Multiscale Variational Imaging

Martin Burger Stanley Osher

May 6, 2014

1 Introduction

Our modern world is dominated by visual communication via digital images and videos, which raise a variety of novel questions to mathematics. Tasks like improving image quality (denoising or superresolution), automatically detecting objects in images or videos (segmentation and tracking), or detecting and analyzing movement in videos become of increasing relevance and need to be automized due to the inflation of image data. Among the techniques used for these problems variational methods (often called energy minimization methods in computer vision) play a particularly prominent role.

In this article we discuss some basic properties as well as recent development of variational methods, in particular highlighting multiscale aspects. We start with some basic motivations for the evolution towards the special type of models marking the current state of the art. We then proceed to iterative refinement of variational approaches and finally provide several examples of real-life applications.

2 From Filters to Nonlinear Variational Methods

Let us start our exposition with the classical task of denoising a grey-scale image, modelled as a function $f : \Omega \to \mathbb{R}$. The most classical technique to denoise respectively smoothing the image is a diffusion filtering, i.e. f is used as the initial value for the heat equation

$$\partial_t u = \Delta u \tag{2.1}$$

and $u(T)$ at some final time is used as the denoising result. The diffusion filter constitutes a multiscale method: for small stopping time T mainly small scales in the image are damped out, while for increasing time larger and larger scales are damped, which can be seen easily from a decomposition into eigenfunctions of the Laplace operator. The decomposition is suboptimal for two reasons: First of all the eigenfunctions are smooth wave functions and hence are not well-adapted to edges in an image, which correspond to discontinuities in the grey-value function. Secondly, the scales represented by the coefficients in an eigenfunction expansion are not eliminated one after the other, but are all damped with different exponential factors.

In order to deal with edges nonlinear diffusion filters have been proposed, which use a gradient-dependent diffusivity (cf. [40])

$$\partial_t u = \nabla \cdot (g(|\nabla u|)\nabla u), \tag{2.2}$$

particularly relevant cases being the Perona-Malik filter $g(s) = \frac{1}{1+s^2}$ and the total variation flow $g(s) = \frac{1}{s}$. The decrease of the diffusivity for larger image gradients means that edges are not necessarily removed from the image.

The first link with variational methods is established with a backward Euler discretization of the diffusion equation, the first time step ($\alpha > 0$ small) can be computed from

$$u \in \arg\min_u \left(\frac{1}{2} \int_\Omega (u - f)^2 \, dx + \alpha \int_\Omega G(|\nabla u|) \, dx \right) \tag{2.3}$$

with G defined via $G'(s) = s \; g(s)$. A prominent example is the ROF-model obtained for $G(s) = s$ (cf. [36]). In the simple denoising case the behaviour of such variational models is very similar to diffusion filters, in some relevant cases one can even prove that the time discretization yields the same solution as the flow (cf. [13, 32]). However, variational methods became popular since they are highly flexible compared to filters or others. If one wants to solve a different imaging task one can easily adapt the variational model, e.g. by changing the first - so-called data fidelity term. Examples are inverse problems in imaging where f rather corresponds to an indirect measurement Ku with an operator K such as deblurring or the Radon transform, or statistical noise models, which yield a data fidelity as a negative log likelihood. Overall, a variational approach is of the form.

$$u \in \arg\min_u \left(D(u, f) + \alpha R(u) \right), \tag{2.4}$$

where D is a data fidelity term and R a regularization functional, weighted by a positive parameter $\alpha > 0$. In this way the influence of the data and the model for their generation are incorporated solely in D, while the a-priori knowledge about the image and its structure are incorporated in R, which is to be designed such that likely images yield lower values of the functional. This clear separation allows to adapt variational methods to different imaging tasks and different kinds of images by changing D or R. It is this fact that made variational methods particularly attractive and popular, since analytical insights and computational methods for certain problems can be transfered various other ones in an efficient way.

While the data term can usually derived from physical modelling, e.g. of the image formation process or motion of objects, and statistical modelling, e.g. of noise properties, there are several different approaches to the regularization functional. Several different approaches have evolved in the last two decades:

- Penalization of gradients as above, respectively also of higher derivatives, examples being the total variation or infimal convolutions of total variation with some higher-order functional to obtain decompositions into smooth and discontinuous components.
- Penalization of coefficients in a multiscale basis or frame system like wavelets or shearlets. A prominent example is the norm in the Besov space $B^1_{1,1}$ realized by the ℓ^1-norm wavelet coefficients.
- Penalization of nonlocal derivatives between patches, where patches are equipped with a weighted graph structures. The weights are derived from the low frequency properties of the patches and measure their similarity. A prominent example is the nonlocal means filter, which can be rewritten as a variational problem with quadratic penalties. Recent approaches generalize nonlocal means to variational problems with nonlocal total variation type functionals.

