Acta Numerica (2018), pp. 1–111
doi:10.1017/S0962492918000016

Modern regularization methods for inverse problems

Martin Benning
DAMTP, Centre for Mathematical Sciences,
Wilberforce Road, Cambridge CB3 0WA, UK
E-mail: mb941@cam.ac.uk

Martin Burger
Institute for Computational and Applied Mathematics,
University of Münster, Einsteinstrasse 62,
D-48149 Münster, Germany
E-mail: martin.burger@wwu.de

Regularization methods are a key tool in the solution of inverse problems. They are used to introduce prior knowledge and allow a robust approximation of ill-posed (pseudo-) inverses. In the last two decades interest has shifted from linear to nonlinear regularization methods, even for linear inverse problems. The aim of this paper is to provide a reasonably comprehensive overview of this shift towards modern nonlinear regularization methods, including their analysis, applications and issues for future research.

In particular we will discuss variational methods and techniques derived from them, since they have attracted much recent interest and link to other fields, such as image processing and compressed sensing. We further point to developments related to statistical inverse problems, multiscale decompositions and learning theory.

CONTENTS

1. Introduction

Starting from the development of tomography and related techniques, the last fifty years have seen a constant rise in interest in the development of *inverse problems* as a research field, in mathematics as well as applied fields such as medical imaging, geophysics and the oil industry, or the steel industry, to mention only a few: see, for example, Bertero and Boccacci (1998), Cakoni and Colton (2005), Chadan, Colton, Päivärinta and Rundell (1997), Colton and Kress (2012), Colton *et al.* (2012), Engl, Louis and Rundell (2012), Groetsch (1993), Isakov (2006, 2008), Natterer (2001), Natterer and Wübbeling (2001), Tarantola and Valette (1982) and Tarantola (2005). Connected with the rise of interest in inverse problems is the development and analysis of *regularization methods*, which are a necessity in most inverse problems due to their ill-posedness: see, for example, Tikhonov, Goncharsky and Bloch (1987) and Engl, Hanke and Neubauer (1996). In particular there is usually no continuous dependence between the data and the solution of the inverse problem, hence in the presence of measurement errors one solves approximate problems with stable dependence instead. The controlled construction and analysis of such modified problems is called regularization, usually with a regularization parameter encoding the level of the approximation.

The canonical example of an ill-posed inverse problem at the abstract level is the linear operator equation

$$Ku = f, \tag{1.1}$$

with a linear operator K between Banach spaces, whose generalized inverse $K^\dagger$ is unbounded. A regularization method is then some parametric approximation R_α of $K^\dagger$, which has better stability properties. In the case of linear regularization methods, R_α is a family of bounded linear operators

converging pointwise to $K^\dagger$ on the domain of the latter as $\alpha \to 0$. A key question in this respect is the convergence for noisy data, related to the choice of the regularization parameter α in dependence on the noise level δ, the latter being a bound for the noise in the deterministic setting or some kind of variance in a stochastic setting.

At the end of the twentieth century, when rather complete understanding of such linear regularization methods was available (based on spectral decompositions of the operators), nonlinear regularization methods, *i.e.* nonlinear maps R_α (possibly even multivalued), were becoming a field of intensive study. This was driven in particular by developments related to variational methods such as total variation techniques (Rudin, Osher and Fatemi 1992, Acar and Vogel 1994, Burger and Osher 2013) or sparsity and compressed sensing (Donoho 2006, Donoho, Elad and Temlyakov 2006, Candès and Donoho 2002), but also by statistical approaches such as advanced Bayesian prior models (Lassas, Saksman and Siltanen 2009, Helin and Lassas 2011, Kolehmainen, Lassas, Niinimäki and Siltanen 2012). Due to the rise of big data and learning techniques, in recent years there has been further interest in applying such paradigms to inverse problems. This is a somewhat delicate task, since in most inverse problems there are no ground truth data, but only results that have been reconstructed via a certain regularization method and specific noise. Hence there are many challenges for future research.

In this paper we will provide a survey of developments in the analysis and applications of modern (nonlinear) regularization methods during the last few decades. Moreover, we will try to provide a fairly structured overview of this field, including some fundamentals of nonlinear regularization methods. In particular we will give clear definitions as to what to expect from a regularization method and its convergence, reminiscent of the rather complete treatment of linear regularization methods in the seminal book by Engl, Hanke and Neubauer (1996), now dating back more than twenty years.

Throughout the paper we assume that $K : \mathcal{U} \to \mathcal{V}$ is a bounded linear operator on Banach spaces $\mathcal{U}$ and $\mathcal{V}$. In many parts there are obvious extensions to nonlinear operators and even metric spaces, but we mainly leave them out in order to increase readability; some links to such extensions are given at the end of the paper.

We will start in Section 2 with a historical exposition on regularization methods, and then proceed in Section 3 to nonlinear variational models, which are the class of methods driving most development in nonlinear regularizations. Section 4 will discuss some basic properties of and requirements on regularization methods, which are then discussed in detail for variational regularization in Section 5. Subsequently we turn to iterative regularization methods in Section 6. As a result of certain insights in these sections we

are led to a discussion on bias and scales in regularization methods in Section 7, and Section 8 will provide some examples of applications. Section 9 will discuss advanced aspects such as nonlinear regularization methods for nonlinear inverse problems and links to machine learning. Finally we conclude and provide an outlook to relevant future topics in Section 10.

2. A little history of regularization methods

It seems difficult to date the origin of regularization methods, but it is now common to identify it with the pioneering work of Tikhonov (1943, 1963, 1966) and the subsequent strong developments in the Russian community in the 1960s (*e.g.* Ivanov 1962, Bakushinskii 1967). The starting motivation obviously comes from the concept of *ill-posedness*, negating the definition of a well-posed problem. The latter, consisting of existence, uniqueness and stable dependence upon the input data, is usually attributed to the work of Hadamard in the context of partial differential equations (Hadamard 1902, 1923). However, the third condition was not clearly formulated in those problems, and only later found its true place of importance, for example in the work of John (1960). As a motivation for regularization theory and in particular for convergence theory, however, the lack of stability seems to be the most crucial issue.

It was already understood in early work that in order to have any chance of computing meaningful solutions, the problem needs to be approximated by well-posed ones, usually a family parametrized by the regularization parameter. The obvious first answer of a topologist such as Tikhonov was to restrict the domain to a compact set in some topology (or some kind of family thereof), leading to the concept of conditional well-posedness. A natural choice in a Hilbert space is to use norm balls centred at zero, which are compact in the weak topology. The radius of the ball (or its inverse) can naturally serve as a regularization parameter. This was also called the *selection method*, and the corresponding solutions were termed *quasi-solutions*. Given a minimization problem, *e.g.* least-squares $\|Ku-f\|^2$ for (1.1) in Hilbert spaces, it is a shortcut to the variational formulation (see Section 3 for a detailed discussion of variational models) of what is now called *Tikhonov* or *Tikhonov–Phillips* regularization. Indeed, with an appropriate Lagrange parameter α, this is equivalent to the variational problem

$$\hat{u} = \arg\min_{u\in\mathcal{U}} \frac{1}{2}\|Ku-f\|^2 + \frac{\alpha}{2}\|u\|^2. \tag{2.1}$$

Some of the early work in the Soviet community was already formulated in a much more general variational way, replacing the least-squares term with some discrepancy measure and the regularization with an appropriate functional, in some sense a precursor of the modern theory. At this time

the study was restricted to a rather abstract way of focusing on convergence proofs: strong motivations for other functionals in inverse problems and further methods for quantitative estimates were not available. The concepts and methods were further developed in the Soviet literature, including the question of the regularization parameter choice in dependence on the noise level. Instead of giving a detailed overview we refer to the influential book by Tikhonov and Arsenin (1977), which also made the results more broadly accessible.

As an alternative approach much work also considered what Tikhonov called the *regularization method* (and what seems to be the first appearance of this term in the literature), namely the approximation of K by regular operators and of its generalized inverse by bounded operators. There was parallel development in the West: Phillips (1962) developed an approach similar to Tikhonov's conditional well-posedness for integral equations of the first kind, and consequently the term *Tikhonov–Phillips regularization* is also used in the literature. In a discrete setting of statistical regression, a similar idea for dealing with ill-conditioned problems was developed under the term *ridge regression* (Hoerl 1959, Hoerl and Kennard 1970). A related approach for solving ill-posed problems for partial differential equations was the quasi-reversibility method (Lattès and Lions 1967), although barely analysed in the setting of a regularization method.

A different route to the construction of regularization methods was taken by Backus and Gilbert (1968) from a very applied perspective. Using linear filters, the noisy data were smoothed to be in the range of the forward operator K, followed by direct inversion (or a generalized inverse). It took quite a while for such methods to be understood in a unified way with other regularizations such as Tikhonov regularization (Engl *et al.* 1996). The key step was to relate the smoothing action of the filters to the operator K and its adjoint. This was made clear later in the linear functional strategy by Anderssen (1986) and also in the development of the approximate inverse method by Louis (1996), which turned out to be highly useful in tomography problems, where explicit reconstruction formulas and fast methods for the computation of the inverse are available.

In the 1970s and 1980s the study of linear regularization methods progressed further, dealing with many different regularization techniques such as iterative regularization by early stopping of stable iteration methods, truncated singular value decompositions, and regularization by discretization and projection (*e.g.* Nashed and Wahba 1974*b*, Nashed and Wahba 1974*c*, Wahba 1977, Eldén 1977, Bakushinskii 1977, Bakushinskii 1979, Bates and Wahba 1983, Engl 1987*b*, Hansen 1987). Most work was based on using spectral methods for the construction and detailed analysis of regularization methods. This includes the basic analysis of linear regularization methods in Hilbert spaces, their convergence as the noise level and the

regularization parameter tend to zero, as well as the first error estimates giving dependence on the noise level (*e.g.* Nashed and Wahba 1974*a*, Groetsch and King 1979, Natterer 1984, Neubauer 1988*a*). Moreover, various asymptotic parameter choice rules were suggested and investigated, either founded by theory such as the discrepancy principle or other *a posteriori* rules using the noise level (Morozov 1966, Bakushinskii 1973, Raus 1984, Engl and Neubauer 1985, Engl 1987*a*, Engl and Neubauer 1987, Gfrerer 1987, Engl and Gfrerer 1988, Raus 1992), or heuristic ones such as quasi-optimality or the L-curve method (Tikhonov and Arsenin 1977, Bakushinskii 1984, Thompson, Brown, Kay and Titterington 1991, Hansen 1992). The development of linear regularization methods in the early 1990s was rather complete, culminating in the seminal book by Engl, Hanke and Neubauer (1996), which provides a unified overview.

From the application point of view, a strong focus was placed on models with integral equations of the first kind, and image reconstruction in tomography drove applications (Natterer 2001, Natterer and Wübbeling 2001 and references therein). In parallel, various applications of inverse problems in partial differential equations, such as inverse scattering or parameter identifications, became relevant and were tackled by regularization methods (*e.g.* Payne 1975, Kravaris and Seinfeld 1985, Colton and Monk 1988, Banks and Kunisch 1989, Colton *et al.* 1990). This drove the interest in regularization theory from linear towards nonlinear problems.

The end of the 1980s marks the beginning of the systematic analysis of regularization methods for nonlinear inverse problems (replacing K with a nonlinear operator). In particular, the papers by Seidman and Vogel (1989) gave a well-posedness and convergence analysis of Tikhonov regularization for such problems, and Engl, Kunisch and Neubauer (1989) provided the first error estimates and convergence rates. Many techniques had to be developed to avoid spectral theory arguments that are not available for nonlinear operators; it is not surprising that many of those ideas were also influential for nonlinear regularization (of linear inverse problems). In the 1990s there was a surge in studies for nonlinear inverse problems. In particular, a theory of iterative regularization methods was constructed, which is particularly attractive since the nonlinear problems had to be solved with iterative methods in any case. Prominent examples are Landweber and steepest-descent methods (*e.g.* Hanke, Neubauer and Scherzer 1995), regularized Newton methods (*e.g.* Kaltenbacher 1997) and iterated Tikhonov methods (*e.g.* Scherzer 1993). We refer to Kaltenbacher, Schöpfer and Schuster (2009) for a comprehensive overview.

In parallel, another paradigm evolved, particularly in the image processing community, from the seminal papers of Rudin *et al.* (1992) and Mumford and Shah (1989), who proposed nonlinear variational models to solve denoising (and in the second case also segmentation) problems. From

a regularization point of view this means that a nonlinear regularization method is used to solve a linear inverse problem, a rather unusual idea at this time. From a technical point of view it poses the additional challenge of analysing schemes in anisotropic Banach spaces, such as the space of functions of bounded variation, while theory was previously formulated mainly in Hilbert spaces. In the case of variational regularization methods, basic well-posedness and convergence analysis can be carried out using techniques from variational calculus (Acar and Vogel 1994, Eggermont 1993), while quantitative estimates need completely novel approaches. Early progress in this direction was made for maximum entropy regularization (Eggermont 1993, Engl and Landl 1993); in this case the regularization technique could be related directly to regularization of nonlinear inverse problems in Hilbert spaces via a change of variables (Engl and Landl 1993). However, in a more general set-up the convergence rate theory remained quite open until the dawn of the twenty-first century, when strong progress was made by employing techniques from convex analysis to variational regularization methods. We mention at this point that some of these more geometric ideas were also hidden in earlier work on regularization in Hilbert spaces with convex constraints (Neubauer 1988*b*, Eicke 1992). The improved understanding of variational regularization methods in Banach spaces subsequently led to a variety of other techniques and variants, such as derived iterative regularization, which we will discuss in further detail in the course of this paper.

Another driving force for investigating regularization methods in Banach spaces was sparsity, including wavelet shrinkage and the variational counterpart of regularization ℓ^1-type norms, for example in Besov spaces (Donoho 1992, Donoho and Johnstone 1995). This led in parallel to the field of compressed sensing, where the focus was on designing the appropriate measurement set-ups for optimal compression rather than improving reconstructions on a given inverse problem (Donoho 2006, Candès, Romberg and Tao 2006, Candès and Donoho 2002, Candès and Tao 2004*a*, Candès and Tao 2004*b*, Candès and Romberg 2007, Donoho *et al.* 2006). Despite the fact that the usual setting in compressed sensing is a finite-dimensional one, many arguments based on convex analysis are closely related.

In recent years these techniques also evolved into many practical applications, in particular in the image reconstruction community. The whole list of applications where the methods made an impact in different ways might warrant its own survey paper. In order to illustrate the change in the first decade of the twentieth century we provide Table 2.1, which shows the typical state of the art for inverse problems in medical imaging up to or around the year 2000 and that typically used ten years later.

Note that (with the exception of the statistically motivated EM algorithm) all state-of-the-art methods up to 2000 were linear regularization methods. This has now completely changed, with the exception of fully sampled CT,

Table 2.1. State of the art for inverse problems in medical imaging.

Method	Up to 2000	Since 2010
Full CT	Filtered backprojection	Filtered backprojection
Undersampled CT	Filtered backprojection	TV-type / wavelet sparsity
PET / SPECT	Filtered backprojection / EM	EM-TV / dynamic sparsity
Photacoustics	–	TV-type / wavelet sparsity
EEG / MEG	LORETA	Spatial sparsity / Bayesian
ECG-BSPM	L2 Tikhonov	L1 of normal derivative
Microscopy	None, linear filter	TV-type / shearlet sparsity
PET-CT / MR	–	TV-type anatomical priors

where there is neither a null space nor significant noise, and hence regularization plays a minor role. The details of most other methods, mainly based on variational models, will become clear in the next section.

3. Variational modelling

The variational approach to regularization methods has become very popular in the last few decades, since it provides an intuitive approach to modelling and a framework for its basic analysis, and also allows a variety of computational methods to be applied, particularly in the case of convex regularization functionals. The key idea in constructing a variational regularization method for (1.1) consists of finding two functionals: a data fidelity term F measuring the distance between Ku and f (or its noisy version f^δ) and a regularization functional J favouring appropriate minimizers or penalizing potential solutions with undesired structures. Instead of simply fitting u to data, *i.e.* minimizing $F(Ku, f)$, a weighted version is minimized to obtain

$$\hat{u} \in \arg\min_u (F(Ku, f^\delta) + \alpha J(u)), \tag{3.1}$$

where $\alpha > 0$ is the regularization parameter controlling the influence of the two terms on the minimizer. Since the problem should approach the minimization of only the data fidelity term in the noiseless case, it is natural to think of α as a small parameter.

The choice of the data fidelity term is often straightforward, for example as some kind of least-squares term (squared norm distance in a Hilbert space), or motivated from statistical arguments by some likelihood functional for the noise. In the latter case the variational model can be interpreted as a regularized likelihood model, where the data term usually corresponds to the negative log likelihood of the noise model. A prominent example is the case of additive Gaussian noise, which leads to a least-squares

data term $\frac{1}{2}\|Ku - f\|^2$, where the specific Hilbert space norm to be used is determined by the covariance operator of the noise. Appropriate choices for the latter can have a significant impact: for example, choosing likelihoods for Poisson noise appearing in photon count data leads to strong improvement over least-squares terms, particularly in large noise regimes (*e.g.* Sawatzky *et al.* 2013, Brune, Sawatzky and Burger 2009*c*). Throughout this paper we will assume that F is Fréchet-differentiable on $\mathcal{V}$ unless otherwise stated.

The choice of a regularization functional seems less natural at first glance. Based on the original ideas by Tikhonov, the key ingredient for a successful regularization is its topological properties, so the regularization functional is frequently chosen as some power of a norm (or seminorm) in a Banach space. Classical examples are Tikhonov–Phillips in Hilbert spaces such as $L^2(\Omega)$, $H^1(\Omega)$, or in some sequence space $\ell^2(\mathbb{N})$. As a generalization in function spaces, regularization functionals depending on the gradient (or higher-order derivatives) of u became popular. These correspond to the rather direct intuition that smooth solutions are preferable due to prior knowledge. Non-smooth and oscillatory functions will lead to large or even infinite values of the derivatives, and thus very high values of the regularization functionals. Hence, they are not suitable candidates as a minimizer of (3.1).

In many cases in inverse problems such as image reconstruction, one is instead interested in non-smooth solutions, and in particular their discontinuity sets. One example is that of piecewise constant functions with reasonable edge sets, which are not contained in any Sobolev space $W^{k,p}(\Omega)$ for $k, p \geq 1$, since their gradient is already a concentrated measure (Ambrosio, Fusco and Pallara 2000, Evans and Gariepy 1992). This motivates use of the space of functions of bounded variations $BV(\Omega)$, which consists of all functions in $L^1(\Omega)$ whose distributional gradients are vectorial Radon measures. The regularization with the total variation, that is,

$$TV(u) = |u|_{BV} = \int_\Omega \mathrm{d}|Du|, \tag{3.2}$$

where Du is the gradient measure of u, proposed for denoising by Rudin *et al.* (1992), and the subsequent popularity of investigating such methods can be seen as the advent of modern regularization methods.

The details of reconstructions to be achieved, however, strongly depend on the specific norm used. It is common folklore that the regularization functional is chosen such that desired solutions matching prior knowledge have a small value of J and are thus preferred as the appropriate solutions. This is true only to some extent, but the overall effect of a regularization functional is determined by the effect it has on possible minimizers rather than purely a comparison of functional values. Consider as a simple example one-dimensional total variation regularization. It will of course prefer solutions with small total variation over oscillatory functions with high variation. On

the other hand, it still selects among functions with the same total variation. Structural results on the solution of total variation regularization problems show that canonical solutions for noisy data are piecewise constant, even if the exact solution is not (Ring 2000, Chambolle *et al.* 2010, Jalalzai 2016). This means that total variation actively selects piecewise constant solutions over smooth solutions that have the same total variation, that is, they are *a priori* indistinguishable by the regularization functional. The reason for this behaviour can be seen by inspecting the optimality condition, given by (assuming F to be Fréchet-differentiable)

$$K^* \partial_x F(Ku, f) + \alpha p = 0, \quad p \in \partial J(u). \tag{3.3}$$

Here ∂_x denotes the (partial) Fréchet derivative in the first argument, and $\partial J(u)$ is the subdifferential of J at position u: see Rockafellar (1972, Section 23) or Ekeland and Temam (1999). Solving for the subgradient p, we always obtain a relation of the form $p = K^* \tilde{w}$ for some $\tilde{w} \in \mathcal{V}$, that is, the variational method will select smooth subgradients due to the smoothing properties of the operator K and its adjoint. We will detail the relation between the properties of the subgradients of the solution for total variation and other examples of regularization in the next sections.

In a stochastic set-up, the variational approach is often formulated from Bayesian estimation (Kaipio and Somersalo 2006, Stuart 2010), in particular *maximum a posteriori probability* (MAP) estimators. For the sake of simpler presentation, assume that we are in a finite-dimensional setting for the inverse problem $Ku = f$ and can write down probability densities for the prior $\pi_0(u)$ and the likelihood $\pi(f|u)$ of measuring the data f given the true solution u. Then Bayes' theorem provides the posterior probability density via

$$\pi(u|f) = \frac{1}{\pi_*(f)} \pi(f|u) \pi_0(u), \tag{3.4}$$

where

$$\pi_*(f) = \int \pi(f|u) \pi_0(u) \, \mathrm{d}u \tag{3.5}$$

is the effective prior probability on the data. A MAP estimate $\hat{u}$ is defined as a maximizer of the posterior probability density, or a minimizer of its negative logarithm. Since the part $\pi_*(f)$ independent of u is irrelevant for the minimizer, we thus have

$$\hat{u} \in \arg\min_u (-\log \pi(f|u) - \log \pi_0(u)). \tag{3.6}$$

This formulation is closely related to the variational modelling point of view when interpreting $-\log \pi(f|u)$ as a data fidelity term and $-\log \pi_0(u)$ as the regularization term. Indeed, for many standard stochastic (noise) models

one obtains

$$\pi(f|u) \sim \exp(-F(Ku, f)). \tag{3.7}$$

Examples are additive Gaussian noise leading to a least-squares fidelity and Poisson noise leading to the Kullback–Leibler divergence. Assuming further that the prior is related to some regularization functional J,

$$\pi_0(u) \sim \Phi(-J(u)), \tag{3.8}$$

for some monotone function Φ, we see that the MAP estimation problem becomes

$$\hat{u} \in \arg\min_u F(Ku, f) - \log(\Phi(-J(u))). \tag{3.9}$$

This problem can be reformulated in a more conventional form, even if the prior Φ is not exactly specified. By a standard argument we see that there exists $\gamma > 0$ such that

$$\hat{u} \in \arg\min_{u, J(u) \le \gamma} F(Ku, f),$$

and with the existence of a Lagrange parameter $\alpha > 0$ for the constraint $J(u) \le \gamma$ (which is easily verified for a scalar constraint) we obtain

$$\hat{u} \in \arg\min_u F(Ku, f) + \alpha J(u). \tag{3.10}$$

We mention that similar reasoning in infinite dimensions is not as straightforward: even the definition of the MAP estimate is a non-obvious task (Dashti, Law, Stuart and Voss 2013, Helin and Burger 2015). Recent results, however, provide good characterization in many relevant cases (Helin and Burger 2015, Lie and Sullivan 2017, Agapiou, Burger, Dashti and Helin 2018). A relation between Bayesian estimators and the variational approach also exists beyond the MAP estimate by the *Bayes cost* method. Given a cost ψ measuring a distance on the input space, the Bayes cost approach looks for a minimizer of the posterior expectation of ψ, that is,

$$\hat{u} \in \arg\min_u \int \psi(u, v)\ \pi(v|f)\, \mathrm{d}v, \tag{3.11}$$

a functional that depends in a more implicit way on the data and the forward model.

3.1. Total variation and related regularizations

As mentioned above, total variation regularization has been one of the driving examples for development of regularization methods in Banach spaces starting from Rudin *et al.* (1992) and Acar and Vogel (1994). Since then it has been a constant source of motivation for further development of mathematical analysis (*e.g.* Chambolle and Lions 1997, Strong *et al.* 1996,

Scherzer 1998, Chavent and Kunisch 1997, Ring 2000, Strong and Chan 2003, Burger and Osher 2004, Caselles, Chambolle and Novaga 2007, Allard 2007), computational optimization techniques for non-smooth problems (*e.g.* Chan, Golub and Mulet 1999, Vogel 2002, Chambolle 2004, Kunisch and Hintermüller 2004, Chambolle and Pock 2011), and development of advanced models (*e.g.* Scherzer 1998, Osher *et al.* 2005, Burger, Osher, Xu and Gilboa 2005, Burger *et al.* 2006, Bredies and Holler 2015*a*, Hu and Jacob 2012, Lenzen, Becker and Lellmann 2013, Benning, Brune, Burger and Müller 2013).

The key step for modern analysis and computational methods is the (pre-) dual formulation of total variation,

$$TV(u) = |u|_{BV} := \sup_{g \in C_0^\infty(\Omega)^d, g \in \mathcal{C}} \int_\Omega u \nabla \cdot g \, \mathrm{d}x, \tag{3.12}$$

with the convex set

$$\mathcal{C} = \{g \in L^\infty(\Omega) \mid |g(x)| \leq 1 \text{ a.e. in } \Omega\}.$$

This characterization allows us to understand the structure of subgradients as elements of $\mathcal{C}$ absolutely continuous with respect to the gradient measure D such that

$$\int_\Omega g \cdot \, \mathrm{d}Du = |u|_{BV}.$$

The optimality condition (3.3) provides

$$K^* \partial_x F(Ku, f) + \alpha \nabla \cdot g = 0, \tag{3.13}$$

where g is a vector field such that $g|Du|$ is a polar decomposition of the vector measure (Ambrosio *et al.* 2000).

In one spatial dimension the structure of solutions can be understood directly from the optimality condition. If there is an open set where u is not constant, either with positive or negative derivative, then g equals $+1$ or -1, hence its derivative vanishes. Thus, in such regions the generalized residual $K^* \partial_x F(Ku, f)$ vanishes. In the case of noisy data this is usually not happening for larger sets, thus u is typically piecewise constant. In higher spatial dimensions this is not completely true, but still the case $|g(x)| < 1$ is the canonical one, so in many cases solutions are piecewise constant. On the other hand, piecewise constant structures are not optimal in all instances. In particular, total variation methods are well known to exhibit staircasing phenomena, that is, smoothly varying parts in the solution are often approximated by piecewise constant structures with many jumps resembling a stair structure. For this reason many modifications and variants of total variation regularization have been investigated in recent decades. An immediate option is that of higher-order total variation approaches, which

formally replace the one-norm of the gradient with the one-norm of a higher-order derivative such as the Laplacian, the Hessian or the symmetric part of the Hessian (*e.g.* Scherzer 1998, Chan, Esedoglu and Park 2010, Hinterberger and Scherzer 2006, Papafitsoros and Schönlieb 2014). The disadvantage of such an approach is that solutions of the regularization model will be too regular and discontinuity sets (edges) are lost. In view of (3.12) such approaches can be characterized by $\mathcal{C}$ not being a bounded set in $L^\infty(\Omega)$, but rather being derivatives of bounded measurable functions.

An alternative model trying to take advantage of total variation and higher-order total variation is a decomposition into two or more parts, that is, $u = u_1 + u_2$ with u_1 and u_2 being regularized differently. This was proposed for the first time by Chambolle and Lions (1997) as an infimal convolution of first- and second-order total variations; the effective regularization functional is given by

$$J(u) = \inf_{u_1+u_2=u} (|u_1|_{BV} + |\nabla u_2|_{BV}).$$

The TGV-type models, proposed by Bredies, Kunisch and Pock (2010), became a popular alternative. Instead of decomposing u, these instead decompose the gradient measure Du into Du_1 and some vector field u_2. One version of the regularization functional is then given by

$$J(u) = \inf_{Du_1+u_2=Du} (|u_1|_{BV} + |u_2|_{BV}).$$

The fact that the higher-order part is an arbitrary vector field provides additional freedom that can be beneficial compared to the infimal convolution model (Bredies and Holler 2015*a*, Benning, Brune, Burger and Müller 2013, Grah 2017). We also mention that the original TGV model in Bredies *et al.* (2010) does not use a bounded variation model for u_2, but only bounded deformations, that is, the symmetric part of the gradient. Moreover, the approach can be formulated for arbitrary order of regularization. In the dual formulation (3.12) approaches such as infimal convolution or TGV still lead to $\mathcal{C}$ being a subset of the unit ball in $L^\infty(\Omega)$, which implies

$$J(u) \leq |u|_{BV} \quad \text{for all } u \in BV(\Omega).$$

On the other hand, for many of them a lower bound inequality can be shown at least when excluding a low- (finite-) dimensional null space (Benning *et al.* 2013), that is, there exists a positive constant c and some linear functionals ℓ_i such that

$$J(u) \geq c|u|_{BV} \quad \text{for all } u \in BV(\Omega), \quad \text{such that } \ell_i(u) = 0, i = 1, \dots M.$$

Hence, J is an equivalent norm on the subspace of BV excluding the null space. For the combination of first- and second-order derivatives the null space naturally consists of piecewise affine functions (thus $M = d +$

1). For further discussion and advanced aspects we refer to Bredies *et al.* (2010), Benning *et al.* (2013), Ranftl, Pock and Bischof (2013), Bredies and Holler (2014, 2015*a*, 2015*b*), Burger, Papafitsoros, Papoutsellis and Schönlieb (2015*b*, 2016*c*), Bergounioux (2016), Setzer, Steidl and Teuber (2011), Holler and Kunisch (2014), Gao and Bredies (2017) and Bergounioux and Papoutsellis (2018).

In certain cases it is also interesting to use total variation regularization on some transform of the image. Motivated by research in image analysis taking into account orientations via local Radon transforms (Krause, Alles, Burgeth and Weickert 2016), Burger, Müller, Papoutsellis and Schönlieb (2014) investigated total variation regularization on the Radon transform, combined with total variation on the image itself, to promote piecewise constant images with very thin structures resembling lines. Grah (2017) investigated total variation on the spherical Radon transform (equivalent to circular Hough transform in computer vision) in order to reconstruct small circular structures.

Another variant is total variation regularization for vector fields, for example arising for colour images (Bresson and Chan 2008, Blomgren and Chan 1998), flow fields (Hinterberger, Scherzer, Schnörr and Weickert 2002, Zach, Pock and Bischof 2007) or joint reconstruction problems (Knoll *et al.* 2017). While many aspects remain the same as in the scalar case, it is particularly interesting to consider which matrix norm should be used for Du, and which dual norm for g (observing that this becomes a matrix in (3.12)).

3.2. Sparsity regularization

Total variation regularization, in particular its discrete version, can be interpreted as a functional favouring sparsity, in this case of the gradient. The paradigm of sparsity has developed in parallel to the total variation regularization (Donoho 1992). A key insight driving sparsity priors was the (approximate) sparsity of signals and natural images in wavelet bases (Mallat and Zhang 1993, Huang and Mumford 1999, Mallat 2008, Starck, Murtagh and Fadili 2010). Further improvements were made by replacing the orthonormal bases with frames (Christensen 2003) such as curvelets (Candès and Donoho 2000*a*, 2000*b*) or shearlets (Labate, Lim, Kutyniok and Weiss 2005, Guo and Labate 2007, Kutyniok and Labate 2012).

Sparsity is naturally measured by the ℓ^0-norm,[1] the number of non-zero entries. Since the minimization of the ℓ^0-norm is highly non-convex and even NP-complete, it is usually relaxed to the convex ℓ^1-norm. In the *analysis formulation* a frame system ϕ_i is used to test sparsity of $\langle u, \phi_i \rangle$, and the

[1] The ℓ^0-norm is not a norm in the traditional sense, but is often referred to as such because it arises as the limit of the ℓ^p-norm for $p \to 0$.

corresponding regularization functional is given by

$$J(u) = \sum_i |\langle u, \phi_i \rangle|. \tag{3.14}$$

If (ϕ_i) is an orthonormal system, this is equivalent to the *synthesis formulation*, which is based on writing

$$J(u) = \sum_i |c_i| \quad \text{where } u = \sum_i c_i \phi_i. \tag{3.15}$$

Note that in general the two formulations may differ for frames (Elad, Milanfar and Rubinstein 2007).

In the synthesis formulation we can effectively define the variational problem on the coefficient vector c, that is,

$$\tilde{K} : \ell^2(\mathbb{N}) \to \mathcal{V}, \quad c \mapsto \sum_i c_i K \phi_i,$$

and compute

$$\hat{u} = \sum_i \hat{c}_i \phi_i, \quad c_i \in \arg\min_c F(\tilde{K}c, f) + \alpha |c|_1.$$

The corresponding optimality condition is given by

$$(\tilde{K}^* \partial_x F(\tilde{K}c, f))_i + \alpha s_i = 0,$$

where s_i is a multivalued sign of c_i, that is, an element of $[-1, 1]$ for $c_i = 0$. If $\tilde{K}$ is a bounded linear operator on $\ell^2(\mathbb{N})$, then its adjoint maps into the same space, and hence $(s_i) \in \ell^2(\mathbb{N})$. This implies in particular that $|s_i| < 1$, hence $c_i = 0$, for i sufficiently large. Thus, we always obtain some sparsity with this model.

In the analysis formulation the optimality condition is instead given by

$$K^* \partial_x F(Ku, f) + \alpha s_i \phi_i = 0,$$

where s_i is a multivalued sign for $\langle u, \phi_i \rangle$. Here the understanding of the sparsity property is more complicated: the s_i are actually related to the residual via the linear system

$$\sum_j \langle \phi_i, \phi_j \rangle s_j = -\frac{1}{\alpha} \langle K\phi_i, \partial F(Ku, f) \rangle.$$

We refer to Vaiter *et al.* (2013*a*, 2013*b*) for a detailed analysis in this case. Sparsity models for inverse problems have been studied with different frames and applications extensively in the past decade (*e.g.* Cotter, Rao, Engan and Kreutz-Delgado 2005, Chaux, Combettes, Pesquet and Wajs 2007, Colonna, Easley, Guo and Labate 2010, Recht, Fazel and Parrilo 2010).

There are several relevant extensions of sparsity priors to multidimensional systems, in particular via a synthesis-type formulation:

$$u = \sum_{i,j} c_{ij} \phi_i \otimes \psi_j.$$

The different dimensions are often space (characterized by basis functions ϕ_i) and time or frequency (characterized by basis functions ψ_j). Instead of overall sparsity, more detailed prior knowledge can be introduced. The most popular example is joint or collaborative sparsity, which means that only a few of the basis functions, for example in the second dimension, can be used to explain the solution. This means that $c_{\cdot j}$ vanishes for most j, as does the norm of $c_{\cdot j}$. A common regularization for this case is the joint or collaborative sparsity prior

$$J(u) = \sum_j \|c_{\cdot j}\|_{\ell^r},$$

usually with $r = 2$ or $r = \infty$ (Duarte *et al.* 2005, Teschke and Ramlau 2007, Fornasier and Rauhut 2008, Gholami and Siahkoohi 2010, Lee, Kim, Bresler and Ye 2011). An alternative type of prior knowledge is local sparsity, which means that for each i only a few basis functions ψ_j are used. The term *local* is due to an imaging interpretation of the ϕ_i as basis functions local in space (*e.g.* for each pixel). This is a common issue in dynamic or spectral imaging, where one can assume that only a few materials and their characteristic evolutions or spectral curves can be found in each pixel. A regularization functional proposed for this issue (Heins, Moeller and Burger 2015) is

$$J(u) = \max_i \|c_{i\cdot}\|_{\ell^1} + \beta \sum_i \|c_{i\cdot}\|_{\ell^1}.$$

An infinite-dimensional extension of the above sparsity models is sparsity in a space of Radon measures, that is, the regularization functional is given as the total variation norm of the measure u:

$$J(u) = \int_\Omega \mathrm{d}|u| = \sup_{g \in C_0(\Omega), \|g\|_\infty \leq 1} \int_\Omega g \,\mathrm{d}u.$$

This yields a convex regularization functional for reconstructing multiple peaks at unknown locations, and was proposed for inverse problems in Bredies and Pikkarainen (2013) and for super-resolution problems in Candès and Fernandez-Granda (2013, 2014) and Aja-Fernandez, Alberola-Lopez and Westin (2008). The reconstruction properties in deconvolution problems were analysed in Duval and Peyré (2017*a*, 2017*b*) and Denoyelle, Duval and Peyré (2017), and asymptotics from finite-dimensional problems with sparsity priors are found in Heins (2014) and Duval and Peyré (2017*a*, 2017*b*).

3.3. Low-rank regularization

In many applications one seeks a decomposition of the form

$$U = \sum_i \Phi_i \otimes \Psi_i \tag{3.16}$$

with unknown Φ_i, Ψ_i and the additional prior knowledge that there are as few elements as possible in the sum. In a finite-dimensional setting this means that the matrix U has low rank, that is, the rank of U would be the obvious regularization functional. However, since the rank is very far from being convex, several relaxations have been proposed instead. The most popular one, originally proposed for matrix completion problems, is the nuclear norm (Candès and Recht 2009, Recht, Fazel and Parrilo 2010, Cai, Candès and Shen 2010, Cai and Osher 2013, Yang, Ma and Osher 2013)

$$\|U\|_* = \sum \sigma_i, \tag{3.17}$$

where σ_i are the singular values of U.

In many applications the low-rank part alone does not suffice to model the structure of solutions, frequently a low-rank plus sparsity (L+S) model is employed instead (Candès, Li, Ma and Wright 2011, Otazo, Candès and Sodickson 2015), which is again based on the decomposition

$$J(u) = \inf_{u_1+u_2=u} (\|u_1\|_* + \|Tu_2\|_1), \tag{3.18}$$

with a sparsifying transform T (often some derivative as in total variation). Particularly in videos the low-rank part captures background and certain slow dynamics, while the sparse part captures the key changes.

For inverse problems an infinite-dimensional function space setting would be more appropriate, which has not yet been investigated. In particular, a formulation in a space of trace class operators between Hilbert spaces H_1 and H_2 (Reed and Simon 1978) would be natural. Let us mention that the choice of Hilbert spaces H_i opens novel opportunities for improved regularization that have not been exploited as yet, even in the finite-dimensional case.

3.4. Infimal convolutions

As we have seen above, infimal convolution is a versatile tool to combine different regularization approaches, and to define a novel functional that combines their advantages. We want to emphasize this approach in the following by providing formal definitions.

Definition 3.1. Let $J_i : \mathcal{U} \to \mathbb{R} \cup \{+\infty\}$, $i = 1, 2$ be proper convex functionals. Then their *infimal convolution* $J_1 \square J_2 : \mathcal{U} \to \mathbb{R} \cup \{+\infty\}$ is defined by

$$(J_1 \square J_2)(u) = \inf_{v \in \mathcal{U}} (J_1(u - v) + J_2(v)). \tag{3.19}$$

More generally, we can define an infimal convolution for an arbitrary number of convex functionals.

Definition 3.2. Let $J_i : \mathcal{U} \to \mathbb{R} \cup \{+\infty\}$, $i = 1, \ldots, M$ be proper convex functionals. Then their infimal convolution $J : \mathcal{U} \to \mathbb{R} \cup \{+\infty\}$ is defined by

$$J(u) = \inf_{u_i \in \mathcal{U}, \sum u_i = u} \sum_{i=1}^{M} J_i(u_i) \tag{3.20}$$

We mention that *a priori* it is unclear whether the infima above are actually minima. If a minimizer v exists for the infimal convolution of J_1 and J_2, it can be used to deduce optimality conditions, since

$$p \in \partial J(u) \quad \text{if } p \in \partial J_1(u - v) \cap \partial J_2(v).$$

As the above examples for sparsity and in particular higher-order total variation show, there is great freedom in designing infimal convolution models for regularization. Consequently, a lot of options for future research remain open, and interesting results are still to be expected.

3.5. Regularization by denoising

Romano, Elad and Milanfar (2017) recently proposed an approach named *regularization by denoising*, where J is of the form

$$J(u) = \langle u, u - D(u) \rangle.$$

Here D is a potentially nonlinear operator that maps an image u to the output of a black-box image denoising routine. The idea is to make state-of-the-art denoising routines available for the regularization of linear inverse problems.

3.6. Bregman distances

From a single regularization functional several variants can be constructed by using a non-trivial prior u_0 and the so-called Bregman distance (originally introduced in Bregman (1967) for proximal-point-type methods). Instead of shifting the functional directly from $J(u)$ to $J(u - u_0)$, the approach in the Bregman distance performs a shift in the convex conjugate. In the original formulation this amounts to the following.

Definition 3.3. Let $J : \mathcal{U} \to \mathbb{R} \cup \{+\infty\}$ be a convex functional and let $p_0 \in \partial J(u_0)$. Then the Bregman distance between $u \in \mathcal{U}$ and $u_0 \in \mathcal{U}$ with subgradient p_0 is given by

$$D_J^{p_0}(u, u_0) := J(u) - J(u_0) - \langle p_0, u - u_0 \rangle \tag{3.21}$$

Note that the Bregman distance is not a strict distance, that is, it can vanish for $u \neq u_0$ if J is not strictly convex. It is also not symmetric, but

can be made symmetric by taking a sum of one-sided distances (see Burger 2016 for a more detailed discussion). For absolutely one-homogeneous regularization functionals as above, the identity $J(u_0) = \langle p_0, u_0 \rangle$ holds, so the Bregman distance becomes

$$D_J^{p_0}(u, u_0) := J(u) - \langle p_0, u \rangle, \tag{3.22}$$

and thus it is effectively independent of u_0: only the subgradient p_0 matters. This is particularly relevant if the subdifferential of J is not a singleton, or *vice versa* a subgradient $p_0 \in \partial J(u_0)$ can be an element of the subdifferential also at other values of u.

