Weierstraß-Institut
für Angewandte Analysis und Stochastik

Leibniz-Institut im Forschungsverbund Berlin e. V.

Preprint ISSN 2198-5855

Generating structured non-smooth priors and associated primal-dual methods

Michael Hintermüller[1,2], Kostas Papafitsoros[1]

submitted: July 31, 2019

[1] Weierstrass Institute
Mohrenstr. 39
10117 Berlin
Germany
E-Mail: michael.hintermueller@wias-berlin.de
kostas.papafitsoros@wias-berlin.de

[2] Humboldt-Universität zu Berlin
Unter den Linden 6
10099 Berlin
Germany
E-Mail: hint@math.hu-berlin.de

No. 2611
Berlin 2019

2010 *Mathematics Subject Classification.* 94A08, 68U10, 49K20, 49M37, 49M15, 26A45.

Key words and phrases. Non-smooth priors, image processing, total variation, total generalized variation, bilevel optimization, regularization parameter selection.

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Leibniz-Institut im Forschungsverbund Berlin e. V.
Mohrenstraße 39
10117 Berlin
Germany

Fax: +49 30 20372-303
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

Generating structured non-smooth priors and associated primal-dual methods

Michael Hintermüller, Kostas Papafitsoros

Abstract

The purpose of the present chapter is to bind together and extend some recent developments regarding data-driven non-smooth regularization techniques in image processing through the means of a bilevel minimization scheme. The scheme, considered in function space, takes advantage of a dualization framework and it is designed to produce spatially varying regularization parameters adapted to the data for well-known regularizers, e.g. Total Variation and Total Generalized variation, leading to automated (monolithic), image reconstruction workflows. An inclusion of the theory of bilevel optimization and the theoretical background of the dualization framework, as well as a brief review of the aforementioned regularizers and their parameterization, makes this chapter a self-contained one. Aspects of the numerical implementation of the scheme are discussed and numerical examples are provided.

1 Introduction

1.1 Context

Non-smooth regularization functionals have played a central role in the field of mathematical imaging since the introduction of Total Variation (TV) to basic image reconstruction problems in the 1990's [42, 126]. These functionals which, in the function space setting, involve (Radon) norms of distributional derivatives are used as priors in Tikhonov-like variational regularization problems. Their discrete analogues involve versions of ℓ_1 norms thus promoting certain sparsity properties and resulting in noise and artifact reductions as well as in edge-preserving reconstructions. Particularly, the latter property has made this family of regularization functionals a popular tool in the field of image reconstruction and inverse problems.

These inverse problems are typically of the form: Given data

$$f = \mathrm{Noisy}(Tu_{true}), \tag{1.1}$$

where u_{true} represents some *ground truth* image, that we wish to obtain, find a good approximation u of u_{true}. Here, u_{true} as well as u are modelled as functions from some domain $\Omega \subset \mathbb{R}^d$ into $\mathbb{R}^m$. For static black and white images u we have that $d = 2$, i.e., Ω is a two dimensional domain, typically open, connected with Lipschitz boundary (e.g. a rectangle), and u is real-valued, i.e. $m = 1$. In this sense $u(x)$ models the intensity of the image at the point $x \in \Omega$. Generalizations to domains of higher dimension (3D, time-dependent problems) as well as multivalued images (multicolor, multimodal) are also possible. But for the sake of ease of presentation of the essentials, we focus here on real-valued images over a 2D domain, only. We denote the function space which u_{true} belongs to by $\mathcal{X}$. As indicated above, instead of u_{true} typically only some data f are available which correspond to a degraded version of u_{true}. This degradation, comes from two sources as (1.1) depicts. First, the application of

DOI 10.20347/WIAS.PREPRINT.2611 Berlin, July 31, 2019

an operator $T : \mathcal{X} \to \mathcal{Y}$, the *forward operator* reflects the transformation that u_{true} goes through along with possible further limitations. For instance, in Magnetic Resonance Imaging (MRI) T stands for a subsampling of the coefficients associated with the Fourier transform of the magnetization of the tissue; see for instance [60, 100, 107]. In tomography, for example, this transformation is given by the Radon transform, where typically only a relatively small number of line integrals is available [113]. Apart from transformation and incomplete measurements, a second level of degradation comes from the presence of random noise. It is the result of measurement or transmission errors, e.g., due to the heating up of digital sensors, or other sources of inaccuracies that are too complex to be fully quantified or that involve random effects. This noise can be additive, multiplicative, mixed or occur in an even more complicated way in the measurement process. Additive noise in (1.1) leads to

$$f = Tu_{true} + \eta, \tag{1.2}$$

where η represents a highly oscillatory function in some space $\mathcal{Y}$. Typically it is assumed that the mean of η is zero; as otherwise a systematic deterioration is detected and remedied by suitable calibration. In variational image processing, one then aims to recover u_{true} or at least a *sensible* approximation thereof by solving a minimization problem of the type

$$\text{minimize} \quad \Phi(Tu, f) + J_{\boldsymbol{\alpha}}(u) \quad \text{over} \quad u \in \tilde{\mathcal{X}}, \tag{1.3}$$

where $\tilde{\mathcal{X}}$ is a subspace of $\mathcal{X}$ containing functions whose regularity is typically dictated by $J_{\boldsymbol{\alpha}}$. The term $\Phi(Tu, f)$ is a *data fitting* or *fidelity term* term which measures the distance between the data f and the possible reconstruction u after the action of the forward operator. This ensures that the potential reconstruction will be close to u_{true} in a certain sense. The proper choice for Φ is often the consequence of the statistics of the noise. For instance, for Gaussian noise $\frac{1}{2}\|Tu - f\|^2_{L^2(\Omega)}$ is a suitable choice [42, 126], for salt-and-pepper noise $\|Tu - f\|_{L^1(\Omega)}$ is preferable [15, 47, 65, 116], while for Poisson noise the Kullback-Leibler divergence $\int_\Omega u - f \log u \, dx$ is more appropriate [32, 128]. Mixtures of noise can also be treated and are typically addressed by infimal convolutions of the single-noise fidelity terms; see for instance [36, 38] as well as the references therein. From a statistical viewpoint, the choice of the data fidelty is often motivated by maximum likelihood considerations concerning the underlying noise type. For the sake of simplicity, in the examples below we confine ourselves to an L^2 fidelity term.

Minimizing just the fidelity term will not give any sensible results as this corresponds to the direct inversion of T. Due to the presence of noise and potential deficiencies in T this is, however, typically an ill-posed problem, in the sense that small variations in the data f may have an enormous deteriorating effects on u. This adverse behavior is taken care of by an additional term $J_{\boldsymbol{\alpha}}$, the *regularizer*, which makes the overall minimization problem well-posed. Typically it adds some extra regularity requirement to the problem. This leads to a robust reconstruction where ideally the noise is filtered out, as well.

The parameter $\boldsymbol{\alpha}$ is the *regularization parameter* and balances the effect of the two terms in (1.3) thus determining the amount of regularization or filtering in the reconstructed image. It can be one or more (positive) scalar quantities, or spatially varying functions. In the first case, the regularization effect is uniform throughout the image (global effect) while in the second, this effect varies depending on the magnitude of this function in different areas of the domain (local effect). We mention already here that while in simple regularizers where the regularization parameter is included in a multiplicative way the regularization effect scales with the magnitude of $\boldsymbol{\alpha}$, this is not the case for more complex regularizers that involve, e.g., infimal convolutions, where two or more scalar parameters determine the regularization effect in a more complex fashion. This makes their proper selection an even more challenging task.

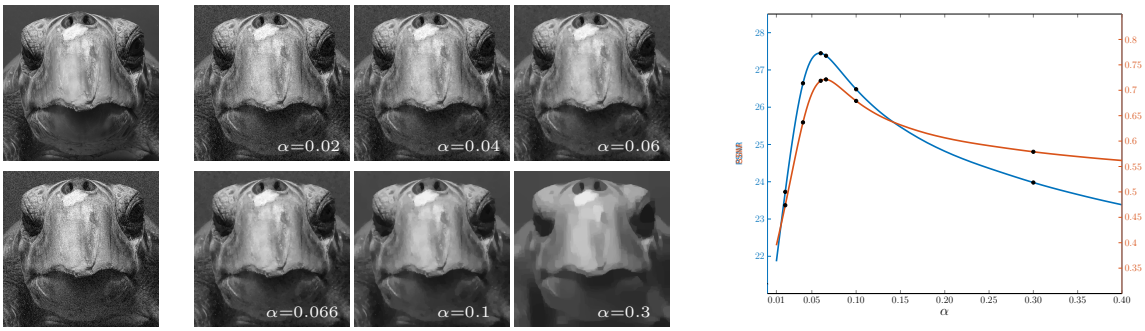

Figure 1: TV denoising examples for different values of the scalar regularization parameter α. Left: Clean and noisy images. Middle: TV denoised images, solutions of the problem (1.4) with $T = Id$. The parameter α is increasing from left to right and top to bottom, corresponding to the marked points in the right figure. Right: Plots that depict the PSNR and SSIM values of the solutions versus the parameter α. The third and fourth marked points correspond to the highest PSNR and SSIM respectively

Indeed, even with a proper selection of both, the fidelity and regularization functionals, the problem (1.3) contains one more degree of freedom, which is connected to a suitable choice of $\boldsymbol{\alpha}$. It aims at producing the best reconstruction result within the capabilities of the selected Φ and J. Establishing automatic regularization parameter selection rules with the help of rich and rigorous mathematical theories that eventually lead to *monolithic* image reconstruction methods approaches lies at the heart of this chapter.

At this point we wish to pin down several of the aforementioned concepts by considering the following classical TV minimization problem

$$\min_{u\in \mathrm{BV}(\Omega)} \frac{1}{2}\|Tu - f\|^2_{L^2(\Omega)} + \alpha |Du|(\Omega). \tag{1.4}$$

Without getting into too many details, we note for the time being only that here $\mathcal{Y} = L^2(\Omega)$, $\tilde{\mathcal{X}} = \mathrm{BV}(\Omega)$ the (non-reflexive Banach) space of *functions of bounded variation* and that $J_{\boldsymbol{\alpha}}(u) = \alpha|Du|(\Omega)$, the total variation of u (also denoted by $\mathrm{TV}(u)$), multiplied by α. Here Du denotes the finite Radon measure that represents the distributional derivative of u and $|Du|$ the corresponding total variation measure of Du. The regularization parameter here is a simple positive scalar, i.e., $\boldsymbol{\alpha} = \alpha > 0$. For precise definitions and properties of TV and $\mathrm{BV}(\Omega)$ we refer to Section 2.1. The problem (1.4) and its variants have been thoroughly used in many applications, like denoising [19, 40, 75, 68, 126], deblurring [69, 142], inpainting [49, 70], zooming [110], image decomposition [48, 145] MRI and PET reconstruction [17, 33, 127], image decompression [6, 24], dejittering [61, 62] to name a few. The solution structure and its dependence on the parameter α are also well-studied and understood [1, 2, 3, 4, 39, 41, 47, 65, 111, 115, 125, 139].

As it is observed in practice, but also proven theoretically in the works above, the strength of regularization due the TV term is proportional to the parameter α, with large values resulting in cartoon-like images of large constant areas and small values having little regularizing effect, meaning that the solution u is such that Tu converges to the data f in an appropriate sense as $\alpha \to 0$. In Figure 1, one can see a simple TV denoising example ($T = Id$) where this behaviour is depicted. Naturally, one is interested in those values of α that produce a (close to) the *best* result with respect to some quality measure. For the peak-signal-to-noise ratio (PSNR) and the Structural Similarity Index (SSIM) see the two plots on the right hand side of Figure 1.

As far as TV minimization with a scalar regularization parameter is concerned, there exists several works in the literature that focus on its automatic selection, see for instance [12, 57, 103, 105] and the

references therein. However these works focus mainly on a scalar α which poses restrictions on the reconstruction quality as we discuss later.

A series of works relevant to this problem as well as to the present survey, was initiated in [55] and at the discrete setting in [102], and further explored in [35, 37]. There, an optimal parameter for the TV minimization is computed by means of variants of the following *bilevel optimization* scheme

$$\begin{cases} \text{minimize} & \frac{1}{2}\|u_\alpha - u_{true}\|^2_{L^2(\Omega)} \quad \text{over } \alpha \geq 0, \\ \text{subject to} & u_\alpha = \underset{u \in H^1(\Omega)}{\operatorname{argmin}} \frac{1}{2}\|u - f\|^2_{L^2(\Omega)} + \alpha\|\nabla u\|_{L^1(\Omega)} + \frac{\epsilon}{2}\|\nabla u\|^2_{L^2(\Omega)}. \end{cases} \tag{1.5}$$

Here it is assumed that the ground truth u_{true} is known and the scheme aims at finding the scalar α that produces the reconstruction u_α that is closest to u_{true} in the L^2 sense (optimal PSNR). The reconstruction u_α is a solution to a smoothed TV reconstruction problem, where smoothing is necessary for the numerical solution. The rationale is then to use this parameter to restore data g that were produced under the same circumstances as f, as this parameter is expected to be near optimal for g, as well.

One of the drawbacks of TV minimization is the *staircasing effect*, i.e., the promotion of piecewise constant (or blocky) structures [41, 125], stemming from the sparsity of the gradient which, in the discrete setting, is due to the use of the ℓ_1 norm. Although local smoothings of the TV functional, e.g. of Huber type [34, 75] can reduce this effect, a reduction or even total elimination is typically achieved via the incorporation of higher order derivatives in the regularization process. For instance see the contributions [42, 46, 74, 104, 108, 109, 121, 130, 132, 133, 135, 146] and also [119] for a more general review. Arguably, one the most successful regularizers of this type is the Total Generalized Variation of second order (TGV) [28], defined as

$$\mathrm{TGV}^2_{\boldsymbol{\alpha}}(u) = \min_{w \in \mathrm{BD}(\Omega)} \alpha_1 |Du - w|(\Omega) + \alpha_0 |\mathcal{E}w|(\Omega), \tag{1.6}$$

where $\mathrm{BD}(\Omega)$ is the space of *functions of bounded deformation* and $\mathcal{E}w$ is the distributional symmetrized gradient of w; see Section 2.2 below for definitions. TGV has the ability to adapt to the regularity of the images resulting in piecewise affine reconstructions where edges are still preserved. It has already been successfully used in a variety of applications; see [22, 25, 30, 95, 100, 141] among others. Note that here the set of regularization parameters consists of two positive scalars $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)$ that balance the effect of the first and higher order terms in (1.6). Here the regularization strength is decided in a more complicated way than in the TV case; see for instance Figure 2. There exist results regarding the influence of these parameters on the structure of solutions, as well as their asymptotic behaviour [29, 120, 122, 123]. These studies have mostly theoretical value providing little practical guidance on how to select the parameters for (near) optimal results. We note that adaptation of the scheme (1.5) was done for TGV in [54, 56], producing optimal parameters with the use of pairs of ground truths and noisy images. Thus, up to now the choice of TGV parameters is mostly based on heuristics only.

So far we have mentioned regularization functionals whose parameters are scalar quantities. As we have already briefly alluded to above, this implies that the regularization strength is uniform across the image. This is unwanted when the amount of noise is non-uniform across the image. Moreover, it is also beneficial to impose weak regularization in image regions containing fine details in order for these to be better preserved and to use strong regularization in homogeneous, i.e., smooth, image areas. This can be achieved by introducing a spatially varying regularization weight, either acting on

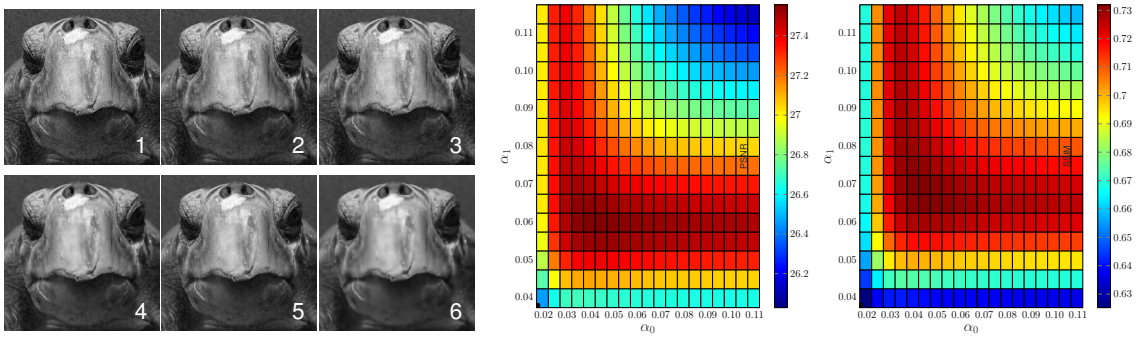

Figure 2: Left: Solutions of the L^2–TGV denoising problem (using the same data as in Figure 1) with combination of the regularization parameters (α_0, α_1) that correspond to the numbers 1-6 on the right plots. The images 2 and 3 correspond to the highest PSNR and SSIM respectively. Right: PSNR and SSIM values of the denoised images with respect to the parameters α_0 and α_1

the fidelity term or the regularizer. For TV minimization, this results in the following two problems:

$$\min_{u\in \mathrm{BV}(\Omega)} \frac{1}{2}\|\sqrt{\lambda}(Tu-f)\|^2_{L^2(\Omega)} + |Du|(\Omega), \quad \lambda \in L^\infty(\Omega),\, \lambda \geq 0 \tag{1.7}$$

$$\min_{u\in \mathrm{BV}(\Omega)} \frac{1}{2}\|Tu-f\|^2_{L^2(\Omega)} + \int_\Omega \alpha\, d|Du|, \quad \alpha \in C(\overline{\Omega}),\, \alpha > 0. \tag{1.8}$$

Here the term $\int_\Omega \alpha\, d|Du|$ denotes the integral of the continuous function α with respect to the measure $|Du|$. The first instance (1.7) and its variants have been studied and applied to image restoration problems in several works. In [63, 64] grayscale and color images subject to blur and Gaussian noise are restored using that model, while in [89] an L^1 fidelity is considered for the case of random-valued impulse noise. In finite dimensions, a technique based on a statistical multiresolution criterion can be found in [67, 96]. In [5] a statistical approach with variance estimators different from the ones in [63, 64] is examined. A deterministic choice rule using a pre-segmentation of the image and a piecewise constant fidelity weight is considered in [18]. As this is related to the present chapter, we briefly mention the basic idea of [63]. Given a normalized weight function $w \in L^\infty(\Omega\times\Omega)$ with $\int_\Omega \int_\Omega w(x,y)\,dxdy = 1$, one defines at every point a localized residual as follows

$$\mathrm{R}u(x) := \int_\Omega w(x,y)(Tu-f)^2(y)\,dy. \tag{1.9}$$

In [63], assuming Gaussian noise of variance σ^2, the parameter λ is updated iteratively in such a way that the localized residual at every point does not significantly exceed σ^2. The method has given satisfactory results for image denoising and deblurring. However, one disadvantage of transforming the regularization parameter into a weight function contained in the fidelity term is related to the fact that it is not immediate clear how to extend the model to data domains, like Fourier or wavelet domains. While analytically and numerically more complex, this observation supports the choice of a second model (1.8) which seems more appropriate in view of such data domains. While these models are equivalent for scalar regularization parameters, for spatially varying ones specific examples can be found yielding singificantly different solution structures for (1.7) and (1.8), respectively; see [83].

The possibility to explore the space spanned by Φ and $J_{\boldsymbol{\alpha}}$ by optimally selecting $\boldsymbol{\alpha}$, immediately leads to the question of whether one can devise a scheme such that the choice of the (best) regularization function α can be decided automatically. We note that as far as problem (1.8) is concerned, in order to be consistent with the functional analytic framework, the function α should be measurable with respect

to the measure $|Du|$ where u is the solution. This is guaranteed when $\alpha \in C(\overline{\Omega})$ as it was ensured in [88] but also see also Proposition 2.9 below. We note that an adaptation of the framework (1.5) to a spatially dependent parameter α can be found in [35, 51]. However this approach is mainly suitable for calibrating TV-based denoising methods for noise of spatially varying intensity and one drawback is the overfitting of the regularization parameter to the training image.

A *monolithic* approach for the automatic selection of the spatially varying function α for problems of the type (1.8) was extensively examined in two recent papers [85, 88]. There, with the aim of combining the reconstruction and α-selection processes in a single mathematical framework, a bilevel optimization approach is adopted with the difference (to the aforementioned approaches) that no ground truth or any training pair is used but rather only the data f, hence the term monolithic. In the spirit of [63], one looks for solutions of the problem (1.8) that force the localized residuals to remain within a variance corridor, close to the variance σ^2 of the noise. This restriction of the localized residuals close to an interval $[\underline{\sigma}^2, \overline{\sigma}^2]$ is imposed by keeping the value of the following upper level objective functional small

$$F(\mathrm{R}u) = \frac{1}{2}\int_\Omega \max(\mathrm{R}u - \overline{\sigma}^2, 0)^2 dx + \frac{1}{2}\int_\Omega \min(\mathrm{R}u - \underline{\sigma}^2, 0)^2 dx. \tag{1.10}$$

In order to get a first understanding, the model considered in [85, 88] aims to approximate

$$\begin{cases} \text{minimize} \quad F(\mathrm{R}u) \quad \text{over } u \in \mathrm{BV}(\Omega), \alpha \in C(\overline{\Omega}) \\ \text{subject to} \;\; u = \underset{u\in \mathrm{BV}(\Omega)}{\operatorname{argmin}} \dfrac{1}{2}\|Tu - f\|^2_{L^2(\Omega)} + \displaystyle\int_\Omega \alpha \, d|Du|. \end{cases} \tag{1.11}$$

Even though in principle the bilevel scheme (1.11) aims to achieve what was discussed above, there exist serious obstacles regarding its functional analytic treatment, well-posedness, and its suitability for numerical realization. No extra regularity is imposed on α, e.g. requiring it to be continuous or at least measurable, and lack of any source of compactness jeopardizes the well-posedness of the scheme. A possible remedy this is to add an extra H^1 regularization term to F together with a boundedness condition requiring $\alpha \in \mathcal{A}_{ad}$ where

$$\mathcal{A}_{ad} = \left\{\alpha \in H^1(\Omega) : \; \underline{\alpha} \leq \alpha \leq \overline{\alpha}, \text{ a.e. in } \Omega\right\}, \quad 0 < \underline{\alpha} < \overline{\alpha}.$$

We further note that the first-order optimality condition of the lower level problem in (1.11) gives rise to a variational inequality (VI) of the second kind. It constitutes a degenerate (non-qualified) constraint (in the sense of the Karush-Kuhn-Tucker theory in Banach spaces) and imposes major challenges in analyzing the dependence $\boldsymbol{\alpha} \to u$. In order to replace the VI of second kind by a more tractable VI of the first kind (but still yielding a degenerate constraint) dualization theory may be employed. Indeed, instead of the lower level problem in (1.11) (primal problem) one considers its Fenchel *predual* [66, 101] which reads

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|\mathrm{div} p + T^* f\|^2_B \quad \text{over } p \in H_0(\mathrm{div}; \Omega), \\ \text{subject to} \;\; & p \in K(\alpha) := \left\{q \in H_0(\mathrm{div}; \Omega) \cap L^\infty(\Omega, \mathbb{R}^d) : \; |q(x)| \leq \alpha(x) \text{ for a.e. } x \in \Omega\right\}, \end{aligned} \tag{1.12}$$

where $\|u\|^2_B = \langle w, B^{-1}u\rangle$, with $B = T^*T$ and T^* the adjoint of T, assuming invertibility; see more details in Section 2.3. Minimizers of the primal and predual problems are connected via the following optimality condition

$$Bu = \mathrm{div} p + T^* f. \tag{1.13}$$

We note that this dualization result was shown in [101] and in [85] for α being a scalar and a continuous function respectively. But as we show in Section 2.3 it also holds for $\alpha \in H^1(\Omega)$. Moreover, the

predual problem (1.12) is more amenable to efficient, image resolution independent, function space based solution algorithms, such as (inexact) *semismooth Newton* methods [101], that can solve the problem up to a very high accuracy fairly quickly. Note that in view of (1.13) the localized residual (1.10) can be written in terms of the dual variable p as follows

$$\mathrm{R}u(x) = R(\mathrm{div}p)(x) = \int_\Omega w(x,y)\left(TB^{-1}\mathrm{div}p - (TB^{-1}T^* - I)f\right)^2 dy. \tag{1.14}$$

This leads to the following bilevel minimization problem which was the starting point in [85]:

$$\begin{cases} \text{minimize} \quad J(p,\alpha) := F(R(\mathrm{div}p)) + \dfrac{\lambda}{2}\|\alpha\|^2_{H^1(\Omega)} \quad \text{over } p \in H_0(\mathrm{div};\Omega), \alpha \in H^1(\Omega), \\ \text{subject to } \ p \in \operatorname{argmin}\left\{\dfrac{1}{2}\|\mathrm{div}p + T^*f\|^2_B : \ p \in K(\alpha)\right\}. \end{cases} \tag{1.15}$$

Existence of solutions of the bilevel problem (1.15) and also the establishment of the dualization framework itself, which links the solution sets of the two problems via (1.13), are closely related to density results of associated convex intersections. In general, the latter establish equivalences of the type

$$\overline{C \cap Y}^X = C, \quad \text{for} \quad \overline{Y} = X, \tag{1.16}$$

where X is a Banach space and C a convex subset, typically represented by pointwise constraints, and Y a dense subset of X. As far as the bilevel problem (1.15) is concerned the density relation

$$\overline{\{\phi \in C_c^\infty(\Omega,\mathbb{R}^d) : \ |\phi(\cdot)| \le \alpha(\cdot), \text{ in } \Omega\}}^{H_0(\mathrm{div};\Omega)} = \{p \in H_0(\mathrm{div};\Omega) : \ |p(\cdot)| \le \alpha(\cdot) \text{ a.e in } \Omega\}$$

is of particular importance. Here, we have $C = K(\alpha)$, $X = H_0(\mathrm{div};\Omega)$ and $Y = C_c^\infty(\Omega,\mathbb{R}^d)$. Establishing such density results can be traced back to [84] and to further extensions in [87]; see the relevant discussion in Section 2.3.

We also note that dualization can be used in order to introduce a rigorous functional analytic framework for a wide class of regularizers that are used in multimodal reconstruction problems. These typically correspond to integrands that are pointwise functions of the gradient, exhibit linear growth

$$J(u) = \int_\Omega j_v(x, \nabla u(x))\, dx, \tag{1.17}$$

and incorporate some a priori knowledge v. In applications, v may correspond to some pre-reconstructed modality. Functionals of the type (1.17) have been used in the discrete setting only and dualization was used in [77] in order to introduce a functional analytic meaning.