The major mathematical issues related to those are to understand the topological, structural, and computational properties of these regularization energies. The obvious questions are existence and possibly uniqueness of minimizers of (2.4), the stable dependence with respect to the data f, and asymptotic behaviour with respect to the regularization parameter α. We mention that those issues appear to be well-understood for most classical image processing models now, but much less is known in a function-space setting when additional dimensions are added. Those can be time (videos) or spectral dimensions, which are inherently different than spatial dimensions. A common approach in the discrete setting is to interpret such as matrices, where the lines relate to different pixels (spatial variable) and the rows to different time steps or spectral values. Regularization functionals are frequently formulated on matrix properties like the nuclear norm (sum of singular values). A natural continuum counterpart would be the understanding of videos as linear operators from the spatial to the time dimension (or vice versa), but then it comes to subtle issues such as chosing the right function spaces and topologies - an issue which appears completely open currently.

The limit $\alpha \to 0$ is a classical topic in regularization theory (cf. [26]), but recently interest shifted towards a better characterization of solutions for $\alpha > 0$, i.e. the fine properties of minimizers of the variational problem (2.4). Those rely on developing novel techniques with the help from diverse fields, examples being optimality conditions for nonsmooth variational problems in nonreflexive Banach spaces or geometric measure theory and differential geometry all playing important roles just in the case of R being the total variation functional. Another strong line of research is the numerical solution of problems like (2.4), which poses particular challenges since often R is nonsmooth and the data term D might include additional nonlinearity and in some inverse problems the evaluation of a complicated nonlocal operator. We refer to [18] for a further discussion.

Models like (2.3) or (2.4) already inherit some kind of scale via the parameter α. Increasing α yields a stronger smoothing effect and hence eliminates smaller and smaller scales, which is apparent from the (formal) optimality conditon of (2.3), the nonlinear partial differential equation

$$-\alpha \nabla \cdot (G'(|\nabla u|)\frac{\nabla u}{|\nabla u|}) + u = f. \tag{2.5}$$

Hence, one might consider the solutions $u(\alpha)$ of the variational problem as (nonlinearly) smoothed versions of f including smaller and smaller scales as α decreases. However, this is not completely true: Unfortunately variational methods like (2.4) include a systematic bias, such that the decomposition into scales is problematic. This can be seen in a theory of ground states and eigenfunctions of the regularizers, which are the natural definitions of solutions at different scale (cf. [10]). Those are reconstructed only up to a multiplicative constant changing continuously with α.

The bias of standard variational methods does not only prevent a clear multiscale decomposition, it is clearly disadvantageous in many image processing tasks and inverse problems, where errors at all scales are introduced and the resolution of the models is limited unnecessarily. A systematic way to cure this issue is introduced by using Bregman iterations as we shall see in the next section.

3 Bregman Iterations and Multiscale Decomposition

The main idea of Bregman iterations is to successively change the regularization via Bregman distances related to the regularization functional. For a convex functional R a (generalized) Bregman distance is given by

$$D_R^p(v,u) = R(v) - R(u) - \langle p, v-u\rangle \tag{3.1}$$

for a subgradient $p \in \partial R(u)$. Bregman distances for convex functionals are nowadays employed frequently in several areas of optimization (cf. [12, 21]), in machine learning (cf. [22]) for error estimates in inverse problems (cf. [17]), and under the name relative entropy also in partial differential equations (cf. [2], albeit usually in situations where the last term vanishes or $p = 0$). In imaging, Bregman iterations are of particular interest for the frequently used degenerate functionals like the total variation or the ℓ^1-norm of coefficients in certain bases or frames.

The Bregman iteration succesively reconstructs u via

$$u^{k+1} \in \arg\min_u \left(D(u,f) + \alpha D_R^{p^k}(u,u^k) \right), \tag{3.2}$$

with subgradient

$$p^k = p^{k-1} - \frac{1}{\alpha}\partial_u D(u^k, f) \in \partial R(u^k). \tag{3.3}$$

