10.1137/S0036142901389165>. (Cited on pages 35 and 39).
- [Nik04] ———, *A variational approach to remove outliers and impulse noise*, *Journal of Mathematical Imaging and Vision* **20** (2004), no. 1, 99–120, <http://dx.doi.org/10.1023/B:JMIV.0000011326.88682.e5>. (Cited on pages 35, 39, 43, 116, 131, 139, 154, and 165).

- [NMS93] M. Nitzberg, D. Mumford, and T. Shiota, *Filtering, segmentation, and depth*, Springer-Verlag New York, Inc., 1993. (Cited on page 38).
- [NQYH07] M. K. Ng, L. Qi, Y.-F. Yang, and Y.-M. Huang, *On semismooth Newton's methods for total variation minimization*, Journal of Mathematical Imaging and Vision **27** (2007), no. 3, 265–276, <http://dx.doi.org/10.1007/s10851-007-0650-0>. (Cited on page 36).
- [NW99] J Nocedal and S. J. Wright, *Numerical optimization*, Springer Series in Operations Research, Springer-Verlag, New York, 1999, <http://dx.doi.org/10.1007/b98874>. (Cited on page 49).
- [Nys29] E. J. Nyström, *Über die praktische auflösung von linearen integralgleichungen mit anwendungen auf randwertaufgaben der potentialtheorie*, Commentationes Physico-Mathematicae **4** (1929), no. 15, 1–52. (Cited on pages 46, 48, 189, and 195).
- [OBG⁺05] S. Osher, M. Burger, D. Goldfarb, J. Xu, and W. Yin, *An iterative regularization method for total variation-based image restoration*, Multiscale Modeling & Simulation **4** (2005), 460–489, <http://dx.doi.org/10.1137/040605412>. (Cited on pages 36 and 137).
- [ODB97] A. Oron, S. H. Davis, and S. G. Bankoff, *Long-scale evolution of thin liquid films*, Reviews of Modern Physics **69** (1997), 931–980, <http://link.aps.org/doi/10.1103/RevModPhys.69.931>. (Cited on page 78).
- [OF96] B. A. Olshausen and D. J. Field, *Emergence of simple-cell receptive field properties by learning a sparse code for natural images*, Nature **381** (1996), 607–609. (Cited on page 146).
- [ORBP15] P. Ochs, R. Ranftl, T. Brox, and T. Pock, *Bilevel optimization with nonsmooth lower level problems*, International Conference on Scale Space and Variational Methods in Computer Vision (SSVM), 2015, to appear. (Cited on page 147).
- [OSV03] S. Osher, A. Solé, and L. A. Vese, *Image decomposition and restoration using total variation minimization and the H^{-1} norm*, Multiscale Modeling & Simulation **1** (2003), no. 3, 349–370, <http://dx.doi.org/10.1137/S1540345902416247>. (Cited on pages 78 and 79).
- [Ott99] F. Otto, *The geometry of dissipative evolution equations the geometry of dissipative evolution equations: the porous medium equation*, Communications in Partial Differential Equations **26** (1999), no. 1-2. (Cited on page 80).

- [PB15] K. Papafitsoros and K. Bredies, *A study of the one dimensional total generalised variation regularisation problem*, Inverse Problems and Imaging **9** (2015), no. 2. (Cited on page 37).
- [Pey] G. Peyré, *Entropic approximation of Wasserstein gradient flows*, to appear in SIAM Journal on Imaging Sciences. (Cited on page 82).
- [Pey11] ———, *The numerical tours of signal processing*, Advanced and Image Processing IEEE Computing in Science and Engineering **13** (2011), no. 4, 94–97. (Cited on page 39).
- [PF11] G. Peyré and J. M. Fadili, *Learning Analysis Sparsity Priors*, SAMPTA 11 (Singapore), 2011, <https://hal.archives-ouvertes.fr/hal-00542016>. (Cited on page 146).
- [PIK92] J. Princen, J. Illingworth, and J. Kittler, *A formal definition of the Hough transform: Properties and relationships*, Journal of Mathematical Imaging and Vision **1** (1992), no. 2, 153–168, <http://dx.doi.org/10.1007/BF00122210>. (Cited on page 204).
- [PJ92] B. T. Polyak and A. B. Juditsky, *Acceleration of stochastic approximation by averaging*, SIAM Journal on Control and Optimization **30** (1992), no. 4, 838–855, <http://dx.doi.org/10.1137/0330046>. (Cited on page 166).
- [PM99] J. Potti and S. Montalvo, *Male arrival and female mate choice in pied flycatchers *ficedula hypoleuca* in central spain*, Ornis Scandinavica **22** (199), no. 1, 45–54. (Cited on pages 207, 208, and 210).
- [PM90] P. Perona and J. Malik, *Scale-space and edge detection using anisotropic diffusion*, IEEE Transactions on Pattern Analysis and Machine Intelligence **12** (1990), no. 7, 629–639. (Cited on page 32).
- [PR55] D. W. Peaceman and H. H. Rachford, Jr., *The numerical solution of parabolic and elliptic differential equations*, Journal of the Society of Industrial and Applied Mathematics **3** (1955), 28–41. (Cited on pages 47, 84, and 85).
- [PS08] C. Pöschl and O. Scherzer, *Characterization of minimisers of convex regularization functionals*, Contemporary mathematics **451** (2008), 219–248. (Cited on page 37).
- [PS14] K. Papafitsoros and C.-B. Schönlieb, *A combined first and second order variational approach for image reconstruction*, Journal of Mathematical Imaging and Vision **48** (2014), no. 2, 308–338, <http://dx.doi.org/10.1007/s10851-013-0445-4>. (Cited on pages 37, 117, 137, and 146).