Further challenges exist when devising numerical solution algorithms for the bilevel problem (1.15). It is known that this problem falls into the class of *Mathematical Programs with Equilibrium Constraints* (MPEC) and suffers from constraint degeneracy. For the derivation of stationarity conditions, it requires advanced non-smooth analysis techniques other than the classical Karush–Kuhn–Tucker (KKT) theory; see [106, 118] and [13, 79, 82] for the finite and the infinite dimension problem setting. An extensive discussion is given in Section 4.1.

The bilevel scheme (1.15) has produced promising results in applications like denoising, deblurring, wavelet inpainting and MRI reconstruction [88], see an illustration in Figure 3. However, the staircasing effect, a characteristic of TV, appears to remain in a reduced form. One may extend the above framework to the weighted TGV case. For this purpose, similar dualization frameworks have to be

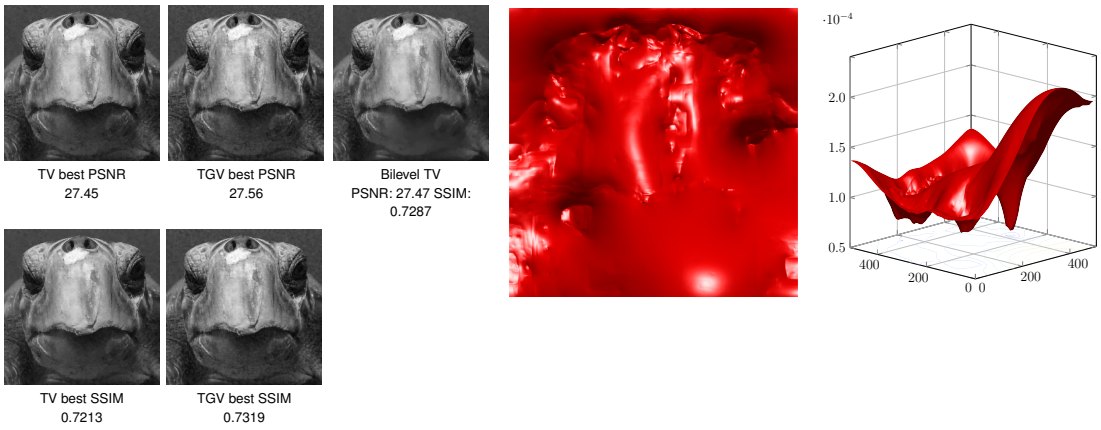

Figure 3: Left: Comparison of the Bilevel weighted TV result with the best scalar TV and TGV denoising results of the Figures 1 and 2. Right: the weighted regularization function α which is determined automatically

established as well as efficient high accuracy algorithms for the solution of the corresponding dual problem need to be developed. We note that an extension of [63] to the TGV case, that is, a study of the problem

$$\min_{u\in \mathrm{BV}(\Omega)} \frac{1}{2}\|\sqrt{\lambda}(Tu-f)\|^2_{L^2(\Omega)} + \mathrm{TGV}^2_{\boldsymbol{\alpha}}(u) \tag{1.18}$$

was done in [21] with good results. But the disadvantages from transforming the regularization parameter into a fidelity weight remain, as it was discussed before.

1.2 Main contributions and organization of this paper

While in its majority, this is a review paper, it also contains some novel elements. In view of this we state clearly in the text which results are new and which are based on earlier works.

In Sections 2.1 and 2.2 we provide an overview of analytical properties of the TV and TGV functionals. We emphasize the roles of the regularization parameters in the reconstruction, like asymptotic, exact solutions as well as structural properties of the solutions of the corresponding variational problems.

In Section 2.3, after briefly reviewing the Fenchel–Rockafellar duality theory, we state the predual problems of TV-based minimizations as well as their weighted versions, and the corresponding dualization frameworks are established. We particularly emphasize the important role that density results of convex intersections play in the establishment of these frameworks.

In Section 3 we discuss algorithms for the solution of the TV variational problems focusing on function space semismooth Newton algorithms for regularized versions of predual problems.

A basic background on the theory of bilevel optimization with emphasis on different notions of stationarity is discussed on Section 4.1. An overview of the bilevel framework for the automatic determination of the regularization parameters, based on the incorporation of localized residuals on the upper objective, is given in Section 4.2. This is done for the TV case, whereas extensions to the TGV regularizer are the current focus of research. We discuss first-order optimality conditions as well as the existence of adjoint states that eventually lead to the derivation of a projected gradient scheme for the algorithmic treatment of the bilevel minimization problem.

Section 5 is dedicated to numerical examples. In particular, in Section 5.1, discretization aspects for

the TV bilevel problem are discussed and in Section 5.2 numerical examples are depicted.

2 Non-smooth priors

2.1 Total Variation

We start with some basic definitions and facts regarding the space of functions of bounded variation $\mathrm{BV}(\Omega)$. For an extensive study, we refer the reader to [8] but also to [11]. Here $\Omega \subset \mathbb{R}^d$ denotes an open, bounded, connected, domain with Lipschitz boundary. The space $\mathrm{BV}(\Omega)$ consists of all functions in $L^1(\Omega)$ whose distributional derivative can be represented by a finite Radon measure which is denoted by $Du = (D_1u, \dots, D_du)$. It can be shown that an L^1 function belongs to $\mathrm{BV}(\Omega)$ if and only if $\mathrm{TV}(u) < \infty$, i.e., it has a finite total variation, where

$$\mathrm{TV}(u) := \sup \left\{ \int_\Omega u \,\mathrm{div}\phi \, dx : \phi \in C_c^\infty(\Omega, \mathbb{R}^d),\ \|\phi\|_{2,\infty} \leq 1 \right\}, \tag{2.1}$$

It can be shown that in this case $\mathrm{TV}(u)$ is equal to the total variation measure of Du evaluated in Ω, i.e., $\mathrm{TV}(u) = |Du|(\Omega)$. In the special case $u \in W^{1,1}(\Omega)$ we have $\mathrm{TV}(u) = \int_\Omega |\nabla u| \, dx$. The TV functional is convex, lower semicontinuous with respect to the strong L^1 convergence and it is a seminorm on the space $\mathrm{BV}(\Omega)$. Moreover $\mathrm{BV}(\Omega)$ is a Banach space when endowed with the norm $\|u\|_{\mathrm{BV}(\Omega)} = \|u\|_{L^1(\Omega)} + |Du|(\Omega)$.

Before we continue we would like to mention a few aspects on the isotropic and the anisotropic version of TV. In the definition (2.1) the finite dimensional $\|\cdot\|_{2,\infty}$ norm is defined as

$$\|\phi\|_{2,\infty} = \operatorname*{ess\,sup}_{x\in\Omega} |\phi(x)|,$$

where $|\cdot|$ is the usual Euclidean norm in $\mathbb{R}^d$. This leads to the isotropic version of TV. If in (2.1), the $\|\cdot\|_{\infty,\infty}$ norm is used instead, where

$$\|\phi\|_{\infty,\infty} = \operatorname*{ess\,sup}_{x\in\Omega} \max \left\{ |\phi_i(x)| :\ i = 1, \dots, d \right\},$$

then one is led to an anisotropic version of TV. In that case it holds that $|Du|(\Omega) = \sum_{i=1}^d |D_iu|(\Omega)$. The resulting norms are equivalent, and while the isotropic version enjoys rotational invariance when used in imaging problems, the anisotropic one is more amenable for our purposes when it comes to the algorithmic treatment of the dual problem in the bilevel framework. For this part of the paper we focus on the isotropic version, but make it explicit when there is an important difference to the anisotropic version.

The measure Du can be decomposed into its absolutely continuous and singular parts with respect to the Lebegue measure $\mathcal{L}^d$, i.e.,

$$Du = \nabla u \mathcal{L}^d + D^s u,$$

with ∇u denoting the Radon-Nikodým density of Du with respect to $\mathcal{L}^d$. Without getting into details, we mention that D^su is further decomposed into the jump part D^ju and the Cantor part of the derivative D^cu. The jump part, is concentrated on the jump set J_u of u which is the set of points where u exhibits (approximate) jump discontinuities. This is the set of points $x \in \Omega$ for which $u^+(x) > u^-(x)$

where

$$u^+(x) = \inf\left\{t \in [-\infty,\infty] : \lim_{r\to 0} \frac{\mathcal{L}^d(\{u > t\} \cap B(x,r))}{r^d} = 0\right\},$$
$$u^-(x) = \sup\left\{t \in [-\infty,\infty] : \lim_{r\to 0} \frac{\mathcal{L}^d(\{u < t\} \cap B(x,r))}{r^d} = 0\right\},$$

are the approximate upper and lower limits of u, respectively. The total variation of $D^j u$ is then written as

$$|D^j u|(\Omega) = \int_{J_u} |u^+ - u^-(x)|\, d\mathcal{H}^{d-1},$$

with $\mathcal{H}^{d-1}$ denoting here the $d-1$-dimensional Hausdorff measure. The fact that BV functions are allowed to have jump discontinuities, in contrast to Sobolev objects, renders them particularly suitable for modeling image intensities as these typically exhibit discontinuities amounting to edges in images.

Besides the strong convergence in BV (norm convergence), two other types of convergence are useful. We say that $(u_n)_{n\in\mathbb{N}}$ converges weakly* in BV to u if it converges strongly in $L^1(\Omega)$ and the measures $(Du_n)_{n\in\mathbb{N}}$ converge weakly* to Du. This type of convergence is useful since it can be shown that any bounded sequence in BV has a weakly* convergent subsequence (compactness in BV). On the other hand we say that $(u_n)_{n\in\mathbb{N}}$ converges strictly to u in BV, if it converges strongly in $L^1(\Omega)$ and also $|Du_n|(\Omega) \to |Du|(\Omega)$. It can be shown that every $u \in \mathrm{BV}(\Omega)$ can be strictly approximated by $C^\infty(\overline{\Omega})$ functions. Regarding embeddings we have that $\mathrm{BV}(\Omega)$ is continuously embedded into $L^{d/d-1}(\Omega)$ if $d \geq 2$ and in $L^\infty(\Omega)$ if $d = 1$.

Returning now to the corresponding TV regularization problem, which in more generality as before reads:

$$\text{minimize } \frac{1}{p}\|Tu - f\|_{\mathcal{Y}}^p + \alpha|Du|(\Omega) \quad \text{over } u \in \mathrm{BV}(\Omega). \tag{2.2}$$

Here $\mathcal{Y}$ is a normed space, $T : L^{d/d-1}(\Omega) \to \mathcal{Y}$ bounded linear, $\alpha > 0$ and $p \geq 1$. Existence and (potential) uniqueness of (2.2) have been shown in different works in the literature. Let us note that in order to show existence, in many of these works, a condition of the type $T(\mathcal{X}_\Omega) \neq 0$ is used, where $\mathcal{X}_A$ denotes the characteristic function of a set $A \subset \Omega$. This condition reflects that T does not annihilate constants. However this property is not necessary for an existence proof, one can show that by using similar techniques as in [26] where the result is shown for TGV instead of TV. We summarize in the following.

Theorem 2.1. *Let $\mathcal{Y}$ be a normed space, $T : L^{d/d-1}(\Omega) \to \mathcal{Y}$ bounded linear, $\alpha > 0$ and $p \geq 1$. Then the minimization problem* (2.2) *has at least one solution. If $p > 1$ and T is injective then the solution is unique.*

We are particularly interested in the parameterization of the problem (2.2), the dependence and structure of solutions with respect to such parameters. Basic asymptotic results are relatively easy to obtain and are summarized in the following:

Proposition 2.2. *The following asymptotic results hold:*

(i) Let $(u_n)_{n\in\mathbb{N}} \subset \mathrm{BV}(\Omega)$ be a sequence of solutions of the problem (2.2) *for an increasing sequence of parameters $(\alpha_n)_{n\in\mathbb{N}}$ such that $\alpha_n \to \infty$. Then there exists a subsequence of $(u_{n_k})_{k\in\mathbb{N}}$ and a constant c such that $u_{n_k} \to c$ weakly* in $\mathrm{BV}(\Omega)$ and it also holds*

$$c \in \operatorname*{argmin}_{u \text{ constant}} \|Tu - f\|_{\mathcal{Y}}.$$

(ii) Let $(\alpha_n)_{n\in\mathbb{N}}$ be a decreasing parameter sequence such that $\alpha_n \to 0$. Then for every sequence $(u_n)_{n\in\mathbb{N}} \subset \mathrm{BV}(\Omega)$ of the corresponding solutions of the problem (2.2) there holds

$$\|Tu_n - f\|_{\mathcal{Y}} \to 0.$$

If in addition there exists u^* such that $Tu^* = f$, then there exists a sequence $(u_n)_{n\in\mathbb{N}}$ of solutions of the problem (2.2), that has a weakly in $\mathrm{BV}(\Omega)$ convergent subsequence to u^*.*

In particular if T maps the set of constant functions to itself then (i) of Proposition 2.2 says that as the regularization parameter grows then the solutions of (2.2) converge up to a subsequence to a constant function which is closest to the data f, with respect to $\|\cdot\|_{\mathcal{Y}}$. If $\mathcal{Y} = L^2(\Omega)$ or $L^1(\Omega)$ for instance, then this constant will be the mean or median value of f, respectively.

The fine structure of the solutions for the TV minimization problem has also been well studied. The solutions are characterized by piecewise constant structures, the *staircasing effect*. For instance it can be shown [29, 125] that in the one dimensional denoising case, i.e., $T = Id$ and $\mathcal{Y} = L^1(\Omega), L^2(\Omega)$, the solution u will be constant in the areas where it is not equal to f (in a certain precise representative sense). In the denoising case and in higher dimensions, in [41] the authors provide a characterization of the area that will exhibit staircasing (extended support). In [99], it is shown that flat areas always occur at global extrema of the data and the extrema of the solution.

Another characteristic of the (isotropic) TV minimization is that no new discontinuities are created in the solution, up to $\mathcal{H}^{d-1}$ measure. This has only been shown so far for the denoising case.

Theorem 2.3. *Let $f \in L^\infty(\Omega) \cap \mathrm{BV}(\Omega)$ and $\alpha > 0$. If u is the solution of the problem*

$$\min_{u\in\mathrm{BV}(\Omega)} \frac{1}{2}\|u - f\|^2_{L^2(\Omega)} + \alpha|Du|(\Omega),$$

then $\mathcal{H}^{d-1}(J_u \setminus J_f) = 0$.

This result was first proven in [39] and was extended in [139], for a wider class of first-order regularizers including Huber-TV. We note that this result does not hold for the anisotropic TV; see [39, Remark 4] nor for the L^1 fidelity. Furthermore, there is a knowledge gap concerning the jump set inclusion when the operator T not the identity.

2.2 Total Generalized Variation

Total generalized variation (TGV) is a higher-order extension of TV, which is tighly related to functions of bounded deformation as pioneered by P.-M. Suquet [134] and then further developed by P. Ciarlet, R. Temam and G. Strang [52, 137, 138]. TGV was introduced in [28] primarily to reduce the often unwanted staircasing effect. For vector of two scalar parameters $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)$, the TGV (of second order) of a function $u \in L^1(\Omega)$ is defined as follows

$$\mathrm{TGV}^2_{\boldsymbol{\alpha}}(u) = \sup\left\{\int_\Omega u\,\mathrm{div}^2\phi\,dx : \phi \in C_c^\infty(\Omega, \mathcal{S}^{d\times d}),\ \|\phi\|_{2,\infty} \le \alpha_0,\ \|\mathrm{div}\phi\|_{2,\infty} \le \alpha_1\right\}. \tag{2.3}$$

Here, $\mathcal{S}^{d\times d}$ denotes the space of symmetric $d \times d$ matrices. For a function $\phi \in C_c^\infty(\Omega, \mathcal{S}^{d\times d})$ the first- and second-order divergences are defined as

$$(\mathrm{div}\phi)_i = \sum_{j=1}^d \frac{\partial\phi_{ij}}{\partial x_j},\ i = 1,\dots,d, \quad \text{and} \quad \mathrm{div}^2\phi = \sum_{i=1}^d \frac{\partial^2\phi_{ii}}{\partial x_i^2} + 2\sum_{i<j}\frac{\partial^2\phi_{ij}}{\partial x_i\partial x_j}$$

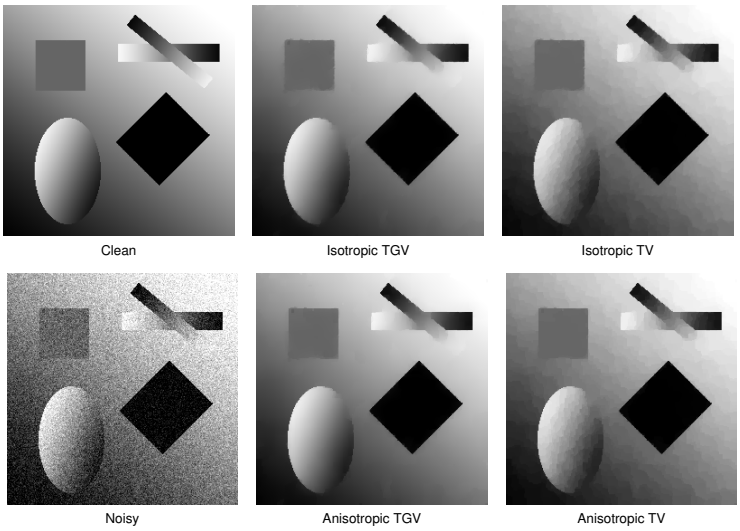

Figure 4: Comparison between scalar TV and TGV isotropic and anisotropic denoising versions. The (scalar) regularization parameters have been manually adjusted for optimal SSIM

Similarly with TV, the anisotropic version of TGV can be defined by substituting the finite dimensional norms $\|\cdot\|_{2,\infty}$ by the corresponding norm $\|\cdot\|_{\infty,\infty}$.

In [30] it was shown that a function $u \in L^1(\Omega)$ has finite TGV value if and only if it belongs to $\mathrm{BV}(\Omega)$, with the norm $\|\cdot\|_{\mathrm{BGV}} := \|\cdot\|_{L^1(\Omega)} + \mathrm{TGV}^2_{\boldsymbol{\alpha}}$ being equivalent to $\|\cdot\|_{\mathrm{BV}(\Omega)}$. Similarly to TV, TGV is a convex functional, lower semicontinuous with respect to the strong L^1 convergence. In [26, 30] it was shown that TGV can be equivalently written as

$$\mathrm{TGV}^2_{\boldsymbol{\alpha}}(u) = \min_{w \in \mathrm{BD}(\Omega)} \alpha_1 |Du - w|(\Omega) + \alpha_0 |\mathcal{E}w|(\Omega). \tag{2.4}$$

Here $\mathcal{E}$ stands for the distributional symmetrized gradient and $\mathrm{BD}(\Omega)$ is the space of functions of bounded deformation, i.e., the space of all functions $w \in L^1(\Omega, \mathbb{R}^d)$ such that $\mathcal{E}w$ is an $\mathcal{S}^{d\times d}$-valued finite Radon measure [134]. The formulation (2.4) of TGV, sometimes called the *minimum* or *primal* representation, or *differentiation cascade*, gives more insights into its regularizing mechanisms. Indeed, in areas where the data is close to an affine function, locally the variable w can be equal to ∇u, resulting in a low value of the first term in (2.4). In this case, also the second term has also low value, since is it roughly equal to $\mathcal{E}(\nabla u) = D^2 u$ (hence the second-order derivative information). On the other hand, setting (locally) $w = 0$ in (2.4) gives back $\alpha_1 |Du|(\Omega)$. Thus, TGV has the possibility to behave locally both like first- and second-order TV. This means that TGV is more suitable to be used in reconstruction of piecewise smooth images. Indeed we depict a simple denoising example in Figure 4.

The asymptotics of TGV with respect to parameters have been studied in [122] with some results also in [26, 140]. In summary, results analogous to Proposition 2.2 hold for TGV as well, with the difference that only convergence of either α_0 or α_1 to zero is enough to give the results of Proposition 2.2 (ii) and the constant functions are substituted by affine ones in (i). The latter family of functions constitutes the kernel of the TGV functional.

Of particular interest is whether there exist combinations of parameters such that $\mathrm{TGV}^2_{\boldsymbol{\alpha}}(u) = \alpha_1 \mathrm{TV}(u)$. This can indeed happen for certain special functions u, and in general one can show that there exist parameters such that TGV is "almost" TV. We summarize this in the following (compare also [120, 122]).

Theorem 2.4. *There exists a constant $C \geq 0$ depending only on the domain Ω such that if with $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)$ satisfies $\alpha_0/\alpha_1 > C$, then*

$$\mathrm{TGV}^2_{\boldsymbol{\alpha}}(u) = \alpha_1 |Du - m_{\mathcal{E}}(\nabla u)|(\Omega), \quad \textit{for all } u \in \mathrm{BV}(\Omega), \tag{2.5}$$

where for a function $g \in L^1(\Omega, \mathbb{R}^d)$ we define

$$m_{\mathcal{E}}(g) := \operatorname*{argmin}_{w \in \mathrm{Ker}\mathcal{E}} \|g - w\|_{L^1(\Omega,\mathbb{R}^d)}. \tag{2.6}$$

Note here that the kernel of the symmetrized gradient $\mathrm{Ker}\,\mathcal{E}$ consists exactly of all the functions of the form $r(x) = Ax + b$ where $b \in \mathbb{R}^d$ and $A \in \mathbb{R}^{d\times d}$ is a skew symmetric matrix. Theorem 2.4 thus states that for a large ratio α_0/α_1 TGV is (almost) equivalent to TV up to an affine correction. In practice this means that for such a combination of parameters the reconstructed images still suffer from a kind of (affine) staircasing effect; see [122].

As far as the structure of solutions is concerned, TGV regularized images have a piecewise affine structure, as expected. This has been studied analytically mainly in dimension one in [14, 29, 120, 123]. Regarding to whether a result analogous to Theorem 2.3 holds for TGV as well is in general unknown. Under some additional regularity assumptions a corresponding result can be shown [140].

In the remainder of this exhibition we primarily focus on TV as our model concept. We mention, however, that the analogous questions, e.g., concerning dualization (next section), bilevel optimization for selecting α_0, α_1 and efficient numerical realization are important for TGV, as well. Some of these problems are in the focus of currently ongoing research in mathematical image processing.

2.3 Dualization

In many circumstances, instead of considering problems of the type (2.2) or its TGV-version it is convenient to rather study their respective (pre)-dual problem. Dualization is here understood in the sense of Fenchel–Rockafellar. In particular, for structured classes of nonsmooth variational problems, duality theory provides a convenient framework for characterising solutions of variational problems through their associated dual quantities. We also point out that the notion of *predualization* arises here due to the non-reflexive nature of $\mathrm{BV}(\Omega)$ and $\mathrm{BD}(\Omega)$, respectively. Dualizing the predual of (2.2) yields (2.2); and analogously for TGV. As we further develop this paper we will address further advantages of this approach, both on the functional analytic as well as on the algorithmic/numerical level. We start by briefly reviewing the main duality theorem; for details see [66].

Consider a general variational problem of the form

$$\inf_{u \in U} F_1(u) + F_2(\Lambda u), \tag{2.7}$$

where U, V are Banach spaces, $\Lambda : U \to V$ is a bounded linear operator, and $F_1 : U \to \overline{\mathbb{R}}$, $F_2 : V \to \overline{\mathbb{R}}$ are proper, convex, lower semicontinuous functions. The problem (2.7) is referred to as the *primal* problem. The *dual* problem of (2.7) is given by

$$\sup_{v^* \in V^*} -F_1^*(\Lambda^* v^*) - F_2^*(-v^*). \tag{2.8}$$

Here, $\Lambda^* : V^* \to U^*$ denotes the adjoint operator of Λ, while $F_1^* : U^* \to \overline{\mathbb{R}}$ denotes the convex conjugate of F_1, i.e.,

$$F_1^*(u^*) = \sup_{u \in U} \langle u^*, u\rangle_{U^*,U} - F_1(u),$$

for $u^* \in U^*$, and similarly for F_2^*. We denote by $\inf \mathcal{P}_{primal}$ and by $\sup \mathcal{P}_{dual}$ the infimum of (2.7) and the supremum of (2.8), respectively. The following proposition states that provided these quantities are equal, the solutions of the two variational problems are connected by an Euler-Lagrange system. From now on ∂F denotes the subdifferential of the convex function F.

Theorem 2.5. *Suppose that both the primal and the dual problems* (2.7) *and* (2.8) *have solutions and that*

$$-\infty < \sup \mathcal{P}_{dual} = \inf \mathcal{P}_{primal} < +\infty \qquad \textit{(zero duality gap).} \tag{2.9}$$

Then all solutions u, v^* of the primal and the dual problem, respectively, are related through the following first-order optimality conditions:

$$\Lambda^* v^* \in \partial F_1(u), \tag{2.10}$$
$$-v^* \in \partial F_2(\Lambda u), \tag{2.11}$$

which are also equivalent to

$$u \in \partial F_1^*(\Lambda^* v^*), \tag{2.12}$$
$$\Lambda u \in \partial F_2^*(-v^*). \tag{2.13}$$

Conversely, if $u \in U$ and $v^* \in V^*$ satisfy (2.10)–(2.11)*, then they are solutions of* (2.7) *and* (2.8)*, respectively, and $-\infty < \sup \mathcal{P}_{dual} = \inf \mathcal{P}_{primal} < +\infty$.*

There are several conditions that guarantee zero duality gap, eventually leading to the characterization (2.10)–(2.11) or (2.12)–(2.13). Below we highlight two of the most useful ones in practice.

Proposition 2.6. *Suppose that one of the following conditions holds:*

(i) [66, Proposition 2.3] There exists $u_0 \in U$ such that $F_1(u_0) < \infty$, $F_2(\Lambda u_0) < \infty$ and F_2 is continuous at Λu_0.

(ii) [10] The set $\bigcup_{\lambda \geq 0} \lambda(\mathrm{dom}(F_2) - \Lambda(\mathrm{dom}(F_1)))$ is a closed subspace of Y.

Then the dual problem (2.8) *has a solution and* (2.9) *holds true.*

Our target is to apply this dualization framework to a class of TV and TGV variational problems and their weighted parameter versions. For that we will first need several definitions.