In the frequently investigated case of a quadratic fidelity

$$D(u,f) = \frac{1}{2}\|Ku - f\|^2,$$

with a linear forward operator K, it can be shown that the Bregman iteration is equivalent to the augmented Lagrangian method for the constrained problem of minimizing $R(u)$ subject to $Ku = f$. The Lagrange parameter is then related to p^k via $p^k = K^*\lambda^k$ (cf. [34]). This equivalence is not true for a general fidelity D (still convex with respect to u), interestingly there the augmented Lagrangian corresponds to a Bregman iteration for a dual problem. In any case one can show that in the case of attainable data $f \in \mathcal{R}(K)$, the Bregman iteration converges to the solution of $Ku = f$ with minimal R. Of higher relevance for image reconstruction is however the semiconvergence property of the Bregman iteration. In the case of noisy data relevant in practice, i.e. f differing from the ideal value $K\hat{u}$ by a noise perturbation, the Bregman iteration yields a sequence such that the Bregman distance between u^k and $\hat{u}$ decreases until the data term becomes to small (compareable to the noise measure $D(K\hat{u}, f)$). The advantageous property of the iteration is that in early iteration steps large scales are reconstructed without bias, while finer and finer scales are introduced in further iterations, which results in strongly improved reconstructions. The scale properties will be made more precise below

When interpreting $\tau = \frac{1}{\alpha}$ as a time step, the Bregman iteration can be interpreted as a backward Euler discretization of the time-continuous flow

$$\partial_t p(t) = -\partial_u D(u(t), f), \qquad p(t) \in \partial R(u(t)), \tag{3.4}$$

which was introduced in [15] and termed nonlinear inverse scale space method due to similarities with earlier algorithms (cf. [37]). From a theoretical point of view, the inverse scale space

has the advantage that no additional parameter (compared to α in the Bregman iteration) is needed.

Let us make the scale decomposition more explicit in the case of the simple data term $D(u,f) = \frac{1}{2}\|u-f\|^2$, restricting ourselves to the inverse scale space method for the moment (similar but more tedious computations are possible for the Bregman iteration). Moreover, we assume that R is a convex and one-homogeneous functional, which is the case for all relevant examples. Integrating the flow in time and using $p(0) = 0$ we have

$$p(t) = tf - \int_0^t u(s)\ ds, \qquad p(t) \in \partial R(u(t)). \tag{3.5}$$

The first observation one can make is that there exists a time interval $(0, t_1)$ such that $u \equiv 0$ and $p(t) = tf$. This is a nice result of duality in convex optimization: For one-homogeneous R with trivial nullspace the subdifferential $\partial R(0)$ is a convex set with $p = 0$ in its interior, hence for small t also tf remains inside this set. This relation can be made quite explicit if R is some Banach space norm, since then $t_1 = \frac{1}{\|f\|_*}$, where $\|\cdot\|_*$ denotes the norm of the dual space. At time t_1 the solution u jumps to a nonzero value, which is obtained by solving the constrained problem

$$\|u-f\|^2 \to \min_u \quad \text{subject to} \quad t_1 f \in \partial R(u). \tag{3.6}$$

The multiscale property of the flow is inherent in the value of t_1, which is related to the largest scale part in f. The smaller the largest scale feature contained in f, the larger t_1. This relation becomes explicit when considering eigenfunctions of the regularization, i.e., $\lambda u \in \partial R(u)$, for which the solution of the inverse scale space flow is given by $u \equiv 0$ for $t < t_1 = \lambda$ and $u \equiv f$ for $t \geq t_1$. Hence, there is an exact scale decomposition by the inverse scale space flow, and the eigenvalue relation also clarifies the scale definition based on the regularization functional. Indeed large scales are related to the low frequencies (eigenfunctions for low eigenvalues of the operator R). Further generalizations are possible, e.g. for quadratic data terms $D(u,f) = \frac{1}{2}\|Ku-f\|^2$ (cf. [10]) and for the behaviour of the flow at $t > t_1$ if f is not an eigenfunction (cf. [16, 32, 33]). Explicit results in the latter case are so far restricted to finite-dimensional cases, detailed studies in function spaces remain an interesting future challenge. Moreover, the analysis suggests to further study in detail eigenvalue problems related to the regularization functional, most of which have been understood in the case of R being the total variation of a function u ([5, 10]). For more complicated functionals, e.g. combinations of total variation of first and higher order [11, 38] only preliminary results exist (cf. [9]). Another important direction for future analysis is the development of a similar theory for D being non-quadratic.

After its introduction to the field of image processing and image reconstruction in [34] Bregman iterations, inverse scale space, and related augmented Lagrangian techniques gave an enormous boost to inverse problems, imaging, and data analysis. Some prominent examples are:

- Novel approaches to iterative regularization in Banach spaces and improved methods without systematic bias (cf. e.g. [3, 34, 39])
- Fast methods for total variation regularization and similar problems (cf. e.g. [29, 43, 19])

- Fast methods for compressed sensing (cf. e.g. [20]), i.e. for solving problems of the form

$$\|x\|_1 \to \min_{x\in\mathbb{R}^n} \quad \text{subject to} \quad Ax = f \tag{3.7}$$

with sensing matrix $A \in \mathbb{R}^{m\times n}$, $m << n$. A particularly interesting result is the fact that the inverse scale space method can be computed exactly for such problems (and a large class of related discrete problems, [16, 33]) and yields an effective methods for computing very sparse solutions.

