

Dualization and Automatic Distributed Parameter Selection of Total Generalized Variation via Bilevel Optimization

MICHAEL HINTERMÜLLER^{1,2}, KOSTAS PAPAFITSOROS², CARLOS N. RAUTENBERG³,
HONGPENG SUN⁴

ABSTRACT. Total Generalized Variation (TGV) regularization in image reconstruction relies on an infimal convolution type combination of generalized first- and second-order derivatives. This helps to avoid the staircasing effect of Total Variation (TV) regularization, while still preserving sharp contrasts in images. The associated regularization effect crucially hinges on two parameters whose proper adjustment represents a challenging task. In this work, a bilevel optimization framework with a suitable statistics-based upper level objective is proposed in order to automatically select these parameters. The framework allows for spatially varying parameters, thus enabling better recovery in high-detail image areas. A rigorous dualization framework is established, and for the numerical solution, a Newton type method for the solution of the lower level problem, i.e. the image reconstruction problem, and a bilevel TGV algorithm are introduced. Denoising tests confirm that automatically selected distributed regularization parameters lead in general to improved reconstructions when compared to results for scalar parameters.

1. INTRODUCTION

In this work we analyze and implement a bilevel optimization framework for automatically selecting spatially varying regularization parameters $\alpha := (\alpha_0, \alpha_1) \in C(\bar{\Omega})^2$, $\alpha > 0$, in the following image reconstruction problem:

$$(1.1) \quad \text{minimize} \quad \frac{1}{2} \int_{\Omega} (Tu - f)^2 dx + \text{TGV}_{\alpha}^2(u) \quad \text{over } u \in \text{BV}(\Omega),$$

where the second-order Total Generalized Variation (TGV) regularization is given by

$$(1.2) \quad \text{TGV}_{\alpha}^2(u) = \sup \left\{ \int_{\Omega} u \operatorname{div}^2 \phi \, dx : \phi \in C_c^{\infty}(\Omega, \mathcal{S}^{d \times d}), |\phi(x)| \leq \alpha_0(x), \right. \\ \left. |\operatorname{div} \phi(x)| \leq \alpha_1(x), \text{ for all } x \in \Omega \right\}.$$

Here, $\Omega \subseteq \mathbb{R}^d$ is a bounded, open image domain with Lipschitz boundary, $\mathcal{S}^{d \times d}$ denotes the space of $d \times d$ symmetric matrices, $T : L^2(\Omega) \rightarrow L^2(\Omega)$ is a bounded linear (output) operator, and f denotes given data which satisfies

$$(1.3) \quad f = Tu_{true} + \eta.$$

In this context, η models a highly oscillatory (random) component with zero mean and known quadratic deviation (variance) σ^2 from the mean. Further, $L^2(\Omega)$ denotes the standard Lebesgue space [1], and $|\cdot|$, represents the Euclidean vector norm or its associated matrix norm. The space of infinitely differentiable functions with compact support in Ω and values in $\mathcal{S}^{d \times d}$ is denoted by $C_c^{\infty}(\Omega, \mathcal{S}^{d \times d})$. Further, we refer to Section 2 for the definition of the first- and second-order divergences div and div^2 , respectively.

¹Humboldt-Universität zu Berlin, Unter den Linden 6, 10999 Berlin, Germany

²Weierstrass Institute for Applied Analysis and Stochastics (WIAS), Mohrenstrasse 39, 10117 Berlin, Germany

³Department of Mathematical Sciences, George Mason University, Fairfax, VA 22030, USA

⁴Institute for Mathematical Sciences, Renmin University of China, 100872 Beijing, People's Republic of China

Emails: hintermueller@wias-berlin.de, papafitsoros@wias-berlin.de, crautenb@gmu.edu, hpsun@amss.ac.cn



FIGURE 1. Gaussian denoising: Typical difference between TV (piecewise constant) and TGV reconstructions (piecewise affine)

Originally, the TGV functional was introduced for scalar parameters $\alpha_0, \alpha_1 > 0$ only; see [15]. It serves as a higher order extension of the well-known Total Variation (TV) regularizer [24, 54], preserves edges (i.e., sharp contrast) [50, 58], and promotes piecewise affine reconstructions while avoiding the often adverse staircasing effect (i.e., piecewise constant structures) of TV [23, 44, 53]; see Figure 1 for an illustration. These properties of TGV have made it a successful regularizer in variational image restoration for a variety of applications [9, 10, 12, 13, 15, 17, 45, 59]. Extensions to manifold-valued data, multimodal and dynamic problems [6, 14, 41, 42, 46, 55] have been proposed as well. In all of these works, the choice of the scalar parameters α_0, α_1 is made “manually” via a direct grid search. Alternatively, selection schemes relying on a known ground truth u_{true} have been studied; see [19, 25, 26, 27]. The latter approach, however, is primarily of interest when investigating the mere capabilities of TGV regularization.