- [PWO89] C. B. Price, P. Wambacq, and A. Oosterlinck, *Applications of reaction-diffusion equations to image processing*, Third International Conference on Image Processing and its Applications, 1989, pp. 49–53. (Cited on page 31).
- [RA07] E. Resmerita and R. S. Anderssen, *Joint additive Kullback-Leibler residual minimization and regularization for linear inverse problems*, Mathematical Methods in the Applied Sciences **30** (2007), no. 13, 1527–1544. (Cited on pages 125 and 126).
- [RB05] S. Roth and M.J. Black, *Fields of experts: a framework for learning image priors*, IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2005, vol. 2, June 2005, pp. 860–867 vol. 2. (Cited on page 146).
- [RLN⁺13] S. Ruuskanen, E. Lehtikoinen, M. Nikinmaa, H. Siitari, W. Waser, and T. Laaksonen, *Long-lasting effects of yolk androgens on phenotype in the pied flycatcher (*ficedula hypoleuca*)*, Behavioural Ecology and Sociobiology **67** (2013), no. 3, 361–372. (Cited on page 208).
- [RM51] H. Robbins and S. Monro, *A stochastic approximation method*, Annals of Mathematical Statistics **22** (1951), 400–407. (Cited on page 166).
- [Roc76] R. T. Rockafellar, *Integral functionals, normal integrands and measurable selections*, Nonlinear operators and the calculus of variations (Summer School, Univ. Libre Bruxelles, Brussels, 1975), Springer, Berlin, 1976, pp. 157–207. Lecture Notes in Math., Vol. 543. (Cited on page 120).
- [ROF92] L.I. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Physica D: Nonlinear Phenomena **60** (1992), no. 1-4, 259–268, [http://dx.doi.org/10.1016/0167-2789\(92\)90242-F](http://dx.doi.org/10.1016/0167-2789(92)90242-F). (Cited on pages 31, 32, 33, 34, 35, 36, 39, 43, 60, 61, 63, 115, 131, 133, 138, 139, 141, 154, and 186).
- [RP14] R. Ranftl and T. Pock, *A deep variational model for image segmentation*, 36th German Conference on Pattern Recognition (GCPR), 2014. (Cited on page 147).
- [Rud87] W. Rudin, *Real and complex analysis*, McGraw-Hill, Boston, 1987. (Cited on pages 47 and 50).
- [Saw11] A. Sawatzky, *(Nonlocal) total variation in medical imaging*, Ph.D. thesis, University of Münster, 2011. (Cited on pages 36, 39, 116, 120, 122, 125, 126, 129, 141, and 143).