The fast convergence of these Bregman type methods is an ℓ_1 related phenomenon. This fact accounts for the resurgence in popularity of these methods in recent years. There are a few technical points which help explain this success.

First we note that if $R(u)$ is homogeneous of degree one, $R(cu) = |c|R(u)$, then

$$D_R^p(v,u) = R(v) - \langle p(u), v\rangle$$

(we use $p(u)$ to denote an element of $\partial R(u)$). In the special case where $R(x) = \|x\|_1,\ x\ \varepsilon\ \mathbb{R}^n$, we have

$$D_R^p(y,x) = \sum_i (|y_i| - y_i p_i(x_i))$$

where we use $p_i(x_k)$ as the i^{th} component of $p(x)$.

The interesting fact is that this Bregman distance vanishes if, for every nonzero component y_i, the corresponding component x_i has the same sign as y_i. So if y is a sparse vector, i.e. has only a few nonzero components, then only a few signs of the components of x have to match up with those of y.

This leads us to an "error forgetting" property of Bregman iteration. Consider (3.2) with

$$D(u,f) = \frac{1}{2}\|Ku - f\|^2.$$

Suppose we arrive at an iterate u^k which has the property that

(i) $D_R^{p^k}(u^+, u^k) = 0$ for some u^+ for which $Ku^+ = f$

(ii) $p(u^k) = K^*v$ for some v, K^* is the adjoint of K.

Then the next iterate, u^{k+1} is a minimizer of $R(u)$ such that $Ku = f$.

The proof is very simple and can be found in [44].

This means, for ℓ_1 regularization, that we need only line up the signs of the components of u^k to agree with the corresponding nonzero components of u^+, and the subgradients to satisfy (ii). The first criterion is very relevant and easily satisfied in compressed sensing type problems.

Note that for strictly convex R, the corresponding Bregman distance vanishes iff $u^* = u^k$ which isn't very interesting . Also note that the rapid convergence phenomenon seems to be valid also for R total variation, indeed for any reasonable R which is homogeneous of degree one, see e.g. [29].

- There has been a lot of interest in the recovery of an unknown low rank or approximately low rank matrices from incomplete information. One example is matrix completion - this arises, for example from partially filled out surveys. The desired unknown matrix is assumed to be of low rank. Let $M \varepsilon R^{m\times n}$ be a low rank matrix whose rank r satisfies $r << \min(m,n)$. It was shown in [45] that most low rank matrices M can be recovered by solving the optimization problem

$$\min_X \|X\|_*$$

such that $X_{ij} = M_{ij}$ for i,j in a given set R. The quantity $\| \ \|_*$ is the nuclear norm, i.e., the sum of the singular values. This is a convex optimization problem.

Another, related matrix decomposition problem involves decomposing M as a sum of a low rank plus a sparse matrix

$$M + L + S.$$

The convex optimization problem is

$$\min \|L\|_* + \lambda \|S\|_1$$

such that $M = L + S$.

Here $\|S\|_1$ is just the sum of absolute values of the entries and λ is a tuning parameter. This problem is of interest in many areas. The $\| \ \|_1$ norm is often replaced by total variation if we are looking for sparse gradients.

Both of these problems can be solved using Bregman methods e.g. [46] with fairly rapid convergence. The slowest part of these algorithms always seems to be computing the nuclear norm of a large matrix at each iteration. Attempts to overcome this difficulty can be found in [47],[48] and elsewhere.

- Statistical learning is widely used for feature selection. The ℓ_1 regularized logistic regression is a popular decoder. The inputs are a set of training data $X = [x_1, \ldots, x_m]^T \varepsilon R^{m\times n}$ where each row of X is a sample. We seek a hyperplane $\{x : w^T x + v = 0\}$ that separates the data belonging to two classes. The ℓ_1 regularized logistic regressor is found via

$$\arg\min_{w,v} \|w\|_1 + \lambda \ell(w,v).$$

For $\theta(z) = \log(1 + \exp(-z)$

$$\ell(w,v) = \frac{1}{m} \sum_{i=1}^{m} \theta(w^T x_i + v) y_i.$$

The regularization parameter λ determines the level of sparsity. One typically determines a path by varying λ_j, solving a sequence of coresponding minimization problems, and choosing an optimal λ via cross validation. Although each of these optimization problems can be solved using split Bregman fairly quickly, there is a bias (analogous to loss of contrast in images) and it seems wasteful to do these computations.