Definition 2.7. *Let $1 \leq q \leq \infty$ and $p \in L^q(\Omega, \mathbb{R}^d)$. We have $\mathrm{div} p \in L^q(\Omega)$ if there exists $w \in L^q(\Omega)$ such that for all $\phi \in C_c^\infty(\Omega)$*

$$\int_\Omega \nabla \phi \cdot p \, dx = -\int_\Omega \phi w \, dx.$$

Then we define

$$W^q(\mathrm{div}; \Omega) := \left\{ p \in L^q(\Omega, \mathbb{R}^d) : \ \mathrm{div} p \in L^q(\Omega) \right\}$$

with the norm $\|p\|^q_{W^q(\mathrm{div};\Omega)} := \|p\|^q_{L^q(\Omega,\mathbb{R}^d)} + \|\mathrm{div} p\|^q_{L^q(\Omega)}$. Similarly we define the space $W^q(\mathrm{div}^2; \Omega)$ as the space of all functions $p \in L^q(\Omega, \mathcal{S}^{d\times d})$ whose first- and second-order divergence belong to L^q, equipped with the norm $\|p\|^q_{W^q(\mathrm{div}^2;\Omega)} := \|p\|^q_{L^q(\Omega)} + \|\mathrm{div} p\|^q_{L^q(\Omega,\mathbb{R}^d)} + \|\mathrm{div}^2 p\|^q_{L^q(\Omega)}$.

Using the density of C_c^∞ in L^q, it can be shown that the first- and second-order divergences are unique. Moreover both spaces above are Banach equipped with the associated norms as stated above. We refer to [27] for a more general definition of these spaces. Note that when $q = 2$ then the standard notation is $H(\mathrm{div};\Omega)$ and $H(\mathrm{div}^2;\Omega)$; see [71] for the former space. The spaces $W_0^q(\mathrm{div};\Omega)$ and $W_0^q(\mathrm{div}^2;\Omega)$ are defined as

$$W_0^q(\mathrm{div};\Omega) = \overline{C_c^\infty(\Omega,\mathbb{R}^d)}^{\|\cdot\|_{W^q(\mathrm{div};\Omega)}},$$
$$W_0^q(\mathrm{div}^2;\Omega) = \overline{C_c^\infty(\Omega,\mathcal{S}^{d\times d})}^{\|\cdot\|_{W^q(\mathrm{div}^2;\Omega)}}$$

Using the definitions above, the following integration by parts formulae hold true:

$$\int_\Omega \nabla\phi\cdot p\,dx = -\int_\Omega \phi\,\mathrm{div}p\,dx, \quad \text{for all } p\in W_0^q(\mathrm{div};\Omega),\ \phi\in C^\infty(\overline{\Omega},\mathbb{R}), \tag{2.14}$$
$$\int_\Omega E\phi\cdot p\,dx = -\int_\Omega \phi\cdot\,\mathrm{div}p\,dx, \quad \text{for all } p\in W_0^q(\mathrm{div}^2;\Omega),\ \phi\in C^\infty(\overline{\Omega},\mathbb{R}^d), \tag{2.15}$$
$$\int_\Omega \nabla\phi\cdot\mathrm{div}p\,dx = -\int_\Omega \phi\,\mathrm{div}^2p\,dx, \quad \text{for all } p\in W_0^q(\mathrm{div}^2;\Omega),\ \phi\in C^\infty(\overline{\Omega},\mathbb{R}), \tag{2.16}$$

with $E\phi$ denoting the symmetrized gradient of ϕ.

With these definitions, we can now state the following result for the scalar version of TV and TGV.

Proposition 2.8. *Let $d>2$, $u\in L^{d/d-1}(\Omega)$ and $\alpha,\alpha_0,\alpha_1>0$*

$$\alpha\mathrm{TV}(u) = \sup\left\{\int_\Omega u\,\mathrm{div}p\,dx : p\in W_0^d(\mathrm{div};\Omega),\ \|p\|_{2,\infty}\le\alpha\right\} \tag{2.17}$$
$$\mathrm{TGV}^2_{\boldsymbol{\alpha}}(u) = \sup\left\{\int_\Omega u\,\mathrm{div}^2p\,dx : p\in W_0^d(\mathrm{div}^2;\Omega),\ \|p\|_{2,\infty}\le\alpha_0,\ \|\mathrm{div}p\|_{2,\infty}\le\alpha_1\right\} \tag{2.18}$$

As we will see later, these equivalent characterizations of the two regularizers are important for the associated dualization framework for the corresponding regularization problems. The proof of Proposition 2.8 is immediate once the following two density results are shown:

$$\overline{C_\alpha}^{L^d(\Omega)} = K_\alpha, \tag{2.19}$$
$$\overline{C_{\boldsymbol{\alpha}}}^{L^d(\Omega)} = K_{\boldsymbol{\alpha}}, \tag{2.20}$$

where

$$C_\alpha := \left\{\mathrm{div}\phi :\ \phi\in C_c^\infty(\Omega,\mathbb{R}^d),\ \|\phi\|_{2,\infty}\le\alpha\right\}, \tag{2.21}$$
$$K_\alpha := \left\{\mathrm{div}p :\ p\in W_0^d(\mathrm{div};\Omega),\ \|p\|_{2,\infty}\le\alpha\right\}, \tag{2.22}$$
$$C_{\boldsymbol{\alpha}} := \left\{\mathrm{div}^2\phi :\ \phi\in C_c^\infty(\Omega,\mathcal{S}^{d\times d}),\ \|\phi\|_{2,\infty}\le\alpha_0,\ \|\mathrm{div}\phi\|_{2,\infty}\le\alpha_1\right\}, \tag{2.23}$$
$$K_{\boldsymbol{\alpha}} := \left\{\mathrm{div}^2p :\ p\in W_0^d(\mathrm{div}^2;\Omega),\ \|p\|_{2,\infty}\le\alpha_0,\ \|\mathrm{div}p\|_{2,\infty}\le\alpha_1\right\}. \tag{2.24}$$

Questions regarding density results of the above type deserve a separate discussion. They relate to a more general category of questions regarding whether for a given Banach space X, a dense subspace $\overline{Y}=X$, and a convex subset $C\subset X$, the following density result holds:

$$\overline{C\cap Y} = C. \tag{2.25}$$

We caution the reader that will such density results may be thought of as being available since $\overline{Y} = X$, there exist striking counterexamples for such a density result (2.25) to fail in general; see [84]. However, among other results, in [84], the density (2.25) was shown for $X = H_0(\mathrm{div}; \Omega)$ endowed with its norm, $Y = C_c^\infty(\Omega, \mathbb{R}^d)$, and

$$C = \left\{p : \ p \in W_0^d(\mathrm{div}; \Omega), \ |p(x)| \leq \alpha(x), \text{ for a.e. } x \in \Omega\right\}, \quad \text{where } \alpha \in C(\overline{\Omega}), \ \alpha > 0.$$

The proof uses the theory of mollifiers, and the result is generalized to $X = W_0^d(\mathrm{div}; \Omega)$. Note that this result is more general than (2.19) where only density for the divergences is required. In fact it is an interesting question if this result is strictly weaker, i.e., whether there are counterexamples where the density holds for the divergence (or another differential operator) but not for the full norm. We note that apart from dualization in variational imaging problems, these type of density results find applications in the study of the limiting behavior of discretized problems, vanishing viscosity problems and others; see [84, 87, 90, 91].

The equalities (2.19)–(2.20), that is, the densities only for the first- and second-order divergences accociated with the TV and TGV functionals were shown in [27, Proposition 3.3] in more generality, i.e., in the context of TGV of arbitrary order; see also [23, Proposition 7]. The associated proof exploits duality and uses scalar weights, only. It turns out that it can be easily be adapted to the case where the weights are continuous functions bounded away from zero.

The case of weighted TGV, for a variety of finite dimensional norms, may also be addressed. It builds on the following result, for which $|\cdot|_r$ denotes the finite dimensional r-norm for $1 \leq r \leq \infty$, and r^* is such that $1/r + 1/r^* = 1$ with the obvious definitions for $r = 1, \infty$. Let $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)$ with $\alpha_0, \alpha_1 \in C(\overline{\Omega})$ and $\alpha_0, \alpha_1 > \underline{\alpha} > 0$, $\underline{\alpha} \in \mathbb{R}$. Define the weighted TGV functional, as follows:

$$\mathrm{TGV}^2_{\boldsymbol{\alpha}}(u) = \sup\left\{\int_\Omega u \,\mathrm{div}^2\phi \, dx : \ \phi \in C_c^\infty(\Omega, \mathcal{S}^{d\times d}), \right. \tag{2.26}$$
$$\left. |\phi(x)|_r \leq \alpha_0(x), \ |\mathrm{div}\phi(x)|_r \leq \alpha_1(x), \text{ for all } x \in \Omega\right\}$$

Then this functional has also the equivalent expression

$$\mathrm{TGV}^2_{\boldsymbol{\alpha}}(u) = \min_{w \in \mathrm{BD}(\Omega)} \int_\Omega \alpha_1 \, d|Du - w|_{r^*} + \int_\Omega \alpha_0 \, d|\mathcal{E}w|_{r^*}. \tag{2.27}$$

Moreover, the weighted TGV functional

$$\mathrm{TGV}^2_{\boldsymbol{\alpha}}(u) = \sup\left\{\int_\Omega u \,\mathrm{div}^2\phi \, dx : \ \phi \in C_c^\infty(\Omega, \mathcal{S}^{d\times d}), \right. \tag{2.28}$$
$$\left. |\phi(x)|_r \leq \alpha_0(x), \ |\mathrm{div}\phi|_r \leq \alpha_1(x), \text{ for all } x \in \Omega\right\}$$

is also equal to

$$\sup\left\{\int_\Omega u \,\mathrm{div}^2 p \, dx : p \in W_0^d(\mathrm{div}^2; \Omega), \ |p(x)|_r \leq \alpha_0(x), \ |\mathrm{div}p(x)|_r \leq \alpha_1(x), \text{ for a.e. } x \in \Omega\right\} \tag{2.29}$$

for all $u \in L^{d/d-1}(\Omega)$.

Recapitulating, we have described how the following equalities can be established when $\alpha, \alpha_0, \alpha_1 \in$

$C(\overline{\Omega})$ bounded away from zero, where $\mathrm{TV}_\alpha(u) = \int_\Omega \alpha d|Du|$:

$$\mathrm{TV}_\alpha(u) = \sup\left\{\int_\Omega u\,\mathrm{div} p\,dx : p \in W_0^d(\mathrm{div};\Omega),\ |p(x)|_r \le \alpha(x),\ \text{for a.e. } x \in \Omega\right\}, \quad (2.30)$$

$$\mathrm{TGV}^2_{\boldsymbol{\alpha}}(u) = \sup\left\{\int_\Omega u\,\mathrm{div}^2 p\,dx : p \in W_0^d(\mathrm{div}^2;\Omega),\right. \quad (2.31)$$
$$\left. |p(x)|_r \le \alpha_0(x),\ |\mathrm{div} p(x)|_r \le \alpha_1(x),\ \text{for a.e. } x \in \Omega\right\}.$$

Regarding more general weight functions, at least for TV, the following proposition follows from [77, Corollary 3.5, Proposition 3.8].

Proposition 2.9 ([77]). *Let $\alpha \in \mathrm{BV}(\Omega)$ with $0 \le \alpha(x) \le C$ for a.e. $x \in \Omega$. Moreover let $j(x,z) = \alpha(x)b(z)$, with $b : \mathbb{R}^d \to \mathbb{R}$ being a convex, positively 1-homogeneous, even function such that there exists $\gamma > 0$ with $b(z) \le \gamma|z|$ for all $z \in \mathbb{R}^d$. Then for every $u \in \mathrm{BV}(\Omega)$ it holds that*

$$\begin{aligned}&\int_\Omega j(x,\nabla u(x))\,dx + \int_\Omega j^-(x,\sigma_{D^c u})\,d|D^c u| + \int_{J_u\cap\Omega}(u^+(x)-u^-(x))j^-(x,\sigma_{D^j u})\,d\mathcal{H}^{d-1}\\ &= \sup\left\{\int_\Omega u\,\mathrm{div} p\,dx : p \in W_0^d(\mathrm{div};\Omega),\ j^\circ(x,p(x)) \le 1,\ \textit{for a.e. } x \in \Omega\right\},\end{aligned} \quad (2.32)$$

where

$$j^\circ(x,z^*) = \sup_{z : j(x,z)\le 1} z^* \cdot z.$$

In particular, if $\alpha \in W^{1,1}(\Omega)$ with $0 \le \alpha(x) \le C$ for a.e. $x \in \Omega$ and $b(z) = |z|_{r^*}$, then

$$\int_\Omega \alpha\, d|Du|_{r^*} = \sup\left\{\int_\Omega u\,\mathrm{div} p\,dx : p \in W_0^d(\mathrm{div};\Omega),\ |p(x)|_r \le \alpha(x),\ \textit{for a.e. } x \in \Omega\right\}. \quad (2.33)$$

The result of (2.32) also holds for more general integrands j that are not of the type $\alpha(x)b(x)$ but nevertheless satisfy some boundedness, convexity, and 1-homogeneity assumptions. The associated proof combines some duality and lower semicontinuity results from [7]. In order to derive (2.33) from (2.32) note first that if $j(x,z) = \alpha(x)b(z)$ with $b(z) = |z|_{r^*}$, then $j^\circ(x,z^*) = |z|_r/\alpha(x)$ which resolves the right hand side. Moreover, when $\alpha \in W^{1,1}(\Omega)$, then for the set of its jump points J_α we have that $\mathcal{H}^{d-1}(J_\alpha) = 0$ which implies $|Du|(J_\alpha) = 0$ for every $u \in \mathrm{BV}(\Omega)$ [8, Lemma 3.76]. Then we can write α instead of α^- and $j^-(x,z) = \alpha(x)b(z)$. Thus, in this case the left hand side of (2.32) equals

$$\begin{aligned}&\int_\Omega \alpha(x)|\nabla u|_{r^*}\,dx + \int_\Omega \alpha(x)\left|\frac{dD^c u}{d|D^c u|}\right|_{r^*} d|D^c u| + \int_{J_u\cap\Omega}(u^+(x)-u^-(x))\alpha(x)\left|\frac{dD^j u}{d|D^j u|}\right|_{r^*} d\mathcal{H}^{d-1},\\ &= \int_\Omega \alpha(x)|\nabla u|_{r^*}\,dx + \int_\Omega \alpha(x)\frac{d|D^c u|_{r^*}}{d|D^c u|}\,d|D^c u| + \int_{J_u\cap\Omega}(u^+(x)-u^-(x))\alpha(x)\frac{d|D^j u|_{r^*}}{d|D^j u|}\,d\mathcal{H}^{d-1},\\ &= \int_\Omega \alpha\, d|Du|_{r^*}\end{aligned}$$

where we have followed the properties of convex functions of measures [58]. We point out that the supremum on the right hand side of (2.32) is finite if and only if $u \in \mathrm{BV}(\Omega)$. Also, it suffices to have $0 < \underline{\alpha} \le \alpha(x) < C$ for a.e. $x \in \Omega$ for some strictly positive constant $\underline{\alpha}$.

2.4 Dualization of the variational regularization problems

The formulation of the TV and TGV regularizers and their weighted versions, of the type (2.30), (2.31) and (2.33) is the key for establishing the duals of the corresponding regularization problems. We will state a general duality result only for the weighted TV case which is connected to Proposition 2.9. The derivation of a corresponding result for TGV is the subject of current research.

The following proposition is a consequence of Proposition 2.9 and [77, Theorem 5.1].

Proposition 2.10 ([77]). *Let $1 \leq p < \infty$, $\alpha \in W^{1,1}(\Omega)$ with $0 < \underline{\alpha} < \alpha(x) < C$ for a.e. $x \in \Omega$. Moreover let $(\mathcal{Y}, \|\cdot\|_{\mathcal{Y}})$ be a Banach space, with $f \in \mathcal{Y}$, and $T \in \mathcal{L}(L^{d/d-1}(\Omega), \mathcal{Y})$. Then, there exists a solution to the primal problem*

$$\min_{u \in \mathrm{BV}(\Omega)} \frac{1}{p}\|Tu - f\|_{\mathcal{Y}}^p + \int_\Omega \alpha d|Du|_{r^*} \tag{2.34}$$

to its predual problem

$$\min\left\{ \left(\frac{1}{p}\|T\cdot - f\|_{\mathcal{Y}}^p\right)^* (\mathrm{div} p) : \ p \in W_0^d(\mathrm{div};\Omega),\ |p(x)|_r \leq \alpha(x) \text{ for a.e. } x \in \Omega \right\}, \tag{2.35}$$

and zero duality gap holds.

Note that the coercivity condition $0 < \underline{\alpha} < \alpha(x) < C$ is not only necessary for the solution of the primal problem to belong to $\mathrm{BV}(\Omega)$ but also for the establishment of the duality result itself. However according to [77, Proposition] it can be omitted provided that the range of T^*, $\mathrm{Rg}(T^*)$ is closed and $\mathrm{Ker}(T)$ is finite dimensional. In that case the solution of the primal problem (2.34) cannot be guaranteed in $\mathrm{BV}(\Omega)$ but only to $L^{d/d-1}(\Omega)$.

In the special case where $B = T^*T$ is invertible and $p = 2$ and $\mathcal{Y} = L^2(\Omega)$, then the predual problem (2.35) takes a more precise form since

$$\left(\frac{1}{2}\|T\cdot - f\|_{L^2(\Omega)}^2\right)^* (u^*) = \frac{1}{2}(u^* + K^*f, B^{-1}(u^* + K^*f)) - \frac{1}{2}\|f\|_{L^2(\Omega)}^2 =: \frac{1}{2}\|u^* + K^*f\|_B^2 - \frac{1}{2}\|f\|_{L^2(\Omega)}^2 \tag{2.36}$$

for every $u^* \in L^d(\Omega)$; see also [101]. We note that if T^*T is not invertible then one can add a quadratic term $\frac{\gamma}{2}\|u\|_{L^2(\Omega)}^2$ for some small $\gamma > 0$ in the primal problem (2.34). In that case $B = \gamma I + T^*T$. Moreover the optimality conditions that correspond to (2.10)–(2.11) are

$$Bu = T^*f + \mathrm{div} p, \tag{2.37}$$

$$\langle(-\mathrm{div})^* u, \tilde{p} - p)\rangle_{W_0^d(\mathrm{div};\Omega)^*, W_0^d(\mathrm{div})} \leq 0, \text{ for all } \tilde{p} \in W_0^d(\mathrm{div}),\ |p(x)|_r \leq \alpha(x) \text{ for a.e. } x \in \Omega, \tag{2.38}$$

where u and p solve (2.34) and (2.35), respectively; see [85, 101].

3 Numerical algorithms

In this section we review some of the most popular numerical algorithms for the solution of the non-smooth variational problems arising in imaging, as these were discussed in the previous section. We should already mention here that the majority of the algorithms are considered in the discretized setting

only, that is, in this case images are not modeled as functions on a continuous domain Ω, but rather as functions defined on a discretized grid $\Omega_h = \{1, 2, \dots n\} \times \{1, 2, \dots m\}$ where $n \times m$ is the pixel resolution of the image. Even though, this seems quite natural and in the end in practice, discretization will inevitably take place anyway, devising a solution algorithm directly in the discrete setting has its disadvantages as we shall explain below. But before we come to this, we start with a brief historical review focusing on the algorithms for the solution of the classical TV minimization problem (1.4).

In the initial paper [126], the TV denoising problem was solved by using a gradient projection method where the step length was related to a (pseudo) time marching scheme with a constant step length over all time steps. This approach however is extremely slow even though some improved time stepping schemes have been studied, e.g., in [143]. Subsequently, several other (improved) algorithmic approaches were developed such as graph cut methods for the anisotropic TV minimization [53], fixed point iterations [98], augmented Lagrangian-based strategies [136, 146, 147], ADMM and Douglas-Rachford splitting [131]. Particular attention has been drawn to Bregman iteration methods [72, 117, 69, 148]. A particular characteristic of this family of methods is the reduction of contrast loss (bias) which is characteristic to these derivative based regularization methods see [31] as well as [16] for a review.

It has become widely accepted that the most efficient and applicable techniques to solve non-smooth variational problems (not only the L^2–TV problem) are based on convex duality. First works can be found in [40, 42] and [45]. A very popular current method is the Primal-Dual hybrid gradient method (PDHGM) or otherwise called Chambolle-Pock algorithm [43] and its extensions; see [44] for a comprehensive review of the corresponding family of methods. These methods are aiming to find solutions to saddle-point problems of the following form

$$\min_{x\in X}\max_{y\in Y}\langle Kx, y\rangle + G(x) - F^*(y) \tag{3.1}$$

where typically X, Y are finite dimensional real vector spaces, G, F are proper, convex, lower semi-continuous functions. and $K : X \to Y$ is a continuous linear operator. Using the Fenchel–Rockafellar duality theory presented in the previous section it can be shown that (x, y) solves (3.1) if and only if x and y solve the following primal and dual problems, respectively, i.e.,

$$\begin{aligned}&\min_{x\in X} \;\; F(Kx) + G(x),\\ &\max_{y\in Y} \;\; -G^*(-K^*y) - F^*(y).\end{aligned}$$

The Chambolle–Pock iteration then reads as follows

$$y^{k+1} = \mathrm{prox}_{\sigma F^*}(y^k + \sigma K\overline{x}^{k+1}), \tag{3.2}$$
$$x^{k+1} = \mathrm{prox}_{\tau G}(x^k - \tau K^* y^k), \tag{3.3}$$
$$\overline{x}^{k+1} = x^{k+1} + \theta(x^{k+1} - x^k). \tag{3.4}$$

Apparently, it consists of an alternation of a (proximal) descent in the primal variable x and an ascent in the dual variable y followed by an extrapolation step. The Chambolle–Pock algorithm is also a popular method of choice when it comes to solving inverse problems involving the TGV functional [20, 24, 95, 100]. Its popularity stems from the fact that the prox operations in (3.2)–(3.3) can typically be computed cheaply leading to an easy implementation. Unfortunately, even though one obtains good reconstructions after relatively few iterations, solving the problem to high accuracy with this and any other first-order method (given some objective stopping rule; other than closeness of successive iterates, which is not objective and may stop the iteration prematurely due to negligible progress of the

chosen iterative solver) typically leads to an excessive number of iterations. Sufficiently high accuracy, however, turns out to be particularly desirable in the context of the bilevel optimization problems that we discuss in the present exposition. A second issue that is often not addressed well by first-order algorithms is (image) resolution independent convergence, which is typically connected to function space algorithms [86]. Resolution independence refers here to the fact that, given an objective stopping rule and an initial iterate common to all resolution levels, the number of iterations to satisfy such a stopping rule is robust under refinements of the resolution of a given image datum (over a bounded domain).

In view of the issues highlighted in the previous paragraph, we now turn our attention to Newton-type methods. In its very basic form, Newton's method aims to solve a nonlinear system

$$F(x) = 0, \tag{3.5}$$

with $F : \mathbb{R}^n \to \mathbb{R}^n$ differentiable, by an iteration which, given some initial guess x^0, computes a root of a current linearization of (3.5) at every step, i.e., one finds for $k = 0, 1, 2, ...$ an update direction d^k with

$$\nabla F(x^k)d^k = -F(x^k), \tag{3.6}$$

$$x^{k+1} := x^k + d^k. \tag{3.7}$$

For x^0 close to x^* with the matrix $\nabla F(x^*)^{-1}$ being (uniformly) invertible along the iterates, the iterates converge at a rate that depends on the order of continuity of ∇F. More specifically, the rate of convergence is (q-)quadratic when ∇F is Lipschitz, and the rate is (q-)superlinear for Hölder continuous ∇F. Relaxations on the regularity of F can still yield convergent Newton methods, with the most relevant regularity here being semismoothness [78, 112, 124]. Of particular interest are semismooth Newtons in Banach spaces [50, 78]. For the rest of this section we give a basic account on such methods and how they can be used for the accurate solution of the TV (and TGV) problems.

Suppose that $F : D \subset X \to Z$, where X, Z are Banach space and D is an open set, and we are interested in finding $x^* \in D$ such that $F(x^*) = 0$. We assume that the map F is generalized (or *Newton*) differentiable on an open set $U \subset D$, that is, there exists a family of mappings $G : U \to \mathcal{L}(X, Z)$ such that

$$\lim_{d\to 0} \frac{1}{\|d\|_X} \|F(x + d) - F(x) - G(x + d)d\|_Z = 0, \tag{3.8}$$

for all $x \in U$. This condition extends a characterization of semismoothness in finite dimensions to possibly infinite dimensional Banach spaces. The operator G is referred to as the generalized (or Newton) derivative of F at x. Note that G need not be unique. However, when F is Fréchet differentiable at X, then $G(x)$ is unique and equals the Fréchet derivative of F at x. For some given $x^0 \in U$, the corresponding semismooth Newton iteration reads

$$d^k = -G(x^k)^{-1}F(x^k),\ k = 0, 1, 2, 3, \dots \tag{3.9}$$

$$x^{k+1} = x^k + d^k, \tag{3.10}$$

where $G(x^k)$ is an arbitrary generalized derivative of F at x^k. Then the following convergence result is available [78].

Proposition 3.1. *Let x^* be a solution of $F(x) = 0$ and F be Newton differentiable in an open neighborhood U containing x^*. Assume furthermore that $G(x)$ is nonsingular for all $x \in U$ and that*

the set $\{\|G(x)^{-1}\| : \ x \in U\}$ *is bounded. Then, provided* $\|x_0 - x^*\|_X$ *is sufficiently small, the semismooth Newton iteration* (3.9)–(3.10) *is well-defined and converges superlinearly to* x^**, i.e.,*

$$\lim_{\substack{x\to x^*\\ x\neq x^*}} \frac{\|x^{k+1} - x^*\|_X}{\|x^k - x^*\|_X} = 0.$$

Under suitable assumptions, it can be shown that semismooth Newton methods exhibit a discretization independent convergence [93, 76]. In fact, for sufficiently small mesh sizes (for discretizing the associated Banach spaces) either the number of iterations of the discretized method is essentially constant with respect to the mesh size, or that the discretized method achieves essentially the same convergence rate independently of the mesh size. We now show how this framework can be applied for the solution of the TV-type variational problem.