Note that in the typical case of $u_0 = 0$ being a minimizer of J, that is, $0 \in \partial J(0)$, the regularization with J can be reinterpreted as penalizing the Bregman distance to $u_0 = 0$. Bleyer and Leitao (2009) carried out a basic analysis of such a variational regularization. The topic has received recent interest, particularly in the context of TV-type regularization in imaging, since it allows us to introduce structural information. The key insight in total variation is that the subgradient encodes information about the discontinuity set, more precisely $p = \nabla \cdot g$, with g being equal to the unit normal vector to the discontinuity set where it is regular. This is again related to (3.22); the total variation does not depend directly on u_0 and in particular the contrast in the image. Instead, it vanishes for all u of the form

$$u(x) = f(u_0(x))$$

with a monotonically increasing function f, that is, a simple contrast change (Resmerita and Scherzer 2006). Assuming that g is a vector field realizing the supremum in the dual definition of the total variation, the Bregman distance becomes

$$D_{TV}^{p_0}(u, u_0) = |u|_{BV} - \int_\Omega (\nabla \cdot g_0) u \, \mathrm{d}x = \int_\Omega (\nabla \cdot (g - g_0)) u \, \mathrm{d}x,$$

and if u is piecewise constant with regular discontinuity set S_u,

$$D_{TV}^{p_0}(u, u_0) = \int_{S_u} [u](g - g_0) \cdot \nu \, \mathrm{d}\sigma = \int_{S_u} [u](1 - g_0 \cdot \nu) \, \mathrm{d}\sigma,$$

where $[u]$ denotes the jump along S_u and ν is the unit normal (oriented such that $[u]$ is positive). One thus observes that the Bregman distance measures differences in the discontinuity set and its orientation, which is perfect for imaging applications with a structural prior (Kaipio, Kolehmainen, Vauhkonen and Somersalo 1999) that mainly yields information about edges, *i.e.* discontinuity sets. One suitable example is that of anatomical priors in medical imaging, where a high-resolution method such as CT or MR is used to obtain information about organ boundaries and other

anatomical features, which are the natural candidates for edge sets in functional methods such as PET, SPECT or MR imaging with special contrast. In some cases a joint reconstruction is also of interest, the most obvious case being colour or hyperspectral images, where naturally intensity changes at the same locations, usually even in the same direction (Moeller, Wittman, Bertozzi and Burger 2012, Moeller, Brinkmann, Burger and Seybold 2014).

In some applications one may find contrast inversion, that is, the jump of the two images along the discontinuity set has a different sign. In such cases the normals are parallel, which means they point in opposite directions and hence lead to large values in the Bregman distance. A potential solution for avoiding such issues is the infimal convolution of Bregman distances, in this case with the two normal fields and thus subgradients of opposite sign (Moeller *et al.* 2014, Rasch *et al.* 2017):

$$J = D_{TV}^{p_0}(\cdot, u_0) \square D_{TV}^{-p_0}(\cdot, -u_0).$$

We also mention some other related approaches to modifying total variation functionals, such as parallel level set models (Ehrhardt and Arridge 2014, Ehrhardt *et al.* 2014, Ehrhardt *et al.* 2016), which can be related to the Bregman distance for total variation (Rasch, Brinkmann and Burger 2018), or directional / structural total variation (Bungert *et al.* 2018, Ehrhardt and Betcke 2016, Hintermüller, Holler and Papafitsoros 2017, Grasmair and Lenzen 2010), formally

$$TV_{g_0}(u) = \int_\Omega |(I - g_0 \otimes g_0)\nabla u| \, \mathrm{d}x.$$

4. Fundamentals of nonlinear regularization

Before discussing the detailed analysis of nonlinear regularization methods, we first aim to provide a suitable basis for understanding regularization methods and their convergence. We start with the case of linear regularization methods in Hilbert spaces, recalling the abstract theory from Engl *et al.* (1996), and then try to work out a suitable analogue for the nonlinear case in Banach spaces.

4.1. Abstract linear regularization methods

We start our exposition with a discussion of possible limits of regularization schemes. In essentially all linear methods such as Tikhonov regularization, truncated SVD or iterative regularization in Hilbert spaces, it is clear which solutions are approximated as the regularization parameter tends to zero, namely those obtained from a generalized inverse. The following definitions are made to characterize these limiting solutions.

Definition 4.1. Let $K : \mathcal{U} \to \mathcal{V}$ be a bounded linear operator between Hilbert spaces and $f \in \mathcal{V}$. We call $\hat{u} \in \mathcal{U}$ a *best approximate solution* of (1.1) if

$$\|K\hat{u} - f\|_{\mathcal{V}} \leq \|Ku - f\|_{\mathcal{V}} \quad \text{for all } u \in \mathcal{U}\,. \tag{4.1}$$

Moreover, we call $\hat{u}$ a *minimal norm solution* if it is a best approximate solution and

$$\|\hat{u}\|_{\mathcal{U}} \leq \|u\|_{\mathcal{U}} \quad \text{for all } u \in \mathcal{U},\ \|K\hat{u} - f\|_{\mathcal{V}} = \|Ku - f\|_{\mathcal{V}}. \tag{4.2}$$

Note that due to the strict convexity of the square of a Hilbert space norm, the minimum solution – being its minimizer on a linear manifold – is a unique object. An abstract regularization method is now a collection of continuous operators approximating the (discontinuous) generalized inverse of K.

Definition 4.2. The bounded linear operators $R_\alpha : \mathcal{V} \to \mathcal{U}$ defined for α in $(0, \alpha_0)$ are called *linear regularization operators*. Together with a parameter choice strategy α depending on the noise level δ and the data f^δ, that is, a function

$$\alpha : (0, \delta_0) \times \mathcal{V} \to (0, \alpha_0), \tag{4.3}$$

it is called the *linear regularization method*.

A linear regularization method is called *convergent* if, for all $f \in \mathcal{R}(K)$, the condition

$$\lim_{\delta \to 0} \sup\{\|R_{\alpha(\delta, f^\delta)}(f^\delta) - u^\dagger\|_{\mathcal{U}} \mid f^\delta \in \mathcal{V}, \|f - f^\delta\|_{\mathcal{V}} \leq \delta\} = 0 \tag{4.4}$$

holds with $u^\dagger$ being the minimum norm solution of (1.1).

For ill-posed problems it is well known that convergence can be arbitrarily slow (Schock 1985). Thus convergence rates can be obtained only on a restricted subset M_ν with the parameter $\nu > 0$ measuring the smoothness or order of convergence. The standard definition is as follows.

Definition 4.3. A regularization method is *convergent at order* ν on a set $\mathcal{M}_\nu$ if, given any $f = Ku^\dagger$, for $u^\dagger \in \mathcal{M}_\nu$, there exists a constant C_ν such that

$$\|R_{\alpha(\delta, f^\delta)}(f^\delta) - u^\dagger\| \leq C_\nu \delta^\nu \tag{4.5}$$

for any data f^δ satisfying $\|f^\delta - f\| \leq \delta$.

It is well known that the set M_ν can be related to the source condition

$$u^\dagger = (K^*K)^\mu w$$

for some $w \in \mathcal{U}$ and appropriate $\mu > 0$ related to ν (Engl *et al.* 1996). The constant C_ν is then related to the norm of w. The simplest cases of source

conditions are $\mu = 1/2$, which can be reformulated as

$$u^\dagger = K^* \tilde{w},$$

for some $\tilde{w} \in \mathcal{V}$, and $\mu = 1$. Source conditions induce conditional well-posedness of the problem, *e.g.* for $\mu = 1/2$, one has for $u_i = K^* \tilde{w}_i$

$$\|u_1 - u_2\|^2 = \langle u_1 - u_2, K^*(\tilde{w}_1 - \tilde{w}_2)\rangle = \langle K(u_1 - u_2), \tilde{w}_1 - \tilde{w}_2\rangle.$$

The Cauchy–Schwarz and triangle inequality then imply the Hölder stability

$$\|u_1 - u_2\| \leq C\sqrt{\|Ku_1 - Ku_2\|},$$

with $C = \sqrt{\|\tilde{w}_1\| + \|\tilde{w}_2\|}$.

4.2. Extension to nonlinear methods

The examples of variational regularization models in the previous section call for a more general theory of nonlinear regularization methods. While the concept of a best-approximate solution is fairly straightforward to generalize, other aspects of convergence and limiting solutions are less obvious. In a general variational regularization, as in the examples discussed above, it would be natural to replace the minimum norm solution with a solution minimizing the regularization functional. The latter is not necessarily unique, however, hence some possible multivaluedness needs to be introduced in the characterization. Similar issues apply to the regularized problem and hence the definition of a regularization operator. In the following we will try to provide a fundamental setting for nonlinear regularization methods. As in the case of linear regularizations we first generalize the possible types of solutions we would like to approximate. The generalization of the first notion is fairly straightforward, so we only allow for more general distance measures, for example functionals related to negative log-likelihoods for non-Gaussian distributions.

Definition 4.4. Given an error measure $F : \mathcal{V} \times \mathcal{V} \to \mathbb{R}_+ \cup \{+\infty\}$, we call $\hat{u} \in \mathcal{U}$ a *best approximate solution* of (1.1) with respect to F if

$$F(K\hat{u}, f) \leq F(Ku, f) \quad \text{for all } u \in \mathcal{U}. \tag{4.6}$$

A suitable generalization of the definition of a minimum norm solution is more involved. In particular, we would like to give a unified concept including the selection by minimizing a regularization functional or maximizing some prior probability. We encode the selection of specific solutions due to prior knowledge in a (multivalued) selection operator.

Definition 4.5. A multivalued operator $\mathcal{S} : \mathcal{R}(K) \rightrightarrows \mathcal{U}$ is called a *selection operator* if $\mathcal{S}(Ku) \subset u + \mathcal{N}(K)$ for all $u \in \mathcal{U}$. A best approximate solution $\hat{u}$ is called a *prior selected solution* of (1.1) if and only if $\hat{u} \in \mathcal{S}(K\hat{u})$.

The general set-valued definition of a selection operator, which we use in order to take care of all the possible cases in regularization methods, also needs to use set-valued ways of convergence. Hence we recall the definition of *Kuratowski convergence* in a metric space.

Definition 4.6. Given a metric space X with metric d, we abuse notation by defining

$$d(u, S) := \inf_{v \in S} d(u, v), \tag{4.7}$$

for $x \in X$ and $S \subset X$. Then the Kuratowski limit inferior and the Kuratowski limit superior of a sequence of sets $S_n \subset X$ are defined as follows:

$$K - \liminf_n(S_n) = \{x \in X \mid \limsup_n d(x, S_n) = 0\}, \tag{4.8}$$

$$K - \limsup_n(S_n) = \{x \in X \mid \liminf_n d(x, S_n) = 0\}. \tag{4.9}$$

For our purposes the lim sup will be of particular interest, because we will use a minimal definition of stability often adopted in the literature on nonlinear methods following Seidman and Vogel (1989) and Engl, Kunisch and Neubauer (1989). Stability is expressed by subsequences of selected solutions having a limit, and each limit of a subsequence being a solution of the limiting problem. The lim inf is less interesting, since there is no reason to ask that any solution of a problem can be the limit of approximate problems. We call an inverse problem stable if, for $f_n \to f$ (usually in terms of norm convergence in $\mathcal{V}$), we have

$$K - \limsup_n \mathcal{S}(f_n) \subset \mathcal{S}(f), \quad K - \limsup_n \mathcal{S}(f_n) \neq \emptyset. \tag{4.10}$$

The metric used for the Kuratowski lim sup will usually be a metrization of some weak or even weak-star convergence in a Banach space; one might also use an extension of the definition to other distance measures.

Having defined the solutions we would like to approximate, the obvious next step is to define a (convergent) regularization method. We start in a deterministic setting, generalizing to a vectorial regularization parameter $\boldsymbol{\alpha} \in \mathbb{R}^M_+$, however, which is useful in many examples, for example the TGV and infimal convolution models with multiple parameters mentioned above. Given an error measure F and $f = Ku^\dagger$ for some exact solution $u^\dagger \in \mathcal{U}$, we call $\delta > 0$ the noise level if it is the best available bound for available data f^δ, that is,

$$F(f, f^\delta) \leq \delta. \tag{4.11}$$

We will be interested in the convergence of regularized solutions to prior selected solutions as the noise level tends to zero. For ease of presentation

and since this is available in almost any known example, we restrict ourselves to convergence with respect to a metric topology τ, which is usually a weak or weak-star topology (on some bounded set in the Banach space).

Definition 4.7. The multivalued operators $R(\cdot, \boldsymbol{\alpha}) : \mathcal{V} \rightrightarrows \mathcal{U}$ defined for $\boldsymbol{\alpha}$ in a subset A of $\mathbb{R}^M$ are called *regularization operators* if, for each $\boldsymbol{\alpha} \in A$, the operator R satisfies the stability property

$$\emptyset \neq K - \liminf_n R(f^{\delta_n}, \boldsymbol{\alpha}) \subset R(f^{\delta}, \boldsymbol{\alpha}) \tag{4.12}$$

for all $f^{\delta} \in \mathcal{V}$ and sequences $f^{\delta_n} \in \mathcal{V}$ converging to f^{δ}. Together with a parameter choice strategy $\boldsymbol{\alpha}$ depending on the noise level δ and the data f^{δ}, that is, a function

$$\boldsymbol{\alpha} : (0, \delta_0) \times \mathcal{V} \to A, \tag{4.13}$$

it is called a *regularization method*.

A regularization method is called *convergent* if, for any sequence $\delta_n \to 0$ and data f^{δ_n} satisfying

$$F(f, f^{\delta_n}) \leq \delta_n, \tag{4.14}$$

we have

$$\emptyset \neq K - \liminf_n R(f^{\delta_n}, \boldsymbol{\alpha}(\delta_n, f^{\delta_n})) \subset \mathcal{S}(f). \tag{4.15}$$

We mention that – besides the very general set-up – our definition of a regularization method deviates from the usual theory, since we do not assume any kind of convergence of the regularization parameter $\boldsymbol{\alpha}$. In the classical theory and most examples, $\boldsymbol{\alpha}$ is a scalar positive value and assumed to converge to zero (or to infinity) as the noise level tends to zero. However, apart from the convenience there seems to be no reason to put such convergence into the definition. Note that in order to approximate a really ill-posed problem, each clustering point of $\boldsymbol{\alpha}(\delta_n, f^{\delta_n})$ will automatically lie outside A. The canonical examples are $A = (0, \alpha_0)$ for variational regularization or $A = \mathbb{N}$ for iterative regularization, where the limiting parameter will converge to zero or infinity. However, we may also consider multi-parameter regularization, where it depends on the formulation whether each component of $\boldsymbol{\alpha}$ has a limit outside the admissible set. Take for example an infimal convolution of two functionals R_1 and R_2. If $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$ are the coefficients of R_1 and R_2, then obviously both should tend to zero in the limit. If, however, α_2 is a relative parameter, that is, α_1 is the coefficient of R_1 and $\alpha_1\alpha_2$ the coefficient of R_2, then it is natural to have a positive limit of α_2. Further motivation for our general definition comes from recent approaches to learning regularization methods for inverse problems, where the $\boldsymbol{\alpha}$ can represent the parameters of the learning scheme. To get a consistent infinite-dimensional theory one could even generalize to non-parametric

learning, which would amount to choosing $\boldsymbol{\alpha}$ in some Banach space. Note that in the remainder of this article we will often write α instead of $\boldsymbol{\alpha}$ if $\boldsymbol{\alpha}$ is only a scalar.

In order to define convergence rates we will also need an error measure $D : \mathcal{U} \times \mathcal{U} \to \mathbb{R}_+ \cup \{+\infty\}$, since there is no natural norm measure as in the Hilbert space. Moreover, we need a restriction to appropriate classes of smoothness, which we denote by M_ν with a parameter $\nu > 0$ measuring the smoothness.

Definition 4.8. A regularization method is called D-convergent if

$$\lim_{\delta \to 0} \sup\{D(u_\delta^{\boldsymbol{\alpha}}, u^\dagger) \mid u_\delta^{\boldsymbol{\alpha}} \in R(f^\delta, \boldsymbol{\alpha}), f^\delta \in \mathcal{V}, F(f, f^\delta) \leq \delta\} = 0. \tag{4.16}$$

A regularization method is called *convergent at order* ν on a set if, for all $f = Ku^\dagger$, $u^\dagger \in \mathcal{M}_\nu$, there exists a constant C_ν such that, for all data g satisfying (4.11), we obtain the estimate

$$D(R(f^\delta, \boldsymbol{\alpha}(\delta, f^\delta)), u^\dagger) \leq C_\nu \delta^\nu. \tag{4.17}$$

Of course the above definition only makes sense for suitable choices of the distance functional and smoothness classes. Remember that in the classical linear Hilbert space theory, these were just norms and spaces obtained by source conditions. We will discuss generalizations of these in the nonlinear setting, in particular related to variational and iterative regularization methods in Banach spaces related to convex regularization functionals. Note also that more general rates than just polynomial ones have been considered in the literature (*e.g.* Hohage 1997, Kaltenbacher 2008).

From an abstract point of view the key insight for generalizing source conditions is the range of the regularization operator. For many linear regularization methods in Hilbert spaces, it is easy to see that the source condition $u^\dagger = K^* \tilde{w}$ means that there exist some data $f^\dagger$ with $u^\dagger = R(f^\dagger, \boldsymbol{\alpha})$. As examples, take Tikhonov regularization:

$$R(\cdot, \boldsymbol{\alpha}) = (K^*K + \alpha I)^{-1} K^* = K^*(KK^* + \alpha I)^{-1}.$$

Due to the invertibility of $(KK^* + \alpha I)^{-1}$, the range of the regularization operator coincides with the range of K^*. Instead of defining source conditions at an abstract level, we thus make the following definition.

Definition 4.9 (range condition). An element $u^\dagger \in \mathcal{S}(f, \boldsymbol{\alpha})$ for $f \in \mathcal{R}(K)$ satisfies the *range condition* if $u^\dagger \in \mathcal{R}(R(\cdot, \boldsymbol{\alpha}))$, that is, there exists $f_\alpha^\dagger$ such that

$$u^\dagger \in R(f_{\boldsymbol{\alpha}}^\dagger, \boldsymbol{\alpha}).$$

We mention that in the case of nonlinear variational methods (with quadratic fidelity), the equivalence of a nonlinear source condition and the range

condition was shown in Burger and Osher (2004), again confirming the appropriateness of this definition.

Roughly speaking, error estimates can now be obtained by some continuity property of the regularization operator, which implies

$$d_{\mathcal{U}}(u_{\boldsymbol{\alpha}}^{\delta}, u^{\dagger}) \leq C(\alpha) d_{\mathcal{V}}(f^{\delta}, f_{\boldsymbol{\alpha}}^{\dagger}),$$

with appropriate distances $d_{\mathcal{U}}$ and $d_{\mathcal{V}}$. With some kind of triangle inequality the right-hand side can be estimated by a distance between f and f^{δ}, which is related to the noise level as well as a distance between f and $f_{\boldsymbol{\alpha}}^{\dagger}$, which is related to the bias of the regularization. This will be discussed in detail for the case of variational regularization methods in Section 5. A weaker concept is that of approximate source conditions (*e.g.* Schuster, Kaltenbacher, Hofmann and Kazimierski 2012, Burger, Helin and Kekkonen 2016*b*) that effectively measure how well the range condition can be approximated. On the other hand stronger conditions can be obtained if $f_{\boldsymbol{\alpha}}^{\dagger}$ above is not arbitrary but in the range of the forward operator K.

4.3. Stochastic approaches

In addition to the deterministic viewpoint, a statistical approach has recently become popular in infinite-dimensional problems as well (Bissantz, Hohage and Munk 2004, Bissantz, Hohage, Munk and Ruymgaart 2007, Cavalier 2008, Kekkonen, Lassas and Siltanen 2014, Giné and Nickl 2015, Hohage and Werner 2016). In such a set-up the data f^{δ} are considered to be random variables drawn from a measure μ_f centred around the exact data f (often representing the expected value and δ being some kind of variance). A regularization operator can then still be applied to each realization and defined in the same way, but we need a different definition of the noise level and the convergence of the regularization method. As a generalization of variance we use the statistical noise level in the mean

$$\mathbb{E}(F(f, f^{\delta})) = \delta. \tag{4.18}$$

Definition 4.10. A regularization operator R with a parameter choice strategy $\boldsymbol{\alpha}$ depending on the statistical noise level δ and the data f^{δ}, that is, a function

$$\boldsymbol{\alpha} : (0, \delta_0) \times \mathcal{V} \to A, \tag{4.19}$$

is called a *statistical regularization method*.

A statistical regularization method is called *convergent* if, for all sequences $\delta_n \to 0$, random variables f^{δ_n} satisfying

$$\mathbb{E}(F(f, f^{\delta_n})) \leq \delta_n, \tag{4.20}$$

and each choice of random variables $u_n \in R_{\boldsymbol{\alpha}_n}(f^{\delta_n})$, there exists a convergent subsequence u_{n_k} in probability in the topology τ, and the limiting random variable $u^\dagger$ satisfies $u^\dagger \in \mathcal{S}(f)$ with probability one.

An extension of this viewpoint is the Bayesian approach to inverse problems, which deals not only with point estimates but with an analogous question for the full posterior distributions. This topic is beyond the scope of this survey: we refer to Kaipio *et al.* (1999), Neubauer and Pikkarainen (2008), Stuart (2010), Kolehmainen *et al.* (2012), Castillo *et al.* (2014), Kekkonen, Lassas and Siltanen (2016), Burger, Helin and Kekkonen (2016*b*) and Nickl *et al.* (2017) for further details.

5. Variational regularization methods

We now return to (3.1) with the viewpoint as in the previous section, we show how variational methods define a regularization operator and then proceed to its further analysis. In this canonical variational regularization method it is apparent how to choose the best approximate and prior selected solution according to Definition 4.4. First of all, the distance measure in the definition of the best approximate solution clearly coincides with the data fidelity. It is just the solution of the variational problem for $\boldsymbol{\alpha}$ in the boundary of A, in the simplest case of a scalar regularization parameter, usually $\alpha = 0$. Of course, the existence of such an element is not obvious, so we define an effective range of the forward operator as

$$\mathcal{R}_F(K) = \left\{ f \in \mathcal{V} \;\middle|\; \operatorname*{arg\,min}_{u \in \mathcal{U}, J(u) < \infty} F(Ku, f) \neq \emptyset \right\}. \tag{5.1}$$

The selection operator is constructed by minimizing the regularization functional on the set of best approximate solutions. Let $f \in \mathcal{R}_F(K)$; then we define

$$\mathcal{S}(f, \boldsymbol{\alpha}) = \operatorname*{arg\,min}_{u \in \mathcal{U}} \left\{ J(u, \boldsymbol{\alpha}) \;\middle|\; u \in \operatorname*{arg\,min}_{\tilde{u} \in \mathcal{U}} F(K\tilde{u}, f) \right\}. \tag{5.2}$$

Remark 5.1. We want to point out that if $\boldsymbol{\alpha} = \alpha$ is just a scalar, the selection operator does not depend on α for regularization functionals of the form $J(u, \alpha) = \alpha J_1(u)$. In this particular case we simply have

$$\mathcal{S}(f) = \operatorname*{arg\,min}_{u \in \mathcal{U}} \left\{ J_1(u) \;\middle|\; u \in \operatorname*{arg\,min}_{\tilde{u} \in \mathcal{U}} F(K\tilde{u}, f) \right\},$$

as the minimizer is not affected by multiplication with a positive scalar. As mentioned above, there are also cases where the selection operator only requires a subset of the parameters as its argument, for example in the case

of infimal convolution regularizations of the form

$$J(u, \boldsymbol{\alpha}) := \inf_v \alpha_1(J_1(u - v) + \alpha_2 J_2(v)),$$

for $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$ and $A = (0, \infty) \times (0, \infty)$. Here $\mathcal{S}(f, \boldsymbol{\alpha}) = \mathcal{S}(f, \alpha_2)$ only depends on α_2.

We will show below that this selection operator is well-defined under standard conditions, which are also used to analyse the variational regularization method.

Following up on variational modelling as described in Section 3, we define a generic variational regularization operator as follows.

Definition 5.2 (variational regularization). Let $F : \mathcal{V} \times \mathcal{V} \to \mathbb{R}_+ \cup \{\infty\}$ be continuous with $F(f, f) = 0$ for all $f \in \mathcal{R}_F(K)$ and $J : \mathcal{U} \times A \to \mathbb{R} \cup \{\infty\}$ be proper, lower semicontinuous and convex functionals, and let $K \in \mathcal{L}(\mathcal{U}, \mathcal{V})$. Then the potentially set-valued operator $R : \mathcal{V} \times A \rightrightarrows \mathcal{U}$ defined as

$$R(f^\delta, \boldsymbol{\alpha}) := \operatorname*{arg\,min}_{u \in \mathcal{U}} \{F(Ku, f^\delta) + J(u, \boldsymbol{\alpha})\} \tag{5.3}$$

is said to be a *variational regularization*, for fixed regularization parameter(s) $\boldsymbol{\alpha} \in A$.

Remark 5.3. We want to emphasize that for convex J, and F that is convex in its first argument, any $u^{\boldsymbol{\alpha}} \in R(f^\delta, \boldsymbol{\alpha})$ can equivalently be characterized via the optimality condition of (5.3), that is,

$$-K^* \partial_x F(Ku^{\boldsymbol{\alpha}}, f^\delta) \in \partial J(u^{\boldsymbol{\alpha}}, \boldsymbol{\alpha}) \tag{5.4}$$

for all $u^{\boldsymbol{\alpha}} \in R(f^\delta, \boldsymbol{\alpha})$.

5.1. Analysis of variational regularization

In the following we will discuss the basic analysis of variational regularization methods; again we try to give a fairly general perspective that covers most of the results in the literature (but due to its generality does not simply reproduce them). Since we focus on the nonlinear regularization we will make the assumption that $\mathcal{V}$ is a separable Hilbert space. A first key issue is the existence of minimizers, which of course depends strongly on the choice of the regularization functional J and possibly also the operator K and the fidelity F. As usual the key issues are lower semicontinuity and compactness in some topology. The latter is always obtained by coercivity in a Banach space norm, which is concluded from the boundedness of the fidelity and in particular the regularization functional. Consequently, the type of compactness is always weak or weak-star, since it is derived from the Banach–Alaoglu theorem (Rudin 2006).

A natural assumption to make for an existence proof is the following.

Assumption 5.4. Let $\mathcal{U} = Z^*$ for some normed space Z and let the weak-star topology on $\mathcal{U}$ be metrizable on bounded sets. Assume moreover

- $K = L^*$ for a bounded linear operator $L : \mathcal{V} \to Z$,
- $J(\cdot, \boldsymbol{\alpha}) = H^*$ for some proper functional $H : Z \to \mathbb{R} \cup \{+\infty\}$ and $J(\cdot, \boldsymbol{\alpha})$ is non-negative,
- F is a proper, non-negative, convex functional in its first argument, and continuous in its second argument, and for every $g \in \mathcal{V}$ there exists u with
$$F(Ku, g) + J(u, \boldsymbol{\alpha}) < \infty.$$
- for each $g \in \mathcal{V}$ and $\boldsymbol{\alpha} \in A$, there exists a constant $c = c(a, b, \|g\|_{\mathcal{V}})$ depending monotonically non-decreasing on all arguments such that
$$\|u\|_{\mathcal{U}} \leq c \quad \text{if } F(Ku, g) \leq a, \quad J(u, \boldsymbol{\alpha}) \leq b.$$

Note that the above assumptions on K and F are reminiscent of the set-up used by Bredies and Pikkarainen (2013) and later by Brinkmann, Burger, Rasch and Sutour (2017). An alternative set-up is to use a compactness assumption on K or some condition on the range of K. Moreover, the assumption on J to be the polar of a proper functional implies convexity, which is predominant in most approaches in regularization theory. With these assumptions we can first verify well-posedness of the selection operator.

Lemma 5.5. Let Assumption 5.4 be satisfied. Then for every $f \in \mathcal{R}_F(K)$ the selection operator $\mathcal{S}$ is well-defined by (5.2) for every $\boldsymbol{\alpha} \in A$.

Proof. If $f \in \mathcal{R}_F(K)$ then there exists a minimizer u^* of $F(Ku, f)$ with $J(u^*, \boldsymbol{\alpha}) < \infty$. Since the minimization in the definition of $\mathcal{S}$ can be restricted to the set of u such that $F(Ku, f) = F(Ku^*, f) =: a$, we obtain an upper bound on the fidelity. On this non-empty set we look for u with $J(u) \leq J(u^*) =: b$. Thus, for the set of such u, the norm in $\mathcal{U}$ is bounded due to Assumption 5.4 and for each minimizing sequence there exists a weak-star convergent subsequence u_n (we can use the metric version of the Banach–Alaoglu theorem due to the assumption of metrizability on bounded sets). Moreover, from our assumptions above it is straightforward to see that $J(\cdot, \boldsymbol{\alpha})$ is sequentially weak-star lower semicontinuous and $F(\cdot, f)$ is weakly lower semicontinuous. From our assumption on K being the adjoint of L, we see that it is continuous from the weak-star topology of $\mathcal{U}$ to the weak topology of $\mathcal{V}$, since for $g \in \mathcal{V}$ we have

$$\langle Ku_n, g \rangle = \langle u_n, Lg \rangle$$

and $Lg \in Z$. As a consequence, the full functional $F(\cdot, f) + J(\cdot, \boldsymbol{\alpha})$ is weak-star lower semicontinuous. Hence, the weak-star limit of u_n is a minimizer, that is, $\mathcal{S}$ is not empty. □

The next step is to verify well-definedness of the regularization operator.

Theorem 5.6. Let Assumption 5.4 be satisfied. Then for every $f \in \mathcal{V}$ the variational regularization model has a minimizer in $\mathcal{U}$ for every $\boldsymbol{\alpha} \in A$, that is, the regularization operator R is well-defined by (5.3). Moreover, $R(f, \boldsymbol{\alpha})$ is a convex set.

Proof. In order to obtain an *a priori* bound, we use the assumption that there exists $\tilde{u}$ with

$$a := F(K\tilde{u}, f) + J(\tilde{u}, \boldsymbol{\alpha}) < \infty.$$

Hence, we can restrict the minimization to those u with functional value less than or equal to a. Setting $b = a$ and using the non-negativity of both terms, we obtain the boundedness of the norm on this subset due to Assumption 5.4. The remaining weak-star compactness and lower semicontinuous arguments to verify the existence of a minimizer are analogous to the proof of Lemma 5.5. The convexity of $R(f, \alpha)$ follows from the convexity of the set of minimizers of a convex functional. □

In order to verify the generalized stability as well as the convergence of the variational regularization, a further condition on F with respect to the second variable is needed. There are several options, the easiest one being satisfied by standard examples such as squared norms is continuity.

Theorem 5.7. Let Assumption 5.4 be satisfied and let F be continuous with respect to the second variable. Then for $\boldsymbol{\alpha} \in A$ and every sequence $f_n \to f \in \mathcal{V}$ there exists a subsequence $u_{n_k} \in R(f_{n_k}, \boldsymbol{\alpha})$ converging to an element $u^* \in R(f, \boldsymbol{\alpha})$ in the weak-star topology.

Proof. By definition of the regularization operator we find for $u_n \in R(f_n, \boldsymbol{\alpha})$ that, for any $u \in \mathcal{U}$,

$$F(Ku_n, f_n) + J(u_n, \boldsymbol{\alpha}) \leq F(Ku, f_n) + J(u, \boldsymbol{\alpha}).$$

Due to the convergence of f_n and the continuity of F in the second argument, the right-hand side in the last estimate is uniformly bounded by some constant a, which again provides uniform bounds for both terms on the left-hand side. Consequently

$$\|u_n\|_{\mathcal{U}} \leq c(a, a, \|f_n\|_{\mathcal{V}}).$$

The boundedness of $\|f_n\|_{\mathcal{V}}$ and monotone dependence of c yields a uniform bound on $\|u_n\|_{\mathcal{U}}$, thus a weakly converging subsequence. Using lower semicontinuity arguments as in the results above and the continuity of F with

respect to the second variable, we see that for the limit u^* the inequality

$$\begin{aligned} F(Ku^*, f) + J(u, \boldsymbol{\alpha}) &\leq \liminf F(Ku_{n_k}, f_{n_k}) + J(u_{n_k}, \boldsymbol{\alpha}) \\ &\leq \lim F(Ku, f_{n_k}) + J(u, \boldsymbol{\alpha}) \\ &= F(Ku, f) + J(u, \boldsymbol{\alpha}). \end{aligned}$$

Hence $u^* \in R(f, \boldsymbol{\alpha})$. □

As mentioned earlier, the type of convergence in Theorem 5.7 corresponds exactly to the type of stability in the Kuratowski limit superior. We finally provide a comment on the convergence of the regularization method. The proof is very similar to the stability result, and an *a priori* bound is obtained by the estimate

$$F(Ku^{\boldsymbol{\alpha}}, f^\delta) + J(u^{\boldsymbol{\alpha}}, \boldsymbol{\alpha}) \leq F(Ku^\dagger, f^\delta) + J(u^\dagger, \boldsymbol{\alpha}) \leq \delta + J(u^\dagger, \boldsymbol{\alpha})$$

for $u^{\boldsymbol{\alpha}} \in R(f^\delta, \boldsymbol{\alpha})$ and any element $u^\dagger \in \mathcal{S}(f, \boldsymbol{\alpha})$. Depending on the specific dependence on $\boldsymbol{\alpha}$, some condition on the interplay of the noise level and the limit of $\boldsymbol{\alpha}$ is needed in order to pass to the limit in

$$J(u^{\boldsymbol{\alpha}}, \boldsymbol{\alpha}) \leq \delta + J(u^\dagger, \boldsymbol{\alpha}).$$

An abstract condition as $\boldsymbol{\alpha}$ converges to $\boldsymbol{\alpha}^*$ outside A is

$$\lim_{\boldsymbol{\alpha} \to \boldsymbol{\alpha}^*} \frac{\delta}{J(u^\dagger, \boldsymbol{\alpha})} = 0,$$

in which case

$$\limsup_{\boldsymbol{\alpha} \to \boldsymbol{\alpha}^*} \frac{J(u^{\boldsymbol{\alpha}}, \boldsymbol{\alpha})}{J(u^\dagger, \boldsymbol{\alpha})} \leq 1.$$

In the standard case $J(u, \boldsymbol{\alpha}) = \alpha J(u)$ the condition is simply $\delta/\alpha \to 0$. Hence, for such parameter choices, variational regularization methods indeed define convergent regularization operators.

5.2. Error estimates

When it comes to the solution of ill-posed, inverse problems, an important question to address is the question of how errors in the measurement data are being propagated in the regularization process; in particular, convergence with respect to the noise level δ and the rate of convergence are of major interest. Following up on Definition 4.8, we look into D-convergence in the case of D being a Bregman distance.

In order to derive error estimates, we restrict ourselves to the following smoothness class $\mathcal{M}_\nu$. Given some unknown ground truth solution $u^\dagger \in \mathcal{S}(f, \boldsymbol{\alpha})$, we ensure $u^\dagger \in \mathcal{R}(R(\cdot, \boldsymbol{\alpha}))$, that is, we have to ensure that there exists data $f^\dagger_{\boldsymbol{\alpha}}$ such that $u^\dagger \in R(f^\dagger_{\boldsymbol{\alpha}}, \boldsymbol{\alpha})$ is a solution of the corresponding variational regularization problem.

Definition 5.8 ((variational) range condition). We say an element $u^\dagger \in \mathcal{S}(f, \boldsymbol{\alpha})$ for $f \in \mathcal{R}_F(K)$ satisfies the *range condition* if $u^\dagger \in \mathcal{R}(R(\cdot, \boldsymbol{\alpha}))$. If $K \in \mathcal{L}(\mathcal{U}, \mathcal{V})$, F is convex and Fréchet-differentiable with respect to its first argument, and $J(\cdot, \boldsymbol{\alpha})$ is proper, convex and lower semicontinuous, then this is equivalent to the existence of $p^\dagger \in \partial J(u^\dagger, \boldsymbol{\alpha})$ and $f^\dagger_{\boldsymbol{\alpha}} \in \mathcal{V}$ such that

$$p^\dagger = -K^* \partial_x F(Ku^\dagger, f^\dagger_{\boldsymbol{\alpha}}). \tag{RC}$$

From now on we assume $K \in \mathcal{L}(\mathcal{U}, \mathcal{V})$, convexity and Fréchet-differentiability of F in its first argument, and properness, convexity and lower semicontinuity of $R(\cdot, \boldsymbol{\alpha})$ for the remainder of this section, which will allow us to use an appropriate optimality condition.

Let us sketch the basic idea in the case of a quadratic fidelity $F(f, g) = \frac{1}{2}\|f - g\|^2$ with some norm in a Hilbert space and $J(u, \boldsymbol{\alpha}) = \alpha J(u)$. The optimality condition (3.1) is given by

$$K^*(Ku^\alpha - f^\delta) + \alpha p^\alpha = 0, \quad p^\alpha \in \partial J(u^\alpha).$$

In order to satisfy the range condition for $u^\dagger$ we need to assume the existence $f^\dagger_\alpha$ such that $p^\dagger \in \partial J(u^\dagger)$ and

$$K^*(Ku^\dagger - f^\dagger_\alpha) + \alpha p^\dagger = 0.$$

We see that this equation implies the condition $p^\dagger = K^* v$ for some v (noting $K^*(Ku^\dagger - f^\dagger) = 0$). On the other hand, if this condition is satisfied we can construct $f^\dagger_\alpha = f - \alpha v$, that is, $p^\dagger = K^* v$ is equivalent to the range condition (RC). An error estimate can then be obtained by subtracting both optimality conditions

$$K^*K(u^\alpha - u^\dagger) + \alpha(p^\alpha - p^\dagger) = K^*(f^\delta - f^\dagger_\alpha).$$

Taking a duality product with $u^\alpha - u^\dagger$ yields

$$\|K(u^\alpha - u^\dagger)\|^2 + \alpha D_J^{p^\alpha}(u^\dagger, u^\alpha) + \alpha D_J^{p^\dagger}(u^\alpha, u^\dagger) = \langle K(u^\alpha - u^\dagger), f^\delta - f^\dagger_\alpha \rangle.$$

Applying Young's inequality on the right-hand side and inserting the special form of $f^\dagger_\alpha$ then immediately yields an error estimate (Burger 2016). Note that we obtain an upper bound on the residual as well as the symmetric Bregman distance

$$D^{\mathrm{symm}}_{J(\cdot,\boldsymbol{\alpha})}(u^\dagger, u^{\boldsymbol{\alpha}}) = D^{p^\alpha}_{J(\cdot,\boldsymbol{\alpha})}(u^\dagger, u^\alpha) + D^{p^\dagger}_{J(\cdot,\boldsymbol{\alpha})}(u^\alpha, u^\dagger). \tag{5.5}$$

For further interpretations of the error estimates see Burger and Osher (2004), Burger, Resmerita and He (2007*b*), Resmerita and Scherzer (2006) and Burger (2016).

We now want to show that (RC) coincides with the well-known source condition (Chavent and Kunisch 1997, Burger and Osher 2004) for a certain class of fidelity functionals. Before we proceed, we have to define this source condition first.

Definition 5.9 (source condition). An element $u^\dagger \in \mathcal{S}(f, \boldsymbol{\alpha})$ for $f \in \mathcal{R}_K(F)$ satisfies the *source condition* if

$$\mathcal{R}(K^*) \cap \partial J(u^\dagger, \boldsymbol{\alpha}) \neq \emptyset.$$

This is equivalent to the existence of $p^\dagger \in \partial J(u^\dagger, \boldsymbol{\alpha})$ and $v \in \mathcal{V}^* \setminus \{0\}$ such that

$$p^\dagger = K^* v. \tag{SC}$$

Remark 5.10. For scalar regularization parameters $\boldsymbol{\alpha} = \alpha$ and regularization functionals of the form $J(u, \alpha) = \alpha J_1(u)$, the source condition for $\alpha = 1$ can be written as $K^* v \in \partial J_1(u^\dagger) = \partial J(u^\dagger, 1)$. Every other potential source condition $K^* v_\alpha \in \partial J(u^\dagger, \alpha)$ can be expressed in terms of v via the relation $v_\alpha = \alpha v$.

It is obvious that (RC) implies (SC). However, we want to go one step further and show that (RC) and (SC) are actually equivalent conditions for fidelity functionals $F(Ku, f^\delta) := G(Ku - f^\delta)$, where G is a *Legendre functional*. Legendre functionals are defined as follows.

Definition 5.11 (Bauschke, Borwein and Combettes 2001, Definition 5.2). Let $G : \mathcal{V} \to \mathbb{R} \cup \{\infty\}$ be a proper, convex and lower semicontinuous functional. We say that G is

- *essentially smooth* if ∂G is both locally bounded and single-valued on its domain,
- *essentially strictly convex* if $(\partial G)^{-1}$ is locally bounded on its domain and G is strictly convex on every convex subset of $\mathrm{dom}(\partial G)$,
- *Legendre* if G is both essentially smooth and essentially strictly convex.

Now we show that (RC) and (SC) are equivalent when G is a Legendre functional.

Theorem 5.12. Let $\mathcal{V}$ be reflexive, and suppose $F(f, f^\delta) := G(f - f^\delta)$ for any $f, f^\delta \in \mathcal{V}$, where $G : \mathcal{V} \to \mathbb{R} \cup \{\infty\}$ is a Legendre functional. Then (RC) and (SC) are equivalent conditions.