An alternative is to use the linearized Bregman algorithm which can be written (for the w minimization) as

$$p(w^{k+1}) - p(w^k) + \frac{1}{\mu}(w^{k+1} - w^k) + \lambda \partial_w \ell(w^k, v) = 0$$

for $\mu > 0$ and as large as makes sense.

This leads to an approximate path for increasing k which has desirable properties. In particular no inversion is needed. There is some connection here between inverse scale space which could be written here (ignoring v) as:

$$\frac{\partial p}{\partial t}(w) + \frac{1}{\mu}\frac{\partial w}{\partial t} = -\partial_w \ell$$

versus solving for a sequence of values $t > 0$

$$\frac{p(w)}{t} + \frac{w}{\mu t} = -\partial_w \ell.$$

See [49]

Other statistical learning algorithms can also benefit from ℓ_1 regularization.

These methods have led to practical improvements in many other applications, a glimpse of which we shall provide in the next section.

4 Impact in Applications

4.1 Biomedical Imaging: Fast Acquisition

A major aim in modern biomedical imaging is to acquire images faster and faster, for several reasons. The first is the development towards dynamic imaging and the need to resolve smaller and smaller time steps. A second is the increase of clinical imaging examinations and the resulting pressure to minimize time per examination. Thus, one naturally deals with problems with few data, but the need to reconstruct major features as with high quality data. Many of the relevant modalities, e.g. in optical devices or emission tomography, are based on counting photons (indirectly) related to the image intensity. The number of photons per detector can be modeled as a Poisson-distributed random variable, whose mean naturally scales with the measurement time. Thus, the data are not undersampled with low additive noise as assumed in the compressed sensing literature, but indeed the signal-dependent noise is the main source of undersampling. In order to deal with the statistical nature of the measurements it is nowadays common to use the negative log-likelihood for the statistical noise model as data term in variational methods. The recent development and analysis of Bregman iteration techniques for such problems and their adaption to practice allows a dramatic reduction of measurement time at compareable quality of reconstruction for the main features. A similar step is not possible with standard variational methods due to their systematic errors (bias). Note that for low count data strong regularization is needed, which yields strong bias

The potential of reducing acquisition time is illustrated in Figure 1 for an example in Positron-Emission-Tomography (PET). Here the forward operator is the Radon transform of the image (with some additional corrections, cf. [41]) and the statistical model with Poisson

distributions is well established. Regular examinations can be quite time consuming, ranging from several minutes up to one hour, and of course it would be desirable to reduce the acquisition time to few minutes or even less. This is driven to the extreme in the example in Figure 1, where the standard measurement interval of 30 minutes used on the left is reduced to 5 seconds using total variation regularization and the Bregman iteration on the right. The main features are reconstructed with compareable quality, which can be made quite precise in studies on software and hardware phantoms (cf. [24]). Examples of other modalities benefiting from such multiscale variational techniques are superresolution microscopy (cf. e.g. [14]), optical tomography (cf. [1]), and of course MR imaging (cf. e.g. [7]).