We start with TV and in particular of the (anisotropic) TV predual problem with L^2 fidelity, (2.35)–(2.36):

$$\begin{aligned} &\min_{p\in W_0^d(\mathrm{div};\Omega)} \frac{1}{2}\|\mathrm{div} p + T^* f\|_B^2 \\ &\text{subject to} \quad |p(x)|_\infty \leq \alpha(x) \text{ for a.e. } x \in \Omega, \end{aligned} \tag{3.11}$$

where $\alpha \in W^{1,1}(\Omega)$ is a weight function. In order to address the constrained minimization problem (3.11) with Newton methods we substitute the box constraints by the following penalty functional (which equals as smoothed version of the Moreau-Yosida regularization with respect to the $L^2(\Omega)$-norm of the indicator of the constraint set in (3.11))

$$\mathcal{P}_\delta(p, \alpha) := \int_\Omega \sum_{i=1}^d (G_\delta(-(p_i + \alpha)) + G_\delta(p_i - \alpha))\, dx \tag{3.12}$$

where $p = (p_1, \ldots, p_d)$ and $G_\delta : \mathbb{R} \to \mathbb{R}$,

$$G_\delta(r) = \begin{cases} \frac{1}{2}r^2 - \frac{\delta}{2}r + \frac{\delta^2}{6}, & \text{if } r \geq \delta, \\ \frac{r^3}{6\delta}, & \text{if } 0 < r < \delta, \\ 0, & \text{if } r \leq 0, \end{cases} \tag{3.13}$$

for $\delta > 0$. The idea is that when the quantity $\mathcal{P}_\delta(p, \alpha)$ is small then p is expected to approximately satisfy the box constraints in (3.12). This is achieved by multiplying this term by $1/\epsilon$ with ϵ sufficiently small. We note that if one is interested only in solving the reconstruction problem via a dual approach, the additional smoothing by $\delta > 0$ is not needed. One may rather directly work with the aforementioned Moreau-Yosida regularization. In the context of bilevel optimization, however, the dual of the reconstruction problem becomes a constraint thus asking for additional smoothing (hereby expressed by G_δ with $\delta > 0$) for a subsequent derivation of stationarity conditions of the bilevel problem.

Note that $\mathcal{P}_\delta(p, \alpha)$ is well-defined if both p and α enjoy L^2 regularity. As $\alpha \in W^{1,1}(\Omega)$, this would be automatically satisfied for α for $d = 2$ by Sobolev embedding; otherwise L^2-regularity must be assumed in addition. As far as p is concerned, L^2 regularity results from the minimization process anyway. For uniqueness of a solution p in (3.11) a multiple of the L^2-norm of p may be added to the objective. The resulting modified predual problem reads

$$\min_{p\in H_0^1(\Omega)^d} \frac{\beta}{2}\|\nabla p\|^2_{L^2(\Omega)^d} + \frac{\gamma}{2}\|p\|^2_{L^2(\Omega)^d} + \frac{1}{2}\|\mathrm{div} p + T^* f\|_B^2 + \frac{1}{\epsilon}\mathcal{P}_\delta(p, \alpha). \tag{3.14}$$

The reason for the additional H^1_0-regularization (typically with $0 < \beta \ll \underline{\alpha} \leq \alpha(x)$ for $x \in \Omega$) for p is the facilitation of the convergence analysis of the Newton method in function space as well as the need for differentiability of the solution map $\alpha \mapsto p(\alpha)$, which is necessary in the context of bilevel optimization. Observe that the homogeneous Dirichlet boundary conditions on p aim to capture the boundary conditions associated with space $W^d_0(\mathrm{div};\Omega)$. It can be shown that the associated Euler–Lagrange system for (3.14) reads for $V = H^1_0(\Omega)^d$

$$\langle -\beta\Delta p + \gamma p - \nabla B^{-1}\mathrm{div} p - \nabla B^{-1}T^* f, v\rangle_{V^*,V} + \frac{1}{\epsilon}\langle P_\delta(p,\alpha), v\rangle_{L^2(\Omega)^d, L^2(\Omega)^d} = 0, \quad \forall v \in V. \tag{3.15}$$

Here the Laplacian $\Delta : H^1_0(\Omega)^d \to [H^1_0(\Omega)^d]^*$ is defined as $\langle \Delta p, v\rangle_{V^*,V} = \sum_{k=1}^d \int_\Omega \nabla p_k \cdot \nabla v_k \, dx$. Moreover $P_\delta : H^1_0(\Omega)^d \times L^2(\Omega) \to L^2(\Omega)^d$ is defined as

$$P_\delta(p,\alpha) := G'_\delta(p - \alpha\mathbf{1}) - G'_\delta(-p - \alpha\mathbf{1}),$$

where $p - \alpha\mathbf{1} = (p_1 - \alpha, \ldots, p_d - \alpha)$ and similarly for $-p - \alpha\mathbf{1}$. Furthermore,

$$G'_\delta(r) = \begin{cases} r - \frac{\delta}{2}, & \text{if } r \geq \delta, \\ \frac{r^2}{2\delta}, & \text{if } 0 < r < \delta, \\ 0, & \text{if } r \leq 0, \end{cases} \tag{3.16}$$

In other words the equation (3.15) is of the type $F(p) = 0$ where $F : H^1_0(\Omega)^d \to [H^1_0(\Omega)^d]^*$ with

$$F(p) = -\beta\Delta p + \gamma p - \nabla B^{-1}\mathrm{div} p - \nabla B^{-1}T^* f + \frac{1}{\epsilon}P_\delta(p,\alpha). \tag{3.17}$$

The regularized version (3.14) is consistent with the original predual problem (3.11) in the sense that when the parameters β, γ and ϵ tend to zero then the divergences of the approximate solutions converge to the divergence of the solution of the original problem in the appropriate sense; see [85, Theorem 5.1]. In the same work (Theorem 5.2), differentiability properties were established for the solution mapping $\alpha \mapsto p(\alpha)$ where $p(\alpha)$ solves (3.15). In particular it was shown that $p(\cdot) : L^{2+\xi}(\Omega) \to H^1_0(\Omega)^d$ is Gâteaux differentiable, for every $\xi \in (0, \overline{\xi}(d)]$, where $\overline{\xi}(d) = +\infty$ for $d = 1$, $\overline{\xi}(d) \in [0,\infty)$ for $d = 2$ and $\overline{\xi}(d) = 2d/(d-2)$ for $d > 2$. While such results are needed for deriving stationarity conditions in bilevel optimization (where α also becomes an optimization variable), we stress once again that the H^1_0-regularity (which lifts the smoothness of p) is crucial for establishing such a result.

As noted above, one may indeed choose $\delta = 0$ [101]. Then we get $G'_0(r) = \max(r, 0)$ which is (still) a semismooth mapping when considered from $L^r(\Omega)$ to $L^2(\Omega)$, $2 < r$; see [78]. Even though, in this case one can still devise a convergent Newton algorithm, which is in fact equivalent to an (efficient to implement) primal-dual active strategy [78, 101], here we prefer to keep $\delta > 0$, as we need the differentiability property of $\alpha \mapsto p(\alpha)$ for the bilevel optimization scheme.

Algorithm 1 displays the corresponding function space version of Newton's method ((3.9)–(3.10), in general terms) specified for solving the equation (3.15).

Before we close this section we would like to make a few remarks regarding the term $\frac{\gamma}{2}\|p\|^2_{L^2}$ in the TV regularised predual problem (3.14). Similar statements apply to the TGV-context. The first remark is that in practice one can set $\gamma = 0$, which suggests that the bilateral constraints impose some type of uniqueness. Alternatively, still for $\gamma = 0$, one may add a projection onto the (nontrivial) kernel of div to the objective of the predual problem in order to guarantee uniqueness of a solution p. The other

Algorithm 1
Function space Newton algorithm for the solution of the regularized TV dual problem (3.14)

Initialise: $p^0 \in H_0^1(\Omega)^d$
while stopping criterion not satisfied **do**
Find $\delta_p^k \in H_0^1(\Omega)^d$ such that

$$-\nabla B^{-1}\mathrm{div}\delta_p^k - \beta\Delta\delta_p^k + \gamma\delta_p^k + \frac{1}{\epsilon}\left(G_\delta''(p^k-\alpha\mathbf{1}) + G_\delta''(-p^k-\alpha\mathbf{1})\right)\delta_p^k = -F(p^k), \quad \text{in } H^{-1}(\Omega)^d$$

Update p^{k+1}:

$$p^{k+1} = p^k + \delta_p^k$$

end while

remark concerns $\gamma > 0$. In this case, this type of L^2-regularization of the predual problem, typically corresponds to a Huber regularization of the regularizing functional in the primal problem. In order to elaborate on this, consider the Huber function $\varphi : \mathbb{R} \to \mathbb{R}^+$ [97]

$$\varphi(t) = \begin{cases} |t| - \frac{\gamma}{2}, & \text{if } |t| \geq \gamma, \\ \frac{1}{2\gamma}t^2, & \text{if } |t| \leq \gamma \end{cases} \tag{3.18}$$

which constitutes a local smoothing of the absolute value function. Then one can define the (isotropic)*Huber* total variation functional as follows

$$\mathrm{TV}_\varphi(u) = \int_\Omega \varphi(|\nabla u|)\,dx + |D^s u|(\Omega), \quad u \in \mathrm{BV}(\Omega) \tag{3.19}$$

as well as the corresponding variational problem

$$\min_{u\in\mathrm{BV}(\Omega)} \frac{1}{2}\|Tu - f\|^2_{L^2(\Omega)} + \alpha\mathrm{TV}_\varphi(u), \tag{3.20}$$

(here written only) for a scalar $\alpha > 0$. Consequently, one can show (see [75]) that the predual of (3.20) is (up to a constant)

$$\begin{aligned} &\min_{p\in W_0^d(\mathrm{div};\Omega)} \frac{1}{2}\|\mathrm{div}p + T^*f\|_B^2 + \frac{\gamma}{2\alpha}\|p\|^2_{L^2(\Omega)} \\ &\text{such that} \quad |p(x)|_\infty \leq \alpha(x) \text{ for a.e. } x \in \Omega. \end{aligned} \tag{3.21}$$

Also in [75] an alternative Newton approach for the solution of the TV minimization problem was proposed. Indeed, instead of using Newton to solve the first-order optimality condition of the regularized predual problem, it was used to solve the primal-dual optimality conditions of the type (2.10)–(2.11) which correspond to the primal and the predual Huber-TV problems (3.20), (3.21) respectively, with an additional small H^1-regularity term $\frac{\mu}{2}\int_\Omega |\nabla u|^2 dx$ in the primal problem. As μ, with $0 < \mu \ll \alpha$, tends to zero a solution of the Huber-TV problem is approached.

4 Bilevel Optimization

In order to construct a structured non-smooth TV- or TVG based prior (or regularization term) we will resort to a bilevel optimization technique. Given some image data f, such an approach aims at

optimizing with respect to both simultaneously, the regularization weight $\boldsymbol{\alpha}$ and the reconstruction u. For this purpose we collect several basic notions and concepts in bilevel programming. For the ease of exposition we tailor the general discussion to the problem structure which is relevant in our subsequent image processing context. For a general introduction to bilevel programming in finite dimensions we refer to [59, 106], and to [13, 114] for a respectively limited access to infinite dimensional settings.

4.1 Background

Consider the following hierarchical optimization problem

$$\begin{aligned} &\text{min} \quad \mathcal{F}(v,\beta) \quad \text{over } (v,\beta) \\ &\text{s.t. } \beta \in B_{\text{ad}} \subset B, \\ &\qquad v \in \operatorname{argmin}\{\mathcal{G}(w;\beta) : w \in V_{\text{ad}} \subset V\}. \end{aligned} \tag{4.1}$$

Here, V and B denote Banach spaces, V_{ad} and B_{ad} associated non-empty, closed and convex subsets, respectively. Further, $\mathcal{F} : V \times B \to \mathbb{R}$ and $\mathcal{G} : V \to \mathbb{R}$ are assumed continuously differentiable functions. Problem (4.1) constitutes a *bilevel* optimization problem for (v,β), where $\mathcal{F}$ represents the upper level objective and B_{ad} the upper level constraint set. Further, $\mathcal{G}$ is the objective of the lower level optimization problem with lower level constraints V_{ad}. Note that the optimization variable β enters the lower level problem as a parameter only, i.e., we may consider $v = v(\beta)$, assuming that the lower level problem admits a solution. Clearly, more general settings of (4.1) with respect to constraints and objective requirements (on both levels, respectively) are conceivable, but would go beyond the scope of this presentation.

Bilevel problems are notoriously delicate. In order to grasp some of the inherent complexities, we next discuss exemplarily some of the structural challenges. First, in case $\mathcal{G}$ is non-convex (or V_{ad} is non-convex), still assuming existence of a solution to the lower level problem, computing a feasible point of (4.1) requires to compute a global (!) solution to a non-convex minimization problem, which resides in the constraint set of the overall problem. This, however, cannot be guaranteed by any of the current state-of-the-art solvers which makes the computation of a feasible point elusive, not even speaking of a minimizer of (4.1). Moreover, typically optimization solvers operate on first-order characterizations (KKT-system or Euler-Lagrange equation) of solutions. Replacing the lower level problem by KKT-conditions yields an overall minimization problem which follows a standard structure of a nonlinear program, but this new problem would not be equivalent to the original bilevel problem. For instance, when $V_{\text{ad}} := \{w \le b\} \subset H_0^1(\Omega)$ with $b \in H_0^1(\Omega)$ the KKT-system associated with the lower level problem reads

$$\mathcal{G}'(v;\beta) + \lambda = 0 \text{ in } H^{-1}(\Omega), \tag{4.2}$$

$$v \le b, \quad \lambda \ge 0 \text{ in } H^{-1}(\Omega), \quad \langle \lambda, v - b\rangle_{H^{-1},H_0^1} = 0, \tag{4.3}$$

where $\langle\cdot,\cdot\rangle_{H^{-1},H_0^1}$ denotes the duality pairing between the Sobolev space $H_0^1(\Omega) =: V$ and its (topological) dual $H^{-1}(\Omega) = V^*$. In this case, using this KKT-system instead of the lower level problem would yield the following *mathematical program with equilibrium constraints (MPEC)* (see, e.g., [106, 79] for rather general accounts of MPECs)

$$\begin{aligned} &\text{min} \quad \mathcal{F}(v,\beta) \quad \text{over } (v,\lambda,\beta) \\ &\text{s.t. } \beta \in B_{\text{ad}} \subset B, \\ &\qquad \mathcal{G}'(v;\beta) + \lambda = 0 \text{ in } H^{-1}(\Omega), \\ &\qquad v \le b, \quad \lambda \ge 0 \text{ in } H^{-1}(\Omega), \quad \langle \lambda, v - b\rangle_{H^{-1},H_0^1} = 0, \end{aligned} \tag{4.4}$$

which is no longer equivalent to (4.1), in general. Equivalence is, however, regained when $\mathcal{G}$ is convex. In this case, the KKT-system is necessary and sufficient for characterizing a solution to the lower level problem. Hence, one may study (4.4) analytically and also numerically. But the reader should be cautioned that (4.4), while appearing to be yet another nonlinear program in Banach space as studied, e.g., in [149], suffers from notoriously degenerate constraints. In order to illustrate this, we focus only on the *complementarity system*

$$v \leq b, \quad \lambda \geq 0 \text{ in } H^{-1}(\Omega), \quad \langle \lambda, v - b \rangle_{H^{-1}, H_0^1} = 0 \tag{4.5}$$

in the constraint set of (4.4). Moreover, to ease the discussion we assume that λ admits $L^2(\Omega)$ regularity. The latter is often realistic under addition regularity requirements on the problem data. In this case we have

$$\langle \lambda, v - b \rangle_{H^{-1}, H_0^1} = 0 \quad \Leftrightarrow \quad \int_\Omega \lambda\,(v - b) dx = 0 \quad \Leftrightarrow \quad \lambda\,(v - b) = 0 \text{ a.e. in } \Omega.$$

From the latter relation it is straightforward to see that the typical surjectivity requirement of linearized constraints for the existence of multipliers (see [149]; Zowe-Kurcyusz-Robinson (ZKR) constraint qualification (CQ)) fails at the presence of a *biactive set*

$$\mathcal{B} := \{x \in \Omega : \lambda(x) = 0,\, v(x) = b(x)\}$$

at a solution (v, λ, β) with $|\mathcal{B}| > 0$ where $|\cdot|$ denotes the d-dimensional Lebesgue measure. This jeopardizes the existence of (bounded) Lagrange multipliers and thus the availability of a KKT-system, which is typically the starting point for numerical solvers.

Assuming that $\mathcal{G}$ is strictly convex (over the non-empty, closed and convex set V_{ad}) yields a unique solution $v = v(\beta)$ of the lower level problem. One may now reduce (4.1) to a problem in β by considering $\beta \mapsto v(\beta)$. This results in

$$\begin{aligned} &\min \quad \hat{\mathcal{F}}(\beta) := \mathcal{F}(v(\beta), \beta) \\ &\text{s.t. } \beta \in B_{\text{ad}} \subset B. \end{aligned} \tag{4.6}$$

In case B_{ad} is simple, this may at first glance appear to be a tractable minimization problem. A closer inspection, however, reveals that while $\beta \mapsto v(\beta)$ is often found to be (locally) Lipschitz continuous, this map is not Fréchet differentiable in general. Consequently, $\hat{\mathcal{F}}$ is typically not Fréchet differentiable, even if $\mathcal{F} : V \times B \to \mathbb{R}$ enjoys Fréchet differentiability. Moreover, even if all constituents in (4.1) are convex, the reduced objective $\hat{\mathcal{F}} : B \to \mathbb{R}$ may still be non-convex only. Summarizing, solving (4.6) amounts to minimizing a non-smooth and non-convex objective over a constraint set B_{ad}, which is a very delicate task. Some of the involved issues relate to computability of generalized derivatives of $\hat{\mathcal{F}}$, the design of an optimization solver guaranteeing descent from one iterate to the next, and the availability of a computable stopping rule. Concerning the latter note that due to the non-smoothness of $\hat{\mathcal{F}}$ one can only expect a set-valued generalized derivative $\partial\hat{\mathcal{F}}(\beta)$ at a solution β. Given the properties of B_{ad}, a stationarity condition of (4.6) then reads

$$0 \in \partial\hat{\mathcal{F}}(\beta) + N_{B_{\text{ad}}}(\beta), \tag{4.7}$$

where we assume that a sum rule for differentiation applies, and $N_{B_{\text{ad}}}(\beta)$ denotes the normal cone to B_{ad} at β. Note that $N_{B_{\text{ad}}}(\beta) = \partial I_{B_{\text{ad}}}(\beta)$ where $I_{B_{\text{ad}}}$ is the indicator function of B_{ad}, i.e., $I_{B_{\text{ad}}}(\beta) = 0$ if $\beta \in B_{\text{ad}}$ and $I_{B_{\text{ad}}}(\beta) = +\infty$ otherwise, and '∂' is the subdifferential of convex analysis. Now, an iterative solver of (4.6) typically has only one generalized gradient of $\hat{\mathcal{F}}$ at its disposal per iteration.

Assuming simple constraints in B_{ad} allowing for an explicit characterization of $N_{B_{\text{ad}}}(\beta)$, this still renders checking of (4.7) at an iterate elusive as information on the entire set $\partial\hat{\mathcal{F}}(\beta)$ would be needed, in general. If, however, $\hat{\mathcal{F}}$ is Fréchet differentiable at β, then (4.7) requires to compute $\hat{\mathcal{F}}'(\beta)$ and to check whether

$$-\hat{\mathcal{F}}'(\beta) \in N_{B_{\text{ad}}}(\beta),$$

which is tractable when $N_{B_{\text{ad}}}(\beta)$ admits an explicit representation.

Let us next assume that B is reflexive, $B \ni \beta \mapsto v(\beta) \in V$ is directionally differentiable, rendering $\hat{\mathcal{F}}$ directionally differentiable as well. Starting from (4.7) and noting that $T_{B_{\text{ad}}}(\beta) = N_{B_{\text{ad}}}(\beta)^\circ := \{d \in B : \langle d, \mu\rangle_{B,B^*} \leq 0 \text{ for all } \mu \in N_{B_{\text{ad}}}(\beta)\}$, we characterize a solution β to (4.6) by

$$\hat{\mathcal{F}}'(\beta; d) \geq 0, \quad \text{for all } d \in T_{B_{\text{ad}}}(\beta). \tag{4.8}$$

Here, $\hat{\mathcal{F}}'(\beta;\cdot)$ denotes the directional derivative of $\hat{\mathcal{F}}$ at β and $T_{B_{\text{ad}}}(\beta)$ is the tangent cone to B_{ad} at β. The latter is the polar cone of $N_{B_{\text{ad}}}(\beta)$, i.e., $N_{B_{\text{ad}}}(\beta)^\circ$. The characterization (4.8) is called *B(ouligand)*-stationarity condition. In the stated form, it is generally difficult to check numerically. With the availability of an explicit characterization of $v'(\beta;\cdot)$, however, this condition may be related to a specific minimization task; see, e.g., [82]. In such cases, often also equivalence to so-called *strong stationarity* may be shown. Through the explicit characterization of $v'(\beta;\cdot)$, strong stationarity (or briefly S-stationarity) is a primal-dual condition (similar to the KKT-system in classical nonlinear programming, with the (Lagrange) multipliers constituting the dual variables [149]), whereas B-stationarity is a purely primal concept. It is well known that in general it is not possible to devise a numerical solution scheme which guarantees to terminate successfully at an S-stationary point. In special instances, e.g., under certain constraint regularity, however, algorihms may discover S-stationary points; see, e.g., [9] for a finite dimensional account and [81] for a specific infinite dimensional instance.

Depending on the problem structure, besides B- and S-stationarity the notions of C(larke)- and M(ordukhovich)-stationarity are relevant. From the viewpoint of numerical computability, C-stationarity is the most attractive format; see, e.g., [92]. For more on stationarity conditions in infinite dimensions we refer to [79, 82] and the references therein, and to [129] for an exposition in finite dimensions.

We end this section by returning to (4.1) under the assumption that $V_{\text{ad}} = V$ and $\mathcal{G}$ is (twice) continuously differentiable and convex. In view of convexity of the lower level problem and (4.4) we may replace (4.1) equivalently by

$$\begin{aligned} &\min \quad \mathcal{F}(v,\beta) \quad \text{over } (v,\beta) \\ &\text{s.t. } \beta \in B_{\text{ad}} \subset B, \\ &\qquad \mathcal{G}'(v;\beta) = 0 \text{ in } V^*. \end{aligned} \tag{4.9}$$

Since $\mathcal{G}'$ is once more differentiable, one may apply classical KKT-theory in Banach space [149] to obtain a primal-dual characterization of a solution to (4.9), provided $\mathcal{G}''(v,\beta)$ along with B_{ad} satisfies the ZKR-CQ. This CQ also allows to locally reduce the problem through the implicitly defined function $\beta \mapsto v(\beta)$ with $\mathcal{G}'(v(\beta),\beta) = 0$. Similar to before, this yields (4.6). Under our smoothness requirements and non-degeneracy assumption on $\mathcal{G}''$ at $(v(\beta),\beta)$ (through ZKR-CQ) the implicit function theorem yields (local) differentiability of $\beta \mapsto v(\beta)$ and, thus, of $\beta \mapsto \hat{\mathcal{F}}(\beta)$. In this case, (4.6) may be solved numerically by a projected gradient method, provided the structure of B_{ad} is sufficiently simple allowing for an efficient computation of the projection of the iterates. The latter viewpoint will be relevant in our numerical approach to computing structured non-smooth priors in the subsequent sections of this paper.

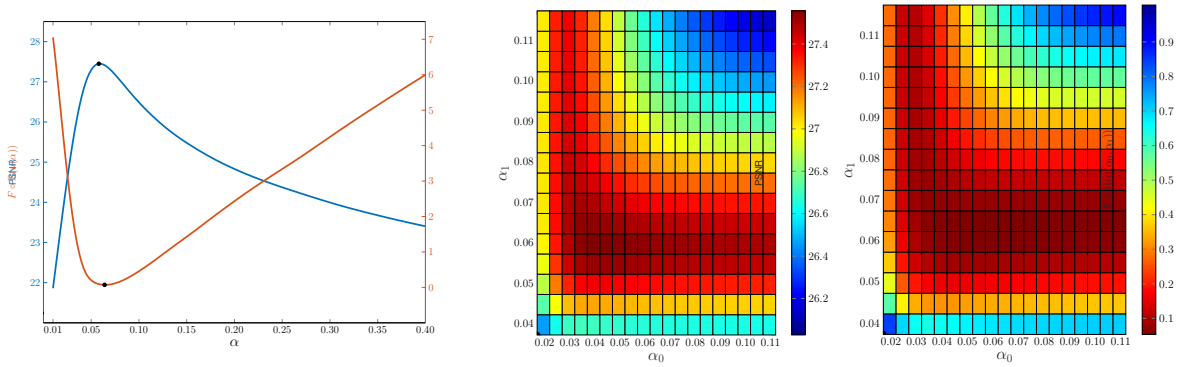

Figure 5: Illustration for the suitability of the functional $F(\mathrm{R}\cdot)$, see (4.10)–(4.11), as an upper level objective. Evaluation of $F(\mathrm{R}u)$ where u solves the (anisotropic, unregularized) primal TV (left) and TGV (right) denoising problems for a variety of scalar regularization parameters α and (α_0, α_1), respectively. The points where this functional is minimal, respectively, is in both cases close to the points that maximize the PSNR. The minimum (for $F(\mathrm{R}\cdot)$) and maximum (for PSNR) points are denoted with a bullet in all graphs. We used the values $\underline{\sigma} = 0.00798$, $\overline{\sigma} = 0.01202$ and w a spatially invariant averaging filter of size 7×7.

4.2 Bilevel optimization–a monolithic approach

The exposition of this section is related to [85, 88] and the references therein which address an automated statistics-based selection of the weight function $\boldsymbol{\alpha}$ in TV minimization problems. In this way, a data-motivated structuring of a non-smooth prior will be obtained. As we mentioned in the introduction as well, the main idea of these approaches is to minimize a suitable upper level objective over both, the image u and the regularization parameter $\boldsymbol{\alpha}$ subject to u being a solution to a (regularized) variational regularization problem with parameter $\boldsymbol{\alpha}$. As a consequence of Fenchel-Rockafellar duality, equivalently the predual problem can be used instead of the primal problem in the lower level.

The upper level objective we discuss here is based on localized residuals. The latter are defined as a function $R : L^2(\Omega) \to L^\infty(\Omega)$ by

$$\mathrm{R}u(x) := \int_\Omega w(x,y)(Tu - f)^2(y)dy \tag{4.10}$$

We note that $\mathrm{R}u(x)$ can be interpreted as a local variance. Indeed, given Gaussian noise of variance σ^2 and zero mean, we have $\int_\Omega (Tu_{true} - f)^2\,dx = \int_\Omega \eta^2\,dx = \sigma^2|\Omega|$. Consequently, if a reconstructed image u is close to u_{true} then it is expected that for every $x \in \Omega$, $\mathrm{R}u(x)$ will be close to σ^2. Hence it is natural to consider an upper level objective which aims to keep Ru within certain tight bounds $\underline{\sigma}, \overline{\sigma}$ where $\underline{\sigma} < \sigma < \overline{\sigma}$. This can be achieved through the function $F : L^2(\Omega) \to \mathbb{R}$ with

$$F(v) := \frac{1}{2}\int_\Omega \max(v - \overline{\sigma}^2, 0)^2 dx + \frac{1}{2}\int_\Omega \min(v - \underline{\sigma}^2, 0)^2 dx. \tag{4.11}$$

Hence, by minimizing $F(\mathrm{R}u)$ one aims to find $\mathrm{R}u$ close to the variance σ^2. In order to illustrate the viability of this choice of an upper level objective we evaluate the upper objective functional $F(\mathrm{R}\cdot)$ for a series of scalar TV denoising restorations obtained from a variety of parameters α for the example image shown in the introduction (Figures 1, 2 and 3). As we see in Figure 5, the minimum value of the functional $F(\mathrm{R}\cdot)$ is achieved for scalar parameter values α (for TV), and (α_0, α_1) (for TGV) that

are not far away from those that are optimizing the PSNR. Note moreover that in order to optimize the PSNR one needs the ground truth image u_{true}! This, however, is not the case for $F(R\cdot)$.