- [SB11] C.-B. Schönlieb and A. Bertozzi, *Unconditionally stable schemes for higher order inpainting*, *Communications in Mathematical Sciences* **9** (2011), no. 2, 413–457. (Cited on pages [32](#), [38](#), [42](#), [48](#), [79](#), [81](#), [199](#), [200](#), and [220](#)).
- [SBBH09] C.-B. Schönlieb, A. Bertozzi, M. Burger, and L. He, *Image inpainting using a fourth-order total variation flow*, SAMPTA'09, International Conference on Sampling Theory and Applications, Marseilles (2009). (Cited on pages [38](#), [42](#), and [78](#)).
- [SBMB09] A. Sawatzky, C. Brune, J. Müller, and M. Burger, *Total variation processing of images with Poisson statistics*, *Computer Analysis of Images and Patterns*, Springer, 2009, http://dx.doi.org/10.1007/978-3-642-03767-2_65, pp. 533–540. (Cited on pages [36](#), [39](#), [43](#), [116](#), [122](#), [133](#), [143](#), and [154](#)).
- [SC02] J. Shen and T. F. Chan, *Mathematical models for local nontexture inpaintings*, *SIAM Journal on Applied Mathematics* **62** (2002), no. 3, 1019–1043, <http://dx.doi.org/10.1137/S0036139900368844>. (Cited on pages [34](#), [36](#), [42](#), and [79](#)).
- [Sch98] O. Scherzer, *Denoising with higher order derivatives of bounded variation and an application to parameter estimation*, *Computing* **60** (1998), no. 1, 1–27, <http://dx.doi.org/10.1007/BF02684327>. (Cited on page [37](#)).
- [Sch09a] C.-B. Schönlieb, *Modern PDE techniques for image inpainting*, Ph.D. thesis, University of Cambridge, 2009. (Cited on pages [37](#) and [42](#)).
- [Sch09b] C.-B. Schönlieb, *Total variation minimization with an H^{-1} constraint*, *Singularities in nonlinear evolution phenomena and applications*, CRM Series, vol. 9, Ed. Norm., Pisa, 2009, pp. 201–234. (Cited on page [81](#)).
- [SL09] P. M. Sirkiä and T. Laaksonen, *Distinguishing between male and territory quality: females choose multiple traits in the pied flycatcher*, *Animal Behaviour* **78** (2009), no. 5, 1051–1060. (Cited on page [208](#)).
- [SM00] J. Shi and J. Malik, *Normalized cuts and image segmentation*, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22** (2000), no. 8, 888–905, <http://dx.doi.org/10.1109/34.868688>. (Cited on pages [31](#), [189](#), [192](#), and [193](#)).
- [SMB94] J.-L. Starck, F. D. Murtagh, and A. Bijaoui, *Image restoration with noise suppression using a wavelet transform and a multiresolution sup-*

- port constraint*, SPIE proceedings, vol. 2302, 1994, pp. 132–143. (Cited on page 116).
- [Sme03] P. Smereka, *Semi-implicit level set methods for curvature and surface diffusion motion*, Journal of Scientific Computing **19** (2003), no. 1-3, 439–456, <http://dx.doi.org/10.1023/A:1025324613450>. (Cited on page 77).
- [SMS⁺97] G.-P. Stre, T. Moum, Bures S., Král M., M. Adamjan, and J. Moreno, *A sexually selected character displacement in flycatchers reinforces pre-mating isolation*, Nature **387** (1997), 589–592. (Cited on page 207).
- [SR14] U. Schmidt and S. Roth, *Shrinkage fields for effective image restoration*, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, 2014, pp. 2774–2781. (Cited on pages 45 and 147).
- [Stu10] A. M. Stuart, *Inverse problems: a Bayesian perspective*, Acta Numerica **19** (2010), 451–559. (Cited on pages 29, 47, 115, and 146).
- [SW97] M. Stoer and F. Wagner, *A simple min-cut algorithm*, Journal of the ACM **44** (1997), no. 4, 585–591, <http://dx.doi.org/10.1145/263867.263872>. (Cited on page 189).
- [Tap07] M. F. Tappen, *Utilizing variational optimization to learn Markov random fields*, IEEE Conference on Computer Vision and Pattern Recognition, 2007, June 2007, pp. 1–8. (Cited on page 146).
- [Tik63] A. Tikhonov, *Solution of incorrectly formulated problems and the regularization method*, Soviet Math. Dokl., vol. 5, 1963, p. 1035. (Cited on pages 30 and 32).
- [Trö10] F. Tröltzsch, *Optimal control of partial differential equations*, Graduate Studies in Mathematics, vol. 112, American Mathematical Society, Providence, RI, 2010, Theory, methods and applications, <http://dx.doi.org/10.1090/gsm/112>. (Cited on pages 47, 65, and 68).
- [TS80] R. Temam and G. Strang, *Functions of bounded deformation*, Archive for Rational Mechanics and Analysis **75** (1980), no. 1, 7–21, <http://dx.doi.org/10.1007/BF00284617>. (Cited on page 150).
- [TSC13] T. Teuber, G. Steidl, and R. H. Chan, *Minimization and parameter estimation for seminorm regularization models with I-divergence constraints*, Inverse Problems **29** (2013), 1–28. (Cited on pages 41 and 154).