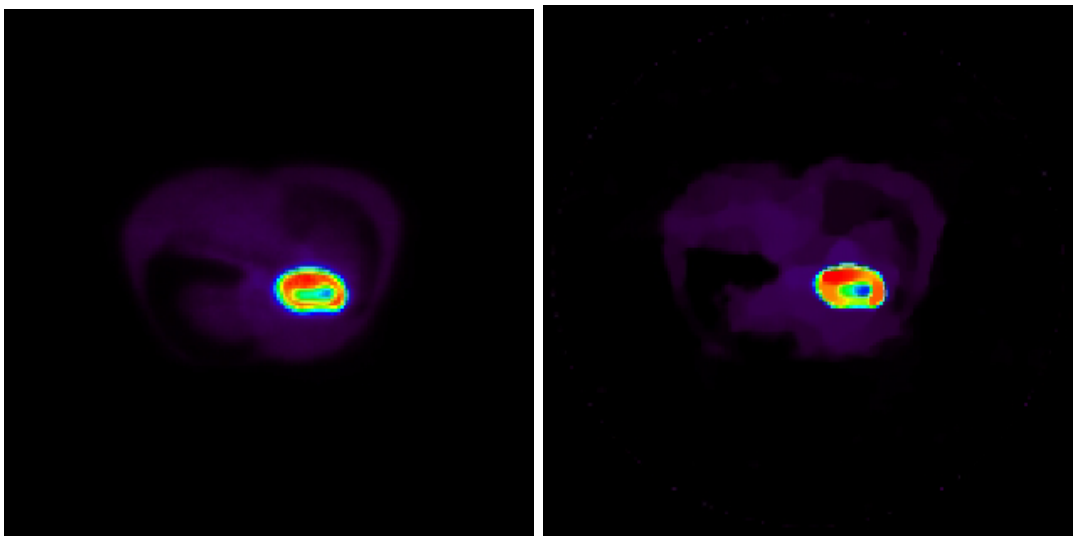

Figure 1: Illustration of image reconstruction from highly undersampled data via Bregmanized total variation regularization in PET. The image on the left is a reconstruction with high photon counts, using the standard measurement time interval of 30 minutes. The image on the right is a reconstruction using only the photon counts of the first 5 seconds, but using total variation regularization and Bregman iteration. The quality of the reconstruction from high count data is restored. Both images show representative slices through of a 3D reconstruction. From [24].

These applications also drive further mathematical research, e.g. they also initiated the need to study those techniques for others than least squares data terms. Key questions remaining for the future are a detailed analysis and understanding of recovery from Poisson-distributed data (cf. [35] for preliminary results) and efficient computational techniques, respectively modifications of the Bregman iteration that avoid too many (costly) evaluations of the forward operator K.

4.2 Hyperspectral Imaging and Unmixing

Hyperspectral imaging (HSI) sensors record up to several hundred different frequencies. The spatial resolution is low and there are multiple materials at a single location. Multiscale variational imaging techniques can be used for clustering, finding anomalies, change detection as well as filling in missing data and, of course, removing noise.

Here we will discuss blind unmixing: The task of determining the abundances of different materials in each pixel is called spectral unmixing. This ill-posed problem usually uses a dictionary with the spectral signatures of the possible materials (denoted as endmembers). In [28] a method was developed for simultaneously detecting the endmembers and computing abundances. One first sets up a matrix $X \varepsilon \mathbb{R}^{m,d}$ where each column of X is the spectral signature of one pixel in the image. Here, m is the number of spectral bands and, d is the number of pixels.

The idea is to look for endmember maps. These can be written as

$$X = AS \quad A, S \geq 0, A \varepsilon R^{m,n} \quad S \varepsilon R^{n,d}$$

We want n to be as small as possible. These columns of A are each endmembers. We require nonnegativity since every pixel contains a nonnegative amount of each column of A.

Once again a relevant matrix optimization problem involving ℓ_1 regularization is used to find A and can be solved using Bregman iteration with good results. The results are illustrated by Figure 2.

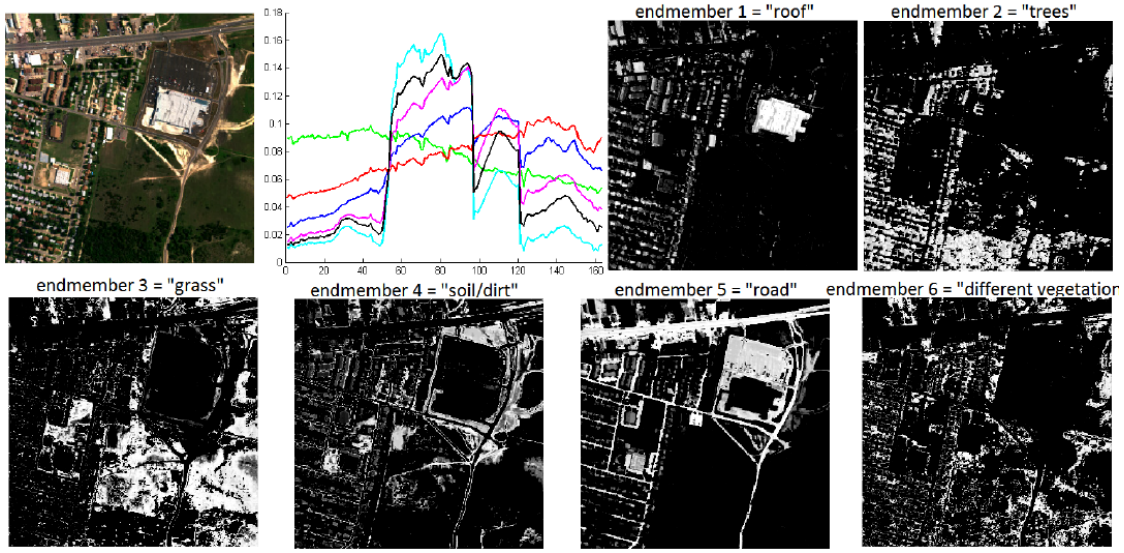

Figure 2: Decomposition of a hyperspectral image of an urban scene. From [28].