As we have also discussed previously we must impose also certain regularity for the regularization parameters $\boldsymbol{\alpha}$ in order for the dualization results to hold. Here we impose H^1-regularity which turns out to also facilitates existence and differentiability analysis, even though the consideration of lower regularity of α is certainly of interest. Furthermore, for reasons we discuss later, we need to constrain these parameters in box sets of the type

$$\mathcal{A}_{ad} = \{\alpha \in H^1(\Omega) : \ \underline{\alpha} \leq \alpha \leq \overline{\alpha}\}, \tag{4.12}$$

$$\begin{aligned} \mathcal{A}^0_{ad} &= \{\alpha_0 \in H^1(\Omega) : \ \underline{\alpha}_0 \leq \alpha_0 \leq \overline{\alpha}_0\}, \\ \mathcal{A}^1_{ad} &= \{\alpha_1 \in H^1(\Omega) : \ \underline{\alpha}_1 \leq \alpha_1 \leq \overline{\alpha}_1\}, \end{aligned} \tag{4.13}$$

for TV and TGV, respectively. Here $\underline{\alpha}, \overline{\alpha} \in L^2(\Omega)$ with $0 < \epsilon_0 \leq \underline{\alpha}(x) < \overline{\alpha}(x) - \epsilon_1$ in Ω for some $\epsilon_1 > 0$ and similary for $\underline{\alpha}_0, \overline{\alpha}_0$ and $\underline{\alpha}_1, \overline{\alpha}_1$.

Recalling that the localized residuals can be expressed in terms of the dual variables p as

$$\mathrm{R}u(x) = R(\mathrm{div}p)(x) = \int_\Omega w(x,y) \left(TB^{-1}\mathrm{div}p - (TB^{-1}T^* - I)f\right)^2 dy, \tag{4.14}$$

$$\mathrm{R}u(x) = R(\mathrm{div}^2 p)(x) = \int_\Omega w(x,y) \left(TB^{-1}\mathrm{div}^2 p - (TB^{-1}T^* - I)f\right)^2 dy, \tag{4.15}$$

for the TV and TGV problems, respectively, the associated bilevel problems (which are related to (4.9) with appropriate settings for the various constituents) are given by

$$\begin{cases} \min J_{\mathrm{TV}}(p,\alpha) := F(R(\mathrm{div}p)) + \dfrac{\lambda}{2}\|\alpha\|^2_{H^1(\Omega)} \qquad \text{over } (p,\alpha) \in H^1_0(\Omega)^d \times \mathcal{A}_{ad} \\ \text{s.t.} \quad p \in \underset{p \in H^1_0(\Omega)^d}{\mathrm{argmin}} \dfrac{\beta}{2}\|\nabla p\|^2_{L^2(\Omega)^d} + \dfrac{\gamma}{2}\|p\|^2_{L^2(\Omega)^d} + \dfrac{1}{2}\|\mathrm{div}p + T^*f\|^2_B + \dfrac{1}{\epsilon}\mathcal{P}_\delta(p,\alpha), \end{cases} \tag{$\mathbb{P}_{\mathrm{TV}}$}$$

and

$$\begin{cases} \min J_{\mathrm{TGV}}(p,\alpha_0,\alpha_1) := F(R(\mathrm{div}^2 p)) + \dfrac{\lambda_0}{2}\|\alpha_0\|^2_{H^1(\Omega)} + \dfrac{\lambda_1}{2}\|\alpha_1\|^2_{H^1(\Omega)} \\ \qquad \text{over } (p,\alpha) \in H^2_0(\Omega, \mathcal{S}^{d\times d}) \times \mathcal{A}^0_{ad} \times \mathcal{A}^1_{ad} \\ \text{s.t.} \quad p \in \underset{p \in H^2_0(\Omega,\mathcal{S}^{d\times d})}{\mathrm{argmin}} \dfrac{\beta}{2}\|\Delta p\|^2_{L^2(\Omega,\mathcal{S}^{d\times d})} + \dfrac{\gamma}{2}\|p\|^2_{L^2(\Omega,\mathcal{S}^{d\times d})} + \dfrac{1}{2}\|T^*f - \mathrm{div}^2 p\|^2_B \\ \qquad\qquad + \dfrac{1}{\epsilon_0}\mathcal{Q}_\delta(p,\alpha_0) + \dfrac{1}{\epsilon_1}\mathcal{P}_\delta(\mathrm{div}p,\alpha_1). \end{cases} \tag{$\mathbb{P}_{\mathrm{TGV}}$}$$

Here, $\mathcal{S}^{d\times d}$ refers to the set of real, symmetric $d \times d$ matrices.

Existence of solutions to the problems ($\mathbb{P}_{\mathrm{TV}}$) and ($\mathbb{P}_{\mathrm{TGV}}$) can be shown using standard arguments; see for instance [85, Theorem 6.1] for the TV-case. We note that as we have shown in the dualization section, H^1-regularity of α suffices to establish the connection between the primal and predual problems in the TV case. In [88], α was further enforced to be a $C(\overline{\Omega})$-function which was guaranteed by a regularity result of H^1-projections onto the set $\mathcal{A}_{ad}$ which was part of the associated solution algorithm. In particular the following result was shown [88, Corollary 2.3].

Proposition 4.1. *Let $\Omega \subset \mathbb{R}^d$, $d \leq 3$, be a bounded convex set and*

$$\mathcal{A} = \{\alpha \in H^1(\Omega) : \ \underline{\alpha} \leq \alpha \leq \overline{\alpha}\}, \quad \textit{with } \underline{\alpha}, \overline{\alpha} \in H^2(\Omega), \quad \frac{\partial \underline{\alpha}}{\partial \nu} = \frac{\partial \overline{\alpha}}{\partial \nu} = 0. \tag{4.16}$$

Let $P_{\mathcal{A}} : H^1(\Omega) \to \mathcal{A} \subset H^1(\Omega)$ denote the projection operator, that is

$$P_{\mathcal{A}}(\omega) := \operatorname*{argmin}_{\alpha \in \mathcal{A}} \frac{1}{2}\|\alpha - \omega\|^2_{H^1(\Omega)}, \quad \omega \in H^1(\Omega).$$

Then, if $\omega^* = P_{\mathcal{A}}(\omega)$, the following implication holds true:

$$\omega \in H^2(\Omega) \quad \textit{and} \quad \frac{\partial \omega}{\partial \nu} = 0 \qquad \Longrightarrow \qquad \omega^* \in H^2(\Omega) \quad \textit{and} \quad \frac{\partial \omega^*}{\partial \nu} = 0.$$

Furthermore,

$$\max(\|\omega^*\|_{H^2(\Omega)}, \|\omega^*\|_{C^{0,r}(\overline{\Omega})}) \leq C\left(\|L\omega\|_{L^2(\Omega)} + \|L\underline{\alpha}\|_{L^2(\Omega)} + \|L\overline{\alpha}\|_{L^2(\Omega)}\right)$$

for some $r \in (0,1)$ and with $L = -\Delta + I$.

Hence, in order for $C(\overline{\Omega})$ regularity for the weight function α to be guaranteed by the projection it suffices that $\underline{\alpha}, \overline{\alpha}$ satisfy the conditions in (4.16). In particular this is satisfied when these are constants. Furthermore, since–as we will see later in the algorithm–this projection is performed in an iterative fashion, the initialization for α must satisfy (4.16).

As we have discussed in the previous section, it is not yet shown that $W^{1,1}$ regularity for α_0 and α_1 suffices to establish a dualization framework for TGV, even though one expects that it can be shown with similar arguments. Hence, for TGV we will follow the H^1-projection regularity result as described above. In fact as we will see in the (preliminary) numerics for TGV, we treat only α_1 as a spatially varying function and α_0 remains a constant.

We make another remark regarding the box constraints for the parameters (4.12) and (4.13). In [56] it was shown that a PSNR-optimizing upper level objective $\tilde{J}(u,\alpha) = \|u(\alpha) - f\|^2_{L^2(\Omega)}$ subject to H^1 and Huber-regularized TV and TGV denoising problems produces under some mild conditions optimal scalar solutions α and (α_0, α_1) that are strictly positive. This, however, appears to require to solve a nonconvex problem to global optimality. Although, as we have already depicted in Figure 5, the upper level objective we discuss here is not far away from optimizing the PSNR, keeping the parameters strictly positive via (4.12) and (4.13) seems indeed necessary.

Let us next address how to treat the bilevel problems ($\mathbb{P}_{\mathrm{TV}}$) and ($\mathbb{P}_{\mathrm{TGV}}$) algorithmically. For this purpose, let $\alpha \mapsto p(\alpha)$ and $(\alpha_0, \alpha_1) \mapsto p(\alpha_0, \alpha_1)$ denote the solution map of the lower level problems, equivalently the solutions of the optimality conditions (3.15) and analogously for the TGV problem. Then the problems ($\mathbb{P}_{\mathrm{TV}}$) and ($\mathbb{P}_{\mathrm{TGV}}$) admit the following reduced versions

$$\min \qquad \hat{J}_{\mathrm{TV}}(\alpha) := J_{\mathrm{TV}}(p(\alpha), \alpha), \qquad \text{over } \alpha \in \mathcal{A}_{ad} \qquad , \tag{4.17}$$

$$\min \quad \hat{J}_{\mathrm{TGV}}(\alpha_0, \alpha_1) := J_{\mathrm{TGV}}(p(\alpha_0, \alpha_1), \alpha_0, \alpha_1), \quad \text{over } \alpha_0 \in \mathcal{A}^0_{ad},\ \alpha_1 \in \mathcal{A}^1_{ad}. \tag{4.18}$$

In view of our general discussion in the preceding subsection we have arrived at the level of (4.6) with corresponding settings for $\hat{\mathcal{F}}$, β, $v(\beta)$ and B_{ad} and a smooth, i.e., differentiable reduced objective. Indeed, the reduced functional $\hat{J}_{\mathrm{TV}} : H^1(\Omega) \to \mathbb{R}$ is differentiable as a composition of differentiable functions. One can show a similar results for $\hat{J}_{\mathrm{TGV}}$, as well. We note however that from now on we will consider only scalar α_0 yielding $\mathcal{A}^0_{ad} = \{\alpha_0 \in \mathbb{R} : \ \underline{\alpha}_0 \leq \alpha_0 \leq \overline{\alpha}_0\}$ where $\underline{\alpha}_0, \overline{\alpha}_0 \in \mathbb{R}$, and also we set $\lambda_0 := 0$ in ($\mathbb{P}_{\mathrm{TGV}}$).

In order to proceed, we now recall some basic facts from optimization (Karush-Kuhn-Tucker) theory in Banach spaces [149] adapted to our purposes. Let $V, \mathcal{A}, Z$ be Banach spaces, $X = V \times \mathcal{A}$,

$\theta : X \to \mathbb{R}$ and $g : X \to Z$ continuously Fréchet differentiable functions. Let, moreover, $\mathcal{C} \subset X$ be a non-empty, closed convex set. Consider the problem

$$\begin{cases} \min\limits_{\mathrm{x}\in X} \theta(\mathrm{x}) \\ \text{s.t.} \quad \mathrm{x} \in \mathcal{C} \text{ and } g(\mathrm{x}) = 0. \end{cases} \tag{4.19}$$

and suppose that the following constraint qualification holds for a solution $\overline{\mathrm{x}}$ of (4.19)

$$g'(\overline{\mathrm{x}})[C(\overline{\mathrm{x}})] + \{\lambda g(\overline{\mathrm{x}}) : \ \lambda \geq 0\} = Z, \quad \text{where } C(\overline{\mathrm{x}}) = \{\lambda(c - \overline{\mathrm{x}}) : \ c \in \mathcal{C}, \ \lambda \geq 0\}. \tag{4.20}$$

Then there exists an adjoint state (Lagrange multiplier) $z^* \in Z^*$ that fulfills the following condition [149]

$$\theta'(\overline{\mathrm{x}}) - z^*(g'(\overline{x})) \in C(\overline{\mathrm{x}})^+, \tag{4.21}$$

where $C(\overline{\mathrm{x}})^+ := \{\mathrm{x}^* \in X^* : \ \mathrm{x}^*(\mathrm{x}) \geq 0, \text{ for all } x \in C(\overline{\mathrm{x}})\}$.

Assume now that for every $\xi \in \mathcal{A}$ the equation $g(\mathrm{x}) = 0$ has a unique solution $v(\xi) \in U$, that is $g(v(\xi), \xi) = 0$. Further assume that $g_v(v(\xi), \xi) \in \mathcal{L}(V, Z)$ is continuously differentiable. Then from the implicit function theorem (see, e.g., [94] for details), we get that $v(\xi)$ is continuously differentiable. Then one can define the reduced problem

$$\begin{cases} \min\limits_{\xi\in\mathcal{A}} \hat{\theta}(\xi) := \theta(v(\xi), \xi) \\ \text{s.t.} \quad \xi \in \hat{\mathcal{C}} := \{\xi \in \mathcal{A} : \ (v(\xi), \xi) \in \mathcal{C}\}, \end{cases} \tag{4.22}$$

where now $\hat{\theta}$ is differentiable as well. The derivative of the reduced functional $\hat{\theta}' \in \mathcal{A}^*$ can be then computed with the help of z^* as follows

$$\hat{\theta}'(\xi) = \theta_\xi(v(\xi), \xi) + g_\xi(v(\xi), \xi)^* z^*, \tag{4.23}$$

where $g_\xi(v(\xi), \xi)^* \in \mathcal{L}(Z^*, \mathcal{A}^*)$ is the adjoint operator of $g_\xi(v(\xi), \xi) \in \mathcal{L}(\mathcal{A}, Z)$.

Coming back to our bilevel problems, we can easily see that the above abstract framework can be adjusted to the problems ($\mathbb{P}_{\mathrm{TV}}$), ($\mathbb{P}_{\mathrm{TGV}}$) and their corresponding reduced problems (4.17) and (4.18). Indeed considering ($\mathbb{P}_{\mathrm{TV}}$) first, we set $V = H_0^1(\Omega)^d$, $\mathcal{A} = H^1(\Omega)$, $Z = V^*$, $\mathcal{C} = V \times \mathcal{A}_{ad}$, $T(\mathrm{x}) = J_{\mathrm{TV}}(p, \alpha)$ and

$$g_{\mathrm{TV}}(\mathrm{x}) = -\beta\Delta p + \gamma p - \nabla B^{-1}\mathrm{div} p - \nabla B^{-1} T^* f + \frac{1}{\epsilon} P_\delta(p, \alpha)$$

Then it can be shown that all necessary differentiability results hold; see [85] for details, and the corresponding condition to (4.21) read for the adjoint variable $q \in V$ and an optimal pair $(\overline{p}, \overline{\alpha}) \in V \times \mathcal{A}$

$$\langle \mathrm{div}^* J_0'(\mathrm{div}\overline{p}), p\rangle_{V^*,V} + \langle -\beta\Delta q + \gamma q - \nabla B^{-1}\mathrm{div} q + \frac{1}{\epsilon} D_1 P_\delta(\overline{p}, \overline{\alpha}) q, p\rangle_{V^*,V} = 0, \tag{4.24}$$

$$\langle \lambda(-\Delta + I)\overline{\alpha} + \frac{1}{\epsilon}(D_2 P_\delta(\overline{p}, \overline{\alpha}))^* q, \alpha - \overline{\alpha}\rangle_{\mathcal{A}^*,\mathcal{A}} \geq 0, \tag{4.25}$$

for all $p \in V$ and for all $\alpha \in \mathcal{A}_{ad}$, where $J_0 := F(R\cdot)$. Moreover D_1 and D_2 denote derivatives with respect to the variables p and α, respectively.

The reduced derivative of $\hat{J}_{\mathrm{TV}}$ can be computed as

$$\hat{J}'_{\mathrm{TV}}(\alpha) = \lambda(-\Delta + I)\alpha + \frac{1}{\epsilon}(D_2 P_\delta(p(\alpha), \alpha))^* q(\alpha) \tag{4.26}$$

where $q(\alpha)$ solves (4.24) for $\overline{\alpha} = \alpha$ and $\overline{p} = p(\alpha)$.

Turning now to the bilevel TGV-problem ($\mathbb{P}_{\mathrm{TGV}}$), the corresponding, spaces, sets and functions will be $V = H_0^2(\Omega, \mathcal{S}^{d\times d})$, $\mathcal{A} = \mathbb{R} \times H^1(\Omega)$, $Z = V^*$, $\mathcal{C} = V \times \mathcal{A}_{ad}^0 \times \mathcal{A}_{ad}^1$, $T(\mathrm{x}) = J_{\mathrm{TGV}}(p, \alpha_0, \alpha_1)$ and

$$g(\mathrm{x}) = \beta\Delta^2 p + \gamma p + \nabla^2 B^{-1}\mathrm{div}^2 p - \nabla^2 B^{-1} T^* f + \frac{1}{\epsilon_0} Q_\delta(p, \alpha_0) - \frac{1}{\epsilon_1}\nabla P_\delta(\mathrm{div} p, \alpha_1)$$

The adjoint variable $q \in V$ satisfies for an optimal triplet $(\overline{p}, \overline{\alpha}_0, \overline{\alpha}_1)$

$$\begin{aligned}\langle(\mathrm{div}^2)^* J_0'(\mathrm{div}^2\overline{p}, p)\rangle_{V^*,V} + \langle\beta\Delta q + \gamma p + \nabla^2 B^{-1}\mathrm{div}^2 q + \frac{1}{\epsilon_0} D_1 Q_\delta(\overline{p}, \overline{\alpha}_0) q \\ - \frac{1}{\epsilon_1} D_1 \nabla P_\delta(\overline{p}, \overline{\alpha}_1) q, p\rangle_{V^*,V} = 0,\end{aligned} \tag{4.27}$$

$$\langle\lambda(-\Delta + I)\overline{\alpha}_1 - \frac{1}{\epsilon_1}(D_2\nabla P_\delta(\overline{p}, \overline{\alpha}_1))^* q, \alpha_1 - \overline{\alpha}_1\rangle_{H^1(\Omega)^*, H^1(\Omega)} \geq 0, \tag{4.28}$$

$$\langle\frac{1}{\epsilon_0}(D_2 Q_\delta(\overline{p}, \overline{\alpha}_0))^* q, \alpha_0 - \overline{\alpha}_0\rangle_{\mathbb{R},\mathbb{R}} \geq 0, \tag{4.29}$$

for all $p \in V$, $\alpha_0 \in \mathcal{A}_{ad}^0$ and $\alpha_1 \in \mathcal{A}_{ad}^1$. The derivative of the reduced objective is then computed by

$$\hat{J}'_{\mathrm{TGV}}(\alpha_0, \alpha_1) = \lambda(-\Delta + I)\alpha_1 + \left(\frac{1}{\epsilon_0}(D_2 Q_\delta(p, \alpha_0)), -\frac{1}{\epsilon_1}(D_2\nabla P_\delta(p, \alpha_1))\right)^* q(\alpha_0, \alpha_1) \tag{4.30}$$

where again $q(\alpha_0, \alpha_1)$ solves (4.27) for $\overline{\alpha}_0 = \alpha_0$, $\overline{\alpha}_1 = \alpha_1$ and $\overline{p} = p(\alpha_0, \alpha_1)$.

For the reduced derivatives we have that $\hat{J}'_{\mathrm{TV}}(\alpha) \in H^1(\Omega)^*$ and $\hat{J}'_{\mathrm{TGV}}(\alpha_0, \alpha_1) \in (H^1(\Omega) \times \mathbb{R})^*$. In order to obtain the gradient of these functional which are essential for the design of gradient based descent algorithms we can apply the inverse Riezs maps to the reduced gradients as follows

$$\nabla\hat{J}_{\mathrm{TV}}(\alpha) := \mathcal{R}_{H^1}^{-1}\hat{J}'_{\mathrm{TV}}(\alpha) \in H^1(\Omega), \tag{4.31}$$

$$\nabla\hat{J}_{\mathrm{TGV}}(\alpha_0, \alpha_1) := \left(\mathcal{R}_{H^1}^{-1} P_1 \hat{J}'_{\mathrm{TGV}}(\alpha_0, \alpha_1), P_2 \hat{J}'_{\mathrm{TGV}}(\alpha_0, \alpha_1)\right) \in H^1(\Omega) \times \mathbb{R}, \tag{4.32}$$

where for $(r_1, r_2) \in H^1 \times \mathbb{R}$ we have

$$\hat{J}'_{\mathrm{TGV}}(\alpha_0, \alpha_1)[r_1, r_2] = P_1 \hat{J}'_{\mathrm{TGV}}(\alpha_0, \alpha_1)[r_1] + P_2 \hat{J}'_{\mathrm{TGV}}(\alpha_0, \alpha_1)[r_2].$$

The above representation of the derivative and gradients of the reduced functional together with the regularity results for the H^1–projection provide the basis for devising a function space based projected gradient algorithm. Indeed such an algorithm is devised in [88] (compare *Algorithm 1* in that reference) for the case of bilevel TV and it can be similarly done for bilevel TGV.

Original Standard grid Staggered grid

Figure 6: Comparison of the standard versus staggered grid discretization see [86]. The latter produces a sharper result which is closer to the one belonging to the original functional space problem.

5 Numerical examples

In the section we will provide some numerical examples for the bilevel TV and TGV algorithms which aim to solve the discretized versions of ($\mathbb{P}_{\mathrm{TV}}$) and ($\mathbb{P}_{\mathrm{TGV}}$), respectively. We will also discuss some discretization and algorithmic aspects.

In the discrete setting, (grayscale) images are considered as functions from $\Omega_h \to \mathbb{R}$ where $\Omega_h = \{1, 2, \dots, n\} \times \{1, 2, \dots, m\}$ is a discrete cartesian grid that corresponds to the image pixels. We defined the mesh size, that is the distance between the grid points as $h = 1/\sqrt{nm}$. All the discrete differential operators will be computed on these points. However we also emphasize that computing primal, dual variables as well as differential operators and their adjoints in different points has to be done with care. First, one has to devise discretizations which respect, on the discrete level, the relation between a discrete operator (including boundary conditions) and its associated adjoint. Moreover, so-called *inf-sup*-stability of the discretization is relevant. It takes care of a robust relation between the primal and dual problems under resolution refinements. For the predual of the TV problem, it has been found [86] that a staggered grid discretization of the vector field p and the associated differential operators successfully eliminates some discretization artifacts. We depict such an example in Figure 6 where the dual TV denoising problem (isotropic) is solved by using a standard discretization and a staggered grid as it is described in [86]. It is known theoretically [111] that applying the isotropic L^2–TV denoising problem with data f being a characteristic function of a disk yields as result a rescaled version of f (simple constrast lost). However this is not the case with the standard discretization as a certain smoothing is observed on edges not aligned with the grid. By using a staggered grid this effect is essentially eliminated. We would also like to point out that examples like this demonstrate the necessity of studying variational imaging problems in function space setting in order to study the models independently of discretization artifacts. Further uses of staggered grid discretization methods can also be found in [73, 136].

5.1 Discrete operators for ($\mathbb{P}_{\mathrm{TV}}$)

We will generally keep the same notation for the discrete differential operators (than for continuous ones). We set the following discrete function spaces

$$U_h = \mathbb{R}^{\Omega_h}, \quad W_h = U_h \times U_h.$$

For a function $u \in U_h$ the discrete ℓ^2 norm is defined as

$$\|u\|^2_{\ell^2(\Omega_h)} = h^2 \sum_{(i,j)\in\Omega_h} |u_{i,j}|^2 \tag{5.1}$$

We define the discrete gradient $\nabla : U_h \to W_h$ and divergence operator $\mathrm{div} : W_h \to U_h$ such that they satisfy the adjoint relation $\nabla = -\mathrm{div}^\top$. For $p = (p^1, p^2) \in W_h$ the divergence is defined with standard backward differences as

$$(\mathrm{div}p)_{i,j} = \frac{1}{h}\left(p^1_{i,j} - p^1_{i-1,j} + p^2_{i,j} - p^2_{i,j-1}\right), \quad (i,j) \in \Omega_h$$

setting $p^1_{\tilde{i},\tilde{j}} = p^2_{\tilde{i},\tilde{j}} = 0$ if $(\tilde{i},\tilde{j})$ falls outside the grid (ghost grid points). We also need to defined the discrete vectorial Laplacian $\Delta : W_h \to W_h$ where $\Delta p = (\Delta_D p^1, \Delta_D p^1)$ with $\Delta_D : U_h \to U_h$. In order to be consistent with the function space setting where $p \in H^1_0(\Omega)^2$ the discrete Laplacian must impose zero Dirichlet boundary conditions, hence (following good practice in numerical analysis) we avoid defining Δ_D as $\mathrm{div}\nabla$ as the latter would result in (unsuitable) mixed boundary conditions. Thus, we define Δ_D using the discrete five-point Laplacian stencil

$$\frac{1}{h^2} \times \begin{array}{ccc} & 1 & \\ 1 & -4 & 1 \\ & 1 & \end{array}$$

and again setting $p_{\tilde{i},\tilde{j}} = 0$ for ghost grid points. We will also make use of the discrete Laplacian with zero Neumann boundary conditions $\Delta_N : U_h \to U_h$ which will be used to act on the weight function α. These are the desired boundary conditions for α dictated by Proposition 4.1. For this purpose, we use the same stencil but we set the function value of ghost grid points to be the same as the function value of the nearest grid point in Ω_h. We follow a similar approach for the $\nabla\mathrm{div} : W_h \to W_h$ operator for the denoising case where there $T = B = Id$. We discretize $\nabla\mathrm{div}$ directly as follows

$$\begin{aligned}(\nabla\mathrm{div}p)_{i,j} = \frac{1}{h^2}\big(&p^1_{i+1,j} - 2p^1_{i,j} + p^1_{i-1,j} + p^2_{i+1,j} - p^2_{i+1,j-1} - p^2_{i,j} + p^2_{i,j-1} \\ &p^2_{i,j+1} - 2p^2_{i,j} + p^2_{i,j-1} + p^1_{i,j+1} - p^1_{i-1,j+1} - p^1_{i,j} + p^1_{i-1,j}\big)\end{aligned}$$

by setting again zero values outside the grid. Typically, in the implementations the variables are regarded as long vectors resulting from concatenating the columns of the pixel mask based matrix representation.