Proof. '$\Rightarrow$' Condition (RC) trivially implies (SC) if we define

$$v := -\partial_x F(Ku^\dagger, f^\dagger_{\boldsymbol{\alpha}}) = -G'(Ku^\dagger - f^\dagger_{\boldsymbol{\alpha}}).$$

'$\Leftarrow$' The source condition (SC) can be written as

$$\begin{aligned} & 0 = p^\dagger - K^* v, \\ \Leftrightarrow\ & 0 = p^\dagger + K^* G'((G^*)'(-v)), \end{aligned}$$

where $G^* : \mathcal{V}^* \to \mathbb{R} \cup \{\infty\}$ denotes the convex conjugate of G. Note that G^* is also a Legendre functional since $\mathcal{V}$ is reflexive (Bauschke, Borwein and

Combettes 2001, Corollary 5.5), and that the last equality is valid for all $v \in \mathrm{dom}(G)$ due to Bauschke *et al.* (2001, Theorem 5.9). Hence, if we define

$$f^\dagger_{\boldsymbol{\alpha}} := Ku^\dagger - (G^*)'(-v),$$

we ensure that the range condition (RC) is satisfied. □

The range condition (RC) allows us to derive error estimates in a Bregman distance setting for these very generic variational regularization methods. The following lemma builds the basis by estimating Bregman distances between $u^{\boldsymbol{\alpha}}$ and $u^\dagger$ in terms of differences in data fidelity.

Lemma 5.13. Let (RC) be satisfied. Then we observe

$$\begin{aligned} &D_{F(K\cdot,f^\delta)}(u^\dagger, u^{\boldsymbol{\alpha}}) + D_{F(K\cdot,f^\dagger_{\boldsymbol{\alpha}})}(u^{\boldsymbol{\alpha}}, u^\dagger) + D^{\mathrm{symm}}_{J(\cdot,\boldsymbol{\alpha})}(u^\dagger, u^{\boldsymbol{\alpha}}) \\ &\quad = F(Ku^\dagger, f^\delta) - F(Ku^\dagger, f^\dagger_{\boldsymbol{\alpha}}) + F(Ku^{\boldsymbol{\alpha}}, f^\dagger_{\boldsymbol{\alpha}}) - F(Ku^{\boldsymbol{\alpha}}, f^\delta) \end{aligned} \tag{5.6}$$

for every $u^{\boldsymbol{\alpha}} \in R(f^\delta, \boldsymbol{\alpha})$.

Proof. Computing the optimality condition (5.4) of (5.3) and subtracting $p^\dagger \in \partial J(u^\dagger, \boldsymbol{\alpha})$ from both sides of the equality yields

$$p_{\boldsymbol{\alpha}} - p^\dagger = -K^* \partial_x F(Ku^{\boldsymbol{\alpha}}, f^\delta) - p^\dagger,$$

for any $p_{\boldsymbol{\alpha}} \in \partial J(u^{\boldsymbol{\alpha}}, \boldsymbol{\alpha})$. Taking a duality product with $u^{\boldsymbol{\alpha}} - u^\dagger$ then yields

$$D^{\mathrm{symm}}_{J(\cdot,\boldsymbol{\alpha})}(u^{\boldsymbol{\alpha}}, u^\dagger) = \underbrace{\langle K^* \partial_x F(Ku^{\boldsymbol{\alpha}}, f^\delta), u^\dagger - u^{\boldsymbol{\alpha}} \rangle}_{= F(Ku^\dagger, f^\delta) - F(Ku^{\boldsymbol{\alpha}}, f^\delta) - D_{F(K\cdot,f^\delta)}(u^\dagger, u^{\boldsymbol{\alpha}})} - \langle p^\dagger, u^{\boldsymbol{\alpha}} - u^\dagger \rangle.$$

Hence, we conclude

$$\begin{aligned} &D_{F(K\cdot,f^\delta)}(u^\dagger, u^{\boldsymbol{\alpha}}) + D^{\mathrm{symm}}_{J(\cdot,\boldsymbol{\alpha})}(u^{\boldsymbol{\alpha}}, u^\dagger) \\ &\quad = F(Ku^\dagger, f^\delta) - F(Ku^{\boldsymbol{\alpha}}, f^\delta) - \langle p^\dagger, u^{\boldsymbol{\alpha}} - u^\dagger \rangle. \end{aligned} \tag{5.7}$$

If we now choose $p^\dagger = -K^* \partial_x F(Ku^\dagger, f^\dagger_{\boldsymbol{\alpha}})$ – which is possible since (RC) holds true – we obtain the equality

$$\begin{aligned} -\langle p^\dagger, u^{\boldsymbol{\alpha}} - u^\dagger \rangle &= \langle K^* \partial_x F(Ku^\dagger, f^\dagger_{\boldsymbol{\alpha}}), u^{\boldsymbol{\alpha}} - u^\dagger \rangle \\ &= F(Ku^{\boldsymbol{\alpha}}, f^\dagger_{\boldsymbol{\alpha}}) - F(Ku^\dagger, f^\dagger_{\boldsymbol{\alpha}}) - D_{F(K\cdot,f^\dagger_{\boldsymbol{\alpha}})}(u^{\boldsymbol{\alpha}}, u^\dagger). \end{aligned} \tag{5.8}$$

Inserting (5.8) into (5.7) then yields (5.6). □

Before we proceed, we make the following observation for data fidelities F that are also Bregman distances.

Corollary 5.14. Let $F : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ be a Bregman distance, that is,

$$F(f, f^\dagger_{\boldsymbol{\alpha}}) = G(f) - G(f^\dagger_{\boldsymbol{\alpha}}) - \langle G'(f^\dagger_{\boldsymbol{\alpha}}), f - f^\dagger_{\boldsymbol{\alpha}} \rangle \geq 0,$$

for all $f, f_{\boldsymbol{\alpha}}^\dagger \in \mathcal{V}$, and some functional $G : \mathcal{V} \to \mathbb{R}$. Then we already observe

$$D_{F(\cdot, f_{\boldsymbol{\alpha}}^\dagger)}(f, f^\delta) = F(f, f^\delta),$$

for all $f, f_{\boldsymbol{\alpha}}^\dagger, f^\delta \in \mathcal{V}$.

Proof. We simply compute

$$\begin{aligned} D_{F(\cdot, f_{\boldsymbol{\alpha}}^\dagger)}(f, f^\delta) &= F(f, f_{\boldsymbol{\alpha}}^\dagger) - F(f^\delta, f_{\boldsymbol{\alpha}}^\dagger) - \langle \partial_x F(f^\delta, f_{\boldsymbol{\alpha}}^\dagger), f - f^\delta \rangle \\ &= G(f) - G(f_{\boldsymbol{\alpha}}^\dagger) - \langle G'(f_{\boldsymbol{\alpha}}^\dagger), f - f_{\boldsymbol{\alpha}}^\dagger \rangle \\ &\quad - G(f^\delta) + G(f_{\boldsymbol{\alpha}}^\dagger) + \langle G'(f_{\boldsymbol{\alpha}}^\dagger), f^\delta - f_{\boldsymbol{\alpha}}^\dagger \rangle \\ &\quad - \langle G'(f^\delta) - G'(f_{\boldsymbol{\alpha}}^\dagger), f - f^\delta \rangle, \\ &= G(f) - G(f^\delta) - \langle G'(f_{\boldsymbol{\alpha}}^\dagger), f - f^\delta \rangle \\ &\quad - \langle G'(f^\delta) - G'(f_{\boldsymbol{\alpha}}^\dagger), f - f^\delta \rangle, \\ &= D_G(f, f^\delta) = F(f, f^\delta), \end{aligned}$$

and hence prove the result. □

As a consequence, Lemma 5.13 reads as follows for data fidelities that are also Bregman distances.

Lemma 5.15. Let the assumptions of Lemma 5.13 and Corollary 5.14 hold true. Then we have

$$D_{J(\cdot,\boldsymbol{\alpha})}^{\mathrm{symm}}(u^\dagger, u^{\boldsymbol{\alpha}}) = \langle G'(f_{\boldsymbol{\alpha}}^\dagger) - G'(Ku^\dagger) - (G'(f^\delta) - G'(Ku^{\boldsymbol{\alpha}})), Ku^\dagger - Ku^{\boldsymbol{\alpha}} \rangle.$$

Proof. From Corollary 5.14 we know that

$$D_{F(K\cdot, f^\delta)}(u^\dagger, u^{\boldsymbol{\alpha}}) = F(Ku^\dagger, Ku^{\boldsymbol{\alpha}}), \quad D_{F(K\cdot, f_{\boldsymbol{\alpha}}^\dagger)}(u^{\boldsymbol{\alpha}}, u^\dagger) = F(Ku^{\boldsymbol{\alpha}}, Ku^\dagger).$$

Hence, we observe

$$\begin{aligned} &D_{F(K\cdot, f^\delta)}(u^\dagger, u^{\boldsymbol{\alpha}}) + D_{F(K\cdot, f_{\boldsymbol{\alpha}}^\dagger)}(u^{\boldsymbol{\alpha}}, u^\dagger) \\ &\quad = F(Ku^\dagger, Ku^{\boldsymbol{\alpha}}) + F(Ku^{\boldsymbol{\alpha}}, Ku^\dagger) \\ &\quad = D_G^{\mathrm{symm}}(Ku^\dagger, Ku^{\boldsymbol{\alpha}}) \\ &\quad = \langle G'(Ku^\dagger) - G'(Ku^{\boldsymbol{\alpha}}), Ku^\dagger - Ku^{\boldsymbol{\alpha}} \rangle. \end{aligned}$$

We also discover

$$\begin{aligned} &F(Ku^\dagger, f^\delta) - F(Ku^\dagger, f_{\boldsymbol{\alpha}}^\dagger) + F(Ku^{\boldsymbol{\alpha}}, f_{\boldsymbol{\alpha}}^\dagger) - F(Ku^{\boldsymbol{\alpha}}, f^\delta) \\ &\quad = G(Ku^\dagger) - G(f^\delta) - \langle G'(f^\delta), Ku^\dagger - f^\delta \rangle \\ &\qquad - \big(G(Ku^\dagger) - G(f_{\boldsymbol{\alpha}}^\dagger) - \langle G'(f_{\boldsymbol{\alpha}}^\dagger), Ku^\dagger - f_{\boldsymbol{\alpha}}^\dagger \rangle\big) \\ &\qquad + G(Ku^{\boldsymbol{\alpha}}) - G(f_{\boldsymbol{\alpha}}^\dagger) - \langle G'(f_{\boldsymbol{\alpha}}^\dagger), Ku^{\boldsymbol{\alpha}} - f_{\boldsymbol{\alpha}}^\dagger \rangle \\ &\qquad - \big(G(Ku^{\boldsymbol{\alpha}}) - G(f^\delta) - \langle G'(f^\delta), Ku^{\boldsymbol{\alpha}} - f^\delta \rangle\big) \end{aligned}$$

$$= \langle G'(f^\delta), Ku^{\boldsymbol{\alpha}} - Ku^\dagger \rangle + \langle G'(f^\dagger_{\boldsymbol{\alpha}}), Ku^\dagger - Ku^{\boldsymbol{\alpha}} \rangle$$
$$= \langle G'(f^\delta) - G'(f^\dagger_{\boldsymbol{\alpha}}), Ku^{\boldsymbol{\alpha}} - Ku^\dagger \rangle.$$

Combining these two equalities with (5.6) yields the desired result. □

Example 5.16. We can use (5.6) to derive the same error estimates presented in Burger *et al.* (2007*b*) for the choice $F(Ku, f^\delta) = \frac{1}{2}\|Ku - f^\delta\|^2_{\mathcal{H}}$, where $\mathcal{H}$ is a Hilbert space. In this case we observe

$$D_{F(K\cdot, f^\delta)}(u^\dagger, u^{\boldsymbol{\alpha}}) = D_{F(K\cdot, f^\dagger_{\boldsymbol{\alpha}})}(u^{\boldsymbol{\alpha}}, u^\dagger) = \frac{1}{2}\|K(u^\dagger - u^{\boldsymbol{\alpha}})\|^2_{\mathcal{H}}.$$

Hence, equation (5.6) reads as

$$\|K(u^\dagger - u^{\boldsymbol{\alpha}})\|^2_{\mathcal{H}} + D^{\text{symm}}_{J(\cdot,\boldsymbol{\alpha})}(u^\dagger, u^{\boldsymbol{\alpha}}) + \frac{1}{2}\|Ku^{\boldsymbol{\alpha}} - f^\delta\|^2_{\mathcal{H}}$$
$$= \frac{1}{2}\|Ku^\dagger - f^\delta\|^2_{\mathcal{H}} + \frac{1}{2}\|Ku^{\boldsymbol{\alpha}} - f^\dagger_{\boldsymbol{\alpha}}\|^2_{\mathcal{H}} - \frac{1}{2}\|Ku^\dagger - f^\dagger_{\boldsymbol{\alpha}}\|^2_{\mathcal{H}}. \tag{5.9}$$

If we make use of the estimate

$$\frac{1}{2}\|Ku^\dagger - f^\delta\|^2_{\mathcal{H}} \leq \frac{1}{2}\|f - f^\delta\|^2_{\mathcal{H}} \leq \delta,$$

then equality (5.9) becomes the inequality

$$\|K(u^\dagger - u^{\boldsymbol{\alpha}})\|^2_{\mathcal{H}} + D^{\text{symm}}_{J(\cdot,\boldsymbol{\alpha})}(u^\dagger, u^{\boldsymbol{\alpha}}) + \frac{1}{2}\|Ku^{\boldsymbol{\alpha}} - f^\delta\|^2_{\mathcal{H}}$$
$$\leq \delta + \frac{1}{2}\|K(u^{\boldsymbol{\alpha}} - u^\dagger) + (Ku^\dagger - f^\dagger_{\boldsymbol{\alpha}})\|^2_{\mathcal{H}} - \frac{1}{2}\|Ku^\dagger - f^\dagger_{\boldsymbol{\alpha}}\|^2_{\mathcal{H}}$$
$$\leq \delta + \|K(u^\dagger - u^{\boldsymbol{\alpha}})\|^2_{\mathcal{H}} + \|Ku^\dagger - f^\dagger_{\boldsymbol{\alpha}}\|^2_{\mathcal{H}} - \frac{1}{2}\|Ku^\dagger - f^\dagger_{\boldsymbol{\alpha}}\|^2_{\mathcal{H}}.$$

Subtracting $\|K(u^\dagger - u^{\boldsymbol{\alpha}})\|^2_{\mathcal{H}}$ on both sides of the inequality then yields the error estimate

$$D^{\text{symm}}_{J(\cdot,\boldsymbol{\alpha})}(u^\dagger, u^{\boldsymbol{\alpha}}) + \frac{1}{2}\|Ku^{\boldsymbol{\alpha}} - f^\delta\|^2_{\mathcal{H}} \leq \delta + \frac{1}{2}\|Ku^\dagger - f^\dagger_{\boldsymbol{\alpha}}\|^2_{\mathcal{H}}. \tag{5.10}$$

We want to emphasize that the constant $\frac{1}{2}\|Ku^\dagger - f^\dagger_{\boldsymbol{\alpha}}\|^2_{\mathcal{H}}$ on the right-hand-side of the inequality depends on the choice of $\boldsymbol{\alpha}$. From Remark 5.10 and the proof of Theorem 5.12 it follows that if we consider regularizations of the form $J(u, \boldsymbol{\alpha}) = \alpha J(u)$, the source condition (SC) and the range condition (RC) are linked via the relation $f^\dagger_{\boldsymbol{\alpha}} = Ku^\dagger + \alpha v$, where v is the source condition element for $\alpha = 1$, *i.e.* $K^*v \in \partial J(u^\dagger, 1) = \partial J(u^\dagger)$. In this setting, the error estimate (5.10) then reads as

$$D^{\text{symm}}_J(u^\dagger, u^{\boldsymbol{\alpha}}) + \frac{1}{2\alpha}\|Ku^{\boldsymbol{\alpha}} - f^\delta\|^2_{\mathcal{H}} \leq \frac{\delta}{\alpha} + \frac{\alpha}{2}\|v\|^2_{\mathcal{H}}.$$

Hence, choosing $\alpha(\delta) = \sqrt{2\delta}/\|v\|_{\mathcal{H}}$ then yields $D^{\text{symm}}_J(u^\dagger, u^{\boldsymbol{\alpha}}) = O(\sqrt{\delta})$.

There are various routes and generalizations that can be taken from these types of estimates, for example to weaker source conditions with the concepts of approximate or variational source conditions (Schuster, Kaltenbacher, Hofmann and Kazimierski 2012, Flemming and Hofmann 2010, Flemming 2013, Flemming 2017*b*, Hohage and Weidling 2017), improved estimates for stronger conditions (Resmerita 2005, Grasmair 2013), or large noise that is not necessarily in $\mathcal{V}$ (Burger *et al.* 2016*b*). In special cases such as ℓ^1-regularization, improved results can be obtained, because the effective finite-dimensionality implies that this case is almost well-posed (Grasmair, Scherzer and Haltmeier 2011, Grasmair 2011, Burger, Flemming and Hofmann 2013*a*, Flemming, Hofmann and Veselić 2015, Flemming, Hofmann and Veselić 2016, Flemming and Gerth 2017). Converse results have also been recently obtained (Flemming 2017*a*, Hohage and Weidling 2017).

5.3. Variational eigenvalue problems

The standard tool for the analysis of linear regularization methods is singular value decomposition. In the case of nonlinear regularization no analogue of singular values and singular vectors was known for a long time. A generalization for nonlinear variational methods was made in Benning and Burger (2013), which we discuss in the following. We generalize singular vectors as eigenvectors of the variational regularization operator R as defined in Definition 5.2, that is, we look for functions u_λ that satisfy

$$\lambda u_\lambda \in R(\sigma K u_\lambda, \boldsymbol{\alpha}), \tag{5.11}$$

for constants $\lambda, \sigma \in [0, \infty)$, typically $\sigma = 1$. For simplicity we focus on the case where $\boldsymbol{\alpha} = \alpha$ is a scalar, and $F(Ku, f^\delta) = G(Ku - f^\delta)$, where G is a Legendre functional for the remainder of this section. If we consider the optimality condition (5.4) of (5.3), we immediately observe that any u_λ satisfying (5.11) also has to satisfy

$$-K^* G'((\lambda - \sigma) K u_\lambda) \in \partial J(\lambda u_\lambda, \alpha). \tag{5.12}$$

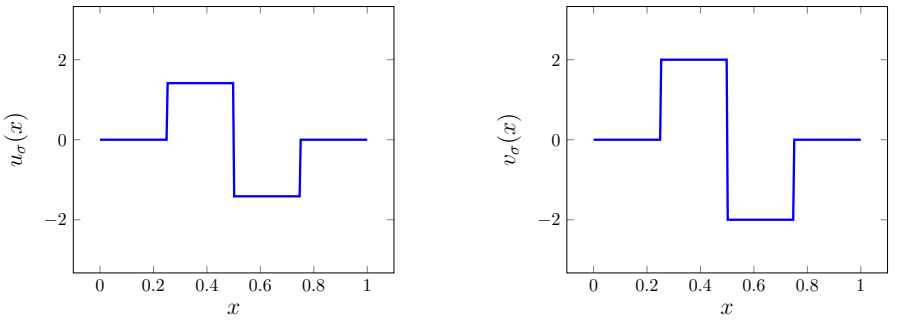
We now assume that both G' and ∂J are homogeneous in the sense that they satisfy $G'(cu) = s_1(c) G'(u)$ and $\partial J(cu, \boldsymbol{\alpha}) = s_2(c, \alpha) \partial J(u)$ for constants $c \in \mathbb{R}$ and functions $s_1, s_2 : \mathbb{R} \to \mathbb{R}$. Then (5.12) simplifies to

$$-\frac{s_1(\lambda - \sigma)}{s_2(\lambda/\sigma, \alpha)} K^* G'(K u_\lambda) \in \partial J(\sigma u_\lambda). \tag{5.13}$$

Equation (5.13) paves the way for the following definition of generalized singular vectors.

Definition 5.17 (generalized singular system). Let $\{u_\sigma, v_\sigma, \sigma\}$ satisfy

$$K u_\sigma = \sigma v_\sigma \quad \text{and} \quad K^* G'(v_\sigma) \in \partial J(\sigma u_\sigma) \tag{5.14}$$



(a) $u_{\sigma_1,1/2}$ (b) $v_{\sigma_1,1/2}$

Figure 5.1. The Haar wavelet $u_{\sigma_1,1/2}$ and its scaled version $v_{\sigma_1,1/2}$. Benning and Burger (2013) have shown that together with $\sigma_1 = 2^{-5/2}$ they form a generalized singular system in the sense of (5.14) with K being the identity in L^2.

for $\sigma > 0$. Then $\{u_\sigma, v_\sigma, \sigma\}$ is called a *generalized singular system*.

Remark 5.18. For

$$G(v) = \frac{1}{2}\|v\|^2_{L^2(\Sigma)} \quad \text{and} \quad J(u,\alpha) = \frac{\alpha}{2}\|u\|^2_{L^2(\Omega)},$$

this definition is consistent with the classical singular vector theory for compact operators.

Example 5.19. Suppose

$$G(v) = \frac{1}{2}\|v\|_{L^2([0,1])^2}, \quad K : \mathrm{BV}([0,1]) \to L^2([0,1])$$

is the embedding operator and $J(u,\alpha) = \alpha\,\mathrm{TV}_*(u)$, where TV_* denotes the (one-dimensional) total variation with Dirichlet-zero boundary conditions. Benning and Burger (2013) have shown that Haar wavelets are generalized singular vectors of TV_*. Precisely, we have $v_{\sigma_n,k} = \sigma_n u_{\sigma_n,k} \in \partial\,\mathrm{TV}_*(\sigma_n u_{\sigma_n,k}) = \partial\,\mathrm{TV}_*(u_{\sigma_n,k})$ for $\sigma_n := 2^{-(n+4)/2}$ and $u_{\sigma_n,k}$ defined by

$$u_{\sigma_n,k}(x) := 2^{n/2}\Psi(2^n x - j) \quad \text{with } \Psi(x) := \begin{cases} 1 & x \in [0,1/2), \\ -1 & x \in [1/2,1), \\ 0 & \text{else.} \end{cases}$$

The singular value σ_n is determined via (5.14). The dual singular vector $v_{\sigma_n,k}$ has to satisfy $v_{\sigma_n,k} \in \partial\,\mathrm{TV}_*(\sigma_n u_{\sigma_n,k}) = \mathrm{TV}_*(u_{\sigma_n,k})$. If we make use of $v_{\sigma_n,k} = u_{\sigma_n,k}/\sigma_n$ and take a dual product with $u_{\sigma_n,k}$, we immediately observe $\sigma_n = \|u_{\sigma_n,k}\|^2_{L^2([0,1])}/\,\mathrm{TV}_*(u_{\sigma_n,k})$. In Figure 5.1 we see the Haar wavelet $u_{\sigma_1,1/2}$ and its scaled version $v_{\sigma_1,1/2} = \sigma_1 u_{\sigma_1,1/2} = 2^{-5/2}u_{\sigma_1,1/2}$.

The generalized singular system is defined so that (5.13) and (5.14) coincide for

$$s_1(\lambda - \sigma) = -s_2(\lambda/\sigma, \alpha). \tag{5.15}$$

Hence, if we choose α and λ such that (5.15) holds true, we already know that (5.11) is satisfied for these particular choices of λ and α.

Example 5.20. For

$$G(v) = \frac{1}{2}\|v\|^2_{L^2(\Sigma)} \quad \text{and} \quad J(u, \alpha) = \frac{\alpha}{2}\|u\|^2_{L^2(\Omega)},$$

with Σ and Ω being domains in $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$, respectively, we observe $s_1(x) = x$ and $s_2(x, \alpha) = \alpha x$. Hence, (5.15) simplifies to $\sigma - \lambda = (\alpha\lambda)/\sigma$. Solving for λ then yields

$$\lambda = \frac{\sigma}{\sigma^2 + \alpha},$$

which perfectly coincides with the singular value decomposition representation of Tikhonov regularization.

Example 5.21. For $G(v) = \frac{1}{2}\|v\|^2_{L^2(\Sigma)}$ and $J(u, \boldsymbol{\alpha}) = \alpha\,\mathrm{TV}(u)$ we have $s_1(x) = x$ and $s_2(x, \alpha) = \alpha$. Consequently, (5.15) solved for λ reads as

$$\lambda = \frac{1 - \alpha}{\sigma}.$$

This eigenvalue of this particular regularization operator is consistent with classical singular value theory in the sense that it satisfies $\lim_{\alpha \downarrow 0} \lambda = 1/\sigma$.

An interesting observation from Examples 5.20 and 5.21 is that $\alpha > 0$ automatically implies $\lambda < 1/\sigma$ (unless $u_\sigma \in \ker(J)$). This implies that there is always a systematic error when it comes to recovering singular vectors with variational regularization methods that have quadratic fidelity. This is also true for input data that are not given in terms of a singular vector: see Benning and Burger (2013, Theorem 7). In the next section we see that iterative regularization methods can overcome this systematic reconstruction bias.

6. Iterative regularization methods

Iterative regularizations are founded on a different paradigm to variational methods, based on the simple observation that most iterative procedures can be applied in a robust fashion to ill-posed problems. The standard example in a Hilbert space is the Landweber iteration (Landweber 1951)

$$u^k = u^{k-1} - \tau K^*(Ku^{k-1} - f^\delta),$$

which merely uses the continuous operators K and K^*. Let us mention again at this point that with standard initial values such as $u^0 = 0$ the iterates satisfy a range condition $u^k \in \mathcal{R}(K^*)$. At an abstract level we construct an iteration procedure

$$u^k = R_I(f^\delta, v^{k-1}, \boldsymbol{\alpha}), \tag{6.1}$$

with some iteration operator R_I and a collection of variables v^{k-1} summarizing the information used in the first $k-1$ steps. In this case the parameter set $\boldsymbol{\alpha}$ will contain the iteration index as well as auxiliary parameters such as the step size τ. In the simplest case of a one-step method like the Landweber iteration, we simply have $v^{k-1} = u^{k-1}$; for multistep methods the variable v^{k-1} could be a collection of several previous iterations. As we shall see in the methods below, v^{k-1} could also collect some auxiliary variables.

For such methods one observes a so-called *semi-convergence* phenomenon. In the case of exact data $f \in \mathcal{R}(K)$ the method is converging, while in the case of noisy data it seems to approximate the exact solution for an initial phase of the iteration and then starts to diverge. This behaviour naturally leads to the idea of achieving a regularizing effect by stopping the iterations early. A standard approach is the so-called discrepancy principle, which monitors the residual during the iteration and compares it with the noise level. Since the exact solution could lead to a residual at this level there is no particular reason to iterate further once the residual is at the size of the noise level.

Definition 6.1 (Morozov's discrepancy principle). Let f and f^δ satisfy $F(f, f^\delta) \leq \delta$, and let $R_I : \mathcal{V} \times \mathcal{U} \times A \to \mathcal{U}$ be a multivalued mapping. If we choose $\eta \geq 1$ and $k^* := k^*(\delta, f^\delta)$ such that

$$F(Ku^{k^*}, f^\delta) \leq \eta\delta < F(Ku^k, f^\delta)$$

is satisfied for $u^{k^*} \in R_I(f^\delta, v^{k^*-1}, \boldsymbol{\alpha})$ and $u^k \in R_I(f^\delta, v^{k-1}, \boldsymbol{\alpha})$ for all $k < k^*$, then u^k is said to satisfy *Morozov's discrepancy principle*.

Given a stopping rule to determine $k^*(\delta, f^\delta)$ such as the discrepancy principle, we can define the full regularizaton operator:

$$R(f^\delta, \boldsymbol{\alpha}) = u^{k_*(\delta, f^\delta)}, \tag{6.2}$$

where for $k = 1, \ldots, k_*(\delta, f^\delta)$ the iterates u^k are determined by (6.1) with some fixed initial value v^0 including u^0.

The semiconvergence behaviour of such a method is then the standard convergence of a nonlinear regularization method. In particular, for consistency we need $u^k \longrightarrow_{\tau_\mathcal{U}} u^\dagger$ as $k \to \infty$ in the case of clean data $f \in \mathcal{R}_F(K)$ and $u^\dagger \in \mathcal{S}(f, \boldsymbol{\alpha})$. A standard tool used to prove the convergence of an iterative regularization method is to find some error measure for the true solution that is decreasing until the stopping index is reached. For the methods

below constructed from a regularization functional J, we will see that this is the case for the Bregman distance, that is,

$$D^{p^{k+1}}_{J(\cdot,\boldsymbol{\alpha})}(u^\dagger, u^{k+1}) \leq D^{p^k}_{J(\cdot,\boldsymbol{\alpha})}(u^\dagger, u^k)$$

for $u^k \in R_I(f^\delta, u^{k-1}, \boldsymbol{\alpha})$ and all $k \leq k^* - 1$, and

$$\lim_{\delta\to 0} D^{p^k}_{J(\cdot,\boldsymbol{\alpha})}(u^k_\delta, u^k) = 0$$

for $u^k_\delta \in R_I(f^\delta, u^{k-1}_\delta, \boldsymbol{\alpha})$ and $u^k \in R_I(f, u^{k-1}, \boldsymbol{\alpha})$. With some further effort one can then conclude the convergence of the regularization method in this sense:

$$\lim_{\delta\to 0} \sup\{ D^{p^{k^*(\delta,f^\delta)}}_{J(\cdot,\boldsymbol{\alpha})}(u^\dagger, u^{k^*(\delta,f^\delta)}) \mid f^\delta \in \mathcal{V}, F(f, f^\delta) \leq \delta \} = 0,$$

for $R(f^\delta, \boldsymbol{\alpha}) = u^{k^*(\delta,f^\delta)}$ (Osher *et al.* 2005, Schuster *et al.* 2012).

As in the case of Banach spaces such as BV there is no immediate analogue of simple iterative procedures in Hilbert spaces, one often resorts to defining an iteration operator R_I by solving a variational problem. This approach will be detailed in the following sections.

6.1. Bregman iteration

The concept of Bregman iteration – also known as proximal minimization algorithm – introduces an iteration into the variational regularization framework by replacing the regularization functional $J(u, \boldsymbol{\alpha})$ with the corresponding generalized Bregman distance $D^p_{J(\cdot,\boldsymbol{\alpha})}(u, v)$, for $v \in \mathcal{U}$ and $p \in \partial J(v)$. For the choice

$$J(u, \boldsymbol{\alpha}) = \frac{\boldsymbol{\alpha}}{2}\|u\|^2_{L^2(\Omega)},$$

it is also known as iterated Tikhonov regularization, which dates back to the works of Kryanev (1974), further analysed in Groetsch (1977) and Thomas King and Chillingworth (1979), for example. The extension to more general choices of Bregman distances was first proposed by Censor and Zenios (1992), shortly followed by Teboulle (1992), and has since been the subject of extensive research (Eckstein 1993, Kiwiel 1997). Notably, Osher *et al.* (2005) have extended it to generalized Bregman distances that allow for subdifferentiable rather than differentiable functionals. Note that in such cases there is no one-to-one relation between u^{k-1} and its subgradient p^{k-1}, so we set $v^{k-1} = (u^{k-1}, p^{k-1})$. With a set-valued iteration

operator, the Bregman iteration can be written as

$$u^k \in R_I(f^\delta, v^{k-1}, \boldsymbol{\alpha}) = \arg\min_{u \in \mathcal{U}} \left\{ F(Ku, f^\delta) + D^{p^{k-1}}_{J(\cdot,\boldsymbol{\alpha})}(u, u^{k-1}) \right\},$$
$$p^k = p^{k-1} - K^* \partial_x F(Ku^k, f^\delta),$$

for $p^0 \in \partial J(u^0, \boldsymbol{\alpha})$. The entire method is summarized in Algorithm 1.

Algorithm 1 Bregman iteration

Initialize $\boldsymbol{\alpha} \in A$, $f^\delta \in \mathcal{V}$, $u^0 \in \mathcal{U}$ and p^0 with $p^0 \in \partial J(u^0, \boldsymbol{\alpha})$
for $k = 1, \ldots, k^*$ **do**
  Compute $R_I(f^\delta, v^{k-1}, \boldsymbol{\alpha}) = \arg\min_{u \in \mathcal{U}} \left\{ F(Ku, f^\delta) + D^{p^{k-1}}_{J(\cdot,\boldsymbol{\alpha})}(u, u^{k-1}) \right\}$
  Pick $R_I(f^\delta, v^{k-1}, \boldsymbol{\alpha})$
  Update $p^k = p^{k-1} - K^* \partial_x F(Ku^k, f^\delta)$
  Set $v^k = (u^k, p^k)$
end for
return u^{k^*}, p^{k^*}

Remark 6.2. The update for the subgradient can also be written as

$$p^k = p^0 - \sum_{n=1}^{k} K^* \partial_x F(Ku^n, f^\delta). \tag{6.3}$$

Hence, we can rewrite the primal update to

$$R_I(f^\delta, \{u^n\}_{n=1}^{k-1}, p^0, \boldsymbol{\alpha}) \tag{6.4}$$
$$= \arg\min_{u \in \mathcal{U}} \left\{ F(Ku, f^\delta) + J(u, \boldsymbol{\alpha}) - \left\langle p^0 - \sum_{n=1}^{k-1} K^* \partial_x F(Ku^n, f^\delta), u \right\rangle \right\}.$$

In the following we want to recall (or derive) a few important properties of Algorithm 1. We start with a trivial result establishing monotonic decrease of the data fidelity.

Corollary 6.3 (monotonic decrease of data fidelity). Suppose that u^0 satisfies $F(Ku^0, f^\delta) < \infty$. Then the iterates of Algorithm 1 satisfy

$$F(Ku^{k+1}, f^\delta) + D^{p^k}_{J(\cdot,\boldsymbol{\alpha})}(u^{k+1}, u^k) \leq F(Ku^k, f^\delta),$$

and

$$\lim_{k \to \infty} D^{p^k}_{J(\cdot,\boldsymbol{\alpha})}(u^{k+1}, u^k) = 0,$$

for $u^k \in R_I(f^\delta, v^{k-1}, \boldsymbol{\alpha})$ and all $k \in \mathbb{N}$.

Proof. The first statement follows trivially from the convexity of F (in its first argument) and J, and the fact that u^{k+1} is a minimizer of $E(u) := F(Ku, f^\delta) + D^{p^k}_{J(\cdot,\boldsymbol{\alpha})}(u, u^k)$. The first statement then implies

$$\sum_{k=0}^{N-1} D^{p^k}_{J(\cdot,\boldsymbol{\alpha})}(u^{k+1}, u^k) \leq F(Ku^0, f^\delta) - F(Ku^N, f^\delta) \leq F(Ku^0, f^\delta) < \infty.$$

Taking the limit $N \to \infty$ then yields the second statement. □

If we want to show that the Bregman iteration is a convergent regularization method in the sense of Definition 4.8, a first step towards this result would be the following monotonicity lemma.

Lemma 6.4 (Fejér monotonicity of Algorithm 1). Let $f \in \mathcal{R}_F(K)$, $u^\dagger \in \mathcal{S}(f, \boldsymbol{\alpha})$ and let $f^\delta \in \mathcal{V}$ with $F(f, f^\delta) \leq \delta$. We further assume that the iterates of Algorithm 1 satisfy Definition 6.1 for $\eta = 1$. Then the iterates also satisfy the strict Fejér monotonicity

$$D^{p^k}_{J(\cdot,\boldsymbol{\alpha})}(u^\dagger, u^k) < D^{p^{k-1}}_{J(\cdot,\boldsymbol{\alpha})}(u^\dagger, u^{k-1}),$$

for $u^k \in R_I(f^\delta, v^{k-1}, \boldsymbol{\alpha})$ and all $k < k^*$.

Proof. Through straightforward computations we obtain

$$\begin{aligned} D^{p^k}_{J(\cdot,\boldsymbol{\alpha})}(u^\dagger, u^k) - D^{p^{k-1}}_{J(\cdot,\boldsymbol{\alpha})}(u^\dagger, u^{k-1}) &= \underbrace{-D^{p^{k-1}}_{J(\cdot,\boldsymbol{\alpha})}(u^k, u^{k-1})}_{<0} \\ &\quad - \langle p^k - p^{k-1}, u^\dagger - u^k \rangle \\ &\leq \langle K^* \partial_x F(Ku^k, f^\delta), u^\dagger - u^k \rangle \\ &\leq (\delta - F(Ku^k, f^\delta)) \\ &< 0 \end{aligned}$$

for $k < k^*$, where we have made use of the convexity of F in its first argument, and $F(Ku^\dagger, f^\delta) \leq F(f, f^\delta) \leq \delta$. □

Corollary 6.5. Let $f \in \mathcal{R}_F(K)$ and $u^\dagger \in \mathcal{S}(f, \boldsymbol{\alpha})$. Then the iterates of Algorithm 1 satisfy

$$\sum_{k=0}^{\infty} F(Ku^k, f) < \infty \tag{6.5}$$

for $\delta = 0$ (and, thus, $f^\delta = f$) and u^0 (with $p^0 \in \partial J(u^0, \boldsymbol{\alpha})$) chosen such that $D^{p^0}_{J(\cdot,\boldsymbol{\alpha})}(u^\dagger, u^0) < \infty$.

Proof. For $\delta = 0$ we conclude

$$F(Ku^k, f) \leq D^{p^{k-1}}_{J(\cdot,\boldsymbol{\alpha})}(u^\dagger, u^{k-1}) - D^{p^k}_{J(\cdot,\boldsymbol{\alpha})}(u^\dagger, u^k)$$

from Lemma 6.4. Summing up from $k = 0$ to some $k = k^*$ therefore yields

$$\sum_{k=0}^{k^*} F(Ku^k, f) \leq D^{p^0}_{J(\cdot,\boldsymbol{\alpha})}(u^\dagger, u^0) - D^{p^{k^*}}_{J(\cdot,\boldsymbol{\alpha})}(u^\dagger, u^{k^*}) \leq D^{p^0}_{J(\cdot,\boldsymbol{\alpha})}(u^\dagger, u^0) < \infty.$$

Taking the limit $k^* \to \infty$ yields the assertion. □

Remark 6.6. Given the continuity of F and $K \in \mathcal{L}(\mathcal{U}, \mathcal{V})$, equation (6.5) already implies

$$Ku^k \underset{\tau_\mathcal{V}}{\longrightarrow} f, \tag{6.6}$$

if $\tau_\mathcal{V}$ is an appropriate topology on $\mathcal{V}$ related to F.

Lemma 6.7. Suppose that after a finite number of iterations the k^*th iterate of Algorithm 1 satisfies $Ku^{k^*} = f$, for $u^{k^*} \in R(f, v^{k^*-1}, \boldsymbol{\alpha})$, $f \in \mathcal{R}_K(F)$ and $p^0 \in \mathcal{R}(K^*)$. Then $u^{k^*} \in \mathcal{S}(f, \boldsymbol{\alpha})$.

Proof. We know $D^{p^{k^*}}_{J(\cdot,\boldsymbol{\alpha})}(u, u^{k^*}) \geq 0$ for all $u \in \mathcal{U}$ and $u^{k^*} \in R_I(f, v^{k^*-1}, \boldsymbol{\alpha})$, since J is convex; this in particular holds true for any $\hat{u} \in \{u \mid Ku = f\}$. Hence, we observe

$$\begin{aligned}
J(u^{k^*}) &\leq J(\hat{u}) - \langle p^{k^*}, \hat{u} - u^{k^*} \rangle \\
&= J(\hat{u}) - \langle p^0, \hat{u} - u^{k^*} \rangle + \sum_{n=1}^{k^*} \langle K^* \partial_x F(Ku^n, f), \hat{u} - u^{k^*} \rangle, \\
&\leq J(\hat{u}) - \left\langle q^0, \underbrace{K\hat{u} - Ku^{k^*}}_{=0} \right\rangle + \sum_{n=1}^{k^*} \left\langle \partial_x F(Ku^n, f), \underbrace{K\hat{u} - Ku^{k^*}}_{=0} \right\rangle, \\
&= J(\hat{u}),
\end{aligned}$$

for the substitution $p^0 := K^* q^0$, which is possible since $p^0 \in \mathcal{R}(K^*)$. Here we have made use of equation (6.3). Consequently, we conclude $u^{k^*} \in \mathcal{S}(f, \boldsymbol{\alpha})$. □

In the limiting case $k^* \to \infty$ the selection is less clear: one cannot prove in general that the limit minimizes J. To make this more apparent consider the case of a least-squares fidelity functional $F(Ku, f) = \frac{1}{2}\|Ku - f\|^2$, with the

initial value being a minimizer of the regularization, *i.e.* $p^0 = 0$. Then the estimate, as in the last proof (at arbitrary index m) with $\hat{u} = u^\dagger$, becomes

$$\begin{aligned} J(u^m) &\leq J(u^\dagger) - \sum_{n=1}^{m} \langle K^*(f^\delta - Ku^n), \hat{u} - u^m \rangle \\ &= J(\hat{u}) - \sum_{n=1}^{m} \langle f^\delta - Ku^n, f - Ku^m \rangle. \end{aligned}$$

Using Young's inequality and monotonicity of the residual ($\|Ku^m - f^\delta\| \leq \|Ku^n - f^\delta\|$), we conclude that

$$J(u^m) \leq J(u^\dagger) + \frac{3}{2} \sum_{n=1}^{m} \|Ku^n - f^\delta\|^2 + m\delta.$$

Summing the estimate in the proof of the Fejér monotonicity, we further find

$$\sum_{n=1}^{m} \|Ku^n - f^\delta\|^2 \leq D_J^{p^0}(u^\dagger, u^0) = J(u^\dagger).$$

Thus, we find

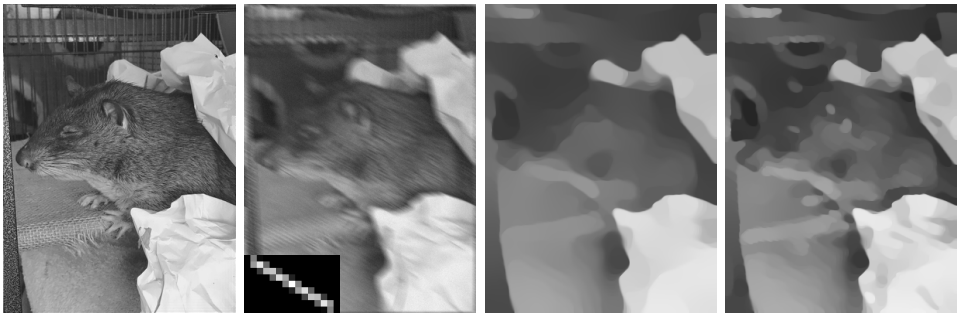
$$J(u^{k^*}) \leq \frac{5}{2} J(u^\dagger) + k^* \delta.$$

Since for the discrepancy principle one can show that $k^* \delta^2 \to 0$ in the limit $\delta \to 0$ (Osher *et al.* 2005), the limit of the regularization has a functional value J bounded by $\frac{5}{2} J(u^\dagger)$. Using a more careful argument based on Young's inequality, this upper bound can be decreased to $2J(u^\dagger)$ but not to $J(u^\dagger)$. On the other hand this might be advantageous, since an estimate of $J(u^{k^*})$ smaller than $J(u^\dagger)$ might mean a bias depending on J, since the value of the regularization functional is actually underestimated. In the case of total variation, for example, this means that the contrast is underestimated by variational methods, which is improved by iterative regularization (Osher *et al.* 2005). To conclude this section, we show numerical results of Bregman-iterative regularization in the context of deconvolution, which demonstrates the effect of total variation regularization.