- [UC97] G. Upton and I. Cook, *Understanding statistics*, 1997. (Cited on page 214).
- [Val04] T. Valkonen, *The jump set under geometric regularization. Part 1: basic technique and first-order denoising*, SIAM Journal on Mathematical Analysis **47** (2004), no. 4, 2587–2629. (Cited on page 62).
- [VBK13] T. Valkonen, K. Bredies, and F. Knoll, *Total generalized variation in diffusion tensor imaging*, SIAM Journal on Imaging Sciences **6** (2013), no. 1, 487–525, <http://dx.doi.org/10.1137/120867172>. (Cited on pages 37, 137, and 150).
- [VC97] L. Vese and T. F. Chan, *Reduced non-convex functional approximations for image restoration & segmentation*, 1997, UCLA CAM Report 97-56. (Cited on pages 185 and 187).
- [vdHV79] P. J. van der Houwen and J. G. Verwer, *One-step splitting methods for semidiscrete parabolic equations*, Computing **22** (1979), no. 4, 291–309, <http://dx.doi.org/10.1007/BF02265311>. (Cited on pages 47, 84, and 85).
- [Ves01] L. Vese, *A study in the BV space of a denoising-deblurring variational problem*, Applied Mathematics and Optimization **44** (2001), no. 2, 131–161. (Cited on pages 35 and 78).
- [VFC12] F. Viola, A. Fitzgibbon, and R. Cipolla, *A unifying resolution-independent formulation for early vision*, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, IEEE, 2012, pp. 494–501. (Cited on page 147).
- [vGB12] Y. van Gennip and A. L. Bertozzi, *Γ -convergence of graph Ginzburg-Landau functionals*, Advances in Differential Equations **17** (2012), no. 11–12, 1115–1180. (Cited on page 192).
- [vGGOB14] Y. van Gennip, N. Guillen, B. Osting, and A. L. Bertozzi, *Mean curvature, threshold dynamics, and phase field theory on finite graphs*, Milan Journal of Mathematics **82** (2014), no. 1, 3–65, <http://dx.doi.org/10.1007/s00032-014-0216-8>. (Cited on pages 192 and 193).
- [Vil03] C. Villani, *Topics in optimal transportation*, Graduate Studies in Mathematics, vol. 58, American Mathematical Society, Providence, RI, 2003. (Cited on page 79).

-
- [vL07] U. von Luxburg, *A tutorial on spectral clustering*, *Statistics and Computing* **17** (2007), no. 4, 395–416, <http://dx.doi.org/10.1007/s11222-007-9033-z>. (Cited on pages 192 and 193).
- [VO96] C. R. Vogel and M. E. Oman, *Iterative methods for total variation denoising*, *SIAM Journal on Scientific Computing* **17** (1996), no. 1, 227–238, <http://dx.doi.org/10.1137/0917016>. (Cited on page 60).
- [Vog02] C. R. Vogel, *Computational methods for inverse problems*, *Frontiers in Applied Mathematics*, vol. 23, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2002, <http://dx.doi.org/10.1137/1.9780898717570>. (Cited on page 60).
- [VSK85] Y. Vardi, L. Shepp, and L. Kaufman, *A statistical model for positron emission tomography*, *Journal of the American Statistical Association* **80** (1985), 8–20. (Cited on page 146).
- [VZ05] M. Varma and A. Zisserman, *A statistical approach to texture classification from single images*, *International Journal of Computer Vision* **62** (2005), no. 1–2, 61–81, <http://dx.doi.org/10.1007/s11263-005-4635-4>. (Cited on pages 201, 202, and 211).
- [Wah77] G. Wahba, *Practical approximate solutions to linear operator equations when the data are noisy*, *SIAM Journal on Numerical Analysis* **14** (1977), no. 4, pp. 651–667. (Cited on pages 41 and 154).
- [WB03] T. P. Witelski and M. Bowen, *ADI schemes for higher-order nonlinear diffusion equations*, *Applied Numerical Mathematics* **45** (2003), no. 2-3, 331–351, [http://dx.doi.org/10.1016/S0168-9274\(02\)00194-0](http://dx.doi.org/10.1016/S0168-9274(02)00194-0). (Cited on pages 86 and 88).
- [Wei98] J. Weickert, *Anisotropic diffusion in image processing*, Teubner, Stuttgart, 1998. (Cited on page 32).
- [WF95] M. Wu and W. J. Fitzgerald, *Analytical approach to changepoint detection in Laplacian noise*, *IEE Proceedings - Vision, Image and Signal Processing*, vol. 142, 1995, pp. 174–180. (Cited on page 119).
- [WRV98] J. Weickert, B. M. T. H. Romeny, and M. A. Viergever, *Efficient and reliable schemes for nonlinear diffusion filtering*, *IEEE Transactions on Image Processing* **7** (1998), 398–410. (Cited on page 87).
- [ZWC10] M. Zhu, S. J. Wright, and T. F. Chan, *Duality-based algorithms for total-variation-regularized image restoration*, *Computational Optimization and*

BIBLIOGRAPHY

Applications **47** (2010), no. 3, 377–400, <http://dx.doi.org/10.1007/s10589-008-9225-2>. (Cited on page 36).