We will also need discrete version of H^1-types of norms and norms of the corresponding duals spaces. For the discrete H^1-norm acting on the weight function we use

$$\|\alpha\|_{H^1(\Omega_h)} = h\sqrt{\alpha^\top(I - \Delta_N)\alpha} \tag{5.2}$$

while the dual norm is defined as

$$\|r\|_{H^1(\Omega_h)^*} = \|(I - \Delta_N)^{-1}r\|_{H^1(\Omega_h)} = h\sqrt{r^\top(I - \Delta_N)^{-1}r} \tag{5.3}$$

based on the $H^1(\Omega) \to H^1(\Omega)^*$ Riezs map $\alpha \mapsto r = (I - \Delta_N)\alpha$. We will also use the discrete dual $H^1_0(\mathrm{div}, \Omega_h)$-norm as

$$\|v\|^*_{H^1_0(\mathrm{div},\Omega_h)} = h\sqrt{v^\top(I - \nabla\mathrm{div})^{-1}v}. \tag{5.4}$$

For the discrete version of the averaging filter in the definition of the localized residuals (4.10) we use a spatially invariant averaging filter of size $n_w \times n_w$, that is, with entries of equal value whose sum is equal to one.

With these definitions the discrete version of ($\mathbb{P}_{\mathrm{TV}}$) is the following

$$\begin{cases} \min \dfrac{1}{2}\|(R(\mathrm{div}p) - \overline{\sigma}^2)^+\|^2_{\ell^2(\Omega_h)} + \dfrac{1}{2}\|(\underline{\sigma}^2 - R(\mathrm{div}p))^+\|^2_{\ell^2(\Omega_h)} + \dfrac{\lambda}{2}\|\alpha\|^2_{H^1(\Omega_h)} \\ \quad \text{over } (p,\alpha) \in W_h \times (\mathcal{A}_{ad})_h \\ \text{such that} \quad -\beta\Delta_D p + \gamma p - \nabla B^{-1}\mathrm{div}p - \nabla B^{-1}T^* f + \dfrac{1}{\epsilon}P_\delta(p,\alpha) = 0. \end{cases} \tag{$\mathbb{P}^h_{\mathrm{TV}}$}$$

Here we have

$$(\mathcal{A}_{ad})_h = \{\alpha \in U_h : \ \underline{\alpha} \le \alpha_{i,j} \le \overline{\alpha}, \text{ for all } (i,j) \in \Omega_h\}, \tag{5.5}$$

and $R(\mathrm{div}p)$ being a discrete convolution, for which here we use periodic boundary conditions. The penalty function $P_\delta : W_h \to W_h$ is also defined straightforwardly in the discrete setting by point-wise application of the function G'_δ.

Regarding the choice of the lower and upper bounds for the local variance $\underline{\sigma}^2$ and $\overline{\sigma}^2$, respectively, we follow here the rules

$$\overline{\sigma}^2 = \sigma^2\left(1 + \frac{\sqrt{2}}{n_w}\right), \qquad \underline{\sigma}^2 = \sigma^2\left(1 - \frac{\sqrt{2}}{n_w}\right) \tag{5.6}$$

where σ^2 is the variance of Gaussian noise which is assumed to be known. The derivation of the formulae (5.6) is done using statistics of the extremes; see [88, Section 4.2.1].

We now proceed with describing the algorithm for the numerical solution of ($\mathbb{P}^h_{\mathrm{TV}}$). In fact, we use a discretized projected gradient method with Armijo line search for globalization. The discretized gradient of the reduced functional is computed with the help of the adjoint equation which is the discrete version of (4.24). We summarize this in Algorithm 2.

We note that in the algorithm above $\mathbf{1}$ denotes the matrix $[Id; Id]$ of size $nm \times 2nm$.

A few remarks on the solution of the lower level problem are in order. The target is to solve the problem for sufficiently small ϵ using Algorithm 1. In order to ensure robustness, we employ a path following approach which, starting for a large ϵ^ℓ, successively solves $g_{\mathrm{TV},\epsilon}(p^\ell, \alpha) = 0$ for $\epsilon = \epsilon_\ell$ until a fixed tolerance

$$g_{\mathrm{TV},\epsilon^\ell}(p^{\ell+1}, \alpha) < \mathrm{tol}_{(l)}$$

is reached. Then $\epsilon^{\ell+1}$ is obtained by decreasing ϵ^ℓ by a factor $0 < \theta_\epsilon < 1$. This yields $\epsilon^{\ell+1} := \max(\theta_\epsilon\epsilon^\ell, \underline{\epsilon})$ for some $0 < \epsilon \ll 1$; compare also [88, Algorithm 3].

The projection $P_{(\mathcal{A}_{Ad})_h}$ is computed by applying a semismooth Newton method to the associated minimization problem. It adapts the path-following method developed in [80] to the projection problem; compare [88, Algorithm 4]. We just mention here that the original discretized H^1–projection problem $P_{(\mathcal{A}_{ad})_h}(\tilde{\alpha})$

$$\begin{cases} \min \dfrac{1}{2}\|\alpha - \tilde{\alpha}\|^2_{H^1(\Omega_h)} := \dfrac{h}{2}(\alpha - \tilde{\alpha})^\top(I - \Delta_N)(\alpha - \tilde{\alpha}) \\ \quad \text{over} \quad \alpha \in (\mathcal{A}_{ad})_h = \{\alpha \in U_h : \ \underline{\alpha} \le \alpha_{i,j} \le \overline{\alpha}\} \end{cases} \tag{5.7}$$

is substituted by the following version where the constraint is treated by a penalty term

$$\min_{\alpha \in U_h} \frac{1}{2}\|\alpha - \tilde{\alpha}\|^2_{H^1(\Omega_h)} + \frac{1}{2\epsilon_\alpha}\left(\|(\alpha - \overline{\alpha})^+\|^2_{\ell^2(\Omega_h)} + \|(\underline{\alpha} - \alpha)^+\|^2_{\ell^2(\Omega_h)}\right) \tag{5.8}$$

for some small $\epsilon_\alpha > 0$.

Algorithm 2
Discretized projected gradient method for the solution of the bilevel TV image reconstruction problem $(\mathbb{P}^h_{\mathrm{TV}})$

Input: f, $\underline{\alpha}$, $\overline{\alpha}$, $\overline{\sigma}$, $\underline{\sigma}$, λ, β, γ, ϵ, δ, n_w τ^0, $0 < c < 1$, $0 < \theta_- < 1 \leq \theta_+$
Initialise: $\alpha^0 \in \mathcal{A}^h_{ad}$ and set $k := 0$.
repeat
Use Algorithm 1 to compute the solution p^k of the lower level problem

$$g_{\mathrm{TV}}(p^k, \alpha^k) := -\beta\Delta_D p^k + \gamma p^k - \nabla B^{-1}\mathrm{div}p - \nabla B^{-1}T^* f + \frac{1}{\epsilon}P_\delta(p^k, \alpha^k) = 0$$

Solve the adjoint equation for q^k

$$-\nabla B^{-1}\mathrm{div}q^k - \beta\Delta q^k + \gamma q^k + \frac{1}{\epsilon}\left(G''_\delta(p^k - \alpha^k\mathbf{1}) + G''_\delta(-p^k - \alpha^k\mathbf{1})\right)q^k$$
$$= 2\nabla B^{-1}T^*\mathrm{div}p^k\left(w * \left((R(\mathrm{div}p^k) - \overline{\sigma}^2)^+ - (\underline{\sigma}^2 - R(\mathrm{div}p^k))^+\right)\right)$$

Compute the reduced derivative

$$\hat{J}'_{\mathrm{TV}}(\alpha^k) = \frac{1}{\epsilon}\mathbf{1}^\top\left(-G''_\delta(p^k - \alpha^k\mathbf{1}) + G''_\delta(-p^k - \alpha^k\mathbf{1})\right)q^k + \lambda(I - \Delta_N)\alpha^k$$

Compute the reduced gradient

$$\nabla\hat{J}_{\mathrm{TV}}(\alpha^k) = (I - \Delta_N)^{-1}\hat{J}'_{TV}(\alpha^k)$$

Compute the trial point $\alpha^{k+1} = P_{(\mathcal{A}_{ad})_h}\left(\alpha^k - \tau^k\nabla\hat{J}_{\mathrm{TV}}(\alpha^k)\right)$
while $\hat{J}_{\mathrm{TV}}(\alpha^{k+1}) > \hat{J}_{\mathrm{TV}}(\alpha^k) + c\hat{J}'_{\mathrm{TV}}(\alpha^k)^\top(\alpha^{k+1} - \alpha^k)$ **do** (Armijo line search)
Set $\tau^k := \theta_-\tau^k$ and re-compute $\alpha^{k+1} = P_{(\mathcal{A}_{ad})_h}\left(\alpha^k - \tau^k\nabla\hat{J}_{\mathrm{TV}}(\alpha^k)\right)$
end while
Update $\tau^{k+1} = \theta_+\tau^k$ and $k := k + 1$
until some stopping condition is satisfied

5.2 Bilevel TV numerical experiments

For the test images depicted in Figure 7 with resolution $n = m = 256$, we now depict some examples where we compare some weighted TV reconstructions, produced by Algorithm 2, with scalar TV ones.

Cameraman Parrot Turtle Hatchling

Figure 7: Test images, resolution 256×256.

We first report the values of the different parameters that we used for these numerical examples. For the lower level problems parameters we used $\beta = 10^{-4}$, $\gamma = 10^{-4}$, $\delta = 10^{-3}$, $\epsilon = 10^{-8}$. Initially the lower level problem is solved for $\epsilon^0 = 10^{-3}$ and subsequently decreased by a factor of $\theta_\epsilon = 0.6$ down to a final values $\underline{\epsilon} = 10^{-8}$. For every ϵ^l each problem is solved up to an accuracy $\mathrm{tol}_{(l)} = 10^{-10}$. The equation in Algorithm 1 corresponds to a linear system in finite dimensions. Here we simply used MATLAB's backslash command, even though preconditioned conjugate gradient methods can be used as well; see [75, 101]. For the H^1-projection we used $\epsilon_\alpha = 10^{-8}$. For the constraint set $(\mathcal{A}_{ad})_h$ we used $\underline{\alpha} = 10^{-8}$ and $\overline{\alpha} = 10^{-2}$. For the upper level objective of the bilevel problem we chose $\lambda = 10^{-9}$ and w to be a $n_w \times n_w$ filter normalized (i.e., with entries $1/n_w^2$), with $n_w = 7$. The local variance barriers $\underline{\sigma}^2$ and $\overline{\sigma}^2$ were set according to (5.6). Since we are depicting examples with $\sigma^2 = 10^2$ and $\sigma^2 = 4 \times 10^{-4}$ the corresponding values for $(\underline{\sigma}, \overline{\sigma})$ are $(0.00798, 0.01202)$ and $(3.2 \times 10^{-4}, 4.8 \times 10^{-4})$. For the Armijo line search parameters we set $\tau^0 = 10^{-3}$, $c = 10^{-8}$, $\theta_- = 0.25$ and $\theta_+ = 2$. The vector p^0 is chosen to be the zero vector.

Note that since we are confronted essentially with a non-convex problem, i.e., the reduced objective is non-convex, initialization of the algorithm with respect to α is crucial. Here we follow [88] and set the initial regularization weight sufficiently large $\alpha^0 = 2.5 \times 10^{-3}$ so that the associated initial image is cartoon-like.

We depict some first examples in Figure 8. There we have corrupted the images in Figure 7 with Gaussian noise of variance $\sigma^2 = 0.01$. In the second row we depict the best scalar TV examples, where the scalar parameter has been manually optimized via a bisection procedure with step 0.125×10^{-4}. We depict here the examples that correspond to the optimal parameters with respect to the Structural Similarity Index SSIM [144]; see also Table 1 for the optimal PSNR values as well. In the third row we depict the weighted TV denoising images u with the corresponding weight functions α in the fourth row. Note that the weighted TV reconstructions are better both visually and with respect to both, PSNR and SSIM. In fact we see that the weight functions get well adapted to the structure of the images, having lower values in areas of high details and larger in homogeneous areas. Thus, we obtained a structured non-smooth prior.

$\sigma^2 = 0.01$	Cameraman	Parrot	Turtle	hatchling
best scalar α for PSNR	$27.54, 0.7857$	$28.88, 0.8119$	$29.27, 0.7924,$	$\mathbf{27.57}, 0.7597$
best scalar α for SSIM	$27.19, 0.8064$	$28.51, 0.8421$	$29.11, 0.8044$	$27.46, 0.7687$
bilevel TV	$\mathbf{27.85}, \mathbf{0.8259}$	$\mathbf{28.96}, \mathbf{0.8477}$	$\mathbf{29.60}, \mathbf{0.8176}$	$27.55, \mathbf{0.7750}$
$\sigma^2 = 4 \times 10^{-4}$	Cameraman	Parrot	Turtle	hatchling
best scalar α for PSNR	$36.26, 0.9364$	$\mathbf{37.29}, 0.9448$	$37.31, 0.9394$	$\mathbf{36.31}, 0.9471$
best scalar α for SSIM	$35.80, 0.9460$	$37.03, \mathbf{0.9492}$	$37.06, 0.9443$	$36.02, \mathbf{0.9521}$
bilevel TV	$\mathbf{36.30}, \mathbf{0.9486}$	$37.01, 0.9476$	$\mathbf{37.50}, \mathbf{0.9481}$	$35.82, 0.9479$

Table 1: PSNR and SSIM comparisons for the images of Figures 8 and 9. Every cell contains the corresponding PSNR and SSIM value.

We depict a second series of examples in Figure 9, where we have used the same example images but with considerable lower Gaussian noise of variance $\sigma^2 = 4 \times 10^{-4}$. There we see that the low level of noise can make the weighting function α to adapt even more to the data. In the "cameraman" and "turtle" images, again the weighted TV result outperforms the best scalar TV one both in SSIM and PSNR; see again Table 1. In the "Parrot" the best scalar results are slightly better in terms of SSIM and PSNR, but perhaps not visually. The same holds for the "hatchling" image. We believe that this is

Figure 8: First row: images corrupted with Gaussian noise of variance $\sigma^2 = 0.01$. Second row: Best scalar TV reconstructions in terms of SSIM. Third row: weighted TV reconstructions. Fourth row: the spatially adapted weight function α

due to the fact that in both of these images, the clean image itself already contains some noise – this is indeed the case in the background of the parrot image – which is still significant given the low level of artificial noise. In the “hatchling” image in fact, oscillatory features of different scale dominate. The bilevel TV also interferes with this type of noise which is naturally present in the images resulting in a lower PSNR and SSIM. Nevertheless the bilevel TV images are still visually more amenable than their best manually optimized scalar TV reconstructions.

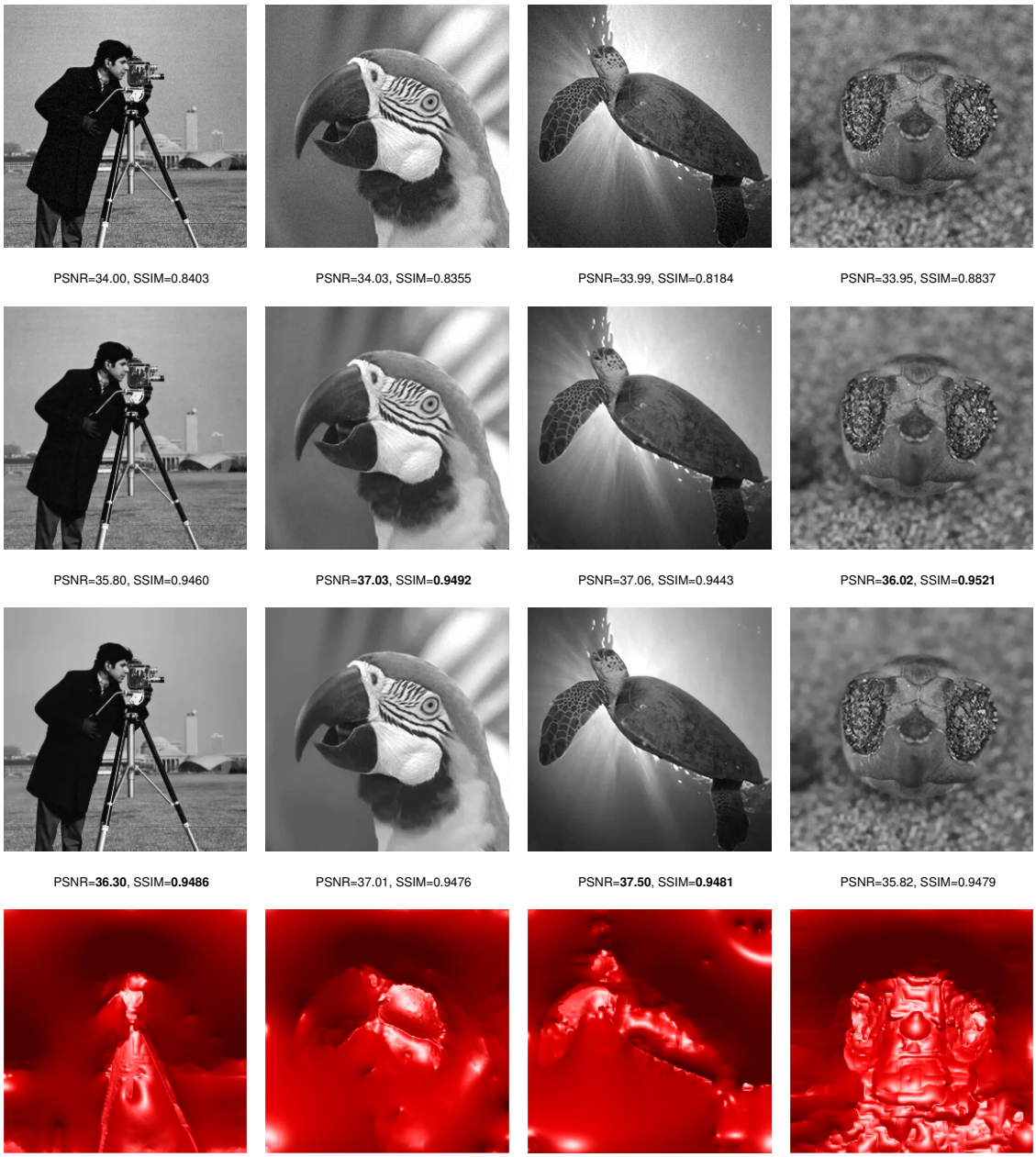

Figure 9: First row: images corrupted with Gaussian noise of variance $\sigma^2 = 4 \times 10^{-4}$. Second row: Best scalar TV reconstructions in terms of SSIM. Third row: weighted TV reconstructions. Fourth row: the spatially adapted weight function α

5.3 Discrete operators for $(\mathbb{P}_{\mathrm{TGV}})$

We now turn our attention to the discretized version of the bilevel TGV problelm $(\mathbb{P}_{\mathrm{TGV}})$. For that we need to defined the discrete function space

$$V_h = U_h \times U_h \times U_h,$$

associated with the predual p, i.e., $p = (p^{11}, p^{12}, p^{22})$ with $p^{11}, p^{12}, p^{22} \in U_h$. In this problem, for the discrete gradient and divergence we have, $\nabla : W_h \to V_h$ and $\mathrm{div} : V_h \to W_h$ satisfying again the

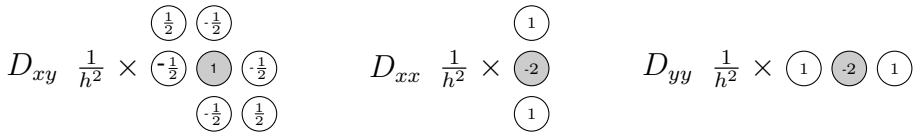

Figure 10: Symmetric finite difference stencil for the second-order derivative operators

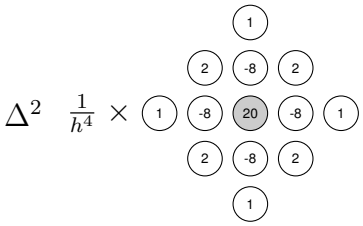

Figure 11: Finite difference stencils that constitute the discrete bi-Laplacian Δ^2

adjoint relation $\nabla = -\mathrm{div}^\top$. For $p \in V_h$, the divergence is defined again as

$$(\mathrm{div} p)^1_{i,j} = \frac{1}{h}(p^{11}_{i,j} - p^{11}_{i-1,j} + p^{12}_{i,j} - p^{12}_{i,j-1}), \quad (i,j) \in \Omega_h,$$
$$(\mathrm{div} p)^2_{i,j} = \frac{1}{h}(p^{12}_{i,j} - p^{12}_{i-1,j} + p^{22}_{i,j} - p^{22}_{i,j-1}), \quad (i,j) \in \Omega_h,$$

setting again zero values at the ghost nodes. For the second-order gradient $\nabla^2 : U_h \to V_h$ we have that $\nabla^2 = (D_{xx}u, D_{xy}u, D_{yy}u)$, where D_{xx}, D_{xy}, D_{xy} are operators $U_h \to V_h$ and are defined using the stencils as shown in Figure 10 with zero values at ghost points. Note that the use of symmetric differences for the mixed derivative results to a symmetric matrix for the matrix representing Dxy. The resulting operators D_{xx}, D_{xy}, D_{yy} are self-adjoint. Hence for the discretized second divergence $\mathrm{div}^2 : V_h \to U_h$, we have $\mathrm{div}^2 p = D_{xx}p^{11} + 2D_{xy}p^{12} + D_{yy}p^{22}$. The vector bi-Laplacian is an operator $V_h \to V_h$ where $p \mapsto (\Delta^2 p^{11}, \Delta^2 p^{12}, \Delta^2 p^{22})$ with $\Delta^2 = D_{xxxx} + D_{yyyy} + D_{xxyy} + D_{yyxx}$. The resulting stencil for the Δ^2 is as shown in Figure 11. In order to reflect the boundary conditions of $H^2_0(\Omega, \mathcal{S}^{2\times 2})$, the bi-Laplacian must be endowed with both, homogeneous Neumann and homogeneous Dirichlet boundary conditions. Again this is enforced by considering at any ghost points (up to two of them in the boundary) zero value. Finally we discuss the discretization of the operator $\nabla^2\mathrm{div}^2 : V_h \to V_h$, which is equal to

$$(\nabla^2\mathrm{div}^2 p)^{11} = Dxxxxp^{11} + 2Dxxxyp^{12} + Dxxyyp^{22}$$
$$(\nabla^2\mathrm{div}^2 p)^{12} = Dxyxxp^{11} + 2Dxyxyp^{12} + Dxyyyp^{22}$$
$$(\nabla^2\mathrm{div}^2 p)^{22} = Dyyxxp^{11} + 2Dyyxyp^{12} + Dyyyyp^{22}$$

where in fact it holds $Dxxxy = Dxyxx$, $Dxxyy = Dxyxy = Dyyxx$ and $Dxyyy = Dyyxy$. For these fourth-order discretized differential operators we use the stencils shown in Figure 12-13: and also using again the same rule to enforce homogeneous Neumann and homogeneous Dirichlet boundary conditions.

We remark that the matrix representing the operator $\nabla^2\mathrm{div}^2$ described above will not be symmetric due to the factor of 2 multiplying the terms concerning p^{12}. This leads to a non-symmetric linear system when employing a Newton-type iteration. However, having a symmetric matrix is beneficial as this can lead to a more efficient and robust solution of the corresponding linear system via iterative solvers, such as, e.g., conjugate gradients. A possible remedy to solve for $(p^{11}, 2p^{12}, p^{22})$ rather than $(p^{11}, 2p^{12}, p^{22})$. This eliminates the factor 2 in the p^{12} part of the matrix that represents $\nabla^2\mathrm{div}^2$. In

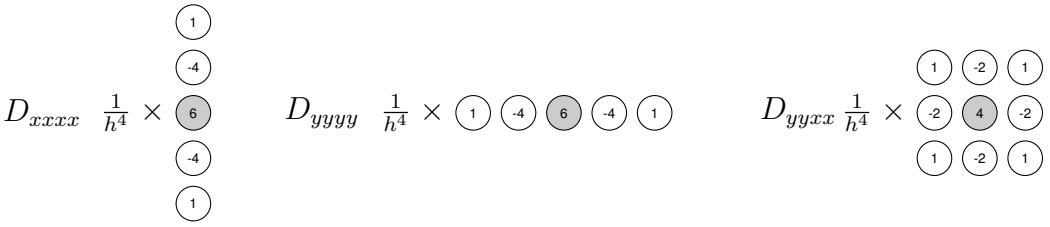

Figure 12: Finite difference stencils that constitute the discrete bi-Laplacian Δ^2

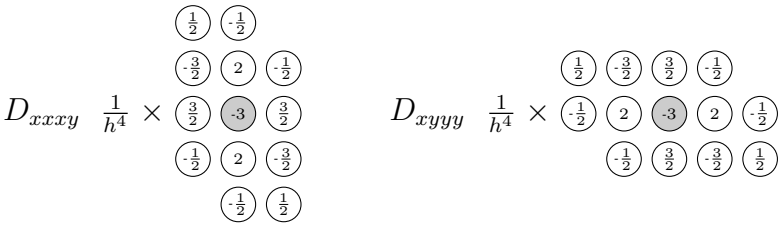

Figure 13: Symmetric finite difference stencil for the fourth order derivative operators

this case, however, other operators must be modified as well. For instance, the vector bi-Laplacian must take the form $(\Delta^2, \frac{1}{2}\Delta^2, \Delta^2)$, similarly for the other differential operators. The functions Q_δ and P_δ must be accordingly modified as well. The following version of the discrete dual $H_0^2(\Omega_h)^*$-norm is used for termination purposes:

$$\|v\|^*_{H_0^2(\Omega_h)} = h\sqrt{v^\top (I+\Delta^2)^{-1} v}. \tag{5.9}$$

We also use the discrete $H^1(\Omega_h)^*$-norm as in (5.3) along with the discrete Riezs map $\alpha_1 \mapsto (I - \Delta_B)\alpha_1$ as in the case of bilevel TV. We are now ready to write down the discrete version of ($\mathbb{P}_{\mathrm{TGV}}$):

$$\begin{cases} \min \dfrac{1}{2}\|(R(\mathrm{div}^2 p) - \overline{\sigma}^2)^+\|^2_{\ell^2(\Omega_h)} + \dfrac{1}{2}\|(\underline{\sigma}^2 - R(\mathrm{div}^2 p))^+\|^2_{\ell^2(\Omega_h)} + \dfrac{\lambda}{2}\|\alpha_1\|^2_{H^1(\Omega_h)} \\ \quad \text{over } (p, \alpha_0, \alpha_1) \in V_h \times (\mathcal{A}^0_{ad})_h \times (\mathcal{A}^1_{ad})_h \\ \text{s.t.} \quad \beta\Delta^2 p + \gamma p + \nabla^2 B^{-1}\mathrm{div}^2 p - \nabla^2 B^{-1} T^* f + \dfrac{1}{\epsilon_0} Q_\delta(p, \alpha_0) - \dfrac{1}{\epsilon_1}\nabla P_\delta(\mathrm{div} p, \alpha_1) = 0. \end{cases} \tag{$\mathbb{P}^h_{\mathrm{TGV}}$}$$

Here we have

$$\begin{aligned} (\mathcal{A}^0_{ad})_h &= \{\alpha_0 \in \mathbb{R} : \underline{\alpha}_0 \leq \alpha_0 \leq \overline{\alpha}_0\}, \\ (\mathcal{A}^1_{ad})_h &= \{\alpha \in U_h : \ \underline{\alpha}_1 \leq (\alpha_1)_{i,j} \leq \overline{\alpha}_1, \ \text{for all } (i,j) \in \Omega_h\}. \end{aligned}$$

The discrete penalty functions $P_\delta : W_h \to W_h$ and $Q_\delta : V_h \to V_h$ are defined in the obvious way. We use the same rule as in (5.6) for the choice of $\underline{\sigma}^2$, $\overline{\sigma}^2$.