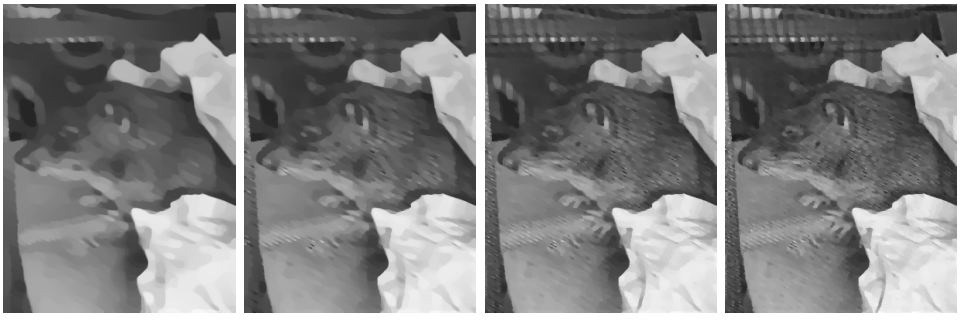
Example 6.8. We consider the inverse problem of the convolution operation, that is, $Ku = f$ with

$$(Ku)(y) := \int_{\mathbb{R}^2} u(x) h(x - y) \, \mathrm{d}x, \tag{6.7}$$

which is therefore also known as *deconvolution*. Here, h denotes the convolution kernel, that we assume to be known *a priori*. Since we cannot expect to know f but just f^δ with $F(f, f^\delta) \leq \delta$, we need to approximate the inverse



(a) original $u^\dagger$ (b) blurred, noisy f^δ (c) iterate 1 (d) iterate 3



(e) iterate 6 (f) iterate 20 (g) iterate 55 (h) iterate 96

Figure 6.1. (a) Image $u^\dagger \in \mathbb{R}^{400\times300}$ of Pixel, a Gambian pouched rat. (b) Degraded and noisy version $f^\delta \in \mathbb{R}^{400\times300}$ of the original image $u^\dagger$. The degradation stems from a discretized version of the convolution (see (6.7)) with periodic boundary conditions and the convolution kernel depicted in the bottom left corner of (b). (c–h) Different iterates of Algorithm 1 for $F(Ku, f^\delta) = \frac{1}{2}\|Ku - f^\delta\|^2_{L^2(\mathbb{R}^2)}$, $J(u, \alpha) = \alpha\,\mathrm{TV}(u)$ and $\alpha = 1/4$. The 96th iterate visualized in (h) is the first that violates Definition 6.1, for $\eta = 1$ and $\delta = 5.95$.

problem solution through regularization. In Figure 6.1 we can see selected iterates of Algorithm 1 for a single parameter $\alpha = 1/4$, the data fidelity term

$$F(Ku, f^\delta) = \frac{1}{2}\|Ku - f^\delta\|^2_{L^2(\mathbb{R}^2)},$$

and the regularization functional $J(u, \alpha) = \alpha\,\mathrm{TV}(u)$. The data $f^\delta = f + n$ are the sum of f, created via a discretized version of the exact forward model (6.7), and noise $n \in \mathcal{N}(0, 0.05)$. For the particular example used here, the fidelity-noise-bound is $F(f, f^\delta) = 5.95$. The inner variational regularization method is solved via the primal–dual hybrid gradient method (PDHGM): see Zhu and Chan (2008), Pock, Cremers, Bischof and Chambolle (2009), Esser, Zhang and Chan (2010) and Chambolle and Pock (2011, 2016). We clearly observe the inverse scale-space nature of the Bregman iteration. The

first iterate only contains features at a very coarse scale, and then ever more features at finer and finer scales are introduced throughout the course of the iteration.

Debiasing generalized eigenfunctions

We want to continue the analysis of the generalized eigenvalue problem introduced in Section 5.3. We have figured out that there is always a systematic bias of variational regularization methods for

$$F(Ku, f^\delta) = \frac{1}{2}\|Ku - f^\delta\|^2_{L^2(\Sigma)},$$

i.e. $\lambda < 1$ in (5.11) for $f^\delta = v_\sigma$. Benning (2011) showed that this systematic bias can be corrected with the help of Bregman iterations in the case of scalar $\boldsymbol{\alpha} = \alpha$ and J with $\partial J(cu, \alpha) = \alpha \partial J(u)$. Assume that α is chosen such that $u^k = 0$ for all $k < k^* - 1$, and $u^{k^*-1} = (1-\alpha)\sigma^{-1}u_\sigma$ for some $k^* \in \mathbb{N}$. Then we can easily conclude from (5.4) and (6.4) that u^{k^*} has to satisfy

$$\frac{1}{\alpha}\left(\lambda - \frac{1-\alpha}{\sigma}\right)K^*Ku^{k^*} \in \partial J(\lambda u^{k^*}).$$

We easily calculate that the above equation simplifies to the singular vector condition (5.14) for the choice $\lambda = 1/\sigma$. Consequently, $u^{k^*} = R(f^\delta, \alpha) = u_\sigma/\sigma$, and we have corrected for the bias of the previous iterate.

These computations demonstrate that Bregman iterations correct for the systematic bias of variational regularization reconstructions of generalized singular vectors for one-homogeneous regularization functionals J. However, the phenomenon is not limited to singular vectors. The following numerical toy example shows that the average reconstruction bias can be significantly reduced with the help of Bregman iterations. Assume the following set-up. Our forward model $K \in \mathbb{R}^{m\times n}$, for $m = 128$ and $n = 512$, is a matrix with its entries drawn randomly from $\mathcal{N}(0,1)$. We define a sparse vector $u^\dagger \in \mathbb{R}^n$ with nine non-zero entries, drawn randomly from $\mathcal{N}(0,1)$, and set $f = Ku^\dagger$. Subsequently, we create one hundred instances of noisy data via $f_j^\delta := f + n_j$, for $n \in \mathcal{N}(0, 0.5)$ and $j \in \{1, \dots, 100\}$. We now compute reconstructions for each of the one hundred instances with the following two regularization methods:

$$R_{\text{Morozov}}(f_j^\delta, \delta_j) = \operatorname*{arg\,min}_{u\in\mathbb{R}^n}\{\|u\|_1 \quad \text{subject to } \|Ku - f_j^\delta\|_2 \le \delta_j\}, \tag{6.8}$$

and

$$\begin{aligned} &R_{\text{Bregman}}(f_j^\delta, \{u_j^n\}_{n=1}^{k-1}, \alpha) \\ &\quad = \operatorname*{arg\,min}_{u\in\mathbb{R}^n}\left\{\frac{1}{2}\left\|Ku - \left(kf^\delta - \sum_{n=1}^{k-1} Ku_j^n\right)\right\|_2^2 + \alpha\|u\|_1\right\}, \end{aligned} \tag{6.9}$$

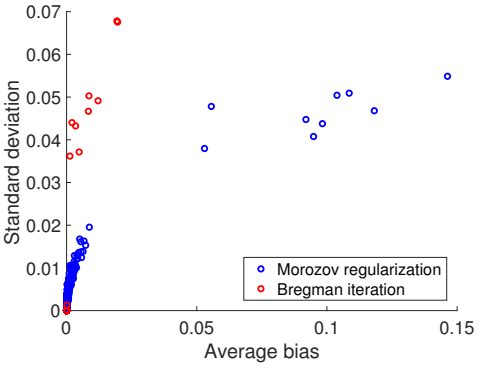


Figure 6.2. Equation (6.10) for the compressed sensing toy example in Section 6.1. The blue circles represent the standard deviation and average absolute bias values for all coefficients recovered with (6.8). The red circles show the same but for quantities for all coefficients recovered with (6.9). It becomes evident that for this example the average bias is significantly reduced, while the standard deviation of the reconstructed coefficients is comparable.

for

$$\begin{aligned} u_j^{\text{Morozov}} &\in R_{\text{Morozov}}(f_j^\delta, \delta_j), \\ \delta_j &:= \frac{1}{2}\|Ku^\dagger - f_j^\delta\|_2^2, \\ u_j^{\text{Bregman}} &\in R_{\text{Bregman}}(f_j^\delta, \{u_j^n\}_{n=1}^{k^*-1}, \alpha), \end{aligned}$$

and k^* chosen according to Definition 6.1 for $\eta = 1$ and $\delta = \delta_j$. We then compute the average absolute bias and the standard deviation of the reconstructions, that is, we compute

$$\left| u^\dagger - \frac{1}{100}\sum_{j=1}^{100} \hat{u}_j \right| \quad \text{and} \quad \sqrt{\frac{1}{99}\sum_{j=1}^{100}\left(\hat{u}_j - \frac{1}{100}\sum_{j=1}^{100}\hat{u}_j\right)^2} \tag{6.10}$$

for $\hat{u}_j \in \{u_j^{\text{Morozov}}, u_j^{\text{Bregman}}\}$. Both average absolute bias and standard deviation are visualized for each of the $n = 512$ coefficients in Figure 6.2. We clearly observe that with similar standard deviation, the average absolute bias is significantly reduced by the Bregman iteration in comparison to the Morozov regularization model.

6.2. Linearized Bregman iteration

As the name suggests, the linearized Bregman iteration can be derived from Algorithm 1 by replacing the term $F(Ku^k, f^\delta)$ with its linearization

$$F(Ku^k, f^\delta) \approx F(Ku^{k-1}, f^\delta) + \langle \partial_x F(Ku^{k-1}, f^\delta), Ku^k - Ku^{k-1} \rangle.$$

Hence, if we replace $F(Ku^k, f^\delta)$ in Algorithm 1 with this linearization multiplied by some constant $\tau > 0$, we obtain

$$\begin{aligned}
&R_I(f^\delta, v^{k-1}, \boldsymbol{\alpha}) \\
&\quad = \arg\min_{u\in\mathcal{U}} \big\{ \tau\big(F(Ku^{k-1}, f^\delta) + \langle \partial_x F(Ku^{k-1}, f^\delta), Ku^k - Ku^{k-1}\rangle\big) \\
&\qquad\quad + D^{p^{k-1}}_{J(\cdot,\alpha)}(u, u^{k-1})\big\}, \\
&\quad = \arg\min_{u\in\mathcal{U}} \big\{ \tau\langle \partial_x F(Ku^{k-1}, f^\delta), Ku^k - Ku^{k-1}\rangle + D^{p^{k-1}}_{J(\cdot,\alpha)}(u, u^{k-1})\big\}, \\
&u^k \in R_I(f^\delta, v^{k-1}, \boldsymbol{\alpha}) \\
&p^k = p^{k-1} - \tau K^* \partial_x F(Ku^{k-1}, f^\delta),
\end{aligned}$$

for $\boldsymbol{\alpha} = (\tau, \alpha)$ and $v^{k-1} := (u^k, p^k)$ for all $k \in \mathbb{N}$. These equations are summarized in Algorithm 2.

The linearized Bregman iteration is a generalization of the Landweber regularization (Landweber 1951) for the choices

$$F(Ku, f^\delta) = \frac{1}{2}\|Ku - f^\delta\|^2_{L^2(\Sigma)} \quad \text{and} \quad J(u, \alpha) = \frac{\alpha}{2}\|u\|^2_{L^2(\Omega)},$$

for some signal domains Ω and Σ. It is also a generalization of the mirror descent algorithm proposed in Nemirovskii and Yudin (1983), where $J(\cdot, \alpha)$ is a Legendre functional in the sense of Definition 5.11. This connection for convex, differentiable F and strongly convex and differentiable $J(\cdot, \alpha)$ was made in Beck and Teboulle (2003). The extension to subdifferentiable convex $J(\cdot, \alpha)$ was first proposed in Darbon and Osher (2007) and has since been studied extensively (Yin, Osher, Goldfarb and Darbon 2008, Cai, Osher and Shen 2009*b*, Cai, Osher and Shen 2009*a*, Yin 2010).

Algorithm 2 Linearized Bregman iteration

Initialize $u^0 \in \mathcal{U}$, p^0 with $p^0 \in \partial J(u^0, \alpha)$, $r^0 = K^* \partial_x F(Ku^0, f^\delta)$, $\boldsymbol{\alpha} = (\alpha, \tau) \in A$

while stopping criterion is not satisfied **do**

Compute $R_I(f^\delta, v^{k-1}, \boldsymbol{\alpha}) = \arg\min_{u\in\mathcal{U}} \big\{ \tau\langle r^{k-1}, u\rangle + D^{p^{k-1}}_{J(\cdot,\alpha)}(u, u^{k-1})\big\}$

Pick $u^k \in R_I(f^\delta, v^{k-1}, \boldsymbol{\alpha})$

Update $p^k = p^{k-1} - \tau\, r^{k-1}$

Compute $r^k = K^* \partial_x F(Ku^k, f^\delta)$

Set $v^k = (u^k, p^k)$

end while

return u^{k^*}, p^{k^*}

As in Remark 6.2, we can rewrite the dual update of the linearized Bregman iteration as

$$p^k = p^0 - \sum_{n=0}^{k-1} K^* \partial_x F(Ku^n, f^\delta), \tag{6.11}$$

and the primal update as

$$\begin{aligned} &R_I(f^\delta, \{u^n\}_{n=0}^{k-1}, p^0, \alpha) \\ &\quad = \arg\min_{u \in \mathcal{U}} \left\{ J(u, \alpha) - \left\langle p^0 - \sum_{n=0}^{k-1} K^* \partial_x F(Ku^n, f^\delta), u \right\rangle \right\}. \end{aligned} \tag{6.12}$$

In order to carry out a convergence analysis similar to the analysis for the standard Bregman iteration, we define the surrogate functional

$$J_\tau(u, \alpha) := J(u, \alpha) - \tau F(Ku, f^\delta). \tag{6.13}$$

We further assume for the remainder of this section that J and τ are chosen such that J_τ is convex. In practice, this requires strong convexity properties of J, which can simply be established by adding a sufficiently strongly convex functional to the original choice of J.

Example 6.9. Let

$$K \in \mathcal{L}(L^2(\Omega), L^2(\Sigma)), \quad F(Ku, f^\delta) = \frac{1}{2}\|Ku - f^\delta\|^2_{L^2(\Sigma)},$$

for domains $\Omega \subset \mathbb{R}^n$ and $\Sigma \subset \mathbb{R}^m$, and let J_1 be a proper, lower semicontinuous and convex functional. Then the functional

$$J_\tau(u, \alpha) := J(u, \alpha) - \frac{\tau}{2}\|Ku - f^\delta\|^2_{L^2(\Sigma)}$$

is convex for the choices

$$J(u, \alpha) := \frac{1}{2}\|u\|^2_{L^2(\Omega)} + J_1(u) \quad \text{and} \quad \tau < \frac{1}{\|K\|^2_{\mathcal{L}(L^2(\Omega), L^2(\Sigma))}}.$$

As for the Bregman iteration analysis, we start with a statement about the monotonic decrease of data fidelity.

Corollary 6.10 (monotonic decrease of data fidelity). Suppose that u^0 satisfies $F(Ku^0, f^\delta) < \infty$. Then the iterates of Algorithm 2 satisfy

$$F(Ku^{k+1}, f^\delta) + \frac{1}{\tau} D^{q^k}_{J_\tau(\cdot,\alpha)}(u^{k+1}, u^k) \leq F(Ku^{k+1}, f^\delta) \tag{6.14}$$

and

$$\lim_{k \to \infty} D^{q^k}_{J_\tau(\cdot,\alpha)}(u^{k+1}, u^k) = 0,$$

for $u^k \in R(f^\delta, v^{k-1}, \boldsymbol{\alpha})$ and $q^k \in \partial J_\tau(u^k, \alpha)$.

Proof. First of all we emphasize that

$$\langle r^{k-1}, u^k - u^{k-1}\rangle = \langle K^*\partial_x F(Ku^{k-1}, f^\delta), u^k - u^{k-1}\rangle$$

can be written as

$$\begin{aligned}&\langle K^*\partial_x F(Ku^{k-1}, f^\delta), u^k - u^{k-1}\rangle\\ &\quad = F(Ku^k, f^\delta) - F(Ku^{k-1}, f^\delta) - D_{F(K\cdot, f^\delta)}(u^k, u^{k-1}),\end{aligned}$$

for all $k \in \mathbb{N}$. Hence, the (primal) update of the linearized Bregman iteration can be rewritten as

$$\begin{aligned}&R_I(f^\delta, v^{k-1}, \boldsymbol{\alpha})\\ &\quad = \arg\min_{u\in\mathcal{U}}\big\{\tau\big(F(Ku^k, f^\delta) - F(Ku^{k-1}, f^\delta)\big) + D^{q^{k-1}}_{J_\tau(\cdot,\alpha)}(u^k, u^{k-1})\big\},\end{aligned}$$

for

$$q^{k-1} = p^{k-1} - \tau K^*\partial_x F(Ku^{k-1}, f^\delta) \in \partial J_\tau(u^{k-1}, \alpha), \quad p^{k-1} \in \partial J(u^{k-1}, \alpha).$$

Hence, we conclude

$$\begin{aligned}&\tau\big(F(Ku^k, f^\delta) - F(Ku^{k-1}, f^\delta)\big) + D^{q^{k-1}}_{J_\tau(\cdot,\alpha)}(u^k, u^{k-1})\\ &\quad \le \underbrace{\tau\big(F(Ku^{k-1}, f^\delta) - F(Ku^{k-1}, f^\delta)\big)}_{=0} + \underbrace{D^{q^{k-1}}_{J_\tau(\cdot,\alpha)}(u^{k-1}, u^{k-1})}_{=0},\end{aligned}$$

and thus equation (6.14). In the same fashion as in the proof of Corollary 6.3, we further conclude $\lim_{k\to\infty} D^{q^k}_{J_\tau(\cdot,\alpha)}(u^{k+1}, u^k) = 0$. □

As in the case of the standard Bregman iteration, the linearized Bregman iteration also satisfies Fejér monotonicity when the discrepancy principle is not violated.

Lemma 6.11 (Fejér monotonicity of Algorithm 2). Let $f \in \mathcal{R}_F(K)$, $u^\dagger \in \mathcal{S}(f, \alpha)$ and let $f^\delta \in \mathcal{V}$ with $F(f, f^\delta) \le \delta$. We further assume that the iterates of Algorithm 2 satisfy Definition 6.1 for $\eta = 1$. Then the iterates also satisfy the strict Fejér monotonicity

$$D^{q^k}_{J_\tau(\cdot,\alpha)}(u^\dagger, u^k) < D^{q^{k-1}}_{J_\tau(\cdot,\alpha)}(u^\dagger, u^{k-1}),$$

for all $u^k \in R(f^\delta, v^{k-1}, \boldsymbol{\alpha})$ and $q^k \in \partial J_\tau(u^k, \alpha)$, for all $k \le k^*$.

Proof. Through straightforward computations we obtain

$$\begin{aligned}&D^{q^k}_{J_\tau(\cdot,\alpha)}(u^\dagger, u^k) - D^{q^{k-1}}_{J_\tau(\cdot,\alpha)}(u^\dagger, u^{k-1})\\ &\quad = \underbrace{-D^{q^{k-1}}_{J_\tau(\cdot,\alpha)}(u^k, u^{k-1})}_{<0} - \langle q^k - q^{k-1}, u^\dagger - u^k\rangle\end{aligned}$$

$$\begin{aligned}
&\leq \langle p^{k-1} - p^k + \tau K^*(\partial_x F(Ku^k, f^\delta) - \partial_x F(Ku^{k-1}, f^\delta)), u^\dagger - u^k \rangle \\
&= \tau \langle K^* \partial_x F(Ku^k, f^\delta), u^\dagger - u^k \rangle \\
&\leq \tau(\delta - F(Ku^k, f^\delta)) \\
&< 0
\end{aligned}$$

for $k \leq k^*$, where we have made use of the convexity of F in its first argument, and $F(Ku^\dagger, f^\delta) \leq \delta$. □

In analogy to Corollary 6.5, we can show the same result for the linearized Bregman iteration.

Corollary 6.12. Let $f \in \mathcal{R}_F(K)$ and $u^\dagger \in \mathcal{S}(f, \alpha)$. Then the iterates of Algorithm 2 satisfy (6.5), for $\delta = 0$ (and thus $f^\delta = f$) and u^0 (with $q^0 \in \partial J_\tau(u^0, \alpha)$) chosen such that $D^{q^0}_{J_\tau(\cdot,\alpha)}(u^\dagger, u^0) < \infty$.

Proof. The proof follows exactly the same steps as the proof of Corollary 6.5. □

As in the case of Bregman iteration, Remark 6.6 follows from this result.

The following result guarantees convergence to a solution in $\mathcal{S}(f, \alpha)$ when $Ku^{k^*} = f$ is satisfied after a finite number k^* of iterations of Algorithm 2.

Lemma 6.13. Suppose that after a finite number of iterations the k^*th iterate of Algorithm 2 satisfies $Ku^{k^*} = f$, for $u^{k^*} = R(f, \boldsymbol{\alpha})$, $f \in \mathcal{R}_K(F)$ and $p^0 \in \mathcal{R}(K^*)$. Then $u^{k^*} \in \mathcal{S}(f, \alpha)$.

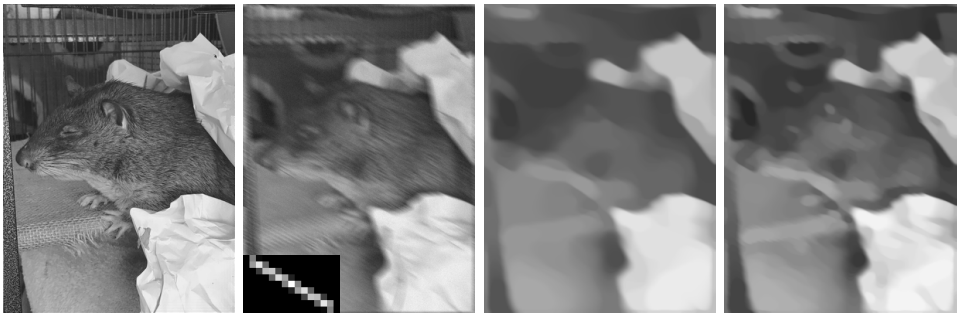
Proof. The proof is almost identical to the proof of Lemma 6.7; the only difference is that we use (6.11) instead of (6.3). □

Remark 6.14. Note that the statements of Lemmas 6.7 and 6.13 look identical, but one needs to remember that the underlying functionals J will most likely not be. This is due to the fact that for the linearized Bregman iteration additional terms have to be added in order to also make J_τ convex.

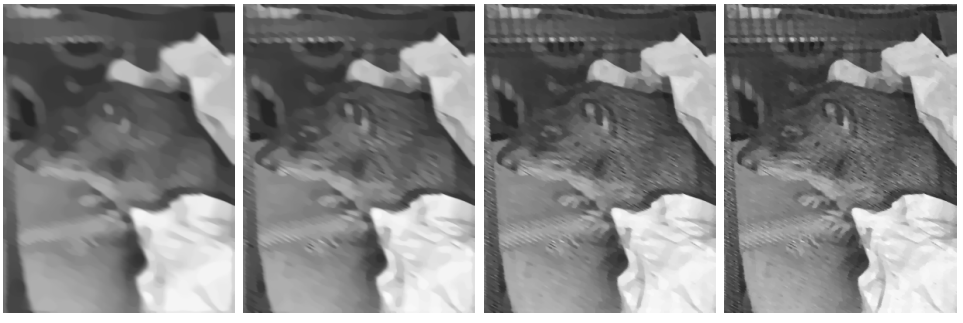
We conclude this section with numerical results for the same deconvolution example introduced in Example 6.8. We observe that with the same choice of regularization parameter and the same initialization, Algorithm 2 requires more iterations in order to converge to a solution that violates the discrepancy principle with the same error bound. See Figure 6.3. On the other hand, the variational subproblems are computationally cheaper to solve compared to the standard Bregman iteration case, at least with the (accelerated) PDHGM used for this example.

6.3. Coupled and modified Bregman iterations

The Bregman iteration (as well as its linearized variant) leave some freedom for modifications. One obvious modification is the choice of the subgradient



(a) original $u^\dagger$ (b) blurred, noisy f^δ (c) iterate 1 (d) iterate 3



(e) iterate 6 (f) iterate 20 (g) iterate 70 (h) iterate 128

Figure 6.3. (a) Image $u^\dagger \in \mathbb{R}^{400\times300}$ of Pixel the Gambian pouched rat, introduced in Figure 6.1(a). (b) The same degraded and noisy version $f^\delta \in \mathbb{R}^{400\times300}$ together with the convolution kernel h as shown in Figure 6.1(b). (c–h) Different iterates of Algorithm 2 for $F(Ku, f^\delta) = \frac{1}{2}\|Ku - f^\delta\|^2_{L^2(\mathbb{R}^2)}$, $J(u, \alpha) = \frac{1}{2}\|u\|^2_{L^2(\mathbb{R}^2)} + \alpha\,\mathrm{TV}(u)$ and $\alpha = 1/4$. The 128th iterate visualized in (h) is the first that violates Definition 6.1, for $\delta = 5.95$.

p^{k-1}. The update from the optimality condition is of course the obvious one and particularly suitable for a convergence proof. However, one may use different ways to determine a subgradient p^k from u^k. As an example one may solve some variational problem

$$p^k \in \arg\min_p \{H(p, p^{k-1}) \mid p \in \partial J(u^k, \alpha)\},$$

with a convex functional H. In the case of ℓ^1-minimization one might choose $H(p, p^{k-1}) = \|p\|_2$, which yields the minimal subgradient, that is, again choosing $\mathrm{sign}_0(p)$ (the single-valued version with $\mathrm{sign}_0(0) = 0$) in the case of a multivalued sign.

Another option for choosing subgradients has been investigated by Moeller *et al.* (2014), when one solves joint reconstruction problems for multiple unknowns $u_1, \ldots, u_M$. Moeller *et al.* (2014) proposed and analysed a coupled

Bregman iteration, which is based on choosing a new subgradient for the Bregman iteration in the ith image u_i from a linear combination of the subgradients in the other channels. In this way a joint subgradient for all the channels is approximated, which means a structural joint sparsity in the case of the ℓ^1-norm or joint edge information in the total variation case. Rasch *et al.* (2018) have investigated an infimal convolution version of the coupled Bregman iteration for an application to PET-MR imaging.

7. Bias and scales

Our earlier arguments relating to eigenfunctions demonstrate that bias and scale are closely related (at least when interpreting scale in terms of eigenfunctions and eigenvalues). The bias of variational regularization methods is larger on small-scale features. Thus, debiasing and multiscale aspects in regularization methods appear to be closely related, as has been worked out very recently. We discuss these ideas below.

7.1. Inverse scale space

For regularization functionals of the form $J(u, \boldsymbol{\alpha}) = \alpha J_1(u)$ we can write the dual Bregman iteration update as

$$\frac{p^k - p^{k-1}}{\Delta t} = -K^* \partial_x F(Ku^k, f^\delta)$$

for $\Delta t := 1/\alpha$ and $p^k \in \partial J_1(u^k)$, for all $k \in \mathbb{N}$. Thus, taking the limit $\alpha \to \infty$, so that $\Delta t \to 0$, yields the following time-continuous formulation of the Bregman iteration, also known as the *inverse scale space flow* (Burger, Osher, Xu and Gilboa 2005, Burger *et al.* 2006, Burger, Frick, Osher and Scherzer 2007*a*),

$$\partial_t p(t) = -K^* \partial_x F(Ku(t), f^\delta), \tag{7.1}$$

for $p(t) \in \partial J_1(u(t))$.

For many typical choices of regularization functionals J_1, it is difficult to numerically compute solutions of (7.1), with the ℓ^1-norm and in general any polyhedral regularization functional being the exception (Burger, Moeller, Benning and Osher 2013*c*, Moeller 2012, Moeller and Burger 2013). Nevertheless, (7.1) is very useful for studying theoretical properties of iterative regularizations in the limiting case.

Unsurprisingly, it is straightforward to carry out an eigenanalysis similar to the one discussed in Sections 5.3 and 6.1 for the regularization operator $R(f^\delta, t) = u(t)$ with $u(t)$ satisfying (7.1) for $F(Ku, f^\delta) = G(Ku - f^\delta)$. The following result is a generalization of Benning and Burger (2013, Theorem 9).

Theorem 7.1. Let (u_σ, v_σ) be a pair of generalized singular vectors with singular value σ, $f = v_\sigma$ and suppose J_1 is (absolutely) one-homogeneous, that is, $J_1(cu) = |c|J_1(u)$ for all $c \in \mathbb{R}$. Then $0 \in R(v_\sigma, t)$ for $0 \leq t < t_*$ and

$$\frac{1}{\sigma} u_\sigma \in R(v_\sigma, t)$$

for $t \geq t_* = 1$.

Proof. First we verify $0 \in R(v_\sigma, t)$ for $0 \in [0, t_*)$. From (5.14) and the absolute one-homogeneity of J_1, we observe $J_1(u_\sigma) = \langle G'(v_\sigma), Ku_\sigma \rangle$. We further see from the definition of the subdifferential that $t \leq 1 = \langle G'(v_\sigma), Ku_\sigma \rangle / J_1(u_\sigma)$ implies $p(t) := tK^*G'(v_\sigma) \in \partial J_1(0)$. Since $\partial_t p(t) = K^*G'(v_\sigma)$ and $p(0) = 0$, we have shown that $u(t) = 0$ is a solution of (7.1).

For $t \geq t_*$ a continuous extension of $p(t)$ is

$$p(t) = p(t_*) + (t_* - t)K^*G'(Ku(t) - v_\sigma).$$

We immediately see that $u(t) = u_\sigma/\sigma$ is a solution for $t \geq t_*$, since $p(t_*) = t_* K^*G'(v_\sigma) \in \partial J_1(u_\sigma/\sigma)$ and $\partial_t p(t) = 0$. □

Hence, the inverse scale space reconstruction also has no bias (for input data v_σ satisfying (5.14)), compared to the variational regularization method.

A similar result can be derived even in the case of noisy data $f^\delta = v_\sigma + n$, where n is an error term that satisfies the specific source condition

$$\mu K^*G'(v_\sigma) + \eta K^* n \in \partial J(\sigma u_\sigma),$$

for constants μ and η. For more details we refer to Benning and Burger (2013, Theorem 10).

In the following we briefly want to discuss reconstruction guarantees for linear combinations of multiple singular vectors. More precisely, we ask when we can guarantee

$$\frac{\gamma_j}{\sigma_j} u_{\sigma_j} \in R\left(\sum_{j=1}^{n} \gamma_j v_{\sigma_j}, t\right),$$

for coefficients $\{\gamma_j\}_{j \in \mathbb{N}}$. Due to the nonlinearity of J_1, in general there is no such decomposition. If we restrict ourselves to the following two conditions, however, such a result can be guaranteed (Schmidt, Benning and Schönlieb 2018, Theorem 3.14). The first condition is K-orthogonality of the singular vectors, that is,

$$\langle Ku_{\sigma_i}, Ku_{\sigma_j} \rangle = \begin{cases} 1 & i = j, \\ 0 & i \neq j, \end{cases} \tag{OC}$$

for $i, j \in \{1, \ldots, n\}$. The second condition is the so-called (SUB0)-condition, which reads as follows.

Definition 7.2 (Schmidt *et al.* 2018, Definition 3.1). Suppose that $(u_{\sigma_1}, u_{\sigma_2}, \dots, u_{\sigma_n})$ are an ordered set of primal singular vectors of J_1 with corresponding dual singular vectors $(v_{\sigma_1}, v_{\sigma_2}, \dots, v_{\sigma_n})$ and singular values $(\sigma_1, \sigma_2, \dots, \sigma_n)$. Then the singular vectors satisfy the (*SUB0*) *condition* if

$$\sum_{j=1}^{k} K^* G'(v_{\sigma_j}) \in \partial J_1(0), \tag{SUB0}$$

for all $k \in \{1, \dots, n\}$.

Given (OC) and (SUB0), we can guarantee the following decomposition result, which is a direct generalization of Schmidt *et al.* (2018, Theorem 3.14).

Theorem 7.3. Let $(u_{\sigma_1}, u_{\sigma_2}, \dots, u_{\sigma_n})$, $(v_{\sigma_1}, v_{\sigma_2}, \dots, v_{\sigma_n})$ be a system of ordered singular vectors, for singular values $(\sigma_1, \sigma_2, \dots, \sigma_n)$, for which the v_j are normalized and (OC) and (SUB0) are satisfied. Then, for data $f = \sum_{j=1}^{n} \gamma_j v_{\sigma_j}$ with positive coefficients $(\gamma_1, \dots, \gamma_n)$ we have $u(t) \in R(f, t)$, with

$$u(t) = \begin{cases} 0 & 0 \le t \le t_1, \\ \sum_{j=1}^{k} \gamma_j {\sigma_j}^{-1} u_{\sigma_j} & t_k \le t < t_{k+1}, \text{ for all } k = 1, \dots, n-1, \\ \sum_{j=1}^{n} \gamma_j {\sigma_j}^{-1} u_{\sigma_j} & t_n \le t, \end{cases}$$

where $t_k = \gamma_k$ and $t_k < t_{k+1}$ for all $k \in \{1, \dots, n\}$.

We refer to Benning and Burger (2013) for more information on individual generalized singular vectors and the inverse scale space flow. For more theoretical results and analytical as well as numerical examples of ordered sets of singular vectors that satisfy (OC) and (SUB0), we refer to Schmidt *et al.* (2018).

7.2. Two-step debiasing

While Bregman iterations and inverse scale space methods perform debiasing in an iterative fashion (and effectively change the variational model), one may also consider two-step procedures that first solve the original variational model and then perform a second step to reduce the bias (Deledalle, Papadakis and Salmon 2015, Deledalle, Papadakis, Salmon and Vaiter 2017). The first and simplest case where this idea was brought up is regularization with the ℓ^1-norm, where a so-called *refitting* strategy (Lederer 2013) is quite natural. After the variational problem

$$u_\delta^\alpha \in \arg\min F(Ku, f^\delta) + \alpha \|u\|_{\ell^1} \tag{7.2}$$

is solved, the second step simply consists in minimizing $F(Ku, f^\delta)$ over the set of all u sharing the support of u_α^δ. Since this procedure throws away

information about the sign of the entries of u, one can further improve to define the regularization operator via

$$R(f^\delta, \alpha) = \arg\min\{F(Ku, f^\delta) \mid \mathrm{sign}_0(u_i) = \mathrm{sign}_0((u_\delta^\alpha)_i), \text{ for all } i\}. \quad (7.3)$$

where $\mathrm{sign}_0(u_i)$ is the single-valued sign (*i.e.* zero for $u_i = 0$). Since the sign corresponds to a subgradient of the ℓ^1-norm, we can reinterpret the debiased regularization operator in a variational way: we minimize the fidelity subject to the constraint of u sharing a subgradient with u_δ^α. This is a key observation towards a generalization for arbitrary convex regularizations, as noted by Brinkmann *et al.* (2017). The general debiasing problem can be rephrased as a two-step procedure,

$$u_\delta^\alpha \in \arg\min_u F(Ku, f^\delta) + \alpha J(u), \quad (7.4)$$

followed by

$$R(f^\delta, \alpha) = \arg\min_u \{F(Ku, f^\delta) \mid p \in \partial J(u) \cap \partial J(u_\delta^\alpha)\}. \quad (7.5)$$

For computational purposes the arbitrary choice of the subgradient $p \in \partial J(u_\delta^\alpha)$ is not suitable, but we can indeed use the subgradient from the first step. Noting that for differentiable fidelities the optimality condition reads

$$p_\delta^\alpha = -\frac{1}{\alpha} K^* \partial F(Ku, f^\delta) \in \partial J(u_\delta^\alpha), \quad (7.6)$$

we can use the debiasing procedure

$$R(f^\delta, \alpha) = \arg\min_u \{F(Ku, f^\delta) \mid p_\delta^\alpha \in \partial J(u)\}. \quad (7.7)$$

The condition $p_\delta^\alpha \in \partial J(u)$ can be reformulated as a vanishing Bregman distance between u and u_α^δ; thus we observe some relations to the Bregman iteration. The second step can be interpreted as a Bregman iteration step in the limit of the regularization parameter to infinity. We refer to Brinkmann *et al.* (2017) for a detailed analysis of this debiasing approach.

The effect of the debiasing is illustrated for the simple case of total variation denoising, that is, the solution of

$$R(f^\delta, \alpha) = \arg\min_{u \in BV(\Omega)} \left(\frac{1}{2} \|u - f^\delta\|^2_{L^2(\Omega)} + \alpha |u|_{BV} \right). \quad (7.8)$$

Figure 7.1 compares the solution of the variational problem in (c) with the one obtained in the two-step debiasing procedure (d) and the Bregman iteration (e). Both methods reduce the contrast loss of the TV regularization (which is difficult to see in the image, but becomes more apparent in the small background buildings). Overall, however, the Bregman iteration seems to restore more of the small details such as the grass structure.

(a) original (b) noisy (c) TV (d) debiasing (e) Bregman

Figure 7.1. Cameraman (256×256): comparison of TV denoising for $\alpha = 0.1$, with the two-step debiasing, and Bregman iterations ($\alpha = 0.5$ and 7 Bregman iterations).

(a) $\alpha = 0.1$ (b) $\alpha = 0.2$ (c) $\alpha = 0.3$

Figure 7.2. Cameraman: TV denoising and debiasing for different values of the regularization parameter.

Figure 7.2 demonstrates the debiasing effect for increasing regularization parameters, where the variational model destroys ever more detail. In particular, for larger α one observes the effect of restoring smaller structures apparently contained in the subgradient but not the primal variable of the variational model.