References

[1] J. F. Abascal, J. Chamorro-Servent, J. Aguirre, J. J. Vaquero, S. Arridge, T. Correia, J. Ripoll amd M. Desco, Fluorescence diffuse optical tomography using the split Bregman method, *Med. Phys.* 38 (2011), 6275.

[2] A. Arnold, et al. Entropies and equilibria of many-particle systems: an essay on recent research. In: *Nonlinear Differential Equation Models*, Springer Vienna, 2004, 35-43.

[3] M. Bachmayr and M. Burger, Iterative total variation methods for nonlinear inverse problems, *Inverse Problems* 25 (2009), 105004.

[4] H. H.Bauschke, J. M. Borwein, P. L. Combettes, Bregman monotone optimization algorithms, *SIAM Journal on control and optimization* 42 (2003), 596-636.

[5] G. Bellettini, V. Caselles and M. Novaga, Explicit solutions of the eigenvalue problem $div\left(\frac{Du}{|Du|}\right) = u$ in R^2, *SIAM J. Math. Anal.* 36 (2005), 10951129.

[6] M. Benning, *Singular Regularization of Inverse Problems*, Ph.D. thesis, Institute for Computational and Applied Mathematics, University of Münster, Germany, may 2011.

[7] M. Benning, L. Gladden, D. Holland, C. Schönlieb, T. Valkonen, Phase reconstruction from velocity-encoded MRI measurements - A survey of sparsity-promoting variational approaches, *Journal of Magnetic Resonance* 238 (2014), 26 - 43.

[8] M. Benning and M. Burger, Error Estimates for General Fidelities, *Electronic Transactions on Numerical Analysis* **38** (2011), 44–68.

[9] M. Benning, C. Brune, M. Burger and J. Müller, Higher-order TV methods enhancement via Bregman iteration, *J. Sci. Comp.* 54 (2013), 269-310.

[10] M. Benning and M. Burger, Ground states and singular vectors of convex variational regularization methods, *Meth. Appl. Analysis* 20 (2013), 295-334.

[11] K. Bredies, K. Kunisch and T. Pock, Total generalized variation, *SIAM J. Img. Sci.* **3** (2010), 492–526.

[12] L. M. Bregman, The relaxation method for finding the common point of convex sets and its application to the solution of problems in convex programming, *USSR Comp. Math. Math. Phys.* 7 (1967), 200217.

[13] A. Briani, A. Chambolle, M. Novaga, G. Orlandi, On the gradient flow of a one-homogeneous functional, *Confluentes Mathematici* 3 (2011), 617-635.

[14] C. Brune, A. Sawatzky, M. Burger, Bregman-EM-TV methods with application to optical nanoscopy, in *Proceedings of the 2nd International Conference on Scale Space and Variational Methods in Computer Vision*. LNC, vol. 5567 (Springer, Berlin, 2009), 235246

[15] M. Burger, G. Gilboa, S. Osher and J. Xu, Nonlinear inverse scale space methods, *Comm. Math. Sci.* 4 (2006), 179-212.

[16] M. Burger, M. Möller, M. Benning and S. Osher, An adaptive inverse scale space method for compressed sensing, *Mathematics of Computation* 82 (2013), 269-299.

[17] M. Burger and S. Osher, Convergence rates of convex variational regularization, *Inverse Problems* **20** (2004), 1411–1421.

[18] M. Burger and S. Osher, *A guide to the TV zoo*, in: M.Burger, S.Osher, eds., Level-Set and PDE-based Reconstruction Methods, Springer, 2013.

[19] J. Cai, S. Osher and Z. Shen, Split Bregman Methods and Frame Based Image Restoration, *Multiscale Model. Simul.* **8** (2010), 337–369.

[20] J. Cai, S. Osher and Z. Shen, Linearized Bregman iterations for compressed sensing. *Math. Comput.* 78 (2008), 15151536.

[21] Y. Censor, *Parallel Optimization: Theory, Algorithms, and Applications*, Oxford University Press, 1997.

[22] M. Collins, R. E. Schapire and Y. Singer, Logistic regression, AdaBoost and Bregman distances, *Machine Learning* 48 (2002), 253-285.

[23] M. A. Davenport, M. F. Duarte, Y. C. Eldar and G. Kutyniok, *Introduction to compressed sensing*, Compressed sensing, Cambridge Univ. Press, Cambridge, 2012, pp. 1–64.

[24] S.Dill, Reconstruction Methods for Positron Emission Tomography with Short Scan Time, BSc Thesis, WWU Münster, 2013.

[25] D. L. Donoho, Compressed sensing, *IEEE Trans. Inform. Theory* **52** (2006), 1289–1306.