The corresponding projected gradient algorithm for the numerical solution of ($\mathbb{P}^h_{\mathrm{TGV}}$) is stated in Algorithm 3.

For the sake of notation, here $\mathbf{1}$ notes a matrix either of form $[Id; Id]$ or $[Id; Id; Id]$ of size $nm \times 2nm$ or $nm \times 3nm$, respectively, depending on whether it is applied to α_1 or α_0. On the other hand, $\mathbb{1}$ denotes a matrix of size $1 \times 3nm$ with all entries equal to one. The projection $P_{(\mathcal{A}^1_{ad})_h}$ is computed as in the bilevel TV algorithm while $P_{(\mathcal{A}^1_{ad})_h}(\alpha_0) = \max(\min(\alpha_0, \overline{\alpha}_0), \underline{\alpha}_0)$. As in the TV case, also here a path following scheme is used to solve $g_{\mathrm{TGV},\epsilon_0,\epsilon_1}(p, \alpha_0, \alpha_1) = 0$. This is done successively for $\epsilon_0 = \epsilon_0^\ell$, $\epsilon_1 = \epsilon_1^\ell$ down to a tolerance

$$g_{\mathrm{TGV},\epsilon_0^\ell,\epsilon_1^\ell}(p^{\ell+1}, \alpha_0, \alpha_1) \leq \mathrm{tol}_{(l)}$$

Algorithm 3

Discretized projected gradient method of the solution of the bilevel TGV image reconstruction problem ($\mathbb{P}^h_{\mathrm{TGV}}$)

Input: f, $\underline{\alpha}_0$, $\overline{\alpha}_0$, $\underline{\alpha}_1$, $\overline{\alpha}_1$, $\overline{\sigma}$, $\underline{\sigma}$, λ, β, γ, ϵ_0, ϵ_1, δ, n_w τ_0^0, τ_1^0, $0 < c < 1$, $0 < \theta_- < 1 \leq \theta_+$

Initialise: $\alpha_0^0 \in (\mathcal{A}_{ad}^0)_h$, $\alpha_1^0 \in (\mathcal{A}_{ad}^1)_h$ and set $k = 0$.

repeat

Use a Newton-type method to compute the solution p^k of the lower level problem

$$g_{\mathrm{TGV}}(p^k, \alpha_0^k, \alpha_1^k) := \beta\Delta^2 p^k + \gamma p^k + \nabla^2 B^{-1}\mathrm{div}^2 p^k - \nabla^2 B^{-1} T^* f + \frac{1}{\epsilon_0} Q_\delta(p^k, \alpha_0^k) - \frac{1}{\epsilon_1}\nabla P_\delta(\mathrm{div} p^k, \alpha_1^k) = 0$$

Solve the adjoint equation for q^k

$$\begin{aligned}\beta\Delta^2 q^k + \gamma q^k + \nabla^2 B^{-1}\mathrm{div}^2 q^k + \frac{1}{\epsilon_0}\left(G''_\delta(p^k - \alpha_0^k\mathbf{1}) + G''_\delta(-p^k - \alpha_0^k\mathbf{1})\right) q^k \\ - \frac{1}{\epsilon_1}\nabla\left(G''_\delta(\mathrm{div} p^k - \alpha_1^k\mathbf{1}) + G''_\delta(-\mathrm{div} p^k - \alpha_1^k\mathbf{1})\right)\mathrm{div} q^k \\ = -2\nabla B^{-1} T^* \mathrm{div}^2 p^k \left(w * \left((R(\mathrm{div}^2 p^k) - \overline{\sigma}^2)^+ - (\underline{\sigma}^2 - R(\mathrm{div}^2 p^k))^+\right)\right)\end{aligned}$$

Compute the reduced derivatives with respect to α_0 and α_1

$$\begin{aligned}\hat{J}'_{\mathrm{TGV},\alpha_0}(\alpha_0^k, \alpha_1^k) &= \frac{1}{\epsilon_0}\mathbb{1}\left(-G''_\delta(p^k - \alpha_0^k\mathbf{1}) + G''_\delta(-p^k - \alpha_0^k\mathbf{1})\right) q^k \\ \hat{J}'_{\mathrm{TGV},\alpha_1}(\alpha_0^k, \alpha_1^k) &= -\frac{1}{\epsilon_1}\nabla\left(-G''_\delta(\mathrm{div} p^k - \alpha_1^k\mathbf{1}) + G''_\delta(-\mathrm{div} p^k - \alpha_1^k\mathbf{1})\right) q^k + \lambda(I - \Delta_N)\alpha_1^k\end{aligned}$$

Compute the reduced gradients

$$\nabla_{\alpha_0}\hat{J}_{\mathrm{TGV}}(\alpha_0^k, \alpha_1^k) = \hat{J}'_{\mathrm{TGV},\alpha_0}(\alpha_0^k, \alpha_1^k), \quad \nabla_{\alpha_1}\hat{J}_{\mathrm{TGV}}(\alpha_0^k, \alpha_1^k) = (I - \Delta_N)^{-1}\hat{J}'_{\mathrm{TV}}(\alpha_0^k, \alpha_1^k)$$

Compute the trial points

$$\alpha_0^{k+1} = P_{(\mathcal{A}_{ad}^0)_h}\left(\alpha_0^k - \tau_0^k\nabla_{\alpha_0}\hat{J}_{\mathrm{TGV}}(\alpha_0^k, \alpha_1^k)\right), \quad \alpha_1^{k+1} = P_{(\mathcal{A}_{ad}^1)_h}\left(\alpha_1^k - \tau_1^k\nabla_{\alpha_1}\hat{J}_{\mathrm{TGV}}(\alpha_0^k, \alpha_1^k)\right)$$

while

$$\begin{aligned}\hat{J}_{\mathrm{TGV}}(\alpha_0^{k+1}, \alpha_1^{k+1}) > \hat{J}_{\mathrm{TGV}}(\alpha_0^k, \alpha_1^k) \\ + c\left(\hat{J}'_{\mathrm{TGV},\alpha_0}(\alpha_0^k, \alpha_1^k)^\top(\alpha_0^{k+1} - \alpha_0^k) + \hat{J}'_{\mathrm{TGV},\alpha_1}(\alpha_0^k, \alpha_1^k)^\top(\alpha_1^{k+1} - \alpha_1^k)\right)\end{aligned}$$

do(Armijo line search)

Set $\tau_0^k := \theta_-\tau_0^k$, $\tau_1^k := \theta_-\tau_1^k$ and re-compute

$$\alpha_0^{k+1} = P_{(\mathcal{A}_{ad}^0)_h}\left(\alpha_0^k - \tau_0^k\nabla_{\alpha_0}\hat{J}_{\mathrm{TGV}}(\alpha_0^k, \alpha_1^k)\right), \quad \alpha_1^{k+1} = P_{(\mathcal{A}_{ad}^1)_h}\left(\alpha_1^k - \tau_1^k\nabla_{\alpha_1}\hat{J}_{\mathrm{TGV}}(\alpha_0^k, \alpha_1^k)\right)$$

end while

Update $\tau_0^{k+1} = \theta_+\tau_0^k$, $\tau_1^{k+1} = \theta_+\tau_1^k$ and $k := k + 1$

until some stopping condition is satisfied

and then setting $\epsilon_0^{\ell+1} := \max(\theta_\epsilon \epsilon_0^\ell, \epsilon_0)$, $\epsilon_1^{\ell+1} := \max(\theta_\epsilon \epsilon_1^\ell, \epsilon_1)$ for some $0 < \theta_\epsilon < 1$.

5.4 Bilevel TGV numerical experiments

Here we will show a few numerical examples produced by Algorithm 3. The same test images as for the bilevel TV case were used; see Figure 7. For the lower level TGV problem we used $\beta = 10^{-3}$, $\gamma = 0$, $\delta = 10^{-6}$, $\epsilon_0 = 10^{-12}$, $\epsilon_1 = 10^{-12}$. Initially the lower problem is solved for $\epsilon_0^0 = 10^3$, $\epsilon_1^0 = 10^3$ and each of these successively decreased by the same factor $\theta_\epsilon = 0.05$ down to finite values $\epsilon_0 = \epsilon_1 = 10^{-12}$.

We again used backslash for the solution of the linear systems, but sophisticated iterative solvers may be employed as well. We set $\underline{\alpha}_0 = 10^{-7}$, $\overline{\alpha}_0 = 10^{-2}$, while for the H^1-projection we used again $\epsilon_\alpha = 10^{-10}$, and $\underline{\alpha}_1 = 10^{-7}$, $\overline{\alpha}_1 = 10^{-2}$. The normalized filter w and the local variance barriers $\underline{\sigma}^2$ and $\overline{\sigma}^2$ were chosen as before. For the Armijo line search we set $\tau_0^0 = 1$, $\tau_1^0 = 10^{-12}$, while c, θ_-, θ_+ were chose as in the TV case.

We note that extra attention must be paid to the initialization of the algorithm. As in the TV case α_0^0 and α_0^1 must be large enough in order to produce cartoon-like images, providing the local variance estimator with useful information. However, if α_0 is initially too large then there is a danger of following into the regime of Theorem 2.4, in which the TGV functional and hence the solution map of (at least the non-regularized) lower level problem does not depend on α_0. This means that there is a danger that the derivative of the reduced functional with respect to α_0 will be close to zero, thus making only negligible progress towards optimality. This behavior was confirmed by intensive numerical experimentation. Note that an analogous phenomenon can hold also in the case where α_0 is much smaller than α_1. Then the effect of α_1 becomes negligible. In fact this has been shown in [122, Proposition 2] for dimension one, but numerical experiments indicate that this can indeed be also a viable scenario in higher dimensions. In our examples we used $\alpha_0^0 =$ and $\alpha_1^0 = 9 \times 10^{-4}$ and 3.125×10^{-6}.

We show the results in Figure 14. Once again we notice in general a clear improvement in image quality in comparison to the best scalar TGV results. The only exception is again the “hatchling” image where the scalar TGV result is slightly better than the weighted TGV one, even though the latter one is able to preserve better the detailed features in the eyes region. Finally in Figure 15 we depict for the sake of comparison the regularization functions α and α_1 for the weighted TV and TGV results respectively, that correspond to the Figures 8 and 14.

References

[1] R. Acar and C.R. Vogel, *Analysis of bounded variation penalty methods for ill-posed problems*, Inverse Problems **10** (1994), no. 6, 1217–1229, http://dx.doi.org/10.1088/0266-5611/10/6/003.

[2] W. Allard, *Total variation regularization for image denoising, I. Geometric theory*, SIAM Journal on Mathematical Analysis **39** (2008), no. 4, 1150–1190, http://dx.doi.org/10.1137/060662617.

[3] ———, *Total variation regularization for image denoising, II. Examples*, SIAM Journal on Imaging Sciences **1** (2008), no. 4, 400–417, http://dx.doi.org/10.1137/070698749.

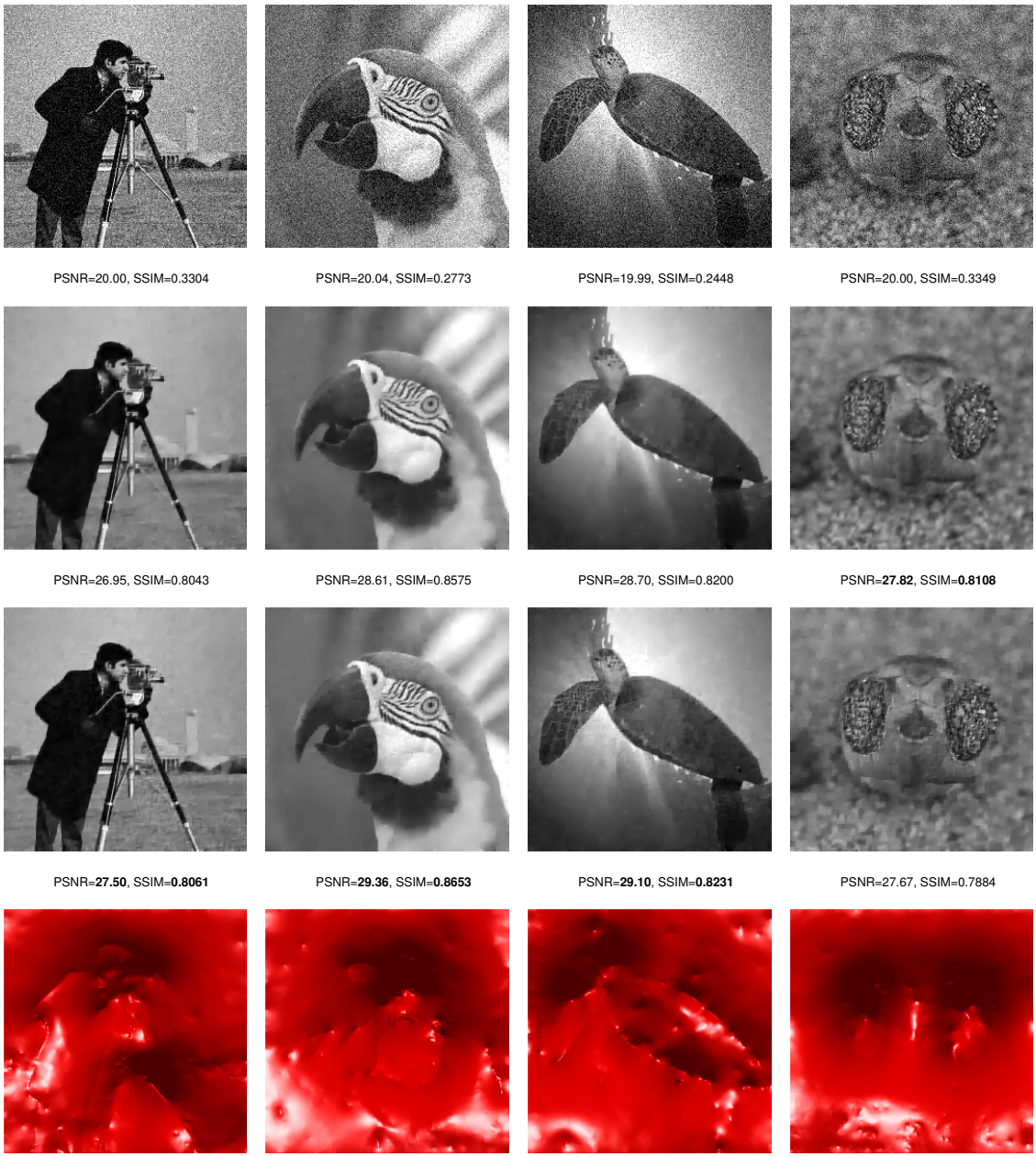

Figure 14: First row: images corrupted with Gaussian noise of variance $\sigma^2 = 0.01$. Second row: Best scalar TGV reconstructions in terms of SSIM. Third row: weighted TGV reconstructions. Fourth row: the spatially adapted weight function α_1 for weighted TGV

[4] ———, *Total variation regularization for image denoising, III. Examples.*, SIAM Journal on Imaging Sciences **2** (2009), no. 2, 532–568, http://dx.doi.org/10.1137/070711128.

[5] A. Almansa, C. Ballester, V. Caselles, and G. Haro, *A TV based restoration model with local constraints*, Journal of Scientific Computing **34** (2008), no. 3, 209–236, https://doi.org/10.1007/s10915-007-9160-x.

[6] F. Alter, S. Durand, and J. Froment, *Adapted total variation for artifact free decompression of JPEG images*, Journal of Mathematical Imaging and Vision **23** (2005), 199–211, http://dx.doi.org/10.1007/s10851-005-6467-9.

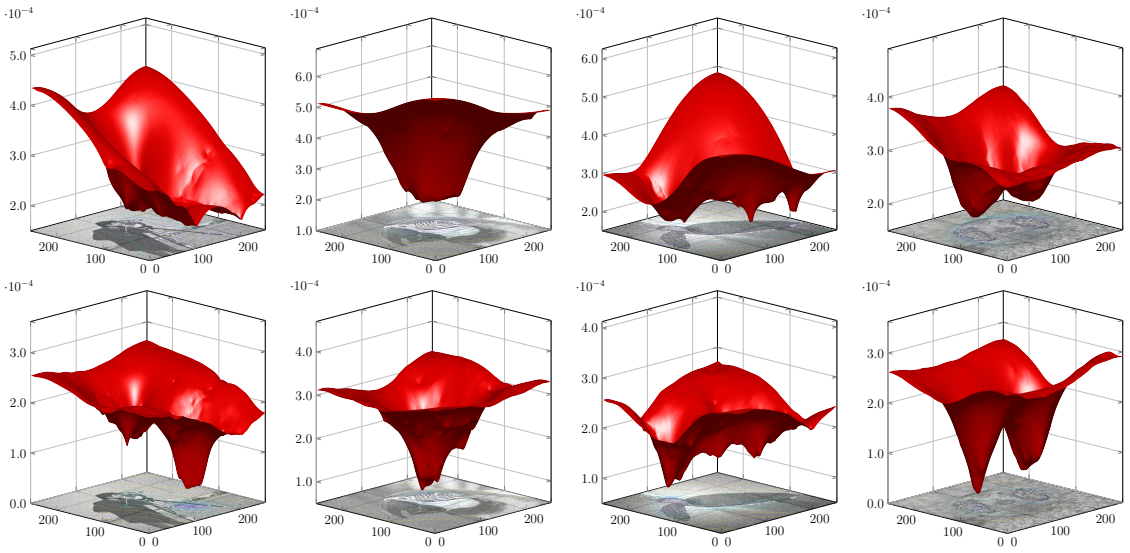

Figure 15: First row: the regularization functions α for the weighted TV results of Figure 8. Second row: the regularization functions α_1 for the weighted TGV results of Figure 14

[7] M. Amar, V. De Cicco, and N. Fusco, *Lower semicontinuity and relaxation results in BV for integral functionals with BV integrants*, ESAIM: Control, Optimization and Calculus of Variations **14** (2008), no. 3, 456–477, https://doi.org/10.1051/cocv:2007061.

[8] L. Ambrosio, N. Fusco, and D. Pallara, *Functions of bounded variation and free discontinuity problems*, Oxford University Press, USA, 2000.

[9] M. Anitescu, P. Tseng, and S.J. Wright, *Elastic-mode algorithms for mathematical programs with equilibrium constraints: global convergence and stationarity properties*, Mathematical Programming **110** (2007), no. 2, 337–371, https://doi.org/10.1007/s10107-006-0005-4.

[10] H. Attouch and H. Brezis, *Duality for the sum of convex functions in general Banach spaces*, North-Holland Mathematical Library **34** (1986), 125–133, https://doi.org/10.1016/S0924-6509(09)70252-1.

[11] H. Attouch, G. Buttazzo, and G. Michaille, *Variational analysis in Sobolev and BV spaces: Applications to PDEs and optimization*, vol. 17, SIAM, 2014.

[12] S.D. Babacan, R. Molina, and Katsaggelos A.K., *Parameter estimation in TV image restoration using variational distribution approximation*, IEEE Transactions on Image Processing **17** (2008), no. 3, 326–339, 10.1109/TIP.2007.916051.

[13] V. Barbu, *Optimal control of variational inequalities*, Research Notes in Mathematics **100** (1984).

[14] M. Benning, C. Brune, M. Burger, and J. Müller, *Higher-order TV methods – Enhancement via Bregman iteration*, J. Sci. Comput. **54** (2013), no. 2-3, 269–310, http://dx.doi.org/10.1007/s10915-012-9650-3.

[15] M. Benning and M. Burger, *Error estimates for general fidelities*, Electronic Transactions on Numerical Analysis **38** (2011), 44–68.

[16] ———, *Modern regularization methods for inverse problems*, Acta Numerica **27** (2018), 1–111, https://doi.org/10.1017/S0962492918000016.

[17] M. Benning, L. Gladden, D. Holland, C.B. Schönlieb, and T. Valkonen, *Phase reconstruction from velocity-encoded MRI measurements – A survey of sparsity-promoting variational approaches*, Journal of Magnetic Resonance **238** (2014), 26–43, http://dx.doi.org/10.1016/j.jmr.2013.10.003.

[18] M. Bertalmio, V. Caselles, B. Rougé, and A. Solé, *TV based image restoration with local constraints*, Journal of Scientific Computing **19**, no. 1, 95–122, http://dx.doi.org/10.1023/A:1025391506181.

[19] P. Blomgren and T. F. Chan, *Color TV: total variation methods for restoration of vector-valued images*, IEEE Transactions on Image Processing **7** (1998), no. 3, 304–309, http://dx.doi.org/10.1109/83.661180.

[20] K. Bredies, *Recovering piecewise smooth multichannel images by minimization of convex functionals with total generalized variation penalty*, Efficient Algorithms for Global Optimization Methods in Computer Vision, Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2014, http://dx.doi.org/10.1007/978-3-642-54774-4_3, pp. 44–77.

[21] K. Bredies, Y. Dong, and M. Hintermüller, *Spatially dependent regularization parameter selection in total generalized variation models for image restoration*, International Journal of Computer Mathematics **90** (2013), no. 1, 109–123, https://doi.org/10.1080/00207160.2012.700400.

[22] K. Bredies and M. Holler, *Artifact-free JPEG decompression with total generalized variation*, VISAP 2012: Proceedings of the International Conference on Computer Vision and Applications, 2012.

[23] ———, *A pointwise characterization of the subdifferential of the total variation functional*, SFB-Report No. 2012-11 (2012), https://imsc.uni-graz.at/mobis/publications/SFB-Report-2012-011_Update2.pdf.

[24] ———, *A total variation-based JPEG decompression model*, SIAM Journal on Imaging Sciences **5** (2012), no. 1, 366–393, http://dx.doi.org/10.1137/110833531.

[25] ———, *A TGV regularized wavelet based zooming model*, Scale Space and Variational Methods in Computer Vision, Springer, 2013, http://dx.doi.org/10.1007/978-3-642-38267-3_13, pp. 149–160.

[26] ———, *Regularization of linear inverse problems with total generalized variation*, Journal of Inverse and Ill-posed Problems **22** (2014), no. 6, 871–913, https://doi.org/10.1515/jip-2013-0068.

[27] ———, *A TGV-based framework for variational image decompression, zooming, and reconstruction. Part I: Analytics*, SIAM Journal on Imaging Sciences **8** (2015), no. 4, 2814–2850, https://doi.org/10.1137/15M1023865.

[28] K. Bredies, K. Kunisch, and T. Pock, *Total generalized variation*, SIAM Journal on Imaging Sciences **3** (2010), no. 3, 492–526, http://dx.doi.org/10.1137/090769521.

[29] K. Bredies, K. Kunisch, and T. Valkonen, *Properties of L^1-TGV^2 : The one-dimensional case*, Journal of Mathematical Analysis and Applications **398** (2013), no. 1, 438 – 454, http://dx.doi.org/10.1016/j.jmaa.2012.08.053.

[30] K. Bredies and T. Valkonen, *Inverse problems with second-order total generalized variation constraints*, Proceedings of SampTA 2011 - 9th International Conference on Sampling Theory and Applications, Singapore, 2011.

[31] E.-M. Brinkmann, M. Burger, J. Rasch, and C. Sutour, *Bias reduction in variational regularization*, Journal of Mathematical Imaging and Vision **59** (2017), no. 3, 534–566, https://doi.org/10.1007/s10851-017-0747-z.

[32] C. Brune, A. Sawatzky, F. Wübbeling, T. Kösters, and M. Burger, *EM-TV methods for inverse problems with poisson noise*, Level Set and PDE Based Reconstruction Methods in Imaging, Lecture Notes in Mathematics, Springer International Publishing, 2013, http://dx.doi.org/10.1007/978-3-319-01712-9_2, pp. 71–142.

[33] M. Burger, J. Müller, E. Papoutsellis, and C.B. Schönlieb, *Total variation regularization in measurement and image space for pet reconstruction*, Inverse Problems **30** (2014), no. 10, 105003, http://stacks.iop.org/0266-5611/30/i=10/a=105003.

[34] M. Burger, K. Papafitsoros, E. Papoutsellis, and C.B. Schönlieb, *Infimal convolution regularisation functionals of BV and L^p spaces. Part I: The finite p case*, Journal of Mathematical Imaging and Vision **55** (2016), no. 3, 343–369, http://dx.doi.org/10.1007/s10851-015-0624-6.

[35] L. Calatroni, C. Chung, J. C. De Los Reyes, C.-B. Schönlieb, and T. Valkonen, *Bilevel approaches for learning of variational imaging models*, RADON book Series on Computational and Applied Mathematics, vol. 18, Berlin, Boston: De Gruyter, 2017, https://www.degruyter.com/view/product/458544.

[36] L. Calatroni, J. De Los Reyes, and C. Schönlieb, *Infimal convolution of data discrepancies for mixed noise removal*, SIAM Journal on Imaging Sciences **10** (2017), no. 3, 1196–1233, https://doi.org/10.1137/16M1101684.

[37] L. Calatroni, J. C. De Los Reyes, and C.-B. Schönlieb, *Dynamic sampling schemes for optimal noise learning under multiple nonsmooth constraints*, System Modeling and Optimization (Christian Pötzsche, Clemens Heuberger, Barbara Kaltenbacher, and Franz Rendl, eds.), IFIP Advances in Information and Communication Technology, vol. 443, Springer Berlin Heidelberg, 2014, http://dx.doi.org/10.1007/978-3-662-45504-3_8, pp. 85–95 (English).