7.3. Nonlinear spectral transform

The iterative regularization methods presented in Section 6 can easily be extended to nonlinear spectral decomposition methods via the following

trivial observation. Every iterate $u^k \in R_I(f^\delta, v^{k-1}, \boldsymbol{\alpha})$ can be represented as the sum of the differences of two subsequent iterates, that is,

$$u^k = u^0 + \sum_{n=1}^{k} u^n - u^{n-1}.$$

If we define $\varphi^0 := u^0$ and $\varphi^n := u^n - u^{n-1}$ for $n > 1$, and equip the sum with coefficients $\{c^n\}_{n=0}^k$, we can write u^k as

$$u^k = \sum_{n=0}^{k} c^n \varphi^n.$$

In the following we are going to motivate why such a decomposition is useful for localizing individual scales if the underlying regularization functional is (absolutely) one-homogeneous and where we have a scalar parameter α. Following up on the bias correction example for generalized singular vectors in Section 6.1, we know that for the Bregman iteration $R_I(f, v^{k-1}, \alpha)$ with $f = v_\sigma$ we observe

$$u^k = \begin{cases} 0 & k < k^*, \\ (k^* - \alpha)\sigma^{-1}\, u_\sigma & k = k^*, \\ \sigma^{-1}\, u_\sigma & k \geq k^* + 1. \end{cases}$$

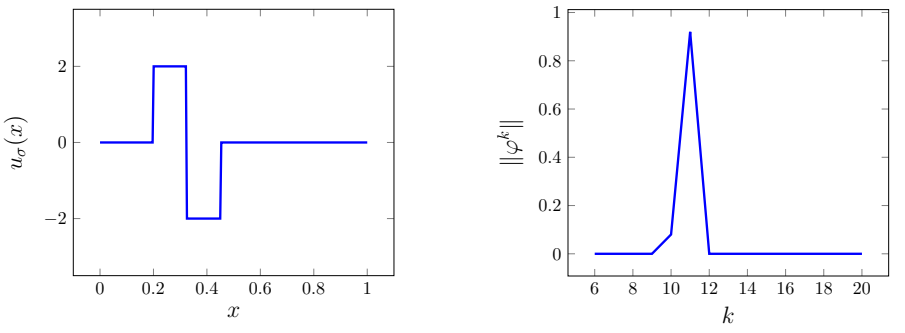
Replacing $f = v_\sigma$ with $f = \sigma v_\sigma = K u_\sigma$ therefore yields

$$u^k = \begin{cases} 0 & k < k^*, \\ (k^* - \alpha\sigma^{-1})\, u_\sigma & k = k^*, \\ u_\sigma & k \geq k^* + 1, \end{cases}$$

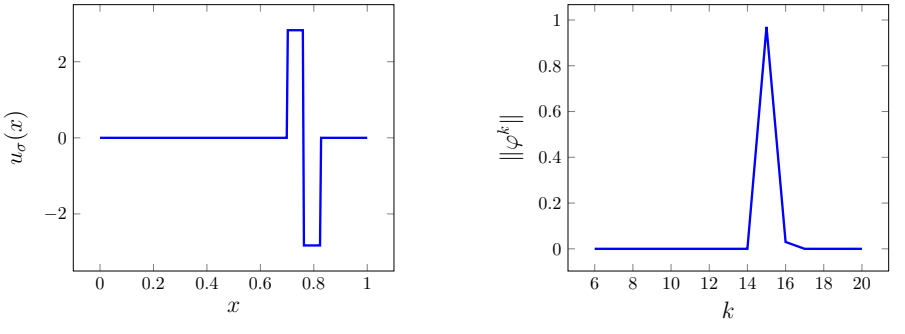
and consequently we observe

$$\varphi^n = \begin{cases} 0 & n \notin \{k^*, k^* + 1\}, \\ (k^* - \alpha\sigma^{-1})\, u_\sigma & n = k^*, \\ (1 + \alpha\sigma^{-1} - k^*)\, u_\sigma & n = k^* + 1. \end{cases}$$

The last equation implies that if the input datum is given in terms of the forward model applied to a (primal) singular vector, this primal singular vector is localized in only two components φ^{k^*} and φ^{k^*+1}. The index k^* depends on the choice of α and on the singular value σ. Hence, singular vectors with different scales, or different values of σ, will be localized in $\varphi^{\hat{k}}$ and $\varphi^{\hat{k}+1}$ for $\hat{k} \neq k^*$. This is visualized in Figures 7.3 and 7.4. It is therefore fair to call $\{\varphi^n\}_{n=1}^k$ a spectrum and the individual φ^n, for $n \in \{1, \ldots, k\}$, the spectral components.



(a) u_σ, $\sigma = 2^{-3}$ (b) $\|\varphi\|$



(c) u_σ, $\sigma = 2^{-7/2}$ (d) $\|\varphi\|$

Figure 7.3. Two singular vectors of $J = \mathrm{TV}_*$ with different σ-values, and excerpts of their corresponding (analytically computed) spectra, for $\alpha = 1.24$. We clearly observe that both vectors are located at different positions of the spectrum. Hence, both singular vectors could be isolated from a sum of the two by applying a band-pass filter to the spectrum.

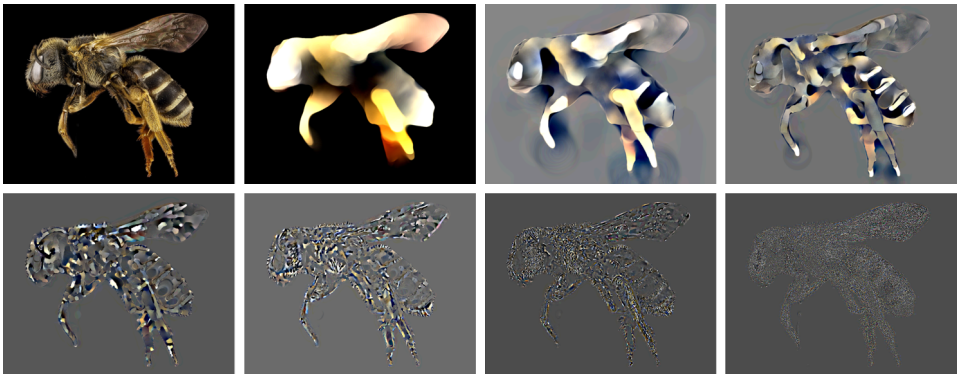


Figure 7.4. Spectral decomposition of the image of a bee. From Benning *et al.* (2017*d*).

Consequently, the operator $\mathcal{S} : \mathcal{V} \times \mathcal{U}^k \times \mathbb{R}^k \times A \to \mathcal{U}$ with

$$\mathcal{S}(f, (u_n)_{n=0}^k, (c_n)_{n=0}^k, \alpha) := \sum_{n=0}^{k} c_n \varphi^n \quad \text{with } \varphi^n := \begin{cases} u^n - u^{n-1} & n > 1, \\ u^0 & n = 1, \end{cases}$$

for $u^n \in R_I(f, v^{n-1}, \alpha)$ can be seen as a spectral transform of the input signal f^δ. For $K : \mathrm{BV}(\Omega) \to L^2(\Omega)$ this type of spectral transform is a discretization of the inverse-scale-space based spectral transform defined in Burger, Eckardt, Gilboa and Moeller (2015*a*) and Burger *et al.* (2016*a*). For $K : \mathrm{BV}(\Omega) \to L^2(\Omega)$ and $J(u, \alpha) = \alpha\,\mathrm{TV}(u)$, the idea of generalized spectral transforms goes back to Gilboa (2014*a*, 2014*b*). For a detailed overview of this form of nonlinear spectral transform we refer to Gilboa, Moeller and Burger (2016). Another interesting recent extension is the spectral transform in the context of image segmentation (Zeune *et al.* 2017).

8. Applications

Obviously modern regularization methods have found applications in all kind of inverse problems and pushed forward the state of the art. As examples, let us mention TV/TGV Bregman iterations for super-resolution (Marquina and Osher 2008), PET reconstruction (Müller *et al.* 2011, Müller 2013) or STED microscopy (Brune, Sawatzky and Burger 2009, Brune, Sawatzky and Burger 2011), as well as TGV reconstructions in MR (Knoll, Bredies, Pock and Stollberger 2011). Providing an overview of the various approaches for well-known imaging methods would far exceed the scope and size of this survey. Hence, in the following we provide some novel examples of applications, which are actually driven by advances in regularization techniques.

8.1. Velocity-encoded magnetic resonance imaging

Magnetic resonance imaging (MRI) is an imaging technique that allows us to visualize the chemical composition of humans/animals or materials. MRI scanners utilize strong magnetic fields and radio waves to excite subatomic particles, such as protons, that subsequently emit radio frequency signals which can be measured with the radio frequency coils that initially excited those radio waves: see for example Callaghan (1993). MRI is often used to measure contrast in tissue. However, due to shear, endless possibilities of radio-frequency pulse sequence design, and programming of the gradient coils, MRI is a versatile imaging tool with capabilities beyond imaging contrast in tissue. A potential, more sophisticated application is phase-encoded magnetic resonance velocity imaging, which in medical imaging is used to study the distribution and variation in blood flow (Gatehouse

et al. 2005). In the physical sciences it is used to study the rheology of complex fluids (Callaghan 1999), liquids and gases flowing through packed beds (Sederman, Johns, Alexander and Gladden 1998, Holland *et al.* 2010), granular flows (Holland *et al.* 2008) and multiphase turbulence flows (Tayler, Holland, Sederman and Gladden 2012). The main advantage of MRI over other methods when it comes to studying flow is that it is possible to image flows non-invasively. However, the main drawback of the technique is the acquisition time of the measurement.

Lustig, Donoho and Pauly (2007) exploited the idea of sub-sampling in the spatial data domain to overcome this limitation and to speed up the MRI acquisition process. Due to fewer measurements being taken, compared to the number of unknowns to be recovered, some form of regularization needs to be integrated into the reconstruction process. Sparsity-promoting variational regularization methods are suitable candidates and, most prominently, total variation regularization has been successfully deployed to increase the temporal resolution of MRI acquisitions. Since measurement noise in MRI data can be modelled as being normally distributed, a standard variational regularization approach is

$$R(f^\delta, \boldsymbol{\alpha}) = \operatorname*{arg\,min}_{u \in \mathcal{U}} \left\{ \frac{1}{2} \|\mathcal{F}u - f^\delta\|_2^2 + J(u, \boldsymbol{\alpha}) \right\}, \tag{8.1}$$

where $\mathcal{F}$ is the operator

$$(\mathcal{F}u)(t^k) := (2\pi)^{-n/2} \int_{\mathbb{R}^n} u(x) \exp\left(-\mathrm{i} \int_{t^{k-1}}^{t^k} x(t) \cdot g(t)\, \mathrm{d}t \right) \mathrm{d}x,$$

and $n \in \{2, 3\}$ denotes the dimension of the signal and $g : [0, T] \to \mathbb{R}^n$ represents the function that controls the gradient coils of the MRI machine. We observe that $\mathcal{F}$ is almost identical to the Fourier transform sampled at discrete locations, if we can approximate

$$\int_{t^{k-1}}^{t^k} x(t) \cdot g(t)\, \mathrm{d}t \approx x \cdot \int_{t^{k-1}}^{t^k} g(t)\, \mathrm{d}t.$$

This can be achieved by adequate programming of the gradient coils. However, $\int_{t^{k-1}}^{t^k} x(t) \cdot g(t)\, \mathrm{d}t$ can be approximated more generally via the Taylor series

$$\int_{t^{k-1}}^{t^k} x(t) \cdot g(t)\, \mathrm{d}t \approx \sum_{r=0}^{\infty} \frac{x^{(r)}(t^{k-1})}{r!} \cdot \int_{t^{k-1}}^{t^k} g(t) t^r\, \mathrm{d}t,$$

and with clever programming of g, other moments such as velocity or acceleration can be so encoded. In the following, we assume that the radio-frequency pulse sequence and the gradient coils are programmed such that

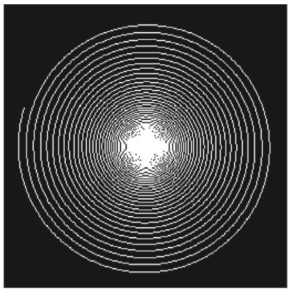
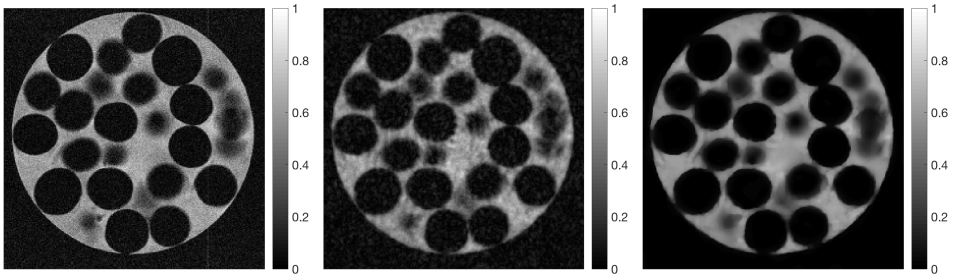
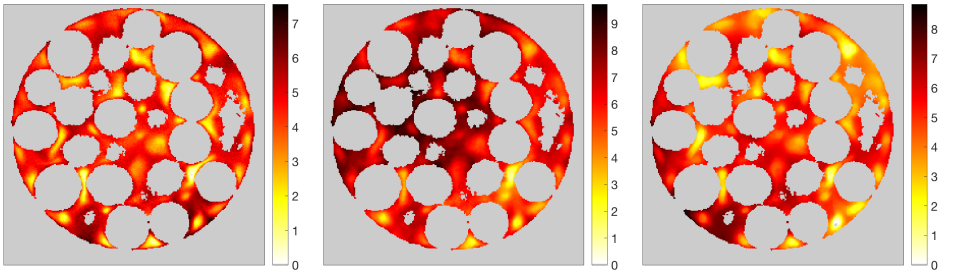


Figure 8.1. A simulated spiral on a Cartesian grid. From Benning *et al.* (2014).



(a) fully sampled u (b) zero-filled u (c) TGV-based u

Figure 8.2. Magnitude images of the velocity dataset used in Benning *et al.* (2014), courtesy of Andrew J. Sederman. (a) Magnitude image derived from applying the inverse of the Fourier transform to the fully sampled Fourier data of the velocity-encoded MRI measurement and subsequently taking the modulus. (b) Magnitude image obtained if we set to zero all Fourier samples that are not part of the spiral visualized in Figure 8.1, and subsequently proceed as for the fully sampled data. (c) Magnitude reconstructions from the TGV^2_β-based variational regularization reconstruction (8.1).



(a) fully sampled (b) zero-filled (c) TGV-based

Figure 8.3. The different velocity reconstructions corresponding to the magnitude reconstructions in Figure 8.2.

we first encode the velocity in the z-direction, that is, for $x = (x_1, x_2, x_3)$ and $g(t) = (g_1(t), g_2(t), g_3(t))$ we have

$$\int_0^{t_0} x(t) \cdot g(t)\,\mathrm{d}t \approx \underbrace{x_3'(0)}_{=:v_z} \int_0^{t_0} g_3(t)\,t\,\mathrm{d}t,$$

in the interval $[0, t_0]$, and then perform the spatial encoding such that

$$\int_{t_{k-1}}^{t^k} x(t) \cdot g(t)\,\mathrm{d}t \approx \begin{pmatrix} x_1(t_{k-1}) \\ x_2(t_{k-1}) \end{pmatrix} \cdot \int_{t^{k-1}}^{t^k} \begin{pmatrix} g_1(t) \\ g_2(t) \end{pmatrix} \mathrm{d}t$$

holds true for $t_0 < t_1 < \cdots < t_m = T$. Then, with $x = (x_1(t_{k-1}), x_2(t_{k-1}))$ and $g = (g_1, g_2)$ as an abuse of notation, $\mathcal{F}$ reads as

$$(\mathcal{F}(u, v_z))(t^k) = \frac{1}{2\pi} \int_{\mathbb{R}^2} u(x) \exp(-\mathrm{i}\sigma v_z(x)) \exp\left(-\mathrm{i}x \cdot \int_{t^{k-1}}^{t^k} g(t)\,\mathrm{d}t\right) \mathrm{d}x, \tag{8.2}$$

for some constant σ. In order to avoid non-linearity of the forward model, we couple u and v_z by simply defining $w := u \exp(-\mathrm{i}\sigma v_z)$. Then the forward model $\mathcal{F}$ simply reduces to the (sub-sampled) Fourier transform.

Benning *et al.* (2014) have investigated three choices for regularization functionals: assuming $\boldsymbol{\alpha} = (\alpha, \beta)$, we have

$$J(w, \alpha, \beta) = \alpha \begin{cases} \mathrm{TV}(w), \\ \mathrm{TGV}^2_\beta(w), \\ \sum_{j=1}^\infty |\langle w, \varphi_j \rangle|. \end{cases} \tag{8.3}$$

Here $\{\varphi_j\}_{j \in \mathbb{Z}}$ denotes a wavelet basis, and TGV^2_β is the second-order total generalized variation in the sense of Bredies and Valkonen (2011, Theorem 3.1). In Figure 8.2 we see computational solutions of (8.1) for the choice $J(w, \alpha, \beta) = \mathrm{TGV}^2_\beta(w)$, a spiral sub-sampling strategy on a Cartesian grid (see Figure 8.1), and the parameter choices $\alpha = 0.1$ and $\beta = 3$. These results have again been computed with the PDHGM. Subsequently, v_z was extracted as the principal value of the reconstruction $w \in R(f^\delta, \alpha, \beta)$. The reconstructed z-velocity v_z is subsequently unwrapped by solving the linear system

$$\Delta \hat{v}_z = \cos(v_z) \Delta \sin(v_z) - \sin(v_z) \Delta \cos(v_z)$$

for $\hat{v}_z$. Here Δ denotes the Laplace operator. The unwrapped reconstructed velocity $\hat{v}_z$ is visualized in Figure 8.3.

In order to demonstrate the capabilities of the Bregman iteration, Benning *et al.* (2014) have qualitatively analysed Algorithm 1 for different sub-sampling strategies and different initial choices of $\alpha > 0$. These comparisons for different sub-sampling strategies are visualized in Figure 8.4.

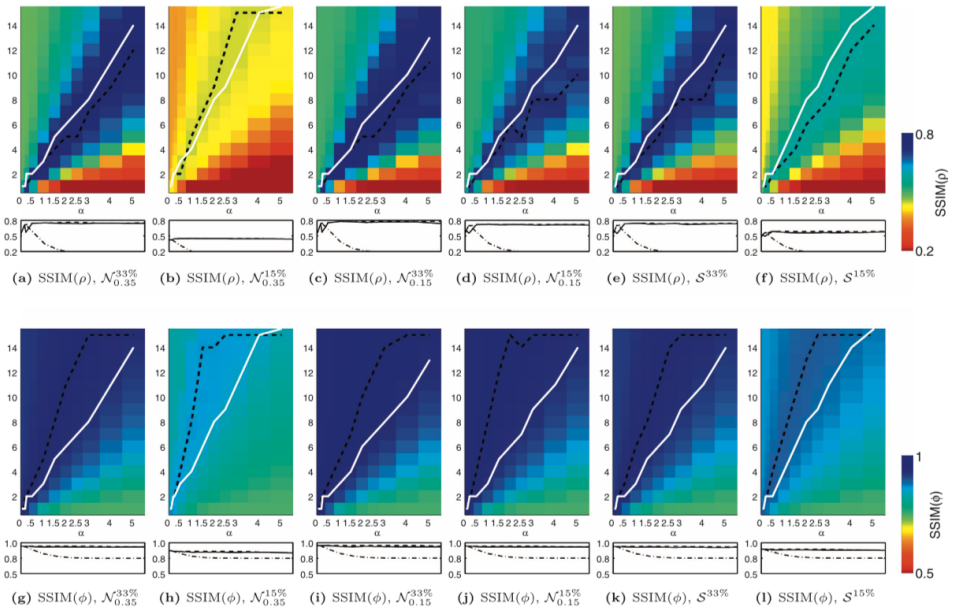


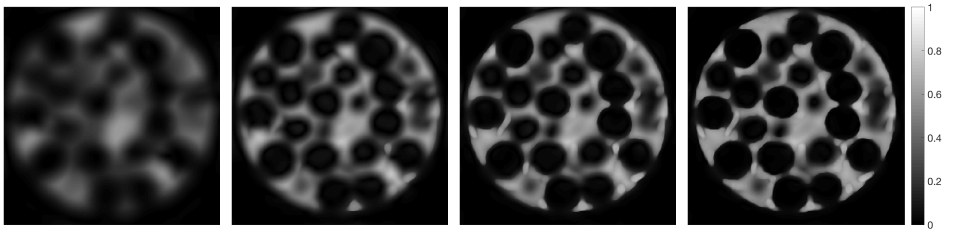
Figure 8.4. The structural similarity index measure (SSIM) (see Wang, Bovik, Sheikh and Simoncelli 2004) of the magnitude images (a–f) and the velocity images (g–l) for Bregmanized TV reconstructions of computer-generated test data with various sampling patterns and noise $\sigma = 0.2$. The parameter α is on the horizontal axis and the Bregman iteration is on the vertical axis. The colours code the SSIM value, also shown in the small lower graph. The continuous line corresponds to violation of the discrepancy principle, and the dashed line to the optimal SSIM. The dash-dotted line in the small graph indicates the SSIM for the first iteration. From Benning *et al.* (2014).

In Figure 8.5 we see the magnitude images of 20 Bregman iterations computed with Algorithm 1 for the same set-up as described earlier, and the parameter choices $\alpha = 1.5$ and $\beta = 3$.

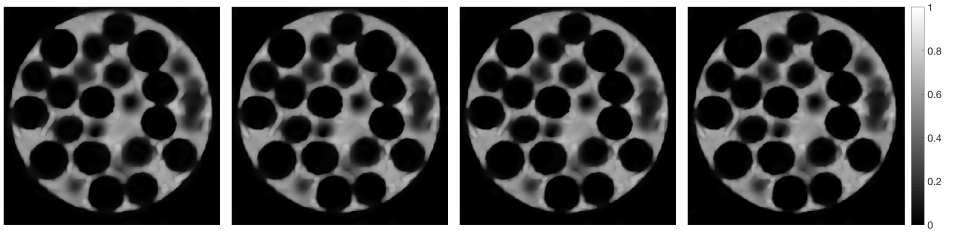
We refer to Benning *et al.* (2014) for more information on iterative regularization in the context of velocity-encoded MRI.

8.2. Dynamic MRI with structural prior

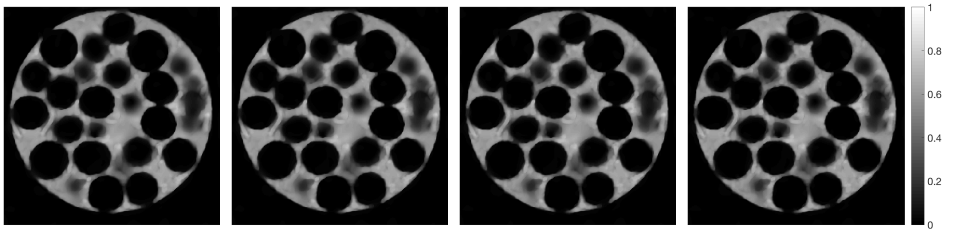
Dynamic MRI is a topic of high current relevance in biomedical imaging, with different techniques such as fMRI or DCE-MRI. The basic issue is to reconstruct a sequence of images $u = (u_1, \dots, u_T)$ from measurements $(K_1 u_1, \dots, K_T u_T)$, with K_t being a sub-sampled Fourier transform (with different sub-sampling at each time step). Due to the significant measurement times in MRI the sub-sampling is necessary to obtain a significant time



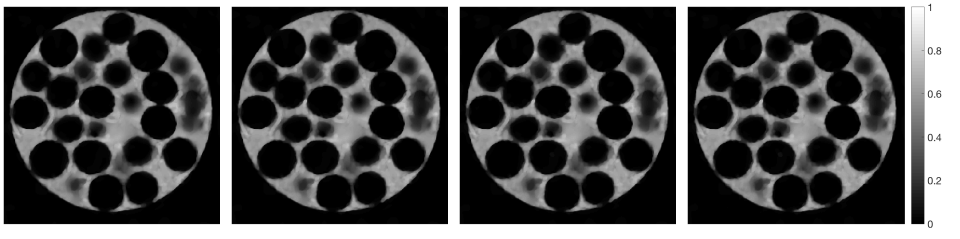
(a) iterate 1 (b) iterate 2 (c) iterate 3 (d) iterate 4



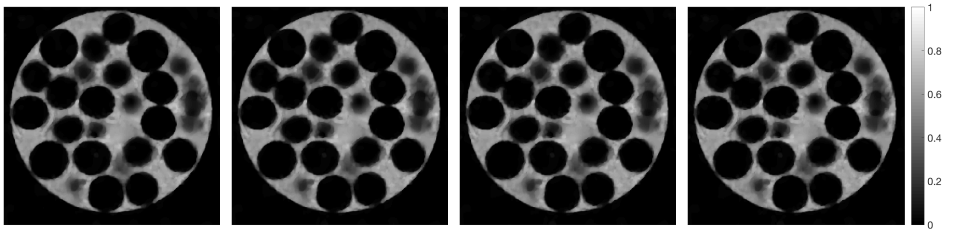
(e) iterate 5 (f) iterate 6 (g) iterate 7 (h) iterate 8



(i) iterate 9 (j) iterate 10 (k) iterate 11 (l) iterate 12



(m) iterate 13 (n) iterate 14 (o) iterate 15 (p) iterate 16



(q) iterate 17 (r) iterate 18 (s) iterate 19 (t) iterate 20

Figure 8.5. Magnitude images of 20 Bregman iterations computed via Algorithm 1, with $\alpha = 1.5$ and $\beta = 3$.

resolution; the time resolution will improve with stronger undersampling (*e.g.* in spokes). The natural data fidelity in this case is thus

$$F(Ku, f) = \frac{1}{2}\sum_{t=1}^{T} \|K_t u_t - f_t\|^2.$$

With substantial undersampling it is hopeless to reconstruct meaningful images from the data at a single time step, hence a regularization in time is needed in order to exploit correlations between close time steps. A natural assumption is smoothness in the time direction, for then a discrete gradient $\|u_{t+1} - u_t\|^2$ can be penalized in a regularization functional. Moreover, in order to take into account the edges it is natural to include some total variation regularization for each u_t. So far, this is an approach that can be used for many dynamic reconstruction problems. A particular feature of such MR investigations, however, is the existence of a structural prior u_0, which is a high-resolution MR image at different contrast (*e.g.* a standard anatomical T1 scan) taken before the start of the dynamic imaging. The prior is reconstructed from a very dense sampling and thus at very high resolution. The important step is to notice that most edges in the images u_t will arise from anatomical structures, and are thus present in u_0. Hence, an additional structural regularization like the infimal convolution of Bregman distances

$$ICBV^{p_0}(\cdot, u_0) = D_{TV}^{p_0}(\cdot, u_0) \square D_{TV}^{-p_0}(\cdot, -u_0)$$

can be used to achieve super-resolution in the dynamic imaging series.

The regularization functional

$$J(u) = \sum_{t=1}^{T} \omega_t |u_t|_{BV} + \sum_{t=1}^{T} (1 - \omega_t) ICBV^{p_0}(u, u_0) + \sum_{t=1}^{T-1} \frac{\gamma_t}{2} \|u_{t+1} - u_t\|^2$$

combining the three parts was proposed and investigated in Rasch *et al.* (2017). The results indicate enormous potential for obtaining reconstructions at high resolution from rather extreme undersamplings in time. These are illustrated in Figure 8.6 for several different time steps of a simulated data set. The first line shows the sampling at different time steps; the last column shows the prior image u_0 instead. The second line provides direct reconstruction without regularization (note that the Fourier transform is continuously invertible, so without undersampling the direct inversion is a standard technique). The third line displays the results with the proposed method, to be compared to the ground truth used for simulating data in the fourth line. These results were obtained on simulated MR data; we refer to Rasch *et al.* (2017) for a further study of real data.

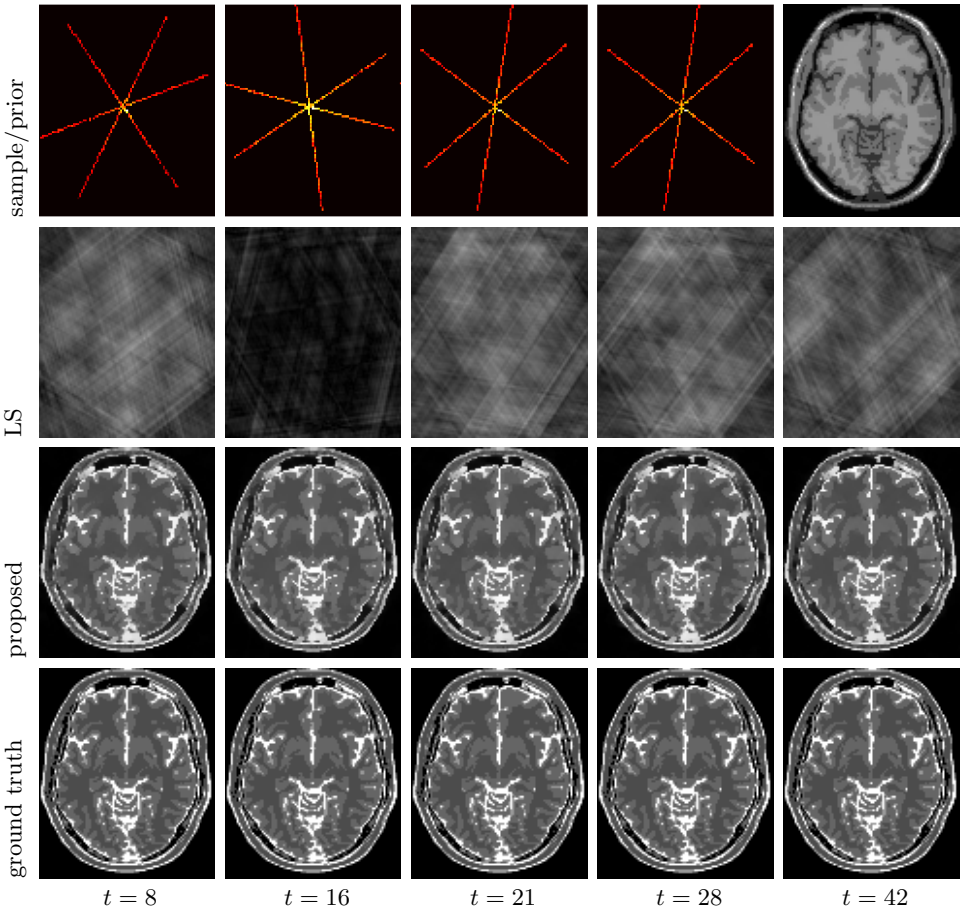
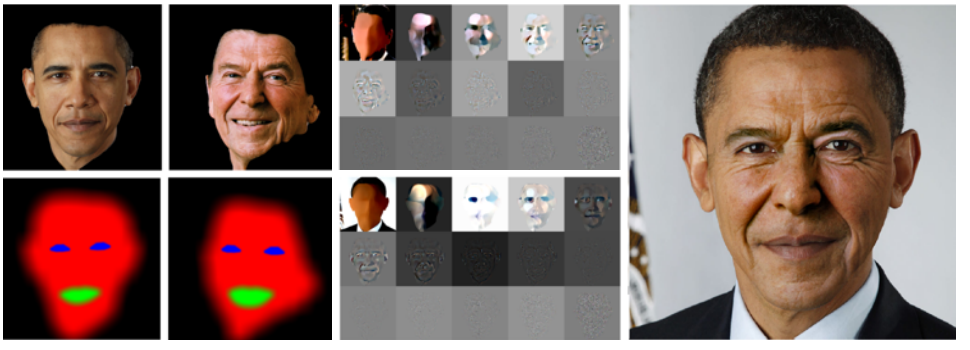


Figure 8.6. Results of undersampled dynamic MRI reconstruction with different methods at five different time steps.

8.3. Nonlinear spectral image fusion

The nonlinear spectral transform as introduced in Section 7.3 can be used to suppress, enhance or extract features of signals at different scales. Benning *et al.* (2017*d*) have used it to fuse features at different scales from two images into a single image, in order to create realistic-looking image fusions. The mathematical procedure is as follows. Given two images, both images are preprocessed such that they are aligned (registered) and that regions within the images are segmented such that the images are fused only in selected regions. Denoting the registered images as f_1 and f_2, they can be

(a) face detection (b) landmark detection (c) registration



(d) face segmentation (e) spectral decomposition (f) image fusion

Figure 8.7. Illustration of the pipeline for facial image fusion using nonlinear spectral decompositions. From Benning *et al.* (2017*d*).

represented via their spectral transforms, that is,

$$u_1 = \mathcal{S}(f_1, (u)_{n=0}^{k^*}, \boldsymbol{c}_1, \alpha) + \underbrace{f_1 - \mathcal{S}(f_1, (u)_{n=0}^{k^*}, \mathbf{1}, \alpha)}_{=:r_1^{\alpha,k^*}},$$

and

$$u_2 = \mathcal{S}(f_2, (u)_{n=0}^{k^*}, \boldsymbol{c}_2, \alpha) + \underbrace{f_2 - \mathcal{S}(f_2, (u)_{n=0}^{k^*}, \mathbf{1}, \alpha)}_{=:r_2^{\alpha,k^*}},$$

for $k^* \geq 1$, $\alpha \in A$ and coefficients $\boldsymbol{c}_1 \in \mathbb{R}^{k^*}$, $\boldsymbol{c}_2 \in \mathbb{R}^{k^*}$ and $\mathbf{1} \in \{1\}^{k^*}$ being the constant one-vector. Obviously we have $u_1 = f_1$ and $u_2 = f_2$ if $\boldsymbol{c}_1 = \mathbf{1}$ and $\boldsymbol{c}_2 = \mathbf{1}$.

In order to incorporate the face segmentation into the image fusion process, we allow the coefficient vectors $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$ to be spatially varying functions $\boldsymbol{c}_1 : \Omega \to \mathbb{R}^{k^*}$ and $\boldsymbol{c}_2 : \Omega \to \mathbb{R}^{k^*}$, respectively. Here Ω denotes the image domain. The image fusion process can then be mathematically described by

$$u_{\text{fused}} := \mathcal{S}(f_1, k^*, \alpha, \boldsymbol{c}_1) + \mathcal{S}(f_2, k^*, \alpha, \boldsymbol{c}_2) + r_1^{\alpha,k^*}.$$

The individual steps of the image fusion pipeline are visualized in Figure 8.7.

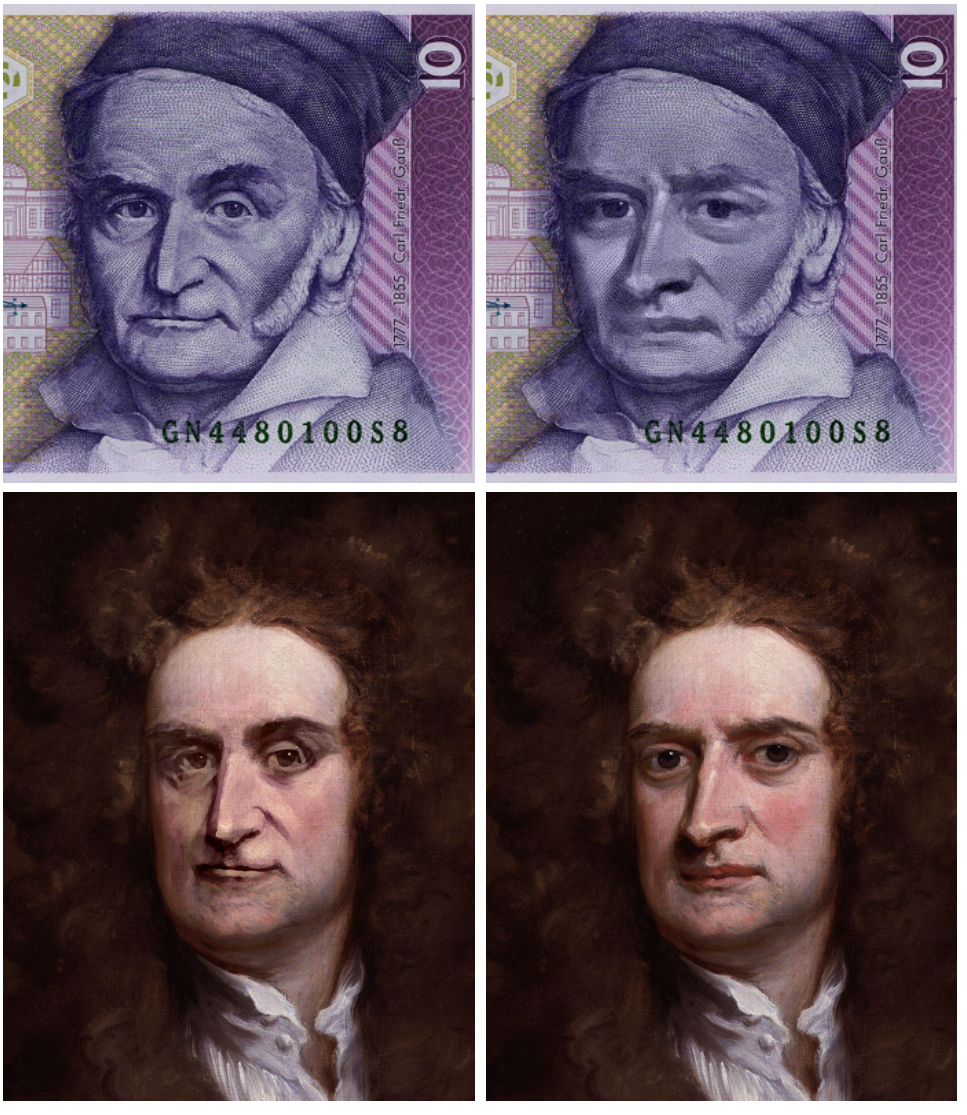


Figure 8.8. Image fusion using the nonlinear spectral TV decomposition for the challenging example of fusing a banknote with a picture of Gauss and a painting of Newton. From the supplementary material of Benning *et al.* (2017*d*).

For challenging examples this automation may very well fail. Nevertheless, the spectral image fusion still works if registration and segmentation are carried out manually, as can be seen in Figure 8.8. For more information on the nonlinear spectral image fusion we refer to Benning *et al.* (2017*d*).

9. Advanced issues

In the following we comment on some advanced issues related to iterative variational methods extending those presented above, namely extensions to non-convex problems, in particular with respect to data fidelities in nonlinear inverse problems, and to modern machine learning approaches.

9.1. Non-convex optimization

In the context of inverse problems one usually deals with data fidelities of the form $F(Ku, f^\delta)$ that measure the deviation between Ku and f^δ in some sense. So far we have always assumed this particular structure, and also that F is convex. Both assumptions can be relaxed. In the following we assume that we simply have some non-convex energy functional $E : \mathcal{U} \to \mathbb{R}$ that is Fréchet-differentiable with gradient ∇E. As there might not exist critical points, or finding them might be unstable due to ill-posedness, it makes sense to generalize (5.3) to

$$R(\boldsymbol{\alpha}) = \operatorname*{arg\,min}_{u \in \mathcal{U}} \{E(u) + J(u, \boldsymbol{\alpha})\}. \tag{9.1}$$

Here we want to emphasize that $R(\boldsymbol{\alpha})$ is not necessarily a regularization operator in the classical sense, as in general we do not make the dependence on some data f^δ explicit. In Section 9.2 we particularly investigate the case in which E is of the form $E(\cdot) = F(K(\cdot), f^\delta)$, where K stems from a nonlinear inverse problem, and where $F(\cdot, f^\delta)$ is potentially non-convex in its first argument.

It is important to emphasize that even for non-smooth, non-convex optimization, there is a vast number of recent publications: forward–backward or proximal-type schemes (Attouch and Bolte 2009, Attouch, Bolte, Redont and Soubeyran 2010, Attouch, Bolte and Svaiter 2013, Bonettini, Loris, Porta and Prato 2016, Bonettini *et al.* 2017), linearized proximal schemes (Xu and Yin 2013, Bolte, Sabach and Teboulle 2014, Xu and Yin 2017, Nikolova and Tan 2017), inertial methods (Ochs, Chen, Brox and Pock 2014, Pock and Sabach 2016), primal–dual algorithms (Valkonen 2014, Li and Pong 2015, Moeller, Benning, Schönlieb and Cremers 2015, Benning, Knoll, Schönlieb and Valkonen 2015), scaled gradient projection methods (Prato *et al.* 2016), non-smooth Gauss–Newton extensions (Drusvyatskiy, Ioffe and Lewis 2016, Ochs, Fadili and Brox 2017), and nonlinear eigenproblems (Hein and Bühler 2010, Bresson, Laurent, Uminsky and Brecht 2012, Benning, Gilboa and Schönlieb 2016, Boţ and Csetnek 2017, Laurent, von Brecht, Bresson and Szlam 2016, Benning, Gilboa, Grah and Schönlieb 2017*c*). Here we focus mainly on recent generalizations of the proximal gradient method and the linearized Bregman iteration for non-convex functionals E; a treatment of all the algorithms mentioned above would be a subject for a survey paper in its own right.

9.1.1. Proximal gradient method

The most basic approach to finding solutions of (9.1) iteratively is via proximal gradient descent, that is, forward–backward splitting (Lions and Mercier 1979). The idea is to linearize the non-convex part E and to add a damping with respect to the previous iterate. If we allow this damping to be carried out via a Bregman distance with respect to a Legendre functional H, we obtain the recently proposed *Bregman proximal gradient method* (Bolte, Sabach, Teboulle and Vaisbourd 2017):

$$\begin{aligned} R(u^{k-1}, \boldsymbol{\alpha}) = \arg\min_{u \in \mathcal{U}} \big\{ & \alpha^{k-1} \langle \nabla E(u^{k-1}), u - u^{k-1} \rangle \\ & + D_H(u, u^{k-1}) + \alpha^{k-1} J(u, \boldsymbol{\alpha}) \big\}, \\ u^k \in R(u^{k-1}, \boldsymbol{\alpha}), & \end{aligned} \tag{9.2}$$

for $\boldsymbol{\alpha} = (\alpha, \alpha^0, \ldots, \alpha^{k-1})$. Here we want to emphasize that $R(u^{k-1}, \boldsymbol{\alpha})$ is no longer a regularization operator in the classical sense as we are no longer necessarily dealing with an inverse problem. Obviously, if E is a (potentially non-convex) data fidelity functional of some nonlinear inverse problem, $R(u^{k-1}, p^{k-1}, \boldsymbol{\alpha})$ depends on some data f^δ and we again deal with a regularization problem, which this time approaches the solution of a (potentially) nonlinear inverse problem. This more specific scenario will be addressed in Section 9.2. Without additional assumptions on E, H and J there is little chance that we can carry out a convergence analysis for (9.2) or even prove existence of the updates. A typical assumption is Lipschitz-continuity of ∇E, that is, we guarantee

$$\|\nabla E(u) - \nabla E(v)\|_{\mathcal{U}^*} \leq L \|u - v\|_{\mathcal{U}}$$

for all $u, v \in \mathcal{U}$ and a constant $L > 0$. A nice aspect of this property is that it implies convexity of the family of functionals

$$\frac{L}{\gamma_i} H_i - E \tag{9.3}$$

(see Bauschke, Bolte and Teboulle 2016, Benning, Betcke, Ehrhardt and Schönlieb 2017*b*, Bolte *et al.* 2017), where $\{H_i\}_{i=1,\ldots}$ is a family of γ_i-strongly convex functionals, that is,

$$\frac{\gamma_i}{2} \|u - v\|_{\mathcal{U}}^2 \leq D_{H_i}(u, v),$$

for all $u, v \in \mathcal{U}$. Let us now assume that H in (9.2) is a member of (9.3) with strong convexity constant γ, that is,

$$H_\gamma(u) := \frac{L}{\gamma} H(u) - E(u) \tag{9.4}$$

is convex for all $u \in \mathcal{U}$. Then this convexity assumption is already enough to ensure a sufficient decrease of the energy $E + J$ in each iteration of (9.2).