[26] H. W. Engl, M. Hanke and A. Neubauer, *Regularization of inverse problems*, Mathematics and its Applications 375, Kluwer Academic Publishers Group, Dordrecht, 1996.

[27] E. Esser, *Primal Dual Algorithms for Convex Models and Applications to Image Restoration, Registration and Nonlocal Inpainting*, PhD thesis, UCLA, 2010.

[28] E. Esser, M. Möller, S. Osher, G. Sapiro, and J. Xin, A Convex Model for Nonnegative Matrix Factorization and Dimensionality Reduction in Physical Space, *Trans. on Image Proc.*, 21 (2012), 3239–3252.

[29] T. Goldstein and S. Osher, The Split Bregman Method for L1-Regularized Problems, *SIAM J. Img. Sci.* **2** (2009), 323–343.

[30] T. Le, R. Chartrand and T. Asaki, A Variational Approach to Reconstructing Images Corrupted by Poisson Noise, *J. Math. Imaging Vision* **27** (2007), 257–263.

[31] Y. Meyer, *Oscillating patterns in image processing and nonlinear evolution equations. The fifteenth Dean Jacqueline B. Lewis memorial lectures*, University Lecture Series 22, American Mathematical Society, Providence, RI, 2001.

[32] M. Möller, *Multiscale Methods for (Generalized) Sparse Recovery and Applications in High Dimensional Imaging*, Ph.D. thesis, Institute for Computational and Applied Mathematics, University of Münster, Germany, july 2012.

[33] M. Möller, M. Burger, Multiscale methods for polyhedral regularizations, *SIAM J. Optim*, 23 (2013), 1424-1456.

[34] S. Osher, M. Burger, D. Goldfarb, J. Xu and W. Yin, An iterative regularization method for total variation-based image restoration, *Multiscale Model. Simul.* **4** (2005), 460–489.

[35] M. Raginsky, R. Willett, Z. Harmany and R. Marcia. Compressed sensing performance bounds under Poisson noise, *IEEE Transactions on Signal Processing* 58 (2010), 3990-4002.

[36] L. I. Rudin, S. Osher and E. Fatemi, Nonlinear total variation based noise removal algorithms, *Physica D* **60** (1992), 259–268.

[37] O. Scherzer, C. Groetsch, Inverse scale space theory for inverse problems, in: *Scale-Space and Morphology in Computer Vision*. Proceedings of the 3rd International Conference Scale-space 2001, ed. by M. Kerckhove (Springer, Berlin, 2001), 317325

[38] S. Setzer, G. Steidl and T. Teuber, Infimal convolution regularizations with discrete ℓ_1-type functionals, *Commun. Math. Sci.* **9** (2011), 797–827.

[39] F. Schoepfer, A. K. Louis and T. Schuster, Nonlinear iterative methods for linear ill-posed problems in Banach spaces, *Inverse Problems* 22 (2006), 311329

[40] J. Weickert, *Anisotropic Diffusion in Image Processing*, Teubner, Stuttgart, 1998.

[41] M. N. Wernick and J. N. Aarsvold (eds.), *Emission Tomography: The Fundamentals of PET and SPECT* (Academic Press, San Diego, 2004)

[42] C. Wu and X. Tai, Augmented Lagrangian Method, Dual Methods, and Split Bregman Iteration for ROF, Vectorial TV, and High Order Models, *SIAM J. Img. Sci.* 3 (2010), 300–339.

[43] X. Zhang, M. Burger and S. Osher, A Unified Primal-Dual Algorithm Framework Based on Bregman Iteration, *Journal of Scientific Computing* 46 (2011), 20–46.

[44] W. Yin and S. Osher, Error Forgetting of Bregman Iteration, *J. Sci. Comput.* 54 (2013), 684–695.

[45] E. Candes and B. Recht, Exact Matric Completion via Convex Optimization, *Foundation of Computational Mathematics* 9 (2009), 717-772.

[46] J.-F. Cai, E.J. Candes, and Z. Shen, A Singular Value Thresholding Algorithm for Matrix Completion, *SIAM J. Optimiz.* 20 (2010), 1956–1982.

[47] J.-F. Cai and S. Osher, Fast Singular Value Thresholding Without Singular Value Decompositions, *UCLA CAM Report 10-24*, (2010).

[48] Y. Xu, W. Yin, Z. Wen, and Y. Zhang, An Alternating Direction Algorithm for Matrix Completion With Nonnegative Factors, *Frontiers Math. China*, 7 (2012), 365–384.

[49] J. Shi, W. Yin, and S. Osher, A Fast Hybrid Algorithm for Large Scale ℓ_1 Regularized Logistic Repression, *J. Mach Learning Research* 11 (2010), 713–741.