[38] L. Calatroni and K. Papafitsoros, *Analysis and automatic parameter selection of a variational model for mixed gaussian and salt & pepper noise removal*, Inverse Problems (2019), http://iopscience.iop.org/10.1088/1361-6420/ab291a.

[39] V. Caselles, A. Chambolle, and M. Novaga, *The discontinuity set of solutions of the TV denoising problem and some extensions*, Multiscale Modeling & Simulation **6** (2007), no. 3, 879–894, http://dx.doi.org/10.1137/070683003.

[40] A. Chambolle, *An algorithm for total variation minimization and applications*, Journal of Mathematical Imaging and Vision **20** (2004), no. 1, 89–97, https://doi.org/10.1023/B:JMIV.0000011325.36760.1e.

[41] A. Chambolle, V. Duval, G. Peyré, and C. Poon, *Geometric properties of solutions to the total variation denoising problem*, Inverse Problems **33** (2017), no. 1, 015002, http://stacks.iop.org/0266-5611/33/i=1/a=015002.

[42] A. Chambolle and P.L. Lions, *Image recovery via total variation minimization and related problems*, Numerische Mathematik **76** (1997), 167–188, http://dx.doi.org/10.1007/s002110050258.

[43] A. Chambolle and T. Pock, *A first-order primal-dual algorithm for convex problems with applications to imaging*, Journal of Mathematical Imaging and Vision **40** (2011), no. 1, 120–145, http://dx.doi.org/10.1007/s10851-010-0251-1.

[44] ———, *An introduction to continuous optimization for imaging*, Acta Numerica **25** (2016), 161–319, https://doi.org/10.1017/S096249291600009X.

[45] T. Chan, G. Golub, and P. Mulet, *A nonlinear primal-dual method for total variation-based image restoration*, SIAM Journal on Scientific Computing **20** (1999), no. 6, 1964–1977, https://doi.org/10.1137/S1064827596299767.

[46] T. Chan, A. Marquina, and P. Mulet, *High-order total variation-based image restoration*, SIAM Journal on Scientific Computing **22** (2001), no. 2, 503–516, http://dx.doi.org/10.1137/S1064827598344169.

[47] T.F. Chan and S. Esedoglu, *Aspects of total variation regularized L^1 function approximation*, SIAM Journal on Applied Mathematics (2005), 1817–1837, http://dx.doi.org/10.1137/040604297.

[48] T.F. Chan, S. Esedoglu, and F.E. Park, *Image decomposition combining staircase reduction and texture extraction*, Journal of Visual Communication and Image Representation **18** (2007), no. 6, 464–486, http://dx.doi.org/10.1016/j.jvcir.2006.12.004.

[49] T.F. Chan and J. Shen, *Variational image inpainting*, Communications on Pure and Applied Mathematics **58** (2005), no. 5, 579–619, http://dx.doi.org/10.1002/cpa.20075.

[50] X. Chen, Z. Nashed, and L. Qi, *Smoothing methods and semismooth methods for nondifferentiable operator equations*, SIAM Journal on Numerical Analysis **38** (2000), 1200–1216, https://doi.org/10.1137/S0036142999356719.

[51] C.V. Chung, J.C. De los Reyes, and C.B. Schönlieb, *Learning optimal spatially-dependent regularization parameters in total variation image denoising*, Inverse Problems **33** (2017), no. 7, 074005, http://stacks.iop.org/0266-5611/33/i=7/a=074005.

[52] P.G. Ciarlet, *Mathematical elasticity: Volume I: three-dimensional elasticity*, North-Holland, 1988.

[53] J. Darbon and M. Sigelle, *A fast and exact algorithm for total variation minimization*, Pattern Recognition and Image Analysis (Jorge S. Marques, Nicolás Pérez de la Blanca, and Pedro Pina, eds.), Springer Berlin Heidelberg, 2005, https://doi.org/10.1007/11492429_43, pp. 351–359.

[54] J. C. De Los Reyes, C.-B. Schönlieb, and T. Valkonen, *Bilevel parameter learning for higher-order Total Variation regularisation models*, Journal of Mathematical Imaging and Vision **57** (2017), no. 1, 1–25, https://doi.org/10.1007/s10851-016-0662-8.

[55] J. C. De los Reyes and C.B. Schönlieb, *Image denoising: learning the noise model via nonsmooth PDE-constrained optimization*, Inverse Problems and Imaging **7** (2013), no. 4, 1183–1214, http://dx.doi.org/10.3934/ipi.2013.7.1183.

[56] J.C. De Los Reyes, C.-B. Schönlieb, and T. Valkonen, *The structure of optimal parameters for image restoration problems*, Journal of Mathematical Analysis and Applications **434** (2016), no. 1, 464–500, https://doi.org/10.1016/j.jmaa.2015.09.023.

[57] C. Deledalle, S. Vaiter, J. Fadili, and G. Peyré, *Stein unbiased gradient estimator of the risk (SUGAR) for multiple parameter selection*, SIAM Journal on Imaging Sciences **7** (2014), no. 4, 2448–2487, https://doi.org/10.1137/140968045.

[58] F. Demengel and R. Temam, *Convex functions of a measure and applications*, Indiana University Mathematics Journal **33** (1984), 673–709.

[59] S. Dempe, *Foundations of bilevel programming*, Springer Berlin / Heidelberg, 2002.

[60] G. Dong, M. Hintermüller, and K. Papafitsoros, *Quantitative magnetic resonance imaging: From fingerprinting to integrated physics-based models*, SIAM Journal on Imaging Sciences **12** (2019), no. 2, 927–971, https://doi.org/10.1137/18M1222211.

[61] G. Dong, A.R. Patrone, O. Scherzer, and O. Öktem, *Infinite dimensional optimization models and PDEs for dejittering*, Scale Space and Variational Methods in Computer Vision (Jean-François Aujol, Mila Nikolova, and Nicolas Papadakis, eds.), Springer International Publishing, 2015, http://dx.doi.org/10.1007/978-3-319-18461-6_54, pp. 678–689.

[62] G. Dong and O. Scherzer, *Nonlinear flows for displacement correction and applications in tomography*, Scale Space and Variational Methods in Computer Vision (François Lauze, Yiqiu Dong, and Anders Bjorholm Dahl, eds.), Springer International Publishing, 2017, http://dx.doi.org/10.1007/978-3-319-58771-4_23, pp. 283–294.

[63] Y. Dong, M. Hintermüller, and M.M. Rincon-Camacho, *Automated regularization parameter selection in multi-scale total variation models for image restoration*, Journal of Mathematical Imaging and Vision **40** (2010), no. 1, 82–104, http://dx.doi.org/10.1007/s10851-010-0248-9.

[64] ———, *A multi-scale vectorial L^τ-TV framework for color image restoration*, International Journal of Computer Vision **92** (2010), no. 3, 296–307, http://dx.doi.org/10.1007/s11263-010-0359-1.

[65] V. Duval, J.F. Aujol, and Y. Gousseau, *The TV$L1$ model: a geometric point of view*, SIAM Journal on Multiscale Modeling & Simulation **8** (2009), no. 1, 154–189, http://dx.doi.org/10.1137/090757083.

[66] I. Ekeland and R. Temam, *Convex analysis and variational problems*, vol. 1, North Holland, 1976.

[67] K. Frick, P. Marnitz, and A. Munk, *Statistical multiresolution Dantzig estimation in imaging: Fundamental concepts and algorithmic framework*, Electronic Journal of Statistics **6** (2012), 231–268, http://dx.doi.org/10.1214/12-EJS671.

[68] P. Getreuer, *Rudin-Osher-Fatemi Total Variation Denoising using Split Bregman*, Image Processing On Line (2012), http://dx.doi.org/10.5201/ipol.2012.g-tvd.

[69] ———, *Total Variation Deconvolution using Split Bregman*, Image Processing On Line (2012), http://dx.doi.org/10.5201/ipol.2012.g-tvdc.

[70] ———, *Total Variation Inpainting using Split Bregman*, Image Processing On Line (2012), http://dx.doi.org/10.5201/ipol.2012.g-tvi.

[71] V. Girault and P.-A. Raviart, *Finite Element Method for Navier-Stokes Equation*, Springer, 1986.

[72] T. Goldstein and S. Osher, *The split Bregman method for L1-regularized problems*, SIAM Journal on Imaging Sciences **2** (2009), no. 2, 323–343, https://doi.org/10.1137/080725891.

[73] J. Hahn, C. Wu, and X.C. Tai, *Augmented lagrangian method for generalized TV-Stokes model*, Journal of Scientific Computing **50** (2012), no. 2, 235–264, https://doi.org/0.1007/s10915-011-9482-6.

[74] W. Hinterberger and O. Scherzer, *Variational methods on the space of functions of bounded Hessian for convexification and denoising*, Computing **76** (2006), no. 1-2, 109–133, http://dx.doi.org/10.1007/s00607-005-0119-1.

[75] M. Hintermüler and G. Stadler, *An infeasible primal-dual algorithm for total bounded variation–based inf-convolution-type image restoration*, SIAM Journal on Scientific Computing **28** (2006), no. 1, 1–23, http://dx.doi.org/10.1137/040613263.

[76] M. Hintermüller, *Mesh-independence and fast local convergence of a primal-dual active-set method for mixed control-state constrained elliptic control problems*, ANZIAM Journal **49** (2007), 1–38, https://doi.org/10.1017/S1446181100012657.

[77] M. Hintermüller, M. Holler, and K. Papafitsoros, *A function space framework for structural total variation regularization with applications in inverse problems*, Inverse Problems **34** (2018), no. 6, 064002, http://stacks.iop.org/0266-5611/34/i=6/a=064002.

[78] M. Hintermüller, K. Ito, and K. Kunisch, *The primal-dual active set strategy as a semismooth Newton method*, SIAM Journal on Optimization **13** (2002), no. 3, 865–888, https://doi.org/10.1137/S1052623401383558.

[79] M. Hintermüller and I. Kopacka, *Mathematical programs with complementarity constraints in function space: C- and strong stationarity and a path-following algorithm*, SIAM Journal on Optimization **20** (2009), no. 2, 868–902, https://doi.org/10.1137/080720681.

[80] M. Hintermüller and K. Kunisch, *Path-following methods for a class of constrained minimization problems in function space*, SIAM Journal on Optimization **17** (2006), no. 1, 159–187, https://doi.org/10.1137/040611598.

[81] M. Hintermüller, C. Löbhard, and H. Tber, *An ℓ_1-penalty scheme for the optimal control of elliptic variational inequalities*, Numerical Analysis and Optimization (M. Al-Baali, L. Grandinetti, and A. Purnama, eds.), Springer Proceedings in Mathematics & Statistics, vol. 134, Springer, 2015, https://doi.org/10.1007/978-3-319-17689-5_7, pp. 151–190.

[82] M. Hintermüller, Boris S. Mordukhovich, and T.M. Surowiec, *Several approaches for the derivation of stationarity conditions for elliptic MPECs with upper-level control constraints*, Mathematical Programming **146** (2014), no. 1, 555–582, https://doi.org/10.1007/s10107-013-0704-6.

[83] M. Hintermüller, K. Papafitsoros, and C.N. Rautenberg, *Analytical aspects of spatially adapted total variation regularisation*, Journal of Mathematical Analysis and Applications **454** (2017), no. 2, 891 – 935, https://doi.org/10.1016/j.jmaa.2017.05.025.

[84] M. Hintermüller and C.N. Rautenberg, *On the density of classes of closed convex sets with pointwise constraints in Sobolev spaces*, Journal of Mathematical Analysis and Applications **426** (2015), no. 1, 585 – 593, http://dx.doi.org/10.1016/j.jmaa.2015.01.060.

[85] M. Hintermüller and C.N. Rautenberg, *Optimal selection of the regularization function in a weighted total variation model. part I: Modelling and theory*, Journal of Mathematical Imaging and Vision **59** (2017), no. 3, 498–514, https://doi.org/10.1007/s10851-017-0744-2.

[86] M. Hintermüller, C.N. Rautenberg, and J. Hahn, *Functional-analytic and numerical issues in splitting methods for total variation-based image reconstruction*, Inverse Problems **30** (2014), no. 5, 055014, http://stacks.iop.org/0266-5611/30/i=5/a=055014.

[87] M. Hintermüller, C.N. Rautenberg, and S. Rösel, *Density of convex intersections and applications*, Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences **473** (2017), no. 2205, http://dx.doi.org/10.1098/rspa.2016.0919.

[88] M. Hintermüller, C.N. Rautenberg, T. Wu, and A. Langer, *Optimal selection of the regularization function in a weighted total variation model. part II: Algorithm, its analysis and numerical tests*, Journal of Mathematical Imaging and Vision **59** (2017), no. 3, 515–533, https://doi.org/10.1007/s10851-017-0736-2.

[89] M. Hintermüller and M.M. Rincon-Camacho, *Expected absolute value estimators for a spatially adapted regularization parameter choice rule in L^1-TV-based image restoration*, Inverse Problems **26** (2010), no. 8, 085005, http://stacks.iop.org/0266-5611/26/i=8/a=085005.

[90] M. Hintermüller and S. Rösel, *A duality-based path-following semismooth Newton method for elasto-plastic contact problems*, Journal of Computational and Applied Mathematics **292** (2016), 150–173, https://doi.org/10.1016/j.cam.2015.06.010.

[91] ———, *Duality results and regularization schemes for Prandtl-Reuss perfect plasticity*, ESAIM: COCV, Forthcoming article (2018), https://doi.org/10.1051/cocv/2018004.

[92] M. Hintermüller and T.M. Surowiec, *A bundle-free implicit programming approach for a class of elliptic MPECs in function space*, Mathematical Programming **160** (2016), no. 1–2, 271–305, https://doi.org/10.1007/s10107-016-0983-9.

[93] M. Hintermüller and M. Ulbrich, *A mesh-independence result for semismooth Newton methods*, Mathematical Programming **101** (2004), no. 1, 154–184, https://doi.org/10.1007/s10107-004-0540-9.

[94] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich, *Optimization with PDE constraints*, vol. 23, Springer Science & Business Media, 2009, https://doi.org/10.1007/978-1-4020-8839-1.

[95] M. Holler and K. Kunisch, *On infimal convolution of TV-type functionals and applications to video and image reconstruction*, SIAM Journal on Imaging Sciences **7** (2014), no. 4, 2258–2300, https://doi.org/10.1137/130948793.

[96] Y. Hotz, P. Marnitz, R. Stichtenoth, L. Davies, Z. Kabluchko, and A. Munk, *Locally adaptive image denoising by a statistical multiresolution criterion*, Computational Statistics & Data Analysis **56** (2012), no. 3, 543 – 558, http://dx.doi.org/10.1016/j.csda.2011.08.018.

[97] P.J. Huber, *Robust estimation of a location parameter*, Annals of Statistics **53** (1964), no. 1, 73–101.

[98] K. Ito and K Kunisch, *BV-type regularization methods for convoluted objects with edge, flat and grey scales*, Inverse Problems **16** (2000), no. 4, 909, http://stacks.iop.org/0266-5611/16/i=4/a=303.

[99] K. Jalalzai, *Some remarks on the staircasing phenomenon in total variation-based image denoising*, Journal of Mathematical Imaging and Vision **54** (2015), no. 2, 256–268, http://dx.doi.org/10.1007/s10851-015-0600-1.

[100] F. Knoll, K. Bredies, T. Pock, and R. Stollberger, *Second order total generalized variation (TGV) for MRI*, Magnetic Resonance in Medicine **65** (2011), no. 2, 480–491, http://dx.doi.org/10.1002/mrm.22595.

[101] K. Kunisch and M. Hintermüller, *Total bounded variation regularization as a bilaterally constrained optimization problem*, SIAM Journal on Applied Mathematics **64** (2004), no. 4, 1311–1333, https://doi.org/10.1137/S0036139903422784.

[102] K. Kunisch and T. Pock, *A bilevel optimization approach for parameter learning in variational models*, SIAM Journal on Imaging Sciences **6** (2013), no. 2, 938–983, http://dx.doi.org/10.1137/120882706.

[103] A. Langer, *Automated parameter selection for total variation minimization in image restoration*, Journal of Mathematical Imaging and Vision **57** (2017), no. 2, 239–268, http://dx.doi.org/10.1007/s10851-016-0676-2.

[104] S. Lefkimmiatis, A. Bourquard, and M. Unser, *Hessian-based norm regularization for image restoration with biomedical applications*, IEEE Transactions on Image Processing **21** (2012), 983–995, http://dx.doi.org/10.1109/TIP.2011.2168232.

[105] Y. Lin, B. Wohlberg, and H. Guo, *UPRE method for total variation parameter selection*, Signal Processing **90** (2010), no. 8, 2546–2551, https://doi.org/10.1016/j.sigpro.2010.02.025.

[106] T. Luo, J.S. Pang, and D. Ralph, *Mathematical programs with equilibrium constraints*, Cambridge University Press, Cambridge, 1996.

[107] M. Lustig, D. L. Donoho, J. M. Santos, and J. M. Pauly, *Compressed sensing MRI*, IEEE Signal Processing Magazine **25** (2008), no. 2, 72–82, https://doi.org/10.1109/MSP.2007.914728.

[108] M. Lysaker, A. Lundervold, and X.C. Tai, *Noise removal using fourth-order partial differential equation with applications to medical magnetic resonance images in space and time*, IEEE Transactions on Image Processing **12** (2003), no. 12, 1579–1590, http://dx.doi.org/10.1109/TIP.2003.819229.

[109] M. Lysaker and X.C. Tai, *Iterative image restoration combining total variation minimization and a second-order functional*, International Journal of Computer Vision **66** (2006), no. 1, 5–18, http://dx.doi.org/10.1007/s11263-005-3219-7.

[110] F. Malgouyres and F. Guichard, *Edge direction preserving image zooming: a mathematical and numerical analysis*, SIAM Journal on Numerical Analysis, no. 1, 1–37, https://doi.org/10.1137/S0036142999362286.

[111] Y. Meyer, *Oscillating patterns in image processing and nonlinear evolution equations: the fifteenth Dean Jacqueline B. Lewis memorial lectures*, vol. 22, American Mathematical Society, 2001.

[112] R. Mifflin, *Semismooth and semiconvex functions in constrained optimization*, SIAM Journal on Control and Optimization **15** (1977), 959–972, https://doi.org/10.1137/0315061.

[113] F. Natterer, *The mathematics of computerized tomography*, Society for Industrial and Applied Mathematics, 2001, https://doi.org/10.1137/1.9780898719284.

[114] P. Neittaanmaki, J. Sprekels, and D. Tiba, *Optimization of elliptic systems. Theory and applications*, Springer, 2006.

[115] M. Nikolova, *Minimizers of cost-functions involving nonsmooth data-fidelity terms. Application to the processing of outliers*, SIAM Journal on Numerical Analysis **40** (2002), no. 3, 965–994, http://dx.doi.org/10.1137/S0036142901389165.

[116] ———, *A variational approach to remove outliers and impulse noise*, Journal of Mathematical Imaging and Vision **20** (2004), no. 1, 99–120, http://dx.doi.org/10.1023/B:JMIV.0000011326.88682.e5.

[117] S. Osher, M. Burger, D. Goldfarb, J. Xu, and W. Yin, *An iterative regularization method for total variation-based image restoration*, Multiscale Modeling & Simulation **4** (2005), no. 2, 460–489, http://dx.doi.org/10.1137/040605412.

[118] J. Outrata, M. Kocvara, and J. Zowe, *Nonsmooth approach to optimization problems with equilibrium constraints*, In: Nonconvex Optimization and its Applications, vol. 28, Kluwer Academic Publishers, Dordrecht, 1998, https://doi.org/10.1007/978-1-4757-2825-5.

[119] K. Papafitsoros, *Novel higher order regularisation methods for image reconstruction*, Ph.D. thesis, University of Cambridge, 2014, https://www.repository.cam.ac.uk/handle/1810/246692.

[120] K. Papafitsoros and K. Bredies, *A study of the one dimensional total generalised variation regularisation problem*, Inverse Problems and Imaging **9** (2015), no. 2, 511–550, http://dx.doi.org/10.3934/ipi.2015.9.511.

[121] K. Papafitsoros and C.B. Schönlieb, *A combined first and second order variational approach for image reconstruction*, Journal of Mathematical Imaging and Vision **48** (2014), no. 2, 308–338, http://dx.doi.org/10.1007/s10851-013-0445-4.

[122] K. Papafitsoros and T. Valkonen, *Asymptotic behaviour of total generalised variation*, Scale Space and Variational Methods in Computer Vision: 5th International Conference, SSVM 2015, Proceedings (Jean-François Aujol, Mila Nikolova, and Nicolas Papadakis, eds.), Springer International Publishing, 2015, http://dx.doi.org/10.1007/978-3-319-18461-6_56, pp. 702–714.

[123] C. Pöschl and O. Scherzer, *Exact solutions of one-dimensional total generalized variation*, Communications in Mathematical Sciences **13** (2015), no. 1, 171–202, http://dx.doi.org/10.4310/CMS.2015.v13.n1.a9.

[124] L. Qi and J. Sun., *A nonsmooth version of Newton's method*, Mathematical Programming **58** (1993), 353–367, https://doi.org/10.1007/BF01581275.

[125] W. Ring, *Structural properties of solutions to total variation regularization problems*, ESAIM: Mathematical Modelling and Numerical Analysis **34** (2000), no. 4, 799–810, http://dx.doi.org/10.1051/m2an:2000104.

[126] L.I. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Physica D: Nonlinear Phenomena **60** (1992), no. 1-4, 259–268, http://dx.doi.org/10.1016/0167-2789(92)90242-F.

[127] A. Sawatzky, *(Nonlocal) total variation in medical imaging*, Ph.D. thesis, University of Münster, 2011.

[128] A. Sawatzky, C. Brune, Müller J., and Burger M., *Total variation processing of images with Poisson statistics*, Computer Analysis of Images and Patterns, Springer, 2009, http://dx.doi.org/10.1007/978-3-642-03767-2_65, pp. 533–540.

[129] H. Scheel and S. Scholtes, *Mathematical programs with complementarity constraints: Stationarity, optimality, and sensitivity*, Mathematics of Operations Research **25** (2000), no. 1, 1–22, https://doi.org/10.1287/moor.25.1.1.15213.

[130] O. Scherzer, *Denoising with higher order derivatives of bounded variation and an application to parameter estimation*, Computing **60** (1998), no. 1, 1–27, http://dx.doi.org/10.1007/BF02684327.

[131] S. Setzer, *Operator splittings, bregman methods and frame shrinkage in image processing*, International Journal of Computer Vision **92** (2011), no. 3, 265–280, https://doi.org/10.1007/s11263-010-0357-3.

[132] S. Setzer and G. Steidl, *Variational methods with higher order derivatives in image processing*, Approximation XII (2008), 360–386.

[133] S. Setzer, G. Steidl, and T. Teuber, *Infimal convolution regularizations with discrete ℓ_1-type functionals*, Communications in Mathematical Sciences **9** (2011), 797–872, http://dx.doi.org/10.4310/CMS.2011.v9.n3.a7.

[134] P.M. Suquet, *Existence et régularité des solutions de la placticité parfaite*, Comptes Rendus de l'Académie des Sciences, Ser. A. **286** (1978), 1201–1204.

[135] X.C. Tai, J. Hahn, and G.J. Chung, *A fast algorithm for Euler's elastica model using augmented Lagrangian method*, SIAM Journal on Imaging Sciences **4** (2011), no. 1, 313–344, http://dx.doi.org/10.1137/100803730.

[136] ———, *A fast algorithm for Euler's elastica model using augmented Lagrangian method*, SIAM Journal on Imaging Sciences **4** (2011), no. 1, 313–344, https://doi.org/10.1137/100803730.

[137] R. Temam, *Mathematical problems in plasticity*, vol. 15, Gauthier-Villars Paris, 1985.

[138] R. Temam and G. Strang, *Functions of bounded deformation*, Archive for Rational Mechanics and Analysis **75** (1980), no. 1, 7–21, https://doi.org/10.1007/BF00284617.

[139] T. Valkonen, *The jump set under geometric regularization. Part 1: Basic technique and first-order denoising*, SIAM Journal on Mathematical Analysis **47** (2015), no. 4, 2587–2629, http://dx.doi.org/10.1137/140976248.

[140] T. Valkonen, *The jump set under geometric regularisation. Part 2: Higher-order approaches*, Journal of Mathematical Analysis and Applications **453** (2017), no. 2, 1044–1085, https://doi.org/10.1016/j.jmaa.2017.04.037.

[141] T. Valkonen, K. Bredies, and F. Knoll, *Total generalized variation in diffusion tensor imaging*, SIAM Journal on Imaging Sciences **6** (2013), no. 1, 487–525, http://dx.doi.org/10.1137/120867172.

[142] L. Vese, *A study in the BV space of a denoising-deblurring variational problem*, Applied Mathematics and Optimization **44** (2001), no. 2, 131–161, https://doi.org/10.1007/s00245-001-0017-7.

[143] C. Vogel and M. Oman, *Iterative methods for total variation denoising*, SIAM Journal on Scientific Computing **17** (1996), no. 1, 227–238, https://doi.org/10.1137/0917016.

[144] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, *Image quality assessment: From error visibility to structural similarity*, IEEE Trans. Image Process. **13** (2004), no. 4, 600–612, http://dx.doi.org/10.1109/TIP.2003.819861.

[145] Y. Wotao, D. Goldfarb, and S. Osher, *The total variation regularized L^1 model for multiscale decomposition*, Multiscale Modeling $\&$ Simulation **6** (2007), no. 1, 190–211, http://dx.doi.org/10.1137/060663027.

[146] C. Wu and X.C. Tai, *Augmented Lagrangian method, dual methods, and split Bregman iteration for ROF, vectorial TV, and high order models*, SIAM Journal on Imaging Sciences **3** (2010), no. 3, 300–339, https://doi.org/10.1137/090767558.

[147] C. Wu, J. Zhang, and X.C. Tai, *Augmented Lagrangian method for total variation restoration with non-quadratic fidelity*, Inverse Problems & Imaging **5** (2011), no. 1, 237–261, https://doi.org/110.3934/ipi.2011.5.237.

[148] W. Yin, S. Osher, D. Goldfarb, and J. Darbon, *Bregman iterative algorithms for ℓ_1-minimisation with applications to compressed sensing*, SIAM Journal on Imaging Sciences **1** (2008), 142–168, http://dx.doi.org/10.1137/070703983.

[149] J. Zowe and S. Kurcyusz, *Regularity and stability for the mathematical programming problem in Banach spaces*, Applied Mathematics and Optimization **5** (1979), no. 1, 49–62, https://doi.org/10.1007/BF01442543.