Lemma 9.1. Suppose E is coercive or has bounded level sets, $\inf_u E(u) > -\infty$ and ∇E is Lipschitz continuous with constant L, and let H be a Legendre functional in the sense of Definition 5.11, which is also γ-strongly convex. Further assume

$$0 < \alpha^{k-1} < \frac{\gamma C^k}{L + \gamma C^k \rho} \quad \text{for} \quad C^k := \frac{D_H^{\text{symm}}(u^k, u^{k-1})}{D_H(u^k, u^{k-1})}, \tag{9.5}$$

for a constant $\rho > 0$, for all $k \in \mathbb{N}$, and that $E + J(\cdot, \alpha)$ has at least one critical point. Then the iterates of (9.2) satisfy

$$E(u^k) + J(u^k, \alpha) + \rho D_H^{\text{symm}}(u^k, u^{k-1}) \leq E(u^{k-1}) + J(u^{k-1}, \alpha), \tag{9.6}$$

for $u^k \in R(u^{k-1}, \boldsymbol{\alpha})$ and all $k \in \mathbb{N}$.

Proof. From the convexity of (9.4) we immediately observe

$$\begin{aligned} 0 \leq D_{H_\gamma}(u^k, u^{k-1}) = \frac{L}{\gamma} D_H(u^k, u^{k-1}) \\ - \big(E(u^k) - E(u^{k-1}) - \langle \nabla E(u^{k-1}), u^k - u^{k-1} \rangle\big). \end{aligned}$$

As a direct consequence, we have derived the estimate

$$E(u^k) + \langle \nabla E(u^{k-1}), u^{k-1} - u^k \rangle - \frac{L}{\gamma} D_H(u^k, u^{k-1}) \leq E(u^{k-1}). \tag{9.7}$$

From the optimality condition of (9.2) we obtain

$$\nabla E(u^{k-1}) = \frac{1}{\alpha^{k-1}} (\nabla H(u^{k-1}) - \nabla H(u^k)) - p^k, \tag{9.8}$$

for $p^k \in \partial J(u^k, \alpha)$. Inserting (9.8) into (9.7) yields

$$\begin{aligned} E(u^k) + \frac{1}{\alpha^{k-1}} D_H^{\text{symm}}(u^k, u^{k-1}) - \frac{L}{\gamma} D_H(u^k, u^{k-1}) \\ \leq E(u^{k-1}) + \langle p^k, u^{k-1} - u^k \rangle. \end{aligned} \tag{9.9}$$

Due to the convexity of $J(\cdot, \alpha)$ we can estimate

$$\langle p^k, u^{k-1} - u^k \rangle \leq J(u^{k-1}, \alpha) - J(u^k, \alpha).$$

Applying this estimate to (9.9) results in

$$\begin{aligned} E(u^k) + J(u^k, \alpha) + \frac{1}{\alpha^{k-1}} D_H^{\text{symm}}(u^k, u^{k-1}) - \frac{L}{\gamma} D_H(u^k, u^{k-1}) \\ \leq E(u^{k-1}) + J(u^{k-1}, \alpha). \end{aligned}$$

Together with the stepsize bound (9.5) this concludes the proof. □

Remark 9.2. Note that we have not made use of the Lipschitz continuity of ∇E, but only of the convexity of (9.4) in order to obtain a sufficient decrease.

Remark 9.3. Due to the γ-strong convexity of H, the estimate (9.6) automatically implies

$$E(u^k) + J(u^k, \alpha) + \rho\gamma\|u^k - u^{k-1}\|_{\mathcal{U}}^2 \leq E(u^{k-1}) + J(u^{k-1}, \alpha). \tag{9.10}$$

If we additionally assume that both ∇E and ∇H are Lipschitz-continuous, we further obtain a bound for the gradient of the energy $E+J$ at iterate u^k.

Lemma 9.4. Suppose the same assumptions hold as in Lemma 9.1. We further assume that ∇E is Lipschitz-continuous with constant L and ∇H is Lipschitz-continuous with constant δ. Then we observe

$$\|\nabla E(u^k) + p^k\|_{\mathcal{U}^*} \leq \left(L + \frac{\delta}{\alpha^{k-1}}\right)\|u^k - u^{k-1}\|_{\mathcal{U}}$$

for all $p^k \in \partial J(u^k, \alpha)$.

Proof. This follows trivially from (9.8) and the Lipschitz-continuity of both ∇E and ∇H. □

In a finite-dimensional setting $\mathcal{U} = \mathbb{R}^n$ it is now sufficient to assume that $E + J$ is a Kurdyka–Łojasiewicz (KL) function (Lojasiewicz 1963, Kurdyka 1998, Bolte, Daniilidis and Lewis 2007) in order to show that the iterates (9.2) converge globally to a critical point of $E + J$.

Theorem 9.5. Let the same assumptions hold true as in Lemma 9.4. Further assume $\mathcal{U} = \mathbb{R}^n$ and that $E + J$ is a KL function that has at least one critical point. Then the iterates (9.2) converge globally to a critical point of the energy $E + J$.

Proof. See the proof of Bolte *et al.* (2017, Theorem 4.1(ii)). □

We refer the reader to Bolte, Daniilidis, Ley and Mazet (2010) for a detailed investigation of the class of KL functions, and to Bolte *et al.* (2017) for more information on the Bregman proximal gradient.

9.1.2. Linearized Bregman iteration for non-convex functionals

The linearized Bregman iteration introduced in Section 6.2 can easily be adapted to tackle general, non-convex optimization problems. Suppose a Fréchet-differentiable functional $E : \mathcal{U} \to \mathbb{R}$ with Fréchet-gradient ∇E; then we can simply modify Algorithm 2 to

$$\begin{aligned} R(u^{k-1}, p^{k-1}, \boldsymbol{\alpha}) &= \operatorname*{arg\,min}_{u \in \mathcal{U}} \left\{ \langle \nabla E(u^{k-1}), u - u^{k-1} \rangle + \alpha^{k-1} D_{J(\cdot,\alpha)}^{p^{k-1}}(u, u^{k-1}) \right\} \\ u^k &\in R(u^{k-1}, p^{k-1}, \boldsymbol{\alpha}) \\ p^k &= p^{k-1} - \frac{1}{\alpha^{k-1}} \nabla E(u^{k-1}), \end{aligned} \tag{9.11}$$

for $\boldsymbol{\alpha} = (\alpha, \alpha^0, \dots, \alpha^{k-1})$ and $p^{k-1} \in \partial J(u^{k-1}, \alpha)$. This method for arbitrary non-convex energies E was introduced in Benning, Betcke, Ehrhardt and Schönlieb (2017*b*) and mathematically analysed in Benning, Betcke, Ehrhardt and Schönlieb (2017*a*). As in Section 9.1.1, $R(u^{k-1}, p^{k-1}, \boldsymbol{\alpha})$ is no longer a regularization operator in the classical sense, unless E is a (potentially non-convex) data fidelity function for some nonlinear inverse problem.

It becomes evident that (9.11) and (9.2) coincide if J in (9.11) is a Legendre functional and if J in (9.2) is zero. Hence, the convergence analysis closely follows the convergence analysis of the proximal gradient method. We assume that $J(\cdot, \alpha)$ is γ-strongly convex and that

$$J_\gamma(u, \alpha) := \frac{L}{\gamma} J(u, \alpha) - E(u) \tag{9.12}$$

is convex. Then we can show the following sufficient decrease of the energy (Benning *et al.* 2017*b*).

Lemma 9.6. Suppose E is coercive or has bounded level sets, $\inf_u E(u) > -\infty$, α^{k-1} satisfies (9.5) with

$$C^k := \frac{D^{\mathrm{symm}}_{J(\cdot,\alpha)}(u^k, u^{k-1})}{D^{p^{k-1}}_{J(\cdot,\alpha)}(u^k, u^{k-1})},$$

and that E has at least one critical point. Then the iterates of (9.11) satisfy

$$E(u^k) + \rho D^{p^{k-1}}_{J(\cdot,\alpha)}(u^k, u^{k-1}) \le E(u^{k-1}). \tag{9.13}$$

Proof. From the convexity of (9.12) we immediately observe

$$\begin{aligned} 0 \le D^{q^{k-1}}_{J_\gamma(\cdot,\alpha)}(u^k, u^{k-1}) = \frac{L}{\gamma} D^{p^{k-1}}_{J(\cdot,\alpha)}(u^k, u^{k-1}) \\ - \big(E(u^k) - E(u^{k-1}) - \langle \nabla E(u^{k-1}), u^k - u^{k-1} \rangle\big), \end{aligned}$$

for $q^{k-1} \in \partial J_\gamma(u, \alpha)$. As a direct consequence, we have derived the estimate

$$E(u^k) + \langle \nabla E(u^{k-1}), u^{k-1} - u^k \rangle - \frac{L}{\gamma} D^{p^{k-1}}_{J(\cdot,\alpha)}(u^k, u^{k-1}) \le E(u^{k-1}). \tag{9.14}$$

Inserting the dual update formula of (9.11) into (9.14) then yields

$$E(u^k) + \frac{1}{\alpha^{k-1}} D^{\mathrm{symm}}_{J(\cdot,\alpha)}(u^k, u^{k-1}) - \frac{L}{\gamma} D^{p^{k-1}}_{J(\cdot,\alpha)}(u^k, u^{k-1}) \le E(u^{k-1}).$$

Together with the stepsize bound (9.5) we conclude (9.13). □

If we further assume that J^* is δ-strongly convex with respect to its first argument, that is,

$$\frac{\delta}{2} \|p - q\|^2_{\mathcal{U}^*} \le D^v_{J^*(\cdot,\alpha)}(p, q),$$

for all $p, q \in \mathcal{U}^*$ and $v \in \partial J^*(q, \alpha)$, and J^* denoting the convex conjugate of J with respect to the first variable, then we can easily derive the following bound for the gradient at each iteration (Benning *et al.* 2017*b*).

Lemma 9.7. Let the same assumptions hold true as in Lemma 9.6, and further assume that J is δ-strongly convex for all arguments and corresponding subgradients. Then the iterates (9.11) satisfy

$$\|\nabla E(u^{k-1})\|_{\mathcal{U}^*} \leq \frac{\alpha^{k-1}}{\delta}\|u^k - u^{k-1}\|_{\mathcal{U}},$$

for all $k \in \mathbb{N}$.

Proof. From the standard duality estimate $\langle p, u\rangle \leq \|u\|_{\mathcal{U}}\|p\|_{\mathcal{U}^*}$, we observe

$$D^{\text{symm}}_{J(\cdot,\alpha)}(p^k, p^{k-1}) = \langle p^k - p^{k-1}, u^k - u^{k-1}\rangle \leq \|p^k - p^{k-1}\|_{\mathcal{U}^*}\|u^k - u^{k-1}\|_{\mathcal{U}}.$$

Together with the strong convexity of $J^*(\cdot, \alpha)$, we therefore estimate

$$\delta\|p^k - p^{k-1}\|_{\mathcal{U}^*} \leq \frac{D^{\text{symm}}_{J(\cdot,\alpha)}(p^k, p^{k-1})}{\|p^k - p^{k-1}\|_{\mathcal{U}^*}} \leq \|u^k - u^{k-1}\|_{\mathcal{U}}.$$

Inserting the dual update formula from (9.11) thus yields

$$\frac{\delta}{\alpha^{k-1}}\|\nabla E(u^{k-1})\|_{\mathcal{U}^*} \leq \|u^k - u^{k-1}\|_{\mathcal{U}}.$$

This concludes the proof. □

Note that we require no Lipschitz-continuity assumptions for ∇E in order for Lemmas 9.6 and 9.7 to apply, but only that (9.12) is convex. As in the case of the proximal gradient method, we can prove global convergence of the iterates (9.11) for finite-dimensional $\mathcal{U} = \mathbb{R}^n$.

Theorem 9.8. Let the same assumptions hold true as in Lemma 9.6. Further assume $\mathcal{U} = \mathbb{R}^n$ and that E is a KL function. Then the iterates (9.11) converge globally to a critical point of the energy E.

Proof. The proof is a special case of the more general proof of Benning *et al.* (2017*a*, Theorem 5.6 & Corollary 5.7). □

We do want to emphasize that we require $J^*(\cdot, \alpha)$ to be strongly convex, which in return implies the restrictive assumption that $J(\cdot, \alpha)$ is a smooth functional with Lipschitz-continuous gradient. In order to get rid of this restrictive condition we split the functional $J(\cdot, \alpha)$ into the two parts

$$J(u, \alpha) = H(u) + \frac{1}{\alpha^{k-1}}G(u, \alpha),$$

and assume that H is γ-strongly convex and has δ-Lipschitz gradient ∇H, and that $G(\cdot, \alpha)$ is proper, lower semicontinuous and convex. Hence, we

modify (9.11) as follows:

$$\begin{aligned} R(u^{k-1}, q^{k-1}, \boldsymbol{\alpha}) = \arg\min_{u \in \mathcal{U}} \big\{ & \langle \nabla E(u^{k-1}), u - u^{k-1} \rangle \\ & + D^{q^{k-1}}_{G(\cdot,\alpha)}(u, u^{k-1}) + \alpha^{k-1} D_H(u, u^{k-1}) \big\}, \\ u^k \in R(u^{k-1}, q^{k-1}, \boldsymbol{\alpha}), & \\ q^k = q^{k-1} - \big(\nabla E(u^{k-1}) + \alpha^{k-1} & (\nabla H(u^k) - \nabla H(u^{k-1})) \big), \end{aligned} \tag{9.15}$$

for $q^0 \in \partial G(u^0, \alpha)$. We then define the surrogate energy

$$E^k(u^k) := E(u^k) + D^{q^{k-1}}_{G(\cdot,\alpha)}(u^k, u^{k-1}), \tag{9.16}$$

for $q^{k-1} \in \partial G(u^{k-1}, \alpha)$. For this surrogate energy we can show the following results.

Lemma 9.9. Suppose E is coercive or has bounded level sets, $\inf_u E(u) > -\infty$ and E has at least one critical point, and assume H is γ-strongly convex with δ-Lipschitz gradient ∇H, and α^{k-1} satisfies (9.5). Then the iterates of (9.15) satisfy

$$E^k(u^k) + \rho D_H(u^k, u^{k-1}) \leq E^{k-1}(u^{k-1}).$$

Proof. The proof follows the same principle as the proofs of Lemmas 9.1 and 9.6. Convexity of $\frac{L}{\gamma} H - E$ implies the estimate in (9.7). Inserting the optimality condition (or the dual update formula) of (9.15), applying (9.5) and adding $D^{q^{k-2}}_{G(\cdot,\alpha)}(u^{k-1}, u^{k-2})$ to both sides of the inequality then yields the desired estimate. □

A bound of the gradient of $E^k(u^k)$ follows from the Lipschitz-continuity of both ∇E and ∇H.

Lemma 9.10. Let the same assumptions hold true as in Lemma 9.9. Then the iterates (9.15) satisfy

$$\|\nabla E(u^k) + q^k - q^{k-1}\|_{\mathcal{U}^*} \leq (L + \delta \alpha^{k-1}) \|u^k - u^{k-1}\|_{\mathcal{U}}.$$

Proof. Using the dual update formula (9.15) and the Lipschitz-continuity of ∇E and ∇H leads to

$$\begin{aligned} & \|\nabla E(u^k) + q^k - q^{k-1}\|_{\mathcal{U}^*} \\ & \quad = \|\nabla E(u^k) - \nabla E(u^{k-1}) + \alpha^{k-1}(\nabla H(u^{k-1}) - \nabla H(u^k))\|_{\mathcal{U}^*} \\ & \quad \leq L \|u^k - u^{k-1}\|_{\mathcal{U}} + \alpha^{k-1} \delta \|u^{k-1} - u^k\|_{\mathcal{U}}, \end{aligned}$$

which proves the conjecture. □

As in the previous case, global convergence can be achieved under the assumption that the domain is finite-dimensional and that $E^k(u)$ is a KL-function.

Theorem 9.11. Let the same assumptions hold true as in Lemma 9.10. Further assume $\mathcal{U} = \mathbb{R}^n$ and that E^k is a KL function for all $k \in \mathbb{N}$. Then the iterates (9.15) converge globally. If, in addition, the sequence $\{q^k\}_{k\in\mathbb{N}}$ is bounded, then the iterates even convergence to a critical point of the energy E.

Proof. The proof is a special case of the more general proof of Benning *et al.* (2017*a*, Theorem 5.10). □

Remark 9.12. Given the structure of the problem, it is also tempting to look at a Fejér-monotonicity with respect to $D^{q^k}_{J_{\alpha^{k-1}}(\cdot,\alpha)}(u^\dagger, u^k)$, for

$$J_{\alpha^{k-1}}(u,\alpha) := \alpha^{k-1} J(u,\alpha) - E(u).$$

If we make the same attempt as in Section 6, we observe

$$\begin{aligned}
&D^{q^k}_{J_{\alpha^{k-1}}(\cdot,\alpha)}(u^\dagger, u^k) - D^{q^{k-1}}_{J_{\alpha^{k-1}}(\cdot,\alpha)}(u^\dagger, u^{k-1}) \\
&\quad = -D^{q^{k-1}}_{J_{\alpha^{k-1}}(\cdot,\alpha)}(u^k, u^{k-1}) - \langle q^k - q^{k-1}, u^\dagger - u^k\rangle, \\
&\quad = -D^{q^{k-1}}_{J_{\alpha^{k-1}}(\cdot,\alpha)}(u^k, u^{k-1}) \\
&\qquad - \alpha^{k-1}\langle p^k - p^{k-1}, u^\dagger - u^k\rangle \\
&\qquad + \langle \nabla E(u^k) - \nabla E(u^{k-1}), u^\dagger - u^k\rangle, \\
&\quad = -D^{q^{k-1}}_{J_{\alpha^{k-1}}(\cdot,\alpha)}(u^k, u^{k-1}) \\
&\qquad + \langle \nabla E(u^k), u^\dagger - u^k\rangle .
\end{aligned}$$

Since we also know that

$$\begin{aligned}
&\langle \nabla E(u^k), u^\dagger - u^k\rangle \\
&\quad = D^{q^k}_{J_{\alpha^{k-1}}(\cdot,\alpha)}(u^\dagger, u^k) - \alpha^{k-1} D^{p^k}_{J(\cdot,\alpha)}(u^\dagger, u^k) + E(u^\dagger) - E(u^k),
\end{aligned}$$

we can combine this equality with the previous one to obtain

$$\begin{aligned}
&D^{q^k}_{J_{\alpha^{k-1}}(\cdot,\alpha)}(u^\dagger, u^k) - D^{q^{k-1}}_{J_{\alpha^{k-1}}(\cdot,\alpha)}(u^\dagger, u^{k-1}) \\
&\quad = -D^{q^{k-1}}_{J_{\alpha^{k-1}}(\cdot,\alpha)}(u^k, u^{k-1}) \\
&\qquad + D^{q^k}_{J_{\alpha^{k-1}}(\cdot,\alpha)}(u^\dagger, u^k) - \alpha^{k-1} D^{p^k}_{J(\cdot,\alpha)}(u^\dagger, u^k) \\
&\qquad + E(u^\dagger) - E(u^k).
\end{aligned}$$

Hence, for $E(u^\dagger) \leq E(u^k)$ we only observe

$$\alpha^{k-1} D^{q^k}_{J(\cdot,\alpha)}(u^\dagger, u^k) \leq D^{q^{k-1}}_{J_{\alpha^{k-1}}(\cdot,\alpha)}(u^\dagger, u^{k-1}),$$

which is not quite sufficient to achieve Fejér-monotonicity.

We mention that non-convex data fidelities find applications in problems with advanced noise models, for example multiplicative noise (Rudin, Lions and Osher 2003, Aubert and Aujol 2008), image registration problems (Modersitzki 2004), or most nonlinear inverse problems. In the next subsection we focus on the special case of E representing a convex data fidelity F of a potentially nonlinear inverse problem, which leads to an overall non-convex problem.

Let us mention that so far no suitable theory of iterative regularization methods in the case of non-convex regularizations is available, although there are several applications such as the Mumford–Shah or Ambrosio–Tortorelli functional (Mumford and Shah 1989, Ambrosio and Tortorelli 1990, Pock *et al.* 2009, Rondi 2008, Klann, Ramlau and Ring 2011, Klann and Ramlau 2013) or polyconvex energies in image registration (Droske, Rumpf and Schaller 2003, Burger, Modersitzki and Ruthotto 2013*b*, Kirisits and Scherzer 2017).

9.2. Nonlinear inverse problems

Nonlinear inverse problems are extensions of (1.1) with nonlinear forward operators $K : \mathcal{U} \to \mathcal{V}$. Given a convex or non-convex data fidelity term $F : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$, we can formulate variational regularizations and iterative regularizations in exactly the same way as in the linear case. As these problems are special cases of the non-convex methodology discussed in Section 9.1, we can further apply the proposed methodologies. In the context of variational regularization (5.3) for nonlinear forward operators and possibly non-convex but Fréchet-differentiable fidelity terms, the kth iterate of the proximal gradient method discussed in Section 9.1.1 reads as

$$\begin{aligned} &R(f^\delta, u^{k-1}, \boldsymbol{\alpha}) \\ &\quad = \arg\min_{u \in \mathcal{U}} \big\{ \alpha^{k-1} \langle \partial_x F(K(u^{k-1}), f^\delta), u - u^{k-1} \rangle \\ &\qquad + D_H(u, u^{k-1}) + \alpha^{k-1} J(u, \alpha) \big\}. \end{aligned}$$

The convergence theory discussed in Section 9.1.1 applies in identical fashion. However, questions of the convergence of the regularization can now also be addressed.

Gauss–Newton methods

The special structure of the non-convex energy functional E for regularizations of nonlinear inverse problems enables different solution strategies compared to arbitrary non-convex functionals. Given a Fréchet-differentiable operator K, one can approximate $K(u^k)$ via a Taylor-approximation around u^{k-1}, that is,

$$K(u^k) \approx K(u^{k-1}) + K'(u^{k-1})(u^k - u^{k-1}).$$

As a consequence, another strategy for solving variational regularization problems for nonlinear inverse problems is via the following iteratively regularized Gauss–Newton approach:

$$R(f^\delta, u^{k-1}, \boldsymbol{\alpha}) = \arg\min_{u\in\mathcal{U}} \big\{ F(K(u^{k-1}) + K'(u^{k-1})(u - u^{k-1}), f^\delta) + \alpha^{k-1} J(u, \alpha) \big\}. \tag{9.17}$$

We refer to Schöpfer, Louis and Schuster (2006), Stück, Burger and Hohage (2011), Bauer, Hohage and Munk (2009), Kaltenbacher *et al.* (2009), Stück *et al.* (2011) and Hohage and Werner (2013) for further discussion.

In the following sections we discuss extensions of the iterative regularization methods presented in Section 6 to nonlinear inverse problems.

9.2.1. Nonlinear Landweber regularization

We easily observe that (9.11) for $E(u) := F(K(u), f^\delta)$ with nonlinear operator K reads as

$$\begin{aligned} R(f^\delta, v^{k-1}, \boldsymbol{\alpha}) = \arg\min_{u\in\mathcal{U}} \big\{ & \langle K'(u^{k-1})^* \partial_x F(K(u^{k-1}), f^\delta), u - u^{k-1} \rangle \\ & + \alpha^{k-1} D^{p^{k-1}}_{J(\cdot,\alpha)}(u, u^{k-1}) \big\}, \\ u^k \in\ & R(f^\delta, v^{k-1}, \boldsymbol{\alpha}), \\ p^k =\ & p^{k-1} - \frac{1}{\alpha^{k-1}} K'(u^{k-1})^* \partial_x F(K(u^{k-1}), f^\delta), \end{aligned}$$

with $v^{k-1} := (u^{k-1}, p^{k-1})$. For

$$F(K(u), f^\delta) = \frac{1}{2}\|K(u) - f^\delta\|^2_{L^2(\Sigma)} \quad \text{and} \quad J(u, \alpha) = \frac{\alpha}{p}\|u\|^p_{L^p(\Omega)}$$

this method was first introduced and analysed in Kaltenbacher *et al.* (2009). General convex regularization functionals $J(\cdot, \alpha)$ with multivalued subdifferential $\partial J(\cdot, \alpha)$ have been considered in Bachmayr and Burger (2009). Both convergence analyses have been carried under additional assumptions on the nonlinear forward operator, such as the tangential cone condition. In a finite-dimensional setting, convergence follows from Theorem 9.11: see Benning *et al.* (2017*a*). However, it is important to point out that although existence of a critical point of $E(u)$ can usually be guaranteed in finite dimensions, ill-conditioning of the problem still requires early stopping of the iterates.

9.2.2. Levenberg–Marquardt regularization

Replacing the regularization functional in the iterative Gauss–Newton regularization with a generalized Bregman distance with respect to the current and the previous iterate yields the following generalized Levenberg–

Marquardt regularization:

$$\begin{aligned} R(f^\delta, v^{k-1}, \boldsymbol{\alpha}) &= \arg\min_{u\in\mathcal{U}} \big\{ F(K(u^{k-1}) + K'(u^{k-1})(u-u^{k-1}), f^\delta) \\ &\qquad + \alpha^{k-1} D^{p^{k-1}}_{J(\cdot,\alpha)}(u, u^{k-1}) \big\}, \\ u^k &\in R(f^\delta, v^{k-1}, \boldsymbol{\alpha}), \\ p^k &= p^{k-1} - \frac{1}{\alpha^{k-1}} K'(u^{k-1})^* \partial_x F(K(u^{k-1}) \\ &\qquad + K'(u^{k-1})(u^k - u^{k-1}), f^\delta), \end{aligned}$$

for $v^{k-1} := (u^{k-1}, p^{k-1})$. This method reduces to the classical Levenberg–Marquardt method (Levenberg 1944, Marquardt 1963) for the choices

$$F(K(u), f^\delta) = \frac{1}{2}\|K(u) - f^\delta\|^2_{L^2(\Sigma)}, \quad J(u,1) = \frac{1}{2}\|u\|^2_{L^2(\Omega)}.$$

For

$$F(K(u), f^\delta) = \frac{1}{2}\|K(u) - f^\delta\|^2_{L^2(\Sigma)}$$

and proper, lower semicontinuous and convex $J(u,\alpha)$ with potentially multi-valued subdifferential $\partial J(u,\alpha)$, this method was introduced and analysed in Bachmayr and Burger (2009).

9.2.3. Examples

In the following we discuss two nonlinear inverse problems that are natural extensions of the linear inverse problems introduced in Example 6.8 and Section 8.1.

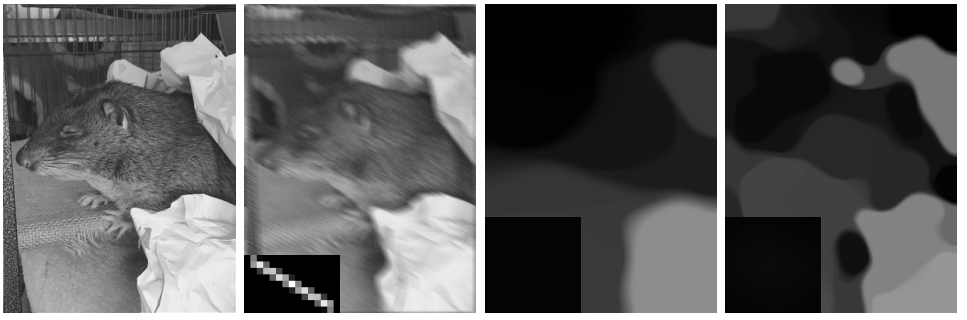
Blind deconvolution

Following up on Example 6.8, an obvious non-convex extension of the problem of deconvolution is blind deconvolution, where the convolution kernel that degrades the image is also unknown (Kundur and Hatzinakos 1996, Chan and Shen 2005, Campisi and Egiazarian 2016). We essentially follow the set-up of Benning *et al.* (2017*a*, Section 6.2), where we assume

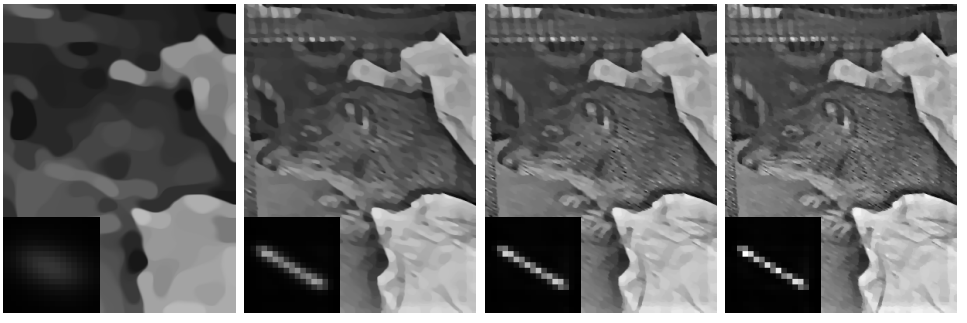
$$\begin{aligned} F(K(u,h), f^\delta) &= \frac{1}{2}\|K(u,h) - f^\delta\|^2_{L^2(\mathbb{R}^2)} \\ &= \frac{1}{2}\|u * h - f^\delta\|^2_{L^2(\mathbb{R}^2)} \end{aligned} \tag{9.18}$$

and apply the nonlinear Landweber regularization as described in Section 9.2.1 with

$$\begin{aligned} &J(u,h,\alpha) \\ &\quad = \frac{1}{2}\|u\|^2_{L^2(\mathbb{R}^2)} + \alpha\,\mathrm{TV}(u) + \int_{\mathbb{R}^2} h(x)\log(h(x)) - h(x)\,\mathrm{d}x + \chi_{P(\mathbb{R}^2)}(h), \end{aligned} \tag{9.19}$$



(a) original $u^\dagger$ (b) blurred, noisy f^δ (c) iterate 1 (d) iterate 11



(e) iterate 49 (f) iterate 999 (g) iterate 2199 (h) iterate 3474

Figure 9.1. (a) Image $u^\dagger \in \mathbb{R}^{400\times 300}$ of Pixel the Gambian pouched rat, introduced in Figure 6.1(a). (b) The same degraded and noisy version $f^\delta \in \mathbb{R}^{400\times 300}$ together with the convolution kernel h as shown in Figures 6.1(b) and 6.3(b). (c–h) Different iterates of Algorithm 2 for $F(K(u,h), f^\delta)$ and $J(u,h)$ as in equations (9.18) and (9.19), respectively, and $\alpha = 10$. The 3474th iterate visualized in (h) is the first that violates Definition 6.1, for $\delta = 5.95$. The reconstructed kernels have been magnified for better visualization.

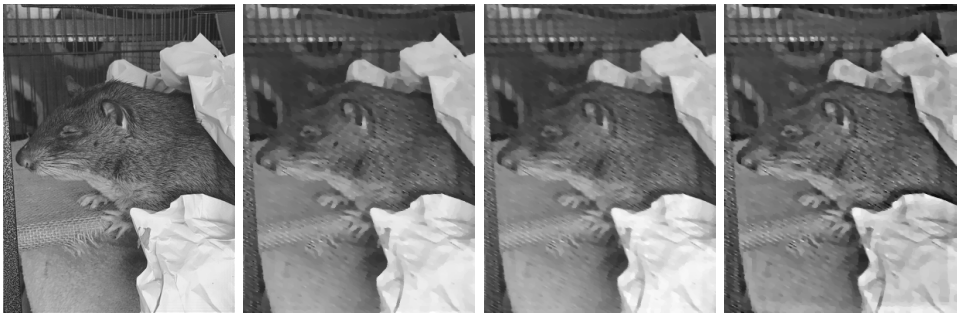
where

$$\chi_{P(\mathbb{R}^2)}(h) = \begin{cases} 0 & h \in P(\mathbb{R}^2), \\ \infty & h \notin P(\mathbb{R}^2) \end{cases}$$

denotes the characteristic functional over the (convex) set of probability distributions

$$P(\mathbb{R}^2) := \left\{ h \in L^2(\mathbb{R}^2) \mid h(x) \geq 0 \text{ a.e.}, \int_{\mathbb{R}^2} h(x)\,\mathrm{d}x = 1 \right\}.$$

The rationale behind this choice of J is that convolution kernels in applications such as motion deblurring are usually non-negative and preserve the mean of the underlying signal. We refer to Benning *et al.* (2017*a*, Sec-



(a) (b) (c) (d)

Figure 9.2. Deconvolution results for the image of Pixel the Gambian pouched rat: (a) the original image; (b) the reconstruction discussed in Example 6.8; (c) the reconstruction with Algorithm 2; (d) the blind deconvolution result computed with the nonlinear Landweber regularization.

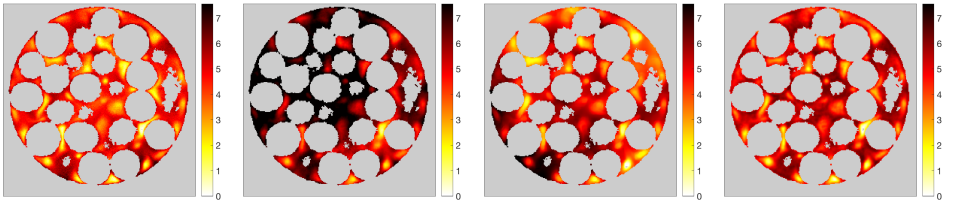
tions 6.2, 7.2) for more information on the discrete formulation of the problem and its numerical realization.

We use $u^\dagger$ and f^δ from Example 6.8, and therefore stop the nonlinear Landweber regularization via the discrepancy principle for $\delta = 5.95$. The parameter α, however, is chosen to be $\alpha = 10$ and is therefore much larger than in Example 6.8 and in Section 6.2. Hence, we require many more iterations in order to reach the same discrepancy. The necessity for this large choice of α stems from the fact that the iterates otherwise converge to unstable solutions with Dirac-delta-like convolution kernels. Several iterates of the nonlinear Landweber regularization are visualized in Figure 9.1.

To conclude, we visually compare the first iterates that violate the discrepancy principle of the Bregman iteration, the linearized Bregman iteration and the nonlinear Landweber regularization in Figure 9.2. Between the reconstructions from the Bregman iteration and the linearized Bregman iteration there are at best small differences in contrast. The reconstruction from the the nonlinear Landweber regularization does have slight artifacts that originate from small imperfections in the reconstructed convolution kernel. Nevertheless, the result is still remarkable given that both image and convolution kernel were unknown and had to be estimated.

Velocity-encoded MRI

We briefly revisit the velocity-encoded MRI problem of Section 8.1. As the original forward problem (8.2) is nonlinear, it is perfectly sensible to recover v_z directly (instead of taking a detour via $w = u \exp(-\mathrm{i}\sigma v_z)$). This idea is not new and has for instance already been addressed in Zhao, Noll, Nielsen and Fessler (2012). We again use the nonlinear Landweber regularization



(a) fully sampled (b) zero-filled (c) TGV recon. (d) nonlin. recon.

Figure 9.3. Comparison of the different z-velocity reconstructions. (a) Unwrapped velocity reconstruction from fully sampled data. (b) Unwrapped velocity reconstruction obtained by filling the missing samples of the sub-sampled data with zero. (c) Unwrapped TGV-based reconstruction of the velocity from sub-sampled data. (d) Nonlinear reconstruction of the velocity, computed via the nonlinear Landweber regularization.

with the functionals

$$F(K(v_z), f^\delta) = \frac{1}{2} \sum_{t=t_0}^{t_m} \|\mathcal{F}(u, v_z) - f_t^\delta\|_2^2,$$

where u is a precomputed spin-proton density, and a scaled H^1-norm

$$J(v_z, \alpha) = \frac{1}{2}\|v_z\|^2_{L^2(\mathbb{R}^2)} + \frac{\alpha}{2}\|\nabla v_z\|^2_{L^2(\mathbb{R}^2)}$$

as the regularization functional of choice.

Figure 9.3 shows the comparison of the velocity reconstruction from the fully sampled data (a), the zero-filled reconstruction from the sub-sampled data (b), the TGV-based reconstruction from the sub-sampled data (c) and a reconstruction from the sub-sampled data via the nonlinear Landweber regularization (d), all clipped to the same intensity range. The latter was initialized with $v_z^0(x) = \pi$ (on some compact domain), $p^0 = u^0$ and $\alpha = 200$. The result shown in Figure 9.3(d) is the first iterate that violates the discrepancy principle for $\eta = 1$ and $\delta = 80$. The inner subproblem was again computed with the PDHGM.

9.3. Learning

A very important question that always pops up when dealing with regularization of inverse problems is the question of how to choose the regularization parameters, that is, how to develop a useful parameter choice strategy. For the iterative regularization strategies discussed in Section 6 we used Morozov's discrepancy principle as an *a posteriori* parameter choice rule to determine when to stop the iteration (which is the regularization parameter for iterative regularizations), based on the noisy data f^δ and the noise level δ. In addition to the standard alternatives, which are *a priori* and heuristic

parameter choice rules, supervised learning strategies have become popular in recent years. The idea is to choose optimal parameters based on pairs $\{(u_j^\dagger, f_j^\delta)\}_{j=1}^m$ of training data by minimizing an empirical risk functional, which is just the empirical expectation of the loss between $u_j^\dagger$ and a $u_j^{\boldsymbol{\alpha}}$ that can be obtained with data f_j^δ. Using our regularization operator notation, a relatively generic approach is to estimate optimal parameters $\hat{\boldsymbol{\alpha}} \in A$ via

$$\hat{\boldsymbol{\alpha}} \in \operatorname*{arg\,min}_{\boldsymbol{\alpha} \in A} \frac{1}{m} \sum_{j=1}^{m} \ell_j(u_j^\dagger, u_j^{\boldsymbol{\alpha}}) + J(\boldsymbol{\alpha}, \boldsymbol{\beta}) \quad \text{subject to } u_j^{\boldsymbol{\alpha}} \in R(f_j^\delta, \boldsymbol{\alpha}), \quad \text{for all } j = 1, \dots, m. \tag{9.20}$$

Here $\{\ell_j\}_{j=1}^m$, with $\ell_j : \mathcal{U} \times \mathcal{U} \to \mathbb{R}$ for all $j \in \{1, \dots, m\}$, denotes a family of loss functionals that measures the deviation between the reconstructions $u_j^{\boldsymbol{\alpha}}$ and the ground truth signals $u_j^\dagger$, and $J : A \times B \to \mathbb{R}$ is a regularization functional that, together with some parameters $\boldsymbol{\beta}$ in some parameter domain B, incorporates prior knowledge to steer the reconstruction of $\hat{\boldsymbol{\alpha}}$ in a certain direction. The operator $R : \mathcal{V} \times A \rightrightarrows \mathcal{U}$ is a regularization operator that takes f_j^δ and $\boldsymbol{\alpha}$ as an input and produces at least one reconstruction $u_j^{\boldsymbol{\alpha}}$ as its output. If $u_j^{\boldsymbol{\alpha}} \in R(f_j^\delta, \boldsymbol{\alpha})$ stems from an optimization problem, then (9.20) is also known as a bilevel optimization problem (Kunisch and Pock 2013, De los Reyes and Schönlieb 2013). It is also quite evident that (9.20) is a regularization problem in itself. An even more generic way to formulate parameter learning would therefore be

$$\hat{\boldsymbol{\alpha}} \in P_R(\{u_j^\dagger\}_{j=1}^m,\ \{f_j^\delta\}_{j=1}^m, \boldsymbol{\beta}),$$

where $P_R : \mathcal{U}^m \times \mathcal{V}^m \times B \rightrightarrows A$ is a regularization operator that also depends on some other regularization operator $R : \mathcal{V} \times A \rightrightarrows \mathcal{U}$. A likely application of this scenario is supervised machine learning with early stopping of, for instance, stochastic gradient descent methods (Johnson and Zhang 2013, Defazio, Bach and Lacoste-Julien 2014, Bertsekas 2011). However, (9.20) is sufficient to explain the majority of current state-of-the-art parameter learning approaches in the context of inverse problems. These cover finite-dimensional Markov random field models (Roth and Black 2005, Tappen 2007, Domke 2012, Chen, Ranftl and Pock 2014*b*, Schmidt and Roth 2014), optimal model design approaches (Haber and Tenorio 2003, Haber, Horesh and Tenorio 2009, Bui-Thanh, Willcox and Ghattas 2008, Biegler *et al.* 2011), optimal regularization parameter estimation in variational regularization (Calatroni, De los Reyes and Schönlieb 2013, Chung, Español and Nguyen 2014, De los Reyes, Schönlieb and Valkonen 2016, De los Reyes, Schönlieb and Valkonen 2017, Calatroni, De los Reyes and Schönlieb 2017, Chung, De los Reyes and Schönlieb 2017), training optimal operators in regularization functionals (Chung, Chung and O'Leary 2011, Chen, Pock,

Ranftl and Bischof 2013, Chen, Pock and Bischof 2014*a*), reaction–diffusion processes (Chen, Yu and Pock 2015, Chen and Pock 2017), so-called variational networks (Hammernik *et al.* 2017, Kobler, Klatzer, Hammernik and Pock 2017, Klatzer *et al.* 2017), and other works related to image processing (Ochs, Ranftl, Brox and Pock 2015, Hintermüller and Wu 2015).

In the following, we want to focus in particular on the connection between modern deep neural network approaches and iterative regularization methods as discussed in Section 6.

9.3.1. Iterative regularization and deep neural networks

In this section we discuss how certain (deep) neural network architectures are closely related (or even equivalent) to the linearized Bregman iteration described in Section 6.2, for a data fidelity term with variable metric. This connection will give insight into how more stable neural network architectures can be learned. For an overview of deep learning and neural network architectures we refer to LeCun, Bengio and Hinton (2015).

We make the assumption that the data fidelity functional is given in terms of

$$F(Ku, f^\delta) = \frac{1}{2}\|Ku - f^\delta\|_{Q_k}^2,$$

for $\|\cdot\|_{Q_k} := \sqrt{\langle Q_k \cdot, \cdot \rangle}$ and some symmetric, positive definite operator Q_k. We now aim to minimize this data fidelity functional with the help of Algorithm 2, but deviate from the standard procedure by allowing the underlying positive definite operator Q_k to vary throughout the iterations.

If we reformulate Algorithm 2 for this particular choice of variable metric data fidelity and linearize around the previous iterate, we obtain the following modification of Algorithm 2:

$$\begin{aligned} R_I(f^\delta, v^{k-1}, \boldsymbol{\alpha}^{k-1}) &= \arg\min_{u \in \mathcal{U}} \big\{ \langle K^* Q_{k-1}(Ku^{k-1} f^\delta), u \rangle + D_\alpha^{p^{k-1}}(u, u^{k-1}) \big\}, \\ u^k &\in R_I(f^\delta, v^{k-1}, \boldsymbol{\alpha}^{k-1}), \\ p^k &= p^{k-1} - K^* Q^{k-1}(Ku^{k-1} - f^\delta). \end{aligned} \tag{9.21}$$

Here we define $\boldsymbol{\alpha}^{k-1} = (\alpha, Q_0, Q_1, \ldots, Q_{k-1})$ and $v^{k-1} = (u^{k-1}, p^{k-1})$. If we now choose J to be of the form

$$J(u, \alpha) = \frac{1}{2}\|u\|_{L^2(\Omega)}^2 + H(u, \alpha),$$

the algorithm simplifies to

$$\begin{aligned} u^k &= (I + \partial H(\cdot, \alpha))^{-1}\big((I - K^* Q_{k-1} K)u^{k-1} + K^* Q_{k-1} f^\delta + q^{k-1}\big), \\ q^k &= u^{k-1} - u^k + q^{k-1} - K^* Q^{k-1}(Ku^k - f^\delta), \end{aligned}$$

for $q^k \in \partial H(u^k, \alpha)$, for all $k \in \mathbb{N}$. Here $(I + \partial H(\cdot, \alpha))^{-1}$ denotes the

proximal mapping of H: see for instance Parikh and Boyd (2014). If we define $A_k := I - K^*Q_kK$ and $b^k := K^*Q_kf^\delta + q^k$ for all $k \in \mathbb{N}$, and choose H to be the pointwise characteristic functional over the convex set of non-negative real numbers, that is,

$$(H(u,\alpha))(x) = (\chi_{\geq 0}(u))(x) = \begin{cases} 0 & u(x) \geq 0, \\ \infty & \text{else}, \end{cases}$$

we obtain the standard rectified linear unit (ReLU: see Nair and Hinton 2010) neural network architecture

$$u^k = \max(0, A_{k-1}u^{k-1} + b^{k-1})$$

for the primal update. However, rather than stopping at this analogy, we want to discuss how the insights of Section 6.2 can help to impose natural conditions on the learning of the parameters A_k and b_k.

Naturally, A_k and b_k have to be of the specific form described above, but we want to look into more detail of what kind of conditions have to be imposed on the free parameters Q_k. We start by defining a surrogate functional that depends on the variable metric data fidelity in the same fashion as we have defined the surrogate functional in Section 6.2, that is, we define

$$J_k(u,\alpha) := J(u,\alpha) - \frac{1}{2}\|Ku - f^\delta\|^2_{Q_k}.$$

If we guarantee convexity of J_k, we can guarantee the following monotonic decrease result.

Corollary 9.13 (monotonic decrease). Suppose u^0 satisfies $\|Ku^0 - f^\delta\|^2_{Q_0} < \infty$. Then the iterates of (9.21) satisfy

$$\frac{1}{2}\|Ku^{k+1} - f^\delta\|^2_{Q_k} + D^{q^k}_{J_k(\cdot,\alpha)}(u^{k+1}, u^k) \leq \|Ku^k - f^\delta\|^2_{Q_k} \tag{9.22}$$

for $u^k \in R_I(f^\delta, v^{k-1}, \boldsymbol{\alpha}^{k-1})$ and $q^k \in \partial J_k(u^k, \alpha)$.

Proof. The proof is identical to the proof of Corollary 6.10. □

If we go back to the assumption $J(u,\alpha) = \frac{1}{2}\|u\|^2_{L^2(\Omega)} + H(u,\alpha)$, we need to ensure that Q_k is chosen such that not just Q_k, but also $I - K^*Q_kK$, is positive (semi-)definite for all k in order to guarantee convexity of J_k. With the next lemma we even observe that this is already enough to ensure Fejér monotonicity of the iterates.

Lemma 9.14. Let $f \in \mathcal{R}_F(K)$, $u^\dagger \in \mathcal{S}(f,\alpha)$ and let $f^\delta \in \mathcal{V}$. Then the iterates satisfy the Fejér monotonicity

$$D^{q^k}_{J_k(\cdot,\alpha)}(u^\dagger, u^k) \leq D^{q^{k-1}}_{J_{k-1}(\cdot,\alpha)}(u^\dagger, u^{k-1}) \tag{9.23}$$

as long as $\|Ku^\dagger - f^\delta\|_{Q_{k-1}} \leq \|Ku^k - f^\delta\|_{Q_{k-1}}$ is satisfied, for all $u^k \in R_I(f^\delta, v^{k-1}, \boldsymbol{\alpha}^{k-1})$ with $q^k \in \partial J_k(u^k, \alpha)$ and $k \in \mathbb{N}$.

Proof. As in our earlier Fejér monotonicity proofs we start by computing

$$\begin{aligned}
&D^{q^k}_{J_k(\cdot,\alpha)}(u^\dagger, u^k) - D^{q^{k-1}}_{J_{k-1}(\cdot,\alpha)}(u^\dagger, u^{k-1}) \\
&\quad = D^{p^k}_{J(\cdot,\alpha)}(u^\dagger, u^k) - D^{p^{k-1}}_{J(\cdot,\alpha)}(u^\dagger, u^{k-1}) \\
&\qquad + D_{\frac{1}{2}\|K\cdot - f^\delta\|^2_{Q_{k-1}}}(u^\dagger, u^{k-1}) - D_{\frac{1}{2}\|K\cdot - f^\delta\|^2_{Q_k}}(u^\dagger, u^k),
\end{aligned}$$

for all $k \in \mathbb{N}$. We further compute

$$\begin{aligned}
&D^{p^k}_{J(\cdot,\alpha)}(u^\dagger, u^k) - D^{p^{k-1}}_{J(\cdot,\alpha)}(u^\dagger, u^{k-1}) \\
&\quad = -D^{p^{k-1}}_{J(\cdot,\alpha)}(u^k, u^{k-1}) - \langle p^k - p^{k-1}, u^\dagger - u^k \rangle \\
&\quad = -D^{p^{k-1}}_{J(\cdot,\alpha)}(u^k, u^{k-1}) + \langle K^* Q_{k-1}(Ku^{k-1} - f^\delta), u^\dagger - u^k \rangle,
\end{aligned}$$

and estimate

$$\begin{aligned}
&D_{\frac{1}{2}\|K\cdot - f^\delta\|^2_{Q_{k-1}}}(u^\dagger, u^{k-1}) - D_{\frac{1}{2}\|K\cdot - f^\delta\|^2_{Q_k}}(u^\dagger, u^k) \\
&\quad \leq D_{\frac{1}{2}\|K\cdot - f^\delta\|^2_{Q_{k-1}}}(u^\dagger, u^{k-1}) \\
&\quad = \frac{1}{2}\|Ku^\dagger - f^\delta\|^2_{Q_{k-1}} - \frac{1}{2}\|Ku^{k-1} - f^\delta\|^2_{Q_{k-1}} \\
&\qquad - \langle K^* Q_{k-1}(Ku^{k-1} - f^\delta), u^\dagger - u^{k-1} \rangle.
\end{aligned}$$

Thus, we observe

$$\begin{aligned}
&D^{q^k}_{J_k(\cdot,\alpha)}(u^\dagger, u^k) - D^{q^{k-1}}_{J_{k-1}(\cdot,\alpha)}(u^\dagger, u^{k-1}) \\
&\quad \leq -D^{p^{k-1}}_{J(\cdot,\alpha)}(u^k, u^{k-1}) \\
&\qquad + \frac{1}{2}\|Ku^\dagger - f^\delta\|^2_{Q_{k-1}} - \frac{1}{2}\|Ku^{k-1} - f^\delta\|^2_{Q_{k-1}} \\
&\qquad - \langle K^* Q_{k-1}(Ku^{k-1} - f^\delta), u^k - u^{k-1} \rangle \\
&\quad = -D^{p^{k-1}}_{J(\cdot,\alpha)}(u^k, u^{k-1}) \\
&\qquad + \frac{1}{2}\|Ku^\dagger - f^\delta\|^2_{Q_{k-1}} - \frac{1}{2}\|Ku^k - f^\delta\|^2_{Q_{k-1}} \\
&\qquad + D_{\frac{1}{2}\|K\cdot - f^\delta\|^2_{Q_{k-1}}}(u^k, u^{k-1})
\end{aligned}$$

$$
\begin{aligned}
&= -D^{p^{k-1}-K^*Q_{k-1}(Ku^{k-1}-f^\delta)}_{J_{k-1}(\cdot,\alpha)}(u^k, u^{k-1}) \\
&\quad + \frac{1}{2}\|Ku^\dagger - f^\delta\|^2_{Q_{k-1}} - \frac{1}{2}\|Ku^k - f^\delta\|^2_{Q_{k-1}} \\
&\leq \frac{1}{2}\|Ku^\dagger - f^\delta\|^2_{Q_{k-1}} - \frac{1}{2}\|Ku^k - f^\delta\|^2_{Q_{k-1}}.
\end{aligned}
$$

Hence, we guarantee Fejér monotonicity as long as

$$\|Ku^\dagger - f^\delta\|_{Q_{k-1}} \leq \|Ku^k - f^\delta\|_{Q_{k-1}}$$

is satisfied. □

Corollary 9.13 and Lemma 9.14 suggest that a sensible model for learning the parameters $\boldsymbol{\alpha}^k$, based on a set of training data pairs $(u_j^\dagger, f_j^\delta)$ for $j = 1, \ldots, m$, is as follows:

$$
\begin{aligned}
\hat{\boldsymbol{\alpha}}^{k^*} = \underset{\boldsymbol{\alpha}^{k^*}}{\arg\min} \sum_{k=1}^{k^*} \left[\sum_{j=1}^{m} D^{q^k}_{J_k(\cdot,\alpha)}(u_j^\dagger, u_j^k) + \chi_{\geq 0}(I - K^*Q_{k-1}K) + \chi_{\geq 0}(Q_{k-1}) \right] \\
\text{subject to } u_j^k \in R_I(f_j^\delta, v_j^{k-1}, \boldsymbol{\alpha}^{k-1}).
\end{aligned}
$$

The minimization problem can either be solved simultaneously for all parameters, or subsequently, keeping all previously computed parameters fixed. Further, the minimization problem can be equipped with additional constraints, such as

$$\|Ku_j^\dagger - f_j^\delta\|_{Q_{k-1}} \leq \|Ku_j^k - f_j^\delta\|_{Q_{k-1}} \quad \text{or} \quad \|Ku_j^{k+1} - f_j^\delta\|_{Q_{k+1}} \leq \|Ku_j^k - f_j^\delta\|_{Q_k},$$

for all $k \in \{0, \ldots, k^* - 1\}$ and $j \in \{1, \ldots, m\}$.

10. Conclusions and outlook

Modern regularization techniques, particularly those based on (non-smooth) convex variational models, are a versatile tool for improved reconstruction in inverse problems when appropriate prior information is available. Further improvements can be made by constructing iterative regularization methods using the same underlying variational model. These can reduce systematic errors and bias, but also yield interesting novel insights into scale properties, spectral and multiscale decompositions, and even link to deep neural network architectures.

Several aspects are expected to play a role in the future development and understanding of regularization methods. A key issue is that of stochastic models and uncertainty quantification, which we have mentioned only superficially in this survey. This topic appears to be at a similar stage to deterministic regularization theory around the year 2000; the Gaussian case (corresponding to linear regularization methods in Hilbert spaces) seems to be reasonably well understood now for linear and nonlinear inverse problems.

Much less is known about non-Gaussian priors in Banach spaces, but there has recently been a surge in papers tackling them. Relevant problems are, for example, the link between Bayesian models and variational approaches, the convergence of posterior distributions, and advanced statistical inference in infinite-dimensional Banach spaces. So far there have been essentially no results on the analysis of iterative regularization methods in a stochastic set-up.

One topic of strong recent interest is that of eigenproblems. While it remains unclear how far they can be pushed for practical purposes, they have already yielded new understanding of the geometry of inverse problems and regularization methods, partly closing the gap with the standard tool of singular value decomposition for linear regularization methods.

A topic that has not yet been investigated from a theoretical point of view, but is often used in engineering practice, is that of methods that effectively compute Nash equilibria instead of minimizers. Such methods arise from problems where two (or more) unknowns are reconstructed in an iterative fashion. Then often one of the variables is frozen and a variational problem with respect to the other one is solved, for example in motion-corrected reconstruction when in alternating iteration images are reconstructed from indirect data with given motion, and motion is estimated directly from image data. Convergence of such procedures is often observed in practice and yields good results, but as yet there is no systematic theory.

From an application point of view, high-dimensional and joint reconstruction problems are a key subject for current and future development, and many aspects of modelling and analysis are still open in this context. Examples of current interest are joint reconstruction of images and motion in many biomedical applications, or reconstructions in dynamic or spectral problems with strong undersampling.

Finally, machine learning is expected to play an important role in regularization methods for inverse problems (as in other disciplines related to processing data). The learning theory will need to be adapted to the special needs of inverse problems due to the aspects of ill-posedness, which cannot be captured by current learning architectures, and the particular difficulties in obtaining meaningful training data for inverse problems.

Acknowledgements

We thank Eva-Maria Brinkmann and Julian Rasch (WWU Münster) for proofreading, for comments that improved the paper, and for providing computational results related to debiasing and dynamic MR reconstruction. MBe acknowledges support from the Leverhulme Trust Early Career Fellowship ‘Learning from mistakes: a supervised feedback-loop for imaging applications’, the Isaac Newton Trust and the Cantab Capital Institute for

the Mathematics of Information. MBu acknowledges support by ERC via Grant EU FP 7 - ERC Consolidator Grant 615216 LifeInverse and by the German Ministry for Science and Education (BMBF) through the project MED4D. The authors would like to thank the Isaac Newton Institute for Mathematical Sciences, Cambridge, for support and hospitality during the programme 'Variational Methods for Imaging and Vision', where work on this paper was undertaken, supported by EPSRC grant no. EP/K032208/1.

REFERENCES[2]

R. Acar and C. R. Vogel (1994), 'Analysis of bounded variation penalty methods for ill-posed problems', *Inverse Problems* **10**, 1217.

S. Agapiou, M. Burger, M. Dashti and T. Helin (2018), 'Sparsity-promoting and edge-preserving maximum *a posteriori* estimators in non-parametric Bayesian inverse problems', *Inverse Problems* **34**, 045002.

S. Aja-Fernandez, C. Alberola-Lopez and C. F. Westin (2008), 'Noise and signal estimation in magnitude MRI and Rician distributed images: A LMMSE approach', *IEEE Trans. Image Process.* **17**, 1383–1398.

W. K. Allard (2007), 'Total variation regularization for image denoising, I: Geometric theory', *SIAM J. Math. Anal.* **39**, 1150–1190.

L. Ambrosio and V. M. Tortorelli (1990), 'Approximation of functional depending on jumps by elliptic functional via t-convergence', *Commun. Pure Appl. Math.* **43**, 999–1036.

L. Ambrosio, N. Fusco and D. Pallara (2000), *Functions of Bounded Variation and Free Discontinuity Problems*, Oxford Mathematical Monographs, Clarendon Press, Oxford University Press.

R. Anderssen (1986), The linear functional strategy for improperly posed problems. In *Inverse Problems* (J. R. Cannon and U. Hornung, eds), Springer, pp. 11–30.

H. Attouch and J. Bolte (2009), 'On the convergence of the proximal algorithm for nonsmooth functions involving analytic features', *Math. Program.* **116**, 5–16.

H. Attouch, J. Bolte and B. F. Svaiter (2013), 'Convergence of descent methods for semi-algebraic and tame problems: Proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods', *Math. Program.* **137**, 91–129.

H. Attouch, J. Bolte, P. Redont and A. Soubeyran (2010), 'Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka–Łojasiewicz inequality', *Math. Oper. Res.* **35**, 438–457.

G. Aubert and J.-F. Aujol (2008), 'A variational approach to removing multiplicative noise', *SIAM J. Appl. Math.* **68**, 925–946.

M. Bachmayr and M. Burger (2009), 'Iterative total variation schemes for nonlinear inverse problems', *Inverse Problems* **25**, 105004.

[2] The URLs cited in this work were correct at the time of going to press, but the publisher and the authors make no undertaking that the citations remain live or are accurate or appropriate.

G. Backus and F. Gilbert (1968), 'The resolving power of gross earth data', *Geophys. J. Internat.* **16**, 169–205.

A. B. Bakushinskii (1967), 'A general method of constructing regularizing algorithms for a linear incorrect equation in Hilbert space', *Zh. Vychisl. Mat. Mat. Fiz.* **7**, 672–677.

A. B. Bakushinskii (1973), 'On the proof of the "discrepancy principle"', *Differential and Integral Equations (Differents. i integr. un-niya), Izd-vo IGU, Irkutsk*.

A. B. Bakushinskii (1977), 'Methods for solving monotonic variational inequalities, based on the principle of iterative regularization', *USSR Comput. Math. Math. Phys.* **17**, 12–24.

A. B. Bakushinskii (1979), 'On the principle of iterative regularization', *USSR Comput. Math. Math. Phys.* **19**, 256–260.

A. B. Bakushinskii (1984), 'Remarks on choosing a regularization parameter using the quasi-optimality and ratio criterion', *USSR Comput. Math. Math. Phys.* **24**, 181–182.

H. Banks and K. Kunisch (1989), *Estimation Techniques for Distributed Parameter Systems*, Birkhäuser.

D. M. Bates and G. Wahba (1983), A truncated singular value decomposition and other methods for generalized cross-validation. Technical report 715, Department of Statistics, University of Wisconsin.

F. Bauer, T. Hohage and A. Munk (2009), 'Iteratively regularized Gauss–Newton method for nonlinear inverse problems with random noise', *SIAM J. Numer. Anal.* **47**, 1827–1846.

H. H. Bauschke, J. Bolte and M. Teboulle (2016), 'A descent lemma beyond Lipschitz gradient continuity: First-order methods revisited and applications', *Math. Oper. Res.* **42**, 330–348.

H. H. Bauschke, J. M. Borwein and P. L. Combettes (2001), 'Essential smoothness, essential strict convexity, and Legendre functions in Banach spaces', *Commun. Contemp. Math.* **3**, 615–647.

A. Beck and M. Teboulle (2003), 'Mirror descent and nonlinear projected subgradient methods for convex optimization', *Oper. Res. Lett.* **31**, 167–175.

M. Benning (2011), Singular regularization of inverse problems: Bregman distances and their applications to variational frameworks with singular regularization energies. PhD thesis, Westfälische Wilhelms-Universität Münster, Germany.

M. Benning and M. Burger (2013), 'Ground states and singular vectors of convex variational regularization methods', *Methods Appl. Anal.* **20**, 295–334.

M. Benning, M. M. Betcke, M. J. Ehrhardt and C.-B. Schönlieb (2017*a*), Choose your path wisely: Gradient descent in a Bregman distance framework. arXiv:1712.04045

M. Benning, M. M. Betcke, M. J. Ehrhardt and C.-B. Schönlieb (2017*b*), Gradient descent in a generalised Bregman distance framework. In *Geometric Numerical Integration and its Applications* (G. R. W. Quispel *et al.*, eds), Vol. 74 of MI Lecture Notes series of Kyushu University, pp. 40–45.

M. Benning, C. Brune, M. Burger and J. Müller (2013), 'Higher-order TV methods: Enhancement via Bregman iteration', *J. Sci. Comput.* **54**, 269–310.

M. Benning, G. Gilboa and C.-B. Schönlieb (2016), 'Learning parametrised regularisation functions via quotient minimisation', *Proc. Appl. Math. Mech.* **16**, 933–936.

M. Benning, G. Gilboa, J. S. Grah and C.-B. Schönlieb (2017*c*), Learning filter functions in regularisers by minimising quotients. In *SSVM 2017: Scale Space and Variational Methods in Computer Vision* (F. Lauze *et al.*, eds), Springer, pp. 511–523.

M. Benning, L. Gladden, D. Holland, C.-B. Schönlieb and T. Valkonen (2014), 'Phase reconstruction from velocity-encoded MRI measurements: A survey of sparsity-promoting variational approaches', *J. Magnetic Resonance* **238**, 26–43.

M. Benning, F. Knoll, C.-B. Schönlieb and T. Valkonen (2015), Preconditioned ADMM with nonlinear operator constraint. In *System Modeling and Optimization* (L. Bociu *et al.*, eds), Springer, pp. 117–126.

M. Benning, M. Möller, R. Z. Nossek, M. Burger, D. Cremers, G. Gilboa and C.-B. Schönlieb (2017*d*), Nonlinear spectral image fusion. In *SSVM 2017: Scale Space and Variational Methods in Computer Vision* (F. Lauze *et al.*, eds), Springer, pp. 41–53.

M. Bergounioux (2016), 'Mathematical analysis of a inf-convolution model for image processing', *J. Optim. Theory Appl.* **168**, 1–21.

M. Bergounioux and E. Papoutsellis (2018), 'An anisotropic inf-convolution BV type model for dynamic reconstruction', *SIAM J. Imaging Sci.* **11**, 129–163.

M. Bertero and P. Boccacci (1998), *Introduction to Inverse Problems in Imaging*, CRC press.

D. P. Bertsekas (2011), Incremental gradient, subgradient, and proximal methods for convex optimization: A survey. In *Optimization for Machine Learning* (S. Sra *et al.*, eds), MIT Press, pp. 85–120.

L. Biegler, G. Biros, O. Ghattas, M. Heinkenschloss, D. Keyes, B. Mallick, L. Tenorio, B. van Bloemen Waanders, K. Willcox and Y. Marzouk (2011), *Large-Scale Inverse Problems and Quantification of Uncertainty*, Wiley.

N. Bissantz, T. Hohage and A. Munk (2004), 'Consistency and rates of convergence of nonlinear Tikhonov regularization with random noise', *Inverse Problems* **20**, 1773.

N. Bissantz, T. Hohage, A. Munk and F. Ruymgaart (2007), 'Convergence rates of general regularization methods for statistical inverse problems and applications', *SIAM J. Numer. Anal.* **45**, 2610–2636.

I. Bleyer and A. Leitao (2009), 'On Tikhonov functionals penalized by Bregman distances', *CUBO* **11**, 99–115.

P. Blomgren and T. F. Chan (1998), 'Color TV: Total variation methods for restoration of vector-valued images', *IEEE Trans. Image Process.* **7**, 304–309.

J. Bolte, A. Daniilidis and A. Lewis (2007), 'The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems', *SIAM J. Optim.* **17**, 1205–1223.

J. Bolte, A. Daniilidis, O. Ley and L. Mazet (2010), 'Characterizations of Łojasiewicz inequalities: Subgradient flows, talweg, convexity', *Trans. Amer. Math. Soc.* **362**, 3319–3363.

J. Bolte, S. Sabach and M. Teboulle (2014), 'Proximal alternating linearized minimization for nonconvex and nonsmooth problems', *Math. Program.* **146**, 459–494.

J. Bolte, S. Sabach, M. Teboulle and Y. Vaisbourd (2017), First order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems. arXiv:1706.06461

S. Bonettini, I. Loris, F. Porta and M. Prato (2016), 'Variable metric inexact line-search based methods for nonsmooth optimization', *SIAM J. Optim.* **26**, 891–921.

S. Bonettini, I. Loris, F. Porta, M. Prato and S. Rebegoldi (2017), 'On the convergence of a linesearch based proximal-gradient method for nonconvex optimization', *Inverse Problems* **33**, 055005.

R. I. Boţ and E. R. Csetnek (2017), 'Proximal-gradient algorithms for fractional programming', *Optimization* **66**, 1383–1396.

K. Bredies and M. Holler (2014), 'Regularization of linear inverse problems with total generalized variation', *J. Inverse Ill-Posed Probl.* **22**, 871–913.

K. Bredies and M. Holler (2015*a*), 'A TGV-based framework for variational image decompression, zooming and reconstruction, I: Analytics', *SIAM J. Imaging Sci.* **8**, 2814–2850.

K. Bredies and M. Holler (2015*b*), 'A TGV-based framework for variational image decompression, zooming, and reconstruction, II: Numerics', *SIAM J. Imaging Sci.* **8**, 2851–2886.

K. Bredies and H. K. Pikkarainen (2013), 'Inverse problems in spaces of measures', *ESAIM Control Optim. Calc. Var.* **19**, 190–218.

K. Bredies and T. Valkonen (2011), Inverse problems with second-order total generalized variation constraints. In *Proceedings of SampTA 2011: 9th International Conference on Sampling Theory and Applications, Singapore*.

K. Bredies, K. Kunisch and T. Pock (2010), 'Total generalized variation', *SIAM J. Imaging Sci.* **3**, 492–526.

L. Bregman (1967), 'The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming', *USSR Comp. Math. Math. Phys.* **7**, 200–217.

X. Bresson and T. F. Chan (2008), 'Fast dual minimization of the vectorial total variation norm and applications to color image processing', *Inverse Probl. Imaging* **2**, 455–484.

X. Bresson, T. Laurent, D. Uminsky and J. V. Brecht (2012), Convergence and energy landscape for Cheeger cut clustering. In *NIPS 2012: Advances in Neural Information Processing Systems 25* (F. Pereira *et al.*, eds), Curran Associates, pp. 1385–1393.

E.-M. Brinkmann, M. Burger, J. Rasch and C. Sutour (2017), 'Bias reduction in variational regularization', *J. Math. Imaging Vision* **59**, 534–566.

C. Brune, A. Sawatzky and M. Burger (2009), Bregman-EM-TV methods with application to optical nanoscopy. In *SSVM 2009: Scale Space and Variational Methods in Computer Vision* (X.-C. Tai *et al.*, eds), Vol. 5567 of Lecture Notes in Computer Science, Springer, pp. 235–246.

C. Brune, A. Sawatzky and M. Burger (2009*c*), Primal and dual Bregman methods with application to optical nanoscopy. CAM Report 09-47, UCLA.

C. Brune, A. Sawatzky and M. Burger (2011), 'Primal and dual Bregman methods with application to optical nanoscopy', *Int. J. Comput. Vis.* **92**, 211–229.

T. Bui-Thanh, K. Willcox and O. Ghattas (2008), 'Model reduction for large-scale systems with high-dimensional parametric input space', *SIAM J. Sci. Comput.* **30**, 3270–3288.

L. Bungert, D. A. Coomes, M. J. Ehrhardt, J. Rasch, R. Reisenhofer and C.-B. Schönlieb (2018), Blind image fusion for hyperspectral imaging with the directional total variation. *Inverse Problems* **34**, 044003.

M. Burger (2016), Bregman distances in inverse problems and partial differential equations. In *Advances in Mathematical Modeling, Optimization and Optimal Control* (J.-B. Hiriart-Urruty *et al.*, eds), Springer, pp. 3–33.

M. Burger and S. Osher (2004), 'Convergence rates of convex variational regularization', *Inverse Problems* **20**, 1411.

M. Burger and S. Osher (2013), A guide to the TV zoo. In *Level Set and PDE Based Reconstruction Methods in Imaging* (M. Burger *et al.*, eds), Springer, pp. 1–70.

M. Burger, L. Eckardt, G. Gilboa and M. Moeller (2015*a*), Spectral representations of one-homogeneous functionals. In *SSVM 2015: Scale Space and Variational Methods in Computer Vision* (J.-F. Aujol *et al.*, eds), Springer, pp. 16–27.

M. Burger, J. Flemming and B. Hofmann (2013*a*), 'Convergence rates in ℓ^1-regularization if the sparsity assumption fails', *Inverse Problems* **29**, 025013.

M. Burger, K. Frick, S. Osher and O. Scherzer (2007*a*), 'Inverse total variation flow', *Multiscale Model. Simul.* **6**, 366–395.

M. Burger, G. Gilboa, M. Moeller, L. Eckardt and D. Cremers (2016*a*), 'Spectral decompositions using one-homogeneous functionals', *SIAM J. Imaging Sci.* **9**, 1374–1408.

M. Burger, G. Gilboa, S. Osher, J. Xu *et al.* (2006), 'Nonlinear inverse scale space methods', *Commun. Math. Sci.* **4**, 179–212.

M. Burger, T. Helin and H. Kekkonen (2016*b*), Large noise in variational regularization. arXiv:1602.00520

M. Burger, J. Modersitzki and L. Ruthotto (2013*b*), 'A hyperelastic regularization energy for image registration', *SIAM J. Sci. Comput.* **35**, B132–B148.

M. Burger, M. Moeller, M. Benning and S. Osher (2013*c*), 'An adaptive inverse scale space method for compressed sensing', **82**, 269–299.

M. Burger, J. Müller, E. Papoutsellis and C.-B. Schönlieb (2014), 'Total variation regularization in measurement and image space for PET reconstruction', *Inverse Problems* **30**, 105003.

M. Burger, S. Osher, J. Xu and G. Gilboa (2005), Nonlinear inverse scale space methods for image restoration. In *VLSM 2005: Variational, Geometric, and Level Set Methods in Computer Vision* (N. Paragios *et al.*, eds), Springer, pp. 25–36.

M. Burger, K. Papafitsoros, E. Papoutsellis and C.-B. Schönlieb (2015*b*), Infimal convolution regularisation functionals of BV and L^p spaces: The case $p = \infty$. In *System Modeling and Optimization* (L. Bociu *et al.*, eds), Springer, pp. 169–179.

M. Burger, K. Papafitsoros, E. Papoutsellis and C.-B. Schönlieb (2016*c*), 'Infimal convolution regularisation functionals of BV and L^p spaces, I: The finite p case', *J. Math. Imaging Vision* **55**, 343–369.

M. Burger, E. Resmerita and L. He (2007*b*), 'Error estimation for Bregman iterations and inverse scale space methods in image restoration', *Computing* **81**, 109–135.

J.-F. Cai and S. Osher (2013), 'Fast singular value thresholding without singular value decomposition', *Methods Appl. Anal.* **20**, 335–352.

J.-F. Cai, E. J. Candès and Z. Shen (2010), 'A singular value thresholding algorithm for matrix completion', *SIAM J. Optim.* **20**, 1956–1982.

J.-F. Cai, S. Osher and Z. Shen (2009*a*), 'Convergence of the linearized Bregman iteration for ℓ_1-norm minimization', *Math. Comp.* **78**, 2127–2136.

J.-F. Cai, S. Osher and Z. Shen (2009*b*), 'Linearized Bregman iterations for compressed sensing', *Math. Comp.* **78**, 1515–1536.

F. Cakoni and D. Colton (2005), 'Open problems in the qualitative approach to inverse electromagnetic scattering theory', *European J. Appl. Math.* **16**, 411–425.

L. Calatroni, J. C. De los Reyes and C.-B. Schönlieb (2013), Dynamic sampling schemes for optimal noise learning under multiple nonsmooth constraints. In *System Modeling and Optimization* (C. Pötzsche *et al.*, eds), Springer, pp. 85–95.

L. Calatroni, J. C. De los Reyes and C.-B. Schönlieb (2017), 'Infimal convolution of data discrepancies for mixed noise removal', *SIAM J. Imaging Sci.* **10**, 1196–1233.

P. T. Callaghan (1993), *Principles of Nuclear Magnetic Resonance Microscopy*, Oxford University Press.

P. T. Callaghan (1999), 'Rheo-NMR: Nuclear magnetic resonance and the rheology of complex fluids', *Rep. Progr. Phys.* **62**, 599.

P. Campisi and K. Egiazarian (2016), *Blind Image Deconvolution: Theory and Applications*, CRC press.

E. J. Candès and D. L. Donoho (2000*a*), Curvelets: A surprisingly effective nonadaptive representation for objects with edges. Technical report, Department of Statistics, Stanford University.

E. J. Candès and D. L. Donoho (2000*b*), Curvelets, multiresolution representation, and scaling laws. In *SPIE Wavelet Applications in Signal and Image Processing VIII*, pp. 1–12.

E. J. Candès and D. L. Donoho (2002), 'Recovering edges in ill-posed inverse problems: Optimality of curvelet frames', *Ann. Statist.* **30**, 784–842.

E. J. Candès and C. Fernandez-Granda (2013), 'Super-resolution from noisy data', *J. Fourier Anal. Appl.* **19**, 1229–1254.

E. J. Candès and C. Fernandez-Granda (2014), 'Towards a mathematical theory of super-resolution', *Commun. Pure Appl. Math.* **67**, 906–956.

E. J. Candès and B. Recht (2009), 'Exact matrix completion via convex optimization', *Found. Comput. Math.* **9**, 717.

E. J. Candès and J. Romberg (2007), 'Sparsity and incoherence in compressive sampling', *Inverse Problems* **23**, 969.

E. J. Candès and T. Tao (2004*a*), 'Decoding by linear programming', *IEEE Trans. Inform. Theory* **51**, 4203–4215.
E. J. Candès and T. Tao (2004*b*), 'Near-optimal signal recovery from random projections: Universal encoding strategies', *IEEE Trans. Inform. Theory* **52**, 5406–5425.
E. J. Candès, X. Li, Y. Ma and J. Wright (2011), 'Robust principal component analysis?', *J. Assoc. Comput. Mach.* **58**, 11.
E. J. Candès, J. Romberg and T. Tao (2006), 'Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information', *IEEE Trans. Inform. Theory* **52**, 489–509.
V. Caselles, A. Chambolle and M. Novaga (2007), 'The discontinuity set of solutions of the TV denoising problem and some extensions', *Multiscale Model. Simul.* **6**, 879–894.
I. Castillo, R. Nickl *et al.* (2014), 'On the Bernstein–von Mises phenomenon for nonparametric Bayes procedures', *Ann. Statist.* **42**, 1941–1969.
L. Cavalier (2008), 'Nonparametric statistical inverse problems', *Inverse Problems* **24**, 034004.
Y. Censor and S. A. Zenios (1992), 'Proximal minimization algorithm withd-functions', *J. Optim. Theory Appl.* **73**, 451–464.
K. Chadan, D. Colton, L. Päivärinta and W. Rundell (1997), *An Introduction to Inverse Scattering and Inverse Spectral Problems*, SIAM.
A. Chambolle (2004), 'An algorithm for total variation minimization and applications', *J. Math. Imaging Vision* **20**, 89–97.
A. Chambolle and P.-L. Lions (1997), 'Image recovery via total variation minimization and related problems', *Numer. Math.* **76**, 167–188.
A. Chambolle and T. Pock (2011), 'A first-order primal–dual algorithm for convex problems with applications to imaging', *J. Math. Imaging Vision* **40**, 120–145.
A. Chambolle and T. Pock (2016), An introduction to continuous optimization for imaging. In *Acta Numerica*, Vol. 25, Cambridge University Press, pp. 161–319.
A. Chambolle, V. Caselles, D. Cremers, M. Novaga and T. Pock (2010), An introduction to total variation for image analysis. In *Theoretical Foundations and Numerical Methods for Sparse Recovery* (M. Fornasier, ed.), Vol. 9 of Radon Series on Computational and Applied Mathematics, De Gruyter, pp. 263–340.
T. F. Chan and J. Shen (2005), *Image Processing and Analysis: Variational, PDE, Wavelet, and Stochastic Methods*, SIAM.
T. F. Chan, S. Esedoglu and F. Park (2010), A fourth order dual method for staircase reduction in texture extraction and image restoration problems. In *ICIP 2010: 17th IEEE International Conference on Image Processing*, pp. 4137–4140.
T. F. Chan, G. H. Golub and P. Mulet (1999), 'A nonlinear primal–dual method for total variation-based image restoration', *SIAM J. Sci. Comput.* **20**, 1964–1977.
C. Chaux, P. L. Combettes, J.-C. Pesquet and V. R. Wajs (2007), 'A variational formulation for frame-based inverse problems', *Inverse Problems* **23**, 1495.
G. Chavent and K. Kunisch (1997), 'Regularization of linear least squares problems by total bounded variation', *ESAIM Control Optim. Calc. Var.* **2**, 359–376.

Y. Chen and T. Pock (2017), 'Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration', *IEEE Trans. Pattern Anal. Machine Intell.* **39**, 1256–1272.

Y. Chen, T. Pock and H. Bischof (2014*a*), Learning ℓ^1-based analysis and synthesis sparsity priors using bi-level optimization. arXiv:1401.4105

Y. Chen, T. Pock, R. Ranftl and H. Bischof (2013), Revisiting loss-specific training of filter-based MRFs for image restoration. In *GCPR 2013: German Conference on Pattern Recognition* (J. Weickert *et al.*, eds), Vol. 8142 of Lecture Notes in Computer Science, Springer, pp. 271–281.

Y. Chen, R. Ranftl and T. Pock (2014*b*), 'Insights into analysis operator learning: From patch-based sparse models to higher order MRFs', *IEEE Trans. Image Process.* **23**, 1060–1072.

Y. Chen, W. Yu and T. Pock (2015), On learning optimized reaction diffusion processes for effective image restoration. In *CVPR 2015: IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5261–5269.

O. Christensen (2003), *An Introduction to Frames and Riesz Bases*, Applied and Numerical Harmonic Analysis, Springer.

C. V. Chung, J. C. De los Reyes and C.-B. Schönlieb (2017), 'Learning optimal spatially-dependent regularization parameters in total variation image denoising', *Inverse Problems* **33**, 074005.

J. Chung, M. Chung and D. P. O'Leary (2011), 'Designing optimal spectral filters for inverse problems', *SIAM J. Sci. Comput.* **33**, 3132–3152.

J. Chung, M. I. Español and T. Nguyen (2014), Optimal regularization parameters for general-form Tikhonov regularization. arXiv:1407.1911

F. Colonna, G. Easley, K. Guo and D. Labate (2010), 'Radon transform inversion using the shearlet representation', *Appl. Comput. Harmon. Anal.* **29**, 232–250.

D. Colton and R. Kress (2012), *Inverse Acoustic and Electromagnetic Scattering Theory*, Vol. 93 of Applied Mathematical Sciences, Springer.

D. Colton and P. Monk (1988), 'The inverse scattering problem for time-harmonic acoustic waves in an inhomogeneous medium', *Quart. J. Mech Appl. Math.* **41**, 97–125.

D. Colton, H. Engl, A. K. Louis, J. McLaughlin and W. Rundell (2012), *Surveys on Solution Methods for Inverse Problems*, Springer.

D. Colton, R. E. Ewing, W. Rundell *et al.* (1990), *Inverse Problems in Partial Differential Equations*, SIAM.

S. F. Cotter, B. D. Rao, K. Engan and K. Kreutz-Delgado (2005), 'Sparse solutions to linear inverse problems with multiple measurement vectors', *IEEE Trans. Signal Process.* **53**, 2477–2488.

J. Darbon and S. Osher (2007), Fast discrete optimization for sparse approximations and deconvolutions. UCLA CAM Report preprint.

M. Dashti, K. J. H. Law, A. M. Stuart and J. Voss (2013), 'MAP estimators and their consistency in Bayesian nonparametric inverse problems', *Inverse Problems* **29**, 095017.

J. C. De los Reyes and C.-B. Schönlieb (2013), 'Image denoising: Learning the noise model via nonsmooth PDE-constrained optimization', *Inverse Probl. Imaging* **7**, 1183–1214.

J. C. De los Reyes, C.-B. Schönlieb and T. Valkonen (2016), 'The structure of optimal parameters for image restoration problems', *J. Math. Anal. Appl.* **434**, 464–500.

J. C. De los Reyes, C.-B. Schönlieb and T. Valkonen (2017), 'Bilevel parameter learning for higher-order total variation regularisation models', *J. Math. Imaging Vision* **57**, 1–25.

A. Defazio, F. Bach and S. Lacoste-Julien (2014), SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *NIPS 2014: Advances in Neural Information Processing Systems 27* (Z. Ghahramani *et al.*, eds), Curran Associates, pp. 1–12.

C.-A. Deledalle, N. Papadakis and J. Salmon (2015), On debiasing restoration algorithms: Applications to total-variation and nonlocal-means. In *SSVM 2015: Scale Space and Variational Methods in Computer Vision* (J.-F. Aujol *et al.*, eds), Springer, pp. 129–141.

C.-A. Deledalle, N. Papadakis, J. Salmon and S. Vaiter (2017), 'CLEAR: Covariant least-square refitting with applications to image restoration', *SIAM J. Imaging Sci.* **10**, 243–284.

Q. Denoyelle, V. Duval and G. Peyré (2017), 'Support recovery for sparse super-resolution of positive measures', *J. Fourier Anal. Appl.* **23**, 1153–1194.

J. Domke (2012), Generic methods for optimization-based modeling. In *Fifteenth International Conference on Artificial Intelligence and Statistics* (N. D. Lawrence and M. Girolami, eds), PMLR, pp. 318–326.

D. L. Donoho (1992), 'Superresolution via sparsity constraints', *SIAM J. Math. Anal.* **23**, 1309–1331.

D. L. Donoho (2006), 'Compressed sensing', *IEEE Trans. Inform. Theory* **52**, 1289–1306.

D. L. Donoho and I. M. Johnstone (1995), 'Adapting to unknown smoothness via wavelet shrinkage', *J. Amer. Statist. Assoc.* **90** (432), 1200–1224.

D. L. Donoho, M. Elad and V. N. Temlyakov (2006), 'Stable recovery of sparse overcomplete representations in the presence of noise', *IEEE Trans. Inform. Theory* **52**, 6–18.

M. Droske, M. Rumpf and C. Schaller (2003), Nonrigid morphological image registration & its practical issues. In *ICIP 2003: IEEE International Conference on Image Processing*, pp. II–699.

D. Drusvyatskiy, A. D. Ioffe and A. S. Lewis (2016), Nonsmooth optimization using Taylor-like models: Error bounds, convergence, and termination criteria. arXiv:1610.03446

M. F. Duarte, S. Sarvotham, M. B. Wakin, D. Baron and R. G. Baraniuk (2005), Joint sparsity models for distributed compressed sensing. In *Proceedings of the Workshop on Signal Processing with Adaptive Sparse Structured Representations*, IEEE.

V. Duval and G. Peyré (2017*a*), 'Sparse regularization on thin grids, I: The Lasso', *Inverse Problems* **33**, 055008.

V. Duval and G. Peyré (2017*b*), 'Sparse spikes deconvolution on thin grids, II: The continuous basis pursuit', *Inverse Problems* **33**, 095008.

J. Eckstein (1993), 'Nonlinear proximal point algorithms using Bregman functions, with applications to convex programming', *Math. Oper. Res.* **18**, 202–226.

P. P. B. Eggermont (1993), 'Maximum entropy regularization for Fredholm integral equations of the first kind', *SIAM J. Math. Anal.* **24**, 1557–1576.
M. J. Ehrhardt and S. R. Arridge (2014), 'Vector-valued image processing by parallel level sets', *IEEE Trans. Image Process.* **23**, 9–18.
M. J. Ehrhardt and M. M. Betcke (2016), 'Multicontrast MRI reconstruction with structure-guided total variation', *SIAM J. Imaging Sci.* **9**, 1084–1106.
M. J. Ehrhardt, P. Markiewicz, M. Liljeroth, A. Barnes, V. Kolehmainen, J. S. Duncan, L. Pizarro, D. Atkinson, B. F. Hutton, S. Ourselin *et al.* (2016), 'PET reconstruction with an anatomical MRI prior using parallel level sets', *IEEE Trans. Medical Imaging* **35**, 2189–2199.
M. J. Ehrhardt, K. Thielemans, L. Pizarro, D. Atkinson, S. Ourselin, B. F. Hutton and S. R. Arridge (2014), 'Joint reconstruction of PET-MRI by exploiting structural similarity', *Inverse Problems* **31**, 015001.
B. Eicke (1992), 'Iteration methods for convexly constrained ill-posed problems in Hilbert space', *Numer. Funct. Anal. Optim.* **13**, 413–429.
I. Ekeland and R. Temam (1999), *Convex Analysis and Variational Problems*, corrected reprint edition, SIAM.
M. Elad, P. Milanfar and R. Rubinstein (2007), 'Analysis versus synthesis in signal priors', *Inverse Problems* **23**, 947.
L. Eldén (1977), 'Algorithms for the regularization of ill-conditioned least squares problems', *BIT Numer. Math.* **17**, 134–145.
H. W. Engl (1987*a*), 'Discrepancy principles for Tikhonov regularization of ill-posed problems leading to optimal convergence rates', *J. Optim. Theory Appl.* **52**, 209–215.
H. W. Engl (1987*b*), 'On the choice of the regularization parameter for iterated Tikhonov regularization of ill-posed problems', *J. Approx. Theory* **49**, 55–63.
H. W. Engl and H. Gfrerer (1988), '*A posteriori* parameter choice for general regularization methods for solving linear ill-posed problems', *Appl. Numer. Math.* **4**, 395–417.
H. W. Engl and G. Landl (1993), 'Convergence rates for maximum entropy regularization', *SIAM J. Numer. Anal.* **30**, 1509–1536.
H. W. Engl and A. Neubauer (1985), Optimal discrepancy principles for the Tikhonov regularization of integral equations of the first kind. In *Constructive Methods for the Practical Treatment of Integral Equations* (G. Hämmerlin and K.-H. Hoffmann, eds), Springer, pp. 120–141.
H. W. Engl and A. Neubauer (1987), Optimal parameter choice for ordinary and iterated Tikhonov regularization. In *Inverse and Ill-Posed Problems* (H. W. Engl and C. W. Groetsch, eds), Elsevier, pp. 97–125.
H. W. Engl, M. Hanke and A. Neubauer (1996), *Regularization of Inverse Problems*, Mathematics and Its Applications, Springer.
H. W. Engl, K. Kunisch and A. Neubauer (1989), 'Convergence rates for Tikhonov regularisation of non-linear ill-posed problems', *Inverse Problems* **5**, 523.
H. W. Engl, A. K. Louis and W. Rundell, eds (2012), *Inverse Problems in Medical Imaging and Nondestructive Testing*, Springer.
E. Esser, X. Zhang and T. F. Chan (2010), 'A general framework for a class of first order primal–dual algorithms for convex optimization in imaging science', *SIAM J. Imaging Sci.* **3**, 1015–1046.

L. Evans and R. Gariepy (1992), *Measure Theory and Fine Properties of Functions*, Studies in Advanced Mathematics, CRC Press.
J. Flemming (2013), 'Variational smoothness assumptions in convergence rate theory: An overview', *J. Inverse Ill-Posed Probl.* **21**, 395–409.
J. Flemming (2017*a*), A converse result for Banach space convergence rates in Tikhonov-type convex regularization of ill-posed linear equations. arXiv:1712.01499
J. Flemming (2017*b*), 'Existence of variational source conditions for nonlinear inverse problems in Banach spaces', *J. Inverse Ill-Posed Probl.* doi:10.1515/jiip-2017-0092
J. Flemming and D. Gerth (2017), 'Injectivity and weak*-to-weak continuity suffice for convergence rates in ℓ^1-regularization', *J. Inverse Ill-Posed Probl.* **26**, 85–94.
J. Flemming and B. Hofmann (2010), 'A new approach to source conditions in regularization with general residual term', *Numer. Funct. Anal. Optim.* **31**, 254–284.
J. Flemming, B. Hofmann and I. Veselić (2015), 'On ℓ^1-regularization in light of Nashed's ill-posedness concept', *Comput. Methods Appl. Math.* **15**, 279–289.
J. Flemming, B. Hofmann and I. Veselić (2016), 'A unified approach to convergence rates for ℓ^1-regularization and lacking sparsity', *J. Inverse Ill-Posed Probl.* **24**, 139–148.
M. Fornasier and H. Rauhut (2008), 'Recovery algorithms for vector-valued data with joint sparsity constraints', *SIAM J. Numer. Anal.* **46**, 577–613.
Y. Gao and K. Bredies (2017), Infimal convolution of oscillation total generalized variation for the recovery of images with structured texture. arXiv:1710.11591
P. D. Gatehouse, J. Keegan, L. A. Crowe, S. Masood, R. H. Mohiaddin, K.-F. Kreitner and D. N. Firmin (2005), 'Applications of phase-contrast flow and velocity imaging in cardiovascular MRI', *European Radiology* **15**, 2172–2184.
H. Gfrerer (1987), 'An *a posteriori* parameter choice for ordinary and iterated Tikhonov regularization of ill-posed problems leading to optimal convergence rates', *Math. Comp.* **49** (180), 507–522.
A. Gholami and H. Siahkoohi (2010), 'Regularization of linear and non-linear geophysical ill-posed problems with joint sparsity constraints', *Geophys. J. Internat.* **180**, 871–882.
G. Gilboa (2014*a*), Nonlinear band-pass filtering using the TV transform. In *EUSIPCO 2014: 22nd European Signal Processing Conference*, IEEE, pp. 1696–1700.
G. Gilboa (2014*b*), 'A total variation spectral framework for scale and texture analysis', *SIAM J. Imaging Sci.* **7**, 1937–1961.
G. Gilboa, M. Moeller and M. Burger (2016), 'Nonlinear spectral analysis via one-homogeneous functionals: Overview and future prospects', *J. Math. Imaging Vision* **56**, 300–319.
E. Giné and R. Nickl (2015), *Mathematical Foundations of Infinite-Dimensional Statistical Models*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.
J. S. Grah (2017), Mathematical imaging tools in cancer research: From mitosis analysis to sparse regularisation. PhD thesis, University of Cambridge.

M. Grasmair (2011), 'Linear convergence rates for Tikhonov regularization with positively homogeneous functionals', *Inverse Problems* **27**, 075014.

M. Grasmair (2013), 'Variational inequalities and higher order convergence rates for Tikhonov regularisation on Banach spaces', *J. Inverse Ill-Posed Probl.* **21**, 379–394.

M. Grasmair and F. Lenzen (2010), 'Anisotropic total variation filtering', *Appl. Math. Optim.* **62**, 323–339.

M. Grasmair, O. Scherzer and M. Haltmeier (2011), 'Necessary and sufficient conditions for linear convergence of ℓ^1-regularization', *Commun. Pure Appl. Math.* **64**, 161–182.

C. W. Groetsch (1977), 'Sequential regularization of ill-posed problems involving unbounded operators', *Comment. Math. Univ. Carolin.* **18**, 489–498.

C. W. Groetsch (1993), *Inverse Problems in the Mathematical Sciences*, Vieweg Mathematics for Scientists and Engineers, Vieweg.

C. W. Groetsch and J. T. King (1979), 'Extrapolation and the method of regularization for generalized inverses', *J. Approx. Theory* **25**, 233–247.

K. Guo and D. Labate (2007), 'Optimally sparse multidimensional representation using shearlets', *SIAM J. Math. Anal.* **39**, 298–318.

E. Haber and L. Tenorio (2003), 'Learning regularization functionals: A supervised training approach', *Inverse Problems* **19**, 611.

E. Haber, L. Horesh and L. Tenorio (2009), 'Numerical methods for the design of large-scale nonlinear discrete ill-posed inverse problems', *Inverse Problems* **26**, 025002.

J. Hadamard (1902), 'Sur les problèmes aux dérivées partielles et leur signification physique', *Princeton University Bulletin* **13**, 49–52.

J. Hadamard (1923), *Lectures on Cauchy's Problem in Linear Partial Differential Equations*, Yale University Press.

K. Hammernik, T. Klatzer, E. Kobler, M. P. Recht, D. K. Sodickson, T. Pock and F. Knoll (2017), 'Learning a variational network for reconstruction of accelerated MRI data', *Magn. Reson. Med.* **79**, 3055–3071.

M. Hanke, A. Neubauer and O. Scherzer (1995), 'A convergence analysis of the Landweber iteration for nonlinear ill-posed problems', *Numer. Math.* **72**, 21–37.

P. C. Hansen (1987), 'The truncated SVD as a method for regularization', *BIT Numer. Math.* **27**, 534–553.

P. C. Hansen (1992), 'Analysis of discrete ill-posed problems by means of the L-curve', *SIAM Review* **34**, 561–580.

M. Hein and T. Bühler (2010), An inverse power method for nonlinear eigenproblems with applications in 1-spectral clustering and sparse PCA. In *NIPS 2010: Advances in Neural Information Processing Systems 23* (J. D. Lafferty *et al.*, eds), Curran Associates, pp. 847–855.

P. Heins (2014), Reconstruction using local sparsity: A novel regularization technique and an asymptotic analysis of spatial sparsity priors. PhD thesis, Westfälische Wilhelms-Universität Münster, Germany.

P. Heins, M. Moeller and M. Burger (2015), 'Locally sparse reconstruction using the $\ell^{1,\infty}$-norm', *Inverse Probl. Imaging* **9**, 1093–1137.

T. Helin and M. Burger (2015), 'Maximum *a posteriori* probability estimates in infinite-dimensional Bayesian inverse problems', *Inverse Problems* **31**, 085009.

T. Helin and M. Lassas (2011), 'Hierarchical models in statistical inverse problems and the Mumford–Shah functional', *Inverse Problems* **27**, 015008.

W. Hinterberger and O. Scherzer (2006), 'Variational methods on the space of functions of bounded Hessian for convexification and denoising', *Computing* **76**, 109–133.

W. Hinterberger, O. Scherzer, C. Schnörr and J. Weickert (2002), 'Analysis of optical flow models in the framework of the calculus of variations', *Numer. Funct. Anal. Optim.* **23**, 69–89.

M. Hintermüller and T. Wu (2015), 'Bilevel optimization for calibrating point spread functions in blind deconvolution', *Inverse Probl. Imaging* **9**, 1139–1169.

M. Hintermüller, M. Holler and K. Papafitsoros (2017), A function space framework for structural total variation regularization with applications in inverse problems. arXiv:1710.01527

A. E. Hoerl (1959), 'Optimum solution of many variables equations', *Chem. Engrg Progr.* **55**, 69–78.

A. E. Hoerl and R. W. Kennard (1970), 'Ridge regression: Biased estimation for nonorthogonal problems', *Technometrics* **12**, 55–67.

T. Hohage (1997), 'Logarithmic convergence rates of the iteratively regularized Gauss–Newton method for an inverse potential and an inverse scattering problem', *Inverse Problems* **13**, 1279.

T. Hohage and F. Weidling (2017), 'Characterizations of variational source conditions, converse results, and maxisets of spectral regularization methods', *SIAM J. Numer. Anal.* **55**, 598–620.

T. Hohage and F. Werner (2013), 'Iteratively regularized Newton-type methods for general data misfit functionals and applications to Poisson data', *Numer. Math.* **123**, 745–779.

T. Hohage and F. Werner (2016), 'Inverse problems with Poisson data: Statistical regularization theory, applications and algorithms', *Inverse Problems* **32**, 093001.

D. Holland, D. Malioutov, A. Blake, A. Sederman and L. Gladden (2010), 'Reducing data acquisition times in phase-encoded velocity imaging using compressed sensing', *J. Magnetic Resonance* **203**, 236–246.

D. Holland, C. Müller, J. Dennis, L. Gladden and A. Sederman (2008), 'Spatially resolved measurement of anisotropic granular temperature in gas-fluidized beds', *Powder Technology* **182**, 171–181.

M. Holler and K. Kunisch (2014), 'On infimal convolution of TV-type functionals and applications to video and image reconstruction', *SIAM J. Imaging Sci.* **7**, 2258–2300.

Y. Hu and M. Jacob (2012), 'Higher degree total variation (HDTV) regularization for image recovery', *IEEE Trans. Image Process.* **21**, 2559–2571.

J. Huang and D. Mumford (1999), Statistics of natural images and models. In *CVPR 1999: IEEE Computer Society Conference On Computer Vision and Pattern Recognition*, pp. 541–547.

V. Isakov (2006), *Inverse Problems for Partial Differential Equations*, Vol. 127 of Applied Mathematical Sciences, Springer.

V. Isakov (2008), 'On inverse problems in secondary oil recovery', *European J. Appl. Math.* **19**, 459–478.

V. K. Ivanov (1962), On linear problems which are not well-posed. *Soviet Math. Dokl.* **3**, 981–983.

K. Jalalzai (2016), 'Some remarks on the staircasing phenomenon in total variation-based image denoising', *J. Math. Imaging Vision* **54**, 256–268.

F. John (1960), 'Continuous dependence on data for solutions of partial differential equations with a prescribed bound', *Commun. Pure Appl. Math.* **13**, 551–585.

R. Johnson and T. Zhang (2013), Accelerating stochastic gradient descent using predictive variance reduction. In *NIPS 2013: Advances in Neural Information Processing Systems 26*, (C. J. C. Burges *et al.*, eds), Curran Associates, pp. 315–323.

J. P. Kaipio and E. Somersalo (2006), *Statistical and Computational Inverse Problems*, Applied Mathematical Sciences, Springer.

J. P. Kaipio, V. Kolehmainen, M. Vauhkonen and E. Somersalo (1999), 'Inverse problems with structural prior information', *Inverse Problems* **15**, 713.

B. Kaltenbacher (1997), 'Some Newton-type methods for the regularization of nonlinear ill-posed problems', *Inverse Problems* **13**, 729.

B. Kaltenbacher (2008), 'A note on logarithmic convergence rates for nonlinear Tikhonov regularization', *J. Inverse Ill-Posed Probl.* **16**, 79–88.

B. Kaltenbacher, F. Schöpfer and T. Schuster (2009), 'Iterative methods for nonlinear ill-posed problems in Banach spaces: Convergence and applications to parameter identification problems', *Inverse Problems* **25**, 065003.

H. Kekkonen, M. Lassas and S. Siltanen (2014), 'Analysis of regularized inversion of data corrupted by white Gaussian noise', *Inverse Problems* **30**, 045009.

H. Kekkonen, M. Lassas and S. Siltanen (2016), 'Posterior consistency and convergence rates for Bayesian inversion with hypoelliptic operators', *Inverse Problems* **32**, 085005.

C. Kirisits and O. Scherzer (2017), 'Convergence rates for regularization functionals with polyconvex integrands', *Inverse Problems* **33**, 085008.

K. C. Kiwiel (1997), 'Proximal minimization methods with generalized Bregman functions', *SIAM J. Control Optim.* **35**, 1142–1168.

E. Klann and R. Ramlau (2013), 'Regularization properties of Mumford–Shah-type functionals with perimeter and norm constraints for linear ill-posed problems', *SIAM J. Imaging Sci.* **6**, 413–436.

E. Klann, R. Ramlau and W. Ring (2011), 'A Mumford–Shah level-set approach for the inversion and segmentation of SPECT/CT data', *Inverse Probl. Imaging* **5**, 137–166.

T. Klatzer, D. Soukup, E. Kobler, K. Hammernik and T. Pock (2017), Trainable regularization for multi-frame superresolution. In *GCPR 2017: German Conference on Pattern Recognition* (V. Roth and T. Vetter, eds), Vol. 10496 of Lecture Notes in Computer Science, Springer, pp. 90–100.

F. Knoll, K. Bredies, T. Pock and R. Stollberger (2011), 'Second order total generalized variation (TGV) for MRI', *Magnetic Resonance Medicine* **65**, 480–491.

F. Knoll, M. Holler, T. Koesters, R. Otazo, K. Bredies and D. K. Sodickson (2017), 'Joint MR-PET reconstruction using a multi-channel image regularizer', *IEEE Trans. Medical Imaging* **36**, 1–16.

E. Kobler, T. Klatzer, K. Hammernik and T. Pock (2017), Variational networks: connecting variational methods and deep learning. In *GCPR 2017: German Conference on Pattern Recognition* (V. Roth and T. Vetter, eds), Vol. 10496 of Lecture Notes in Computer Science, Springer, pp. 281–293.

V. Kolehmainen, M. Lassas, K. Niinimäki and S. Siltanen (2012), 'Sparsity-promoting Bayesian inversion', *Inverse Problems* **28**, 025005.

M. Krause, R. M. Alles, B. Burgeth and J. Weickert (2016), 'Fast retinal vessel analysis', *J. Real-Time Image Processing* **11**, 413–422.

C. Kravaris and J. H. Seinfeld (1985), 'Identification of parameters in distributed parameter systems by regularization', *SIAM J. Control Optim.* **23**, 217–241.

A. Kryanev (1974), 'An iterative method for solving incorrectly posed problems', *USSR Comput. Math. Math. Phys.* **14**, 24–35.

D. Kundur and D. Hatzinakos (1996), 'Blind image deconvolution', *IEEE Signal Processing Magazine* **13**, 43.

K. Kunisch and M. Hintermüller (2004), 'Total bounded variation regularization as a bilaterally constrained optimization problem', *SIAM J. Appl. Math.* **64**, 1311–1333.

K. Kunisch and T. Pock (2013), 'A bilevel optimization approach for parameter learning in variational models', *SIAM J. Imaging Sci.* **6**, 938–983.

K. Kurdyka (1998), 'On gradients of functions definable in o-minimal structures', *Annales de l'Institut Fourier* (Chartres) **48**, 769–784.

G. Kutyniok and D. Labate (2012), Introduction to shearlets. In *Shearlets: Multiscale Analysis for Multivariate Data*, Applied and Numerical Harmonic Analysis, Springer, pp. 1–38.

D. Labate, W.-Q. Lim, G. Kutyniok and G. Weiss (2005), Sparse multidimensional representation using shearlets. In *Optics and Photonics 2005*, Proceedings Vol. 5914, SPIE, 59140U.

L. Landweber (1951), 'An iteration formula for Fredholm integral equations of the first kind', *Amer. J. Math.* **73**, 615–624.

M. Lassas, E. Saksman and S. Siltanen (2009), 'Discretization-invariant Bayesian inversion and Besov space priors', *Inverse Probl. Imaging* **3**, 87–122.

R. Lattès and J.-L. Lions (1967), 'Méthode de quasi-réversibilité et applications'.

T. Laurent, J. von Brecht, X. Bresson and A. Szlam (2016), The product cut. In *NIPS 2016: Advances in Neural Information Processing Systems 29* (D. D. Lee *et al.*, eds), Curran Associates, pp. 3792–3800.

Y. LeCun, Y. Bengio and G. Hinton (2015), 'Deep learning', *Nature* **521** (7553), 436–444.

J. Lederer (2013), Trust, but verify: Benefits and pitfalls of least-squares refitting in high dimensions. arXiv:1306.0113

O. Lee, J. M. Kim, Y. Bresler and J. C. Ye (2011), 'Compressive diffuse optical tomography: Noniterative exact reconstruction using joint sparsity', *IEEE Trans. Medical Imaging* **30**, 1129–1142.

F. Lenzen, F. Becker and J. Lellmann (2013), Adaptive second-order total variation: An approach aware of slope discontinuities. In *SSVM 2015: Scale*

Space and Variational Methods in Computer Vision (J.-F. Aujol *et al.*, eds), Springer, pp. 61–73.
K. Levenberg (1944), 'A method for the solution of certain non-linear problems in least squares', *Quart. Appl. Math.* **2**, 164–168.
G. Li and T. K. Pong (2015), 'Global convergence of splitting methods for nonconvex composite optimization', *SIAM J. Optim.* **25**, 2434–2460.
H. C. Lie and T. Sullivan (2017), Equivalence of weak and strong modes of measures on topological vector spaces. arXiv:1708.02516
P.-L. Lions and B. Mercier (1979), 'Splitting algorithms for the sum of two nonlinear operators', *SIAM J. Numer. Anal.* **16**, 964–979.
S. Lojasiewicz (1963), 'Une propriété topologique des sous-ensembles analytiques réels', *Les Équations aux Dérivées Partielles* **117**, 87–89.
A. Louis (1996), 'Approximate inverse for linear and some nonlinear problems', *Inverse Problems* **12**, 175.
M. Lustig, D. Donoho and J. M. Pauly (2007), 'Sparse MRI: The application of compressed sensing for rapid MR imaging', *Magnetic Resonance Medicine* **58**, 1182–1195.
S. Mallat (2008), *A Wavelet Tour of Signal Processing: The Sparse Way*, Academic Press.
S. Mallat and Z. Zhang (1993), 'Matching pursuits with time-frequency dictionaries', *IEEE Trans. Signal Process.* **12**, 3397–3415.
D. W. Marquardt (1963), 'An algorithm for least-squares estimation of nonlinear parameters', *J. Soc. Indust. Appl. Math.* **11**, 431–441.
A. Marquina and S. J. Osher (2008), 'Image super-resolution by TV-regularization and Bregman iteration', *J. Sci. Comput.* **37**, 367–382.
J. Modersitzki (2004), *Numerical Methods for Image Registration*, Numerical Mathematics and Scientific Computation, Oxford University Press.
M. Moeller (2012), Multiscale methods for polyhedral regularizations and applications in high dimensional imaging. PhD thesis, University of Münster, Germany.
M. Moeller and M. Burger (2013), 'Multiscale methods for polyhedral regularizations', *SIAM J. Optim.* **23**, 1424–1456.
M. Moeller, M. Benning, C. Schönlieb and D. Cremers (2015), 'Variational depth from focus reconstruction', *IEEE Trans. Image Process.* **24**, 5369–5378.
M. Moeller, E. Brinkmann, M. Burger and T. Seybold (2014), 'Color Bregman TV', *SIAM J. Imaging Sci.* **7**, 2771–2806.
M. Moeller, T. Wittman, A. Bertozzi and M. Burger (2012), 'A variational approach for sharpening high dimensional images', *SIAM J. Imaging Sci.* **5**, 150–178.
V. A. Morozov (1966), 'Regularization of incorrectly posed problems and the choice of regularization parameter', *USSR Comput. Math. Math. Phys.* **6**, 242–251.
J. Müller (2013), Advanced image reconstruction and denoising: Bregmanized (higher order) total variation and application in PET. PhD thesis, Westfälische Wilhelms-Universität Münster, Germany.
J. Müller, C. Brune, A. Sawatzky, T. Kösters, K. P. Schäfers and M. Burger (2011), Reconstruction of short time PET scans using Bregman iterations. In *NSS/MIC 2011: IEEE Nuclear Science Symposium and Medical Imaging Conference*, pp. 2383–2385.

D. Mumford and J. Shah (1989), 'Optimal approximations by piecewise smooth functions and associated variational problems', *Commun. Pure Appl. Math.* **42**, 577–685.

V. Nair and G. E. Hinton (2010), Rectified linear units improve restricted Boltzmann machines. In *ICML'10: 27th International Conference on Machine Learning*, pp. 807–814.

M. Z. Nashed and G. Wahba (1974*a*), 'Convergence rates of approximate least squares solutions of linear integral and operator equations of the first kind', *Math. Comp.* **28** (125), 69–80.

M. Z. Nashed and G. Wahba (1974*b*), 'Generalized inverses in reproducing kernel spaces: An approach to regularization of linear operator equations', *SIAM J. Math. Anal.* **5**, 974–987.

M. Z. Nashed and G. Wahba (1974*c*), 'Regularization and approximation of linear operator equations in reproducing kernel spaces', *Bull. Amer. Math. Soc.* **80**, 1213–1218.

F. Natterer (1984), 'Error bounds for Tikhonov regularization in Hilbert scales', *Appl. Anal.* **18**, 29–37.

F. Natterer (2001), *The Mathematics of Computerized Tomography*, SIAM Monographs on Mathematical Modeling and Computation, SIAM.

F. Natterer and F. Wübbeling (2001), *Mathematical Methods in Image Reconstruction*, SIAM.

A. Nemirovskii and D. B. Yudin (1983), *Problem Complexity and Method Efficiency in Optimization*, Wiley-Interscience Series in Discrete Mathematics, Wiley.

A. Neubauer (1988*a*), 'An *a posteriori* parameter choice for Tikhonov regularization in Hilbert scales leading to optimal convergence rates', *SIAM J. Numer. Anal.* **25**, 1313–1326.

A. Neubauer (1988*b*), 'Tikhonov-regularization of ill-posed linear operator equations on closed convex sets', *J. Approx. Theory* **53**, 304–320.

A. Neubauer and H. K. Pikkarainen (2008), 'Convergence results for the Bayesian inversion theory', *J. Inverse Ill-Posed Probl.* **16**, 601–613.

R. Nickl, J. Söhl *et al.* (2017), 'Nonparametric Bayesian posterior contraction rates for discretely observed scalar diffusions', *Ann. Statist.* **45**, 1664–1693.

M. Nikolova and P. Tan (2017), Alternating proximal gradient descent for nonconvex regularised problems with multiconvex coupling terms. hal-01492846v2

P. Ochs, Y. Chen, T. Brox and T. Pock (2014), 'iPiano: Inertial proximal algorithm for nonconvex optimization', *SIAM J. Imaging Sci.* **7**, 1388–1419.

P. Ochs, J. Fadili and T. Brox (2017), Non-smooth non-convex Bregman minimization: Unification and new algorithms. arXiv:1707.02278

P. Ochs, R. Ranftl, T. Brox and T. Pock (2015), Bilevel optimization with nonsmooth lower level problems. In *SSVM 2015: Scale Space and Variational Methods in Computer Vision* (J.-F. Aujol *et al.*, eds), Springer, pp. 654–665.

S. Osher, M. Burger, D. Goldfarb, J. Xu and W. Yin (2005), 'An iterative regularization method for total variation-based image restoration', *Multiscale Model. Simul.* **4**, 460–489.

R. Otazo, E. Candès and D. K. Sodickson (2015), 'Low-rank plus sparse matrix decomposition for accelerated dynamic MRI with separation of background and dynamic components', *Magnetic Resonance Medicine* **73**, 1125–1136.

K. Papafitsoros and C.-B. Schönlieb (2014), 'A combined first and second order variational approach for image reconstruction', *J. Math. Imaging Vision* **48**, 308–338.

N. Parikh and S. Boyd (2014), 'Proximal algorithms', *Found. Trends Optim.* **1**, 127–239.

L. E. Payne (1975), *Improperly Posed Problems in Partial Differential Equations*, Vol. 22 of CBMS-NSF Regional Conference Series in Applied Mathematics, SIAM.

D. L. Phillips (1962), 'A technique for the numerical solution of certain integral equations of the first kind', *J. Assoc. Comput. Mach.* **9**, 84–97.

T. Pock and S. Sabach (2016), 'Inertial proximal alternating linearized minimization (iPALM) for nonconvex and nonsmooth problems', *SIAM J. Imaging Sci.* **9**, 1756–1787.

T. Pock, D. Cremers, H. Bischof and A. Chambolle (2009), An algorithm for minimizing the Mumford–Shah functional. In *ICCV 2009: IEEE 12th International Conference on Computer Vision*, pp. 1133–1140.

M. Prato, S. Bonettini, I. Loris, F. Porta and S. Rebegoldi (2016), 'On the constrained minimization of smooth Kurdyka–Łojasiewicz functions with the scaled gradient projection method', *J. Phys. Conf. Ser.* **756**, 012001.

R. Ranftl, T. Pock and H. Bischof (2013), Minimizing TGV-based variational models with non-convex data terms. In *SSVM 2013: Scale Space and Variational Methods in Computer Vision* (A. Kuijper *et al.*, eds), Springer, pp. 282–293.

J. Rasch, E.-M. Brinkmann and M. Burger (2018), Joint reconstruction via coupled Bregman iterations with applications to PET-MR imaging. *Inverse Problems* **34**, 014001.

J. Rasch, V. Kolehmainen, R. Nivajärvi, M. Kettunen, O. Gröhn, M. Burger and E.-M. Brinkmann (2017), Dynamic MRI reconstruction from undersampled data with an anatomical prescan. arXiv:1712.00099

T. Raus (1984), 'Residue principle for ill-posed problems', *Acta et Comment. Univ. Tartuensis* **672**, 16–26.

T. Raus (1992), 'About regularization parameter choice in case of approximately given error bounds of data', *Acta et Comment. Univ. Tartuensis* **937**, 77–89.

A. J. Reader, J. Matthews, F. C. Sureau, C. Comtat, R. Trébossen and I. Buvat (2007), Fully 4D image reconstruction by estimation of an input function and spectral coefficients. In *IEEE Nuclear Science Symposium Conference*, pp. 3260–3267.

B. Recht, M. Fazel and P. A. Parrilo (2010), 'Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization', *SIAM Review* **52**, 471–501.

M. Reed and B. Simon (1978), *Methods of Mathematical Physics IV: Analysis of Operators*, Elsevier.

E. Resmerita (2005), 'Regularization of ill-posed problems in Banach spaces: Convergence rates', *Inverse Problems* **21**, 1303.

E. Resmerita and O. Scherzer (2006), 'Error estimates for non-quadratic regularization and the relation to enhancement', *Inverse Problems* **22**, 801.

W. Ring (2000), 'Structural properties of solutions to total variation regularization problems', *ESAIM Math. Model. Numer. Anal.* **34**, 799–810.

R. Rockafellar (1972), *Convex Analysis*, Princeton Mathematical Series, Princeton University Press.
Y. Romano, M. Elad and P. Milanfar (2017), 'The little engine that could: Regularization by denoising (RED)', *SIAM J. Imaging Sci.* **10**, 1804–1844.
L. Rondi (2008), 'Reconstruction in the inverse crack problem by variational methods', *European J. Appl. Math.* **19**, 635–660.
S. Roth and M. J. Black (2005), Fields of experts: A framework for learning image priors. In *CVPR 2005: IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 860–867.
L. Rudin, P.-L. Lions and S. Osher (2003), Multiplicative denoising and deblurring: Theory and algorithms. In *Geometric Level Set Methods in Imaging, Vision, and Graphics* (S. Osher and N. Paragios, eds), Springer, pp. 103–119.
L. Rudin, S. Osher and E. Fatemi (1992), 'Nonlinear total variation based noise removal algorithms', *Phys. D: Nonlinear Phenomena* **60**, 259–268.
W. Rudin (2006), *Functional Analysis*, International Series in Pure and Applied Mathematics, McGraw-Hill.
A. Sawatzky, C. Brune, T. Kösters, F. Wübbeling and M. Burger (2013), EM-TV methods for inverse problems with Poisson noise. In *Level Set and PDE Based Reconstruction Methods in Imaging* (M. Burger and S. Osher, eds), Vol. 2090 of Lecture Notes in Mathematics, Springer, pp. 71–142.
O. Scherzer (1993), 'Convergence rates of iterated Tikhonov regularized solutions of nonlinear ill-posed problems', *Numer. Math.* **66**, 259–279.
O. Scherzer (1998), 'Denoising with higher order derivatives of bounded variation and an application to parameter estimation', *Computing* **60**, 1–27.
M. F. Schmidt, M. Benning and C.-B. Schönlieb (2018), Inverse scale space decomposition. *Inverse Problems* **34**, 045008.
U. Schmidt and S. Roth (2014), Shrinkage fields for effective image restoration. In *CVPR 2014: IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2774–2781.
E. Schock (1985), Approximate solution of ill-posed equations: Arbitrarily slow convergence vs. superconvergence. In *Constructive Methods for the Practical Treatment of Integral Equations* (G. Hämmerlin and K. H. Hoffmann, eds), Vol. 73 of International Series of Numerical Mathematics, Springer, pp. 234–243.
F. Schöpfer, A. K. Louis and T. Schuster (2006), 'Nonlinear iterative methods for linear ill-posed problems in Banach spaces', *Inverse Problems* **22**, 311.
T. Schuster, B. Kaltenbacher, B. Hofmann and K. Kazimierski (2012), *Regularization Methods in Banach Spaces*, De Gruyter.
A. Sederman, M. Johns, P. Alexander and L. Gladden (1998), 'Structure-flow correlations in packed beds', *Chem. Engrg Sci.* **53**, 2117–2128.
T. I. Seidman and C. R. Vogel (1989), 'Well posedness and convergence of some regularisation methods for non-linear ill posed problems', *Inverse Problems* **5**, 227.
S. Setzer, G. Steidl and T. Teuber (2011), 'Infimal convolution regularizations with discrete ℓ_1-type functionals', *Comm. Math. Sci* **9**, 797–872.
J.-L. Starck, F. Murtagh and J. M. Fadili (2010), *Sparse Image and Signal Processing: Wavelets, Curvelets, Morphological Diversity*, Cambridge University Press.

D. Strong and T. Chan (2003), 'Edge-preserving and scale-dependent properties of total variation regularization', *Inverse Problems* **19**, S165.

D. Strong, T. Chan *et al.* (1996), Exact solutions to total variation regularization problems. CAM Report 96-41, UCLA.

A. M. Stuart (2010), Inverse problems: A Bayesian perspective. In *Acta Numerica*, Vol. 19, Cambridge University Press, pp. 451–559.

R. Stück, M. Burger and T. Hohage (2011), 'The iteratively regularized Gauss–Newton method with convex constraints and applications in 4Pi microscopy', *Inverse Problems* **28**, 015012.

M. F. Tappen (2007), Utilizing variational optimization to learn Markov random fields. In *CVPR 2007: IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8.

A. Tarantola (2005), *Inverse Problem Theory and Methods for Model Parameter Estimation*, SIAM.

A. Tarantola and B. Valette (1982), 'Inverse problems = quest for information', *J. Geophys.* **50**, 150–170.

A. B. Tayler, D. J. Holland, A. J. Sederman and L. F. Gladden (2012), 'Exploring the origins of turbulence in multiphase flow using compressed sensing MRI', *Phys. Rev. Lett.* **108**, 264505.

M. Teboulle (1992), 'Entropic proximal mappings with applications to nonlinear programming', *Math. Oper. Res.* **17**, 670–690.

G. Teschke and R. Ramlau (2007), 'An iterative algorithm for nonlinear inverse problems with joint sparsity constraints in vector-valued regimes and an application to color image inpainting', *Inverse Problems* **23**, 1851.

J. Thomas King and D. Chillingworth (1979), 'Approximation of generalized inverses by iterated regularization', *Numer. Funct. Anal. Optim.* **1**, 499–513.

A. M. Thompson, J. C. Brown, J. W. Kay and D. M. Titterington (1991), 'A study of methods of choosing the smoothing parameter in image restoration by regularization', *IEEE Trans. Pattern Anal. Machine Intell.* **13**, 326–339.

A. N. Tikhonov (1943), 'On the stability of inverse problems', *Dokl. Akad. Nauk SSSR* **39**, 195–198.

A. N. Tikhonov (1963), 'Solution of incorrectly formulated problems and the regularization method', *Soviet Meth. Dokl.* **4**, 1035–1038.

A. N. Tikhonov (1966), 'On the stability of the functional optimization problem', *USSR Comput. Math. Math. Phys.* **6**, 28–33.

A. N. Tikhonov and V. Y. Arsenin (1977), *Solutions of Ill-Posed Problems*, Winston & Sons.

A. N. Tikhonov, A. Goncharsky and M. Bloch (1987), *Ill-Posed Problems in the Natural Sciences*, Mir.

S. Vaiter, C.-A. Deledalle, G. Peyré, C. Dossal and J. Fadili (2013*a*), 'Local behavior of sparse analysis regularization: Applications to risk estimation', *Appl. Comput. Harmon. Anal.* **35**, 433–451.

S. Vaiter, G. Peyré, C. Dossal and J. Fadili (2013*b*), 'Robust sparse analysis regularization', *IEEE Trans. Inform. Theory* **59**, 2001–2016.

T. Valkonen (2014), 'A primal–dual hybrid gradient method for nonlinear operators with applications to MRI', *Inverse Problems* **30**, 055012.

C. Vogel (2002), *Computational Methods for Inverse Problems*, Frontiers in Applied Mathematics, SIAM.

G. Wahba (1977), 'Practical approximate solutions to linear operator equations when the data are noisy', *SIAM J. Numer. Anal.* **14**, 651–667.

Z. Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli (2004), 'Image quality assessment: From error visibility to structural similarity', *IEEE Trans. Image Process.* **13**, 600–612.

Y. Xu and W. Yin (2013), 'A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion', *SIAM J. Imaging Sci.* **6**, 1758–1789.

Y. Xu and W. Yin (2017), 'A globally convergent algorithm for nonconvex optimization based on block coordinate update', *J. Sci. Comput.* **72**, 700–734.

Y. Yang, J. Ma and S. Osher (2013), 'Seismic data reconstruction via matrix completion', *Inverse Probl. Imaging* **7**, 1379–1392.

W. Yin (2010), 'Analysis and generalizations of the linearized Bregman method', *SIAM J. Imaging Sci.* **3**, 856–877.

W. Yin, S. Osher, D. Goldfarb and J. Darbon (2008), 'Bregman iterative algorithms for ℓ_1-minimization with applications to compressed sensing', *SIAM J. Imaging Sci.* **1**, 143–168.

C. Zach, T. Pock and H. Bischof (2007), 'A duality based approach for realtime TV-L^1 optical flow', *Pattern Recognition: 29th DAGM Symposium* (F. A. Hamprecht *et al.*, eds), Vol. 4713 of Lecture Notes in Computer Science, Springer, pp. 214–223.

L. Zeune, G. van Dalum, L. W. Terstappen, S. A. van Gils and C. Brune (2017), 'Multiscale segmentation via Bregman distances and nonlinear spectral analysis', *SIAM J. Imaging Sci.* **10**, 111–146.

F. Zhao, D. C. Noll, J.-F. Nielsen and J. A. Fessler (2012), 'Separate magnitude and phase regularization via compressed sensing', *IEEE Trans. Medical Imaging* **31**, 1713–1723.

M. Zhu and T. Chan (2008), An efficient primal–dual hybrid gradient algorithm for total variation image restoration. CAM Report 08-34, UCLA.