While there exist automated parameter choice rules for TV regularization, see for instance [37] and the references therein, analogous techniques and results for the TGV parameters are very scarce. One of the very few contributions is [8] where, however, a spatially varying fidelity weight rather than regularization parameter is computed. Compared to the choice of the regularization weight in TV-based models, the infimal convolution type regularization incorporated into the TGV functional significantly complicates the selection; compare the equivalent definition (2.1) below. Further difficulties arise when these parameters are spatially varying as in (1.2). In that case, by appropriately choosing $\alpha = (\alpha_0, \alpha_1)$, one wishes to smoothen homogeneous areas in the image while preserving fine scale details. The overall target is then to not only select the parameters in order to reduce noise while avoiding oversmoothing, as in the TV case, but also to ensure that the interplay of α_0 and α_1 will not produce any staircasing.

For this delicate selection task and inspired by [37, 39] for TV, in this work we propose a bilevel minimization framework for an automated selection of α in the TGV case. Formally, the setting can be characterized as follows:

$$(1.4) \quad \begin{cases} \text{minimize a statistics-based (upper level) objective over } (u, \alpha) \\ \text{subject to } u \text{ solving (1.1) for a regularization weight } \alpha = (\alpha_0, \alpha_1). \end{cases}$$

Note here that the optimization variable α enters the lower level minimization problem (1.1) as a parameter, thus giving rise to $u = u(\alpha)$. We also mention that this optimization format falls into the general framework which is discussed in our review paper [33] where the general opportunities and mathematical as well as algorithmic aspects of bilevel optimization in generating structured non-smooth regularization functionals are discussed in detail.

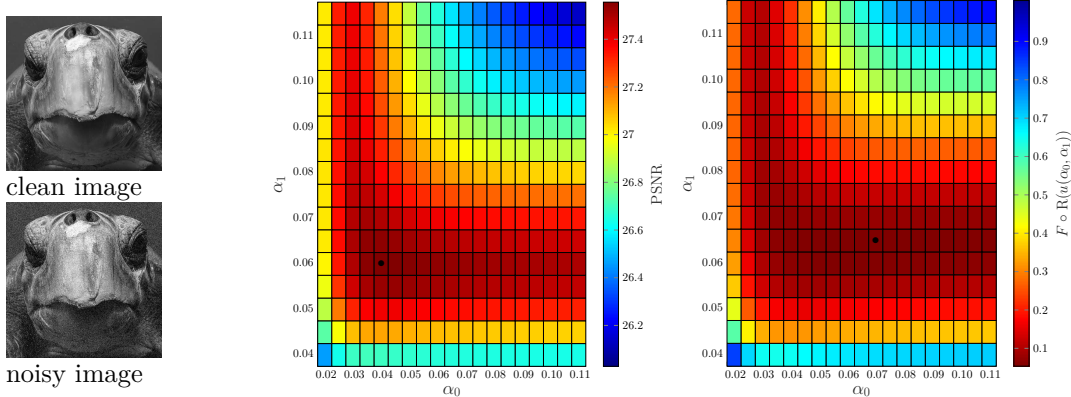


FIGURE 2. Suitability of the functional $F(R\cdot)$ as an upper level objective. Evaluation of $F(Ru)$ where u solves the TGV denoising problem (1.1) ($T = Id$), for a variety of scalar parameters (α_0, α_1)

As our statistical set-up parallels the one in [37, 39], here we resort to the upper level objective proposed in that work. It is based on localized residuals $R : L^2(\Omega) \rightarrow L^\infty(\Omega)$ with

$$(1.5) \quad Ru(x) = \int_{\Omega} w(x, y)(Tu - f)^2(y) dy,$$

where $w \in L^\infty(\Omega \times \Omega)$ with $\int_{\Omega} \int_{\Omega} w(x, y) dx dy = 1$. Note that $Ru(x)$ can be interpreted as a local variance keeping in mind that, assuming Gaussian noise of variance σ^2 , we have that $\int_{\Omega} (Tu_{true} - f)^2 dx = \int_{\Omega} \eta^2 dx = \sigma^2 |\Omega|$. Consequently, if a reconstructed image u is close to u_{true} then it is expected that for every $x \in \Omega$ the value of $Ru(x)$ will be close to σ^2 . Hence it is natural to consider an upper level objective which aims to approximately keep Ru within a corridor $\underline{\sigma}^2 \leq \sigma^2 \leq \bar{\sigma}^2$ with positive bounds $\underline{\sigma}^2, \bar{\sigma}^2$. This can be achieved by utilizing the function $F : L^2(\Omega) \rightarrow \mathbb{R}$ with

$$(1.6) \quad F(v) := \frac{1}{2} \int_{\Omega} \max(v - \bar{\sigma}^2, 0)^2 dx + \frac{1}{2} \int_{\Omega} \min(v - \underline{\sigma}^2, 0)^2 dx.$$

The function $F(R\cdot)$ is then indeed suitable as an upper level objective. This is demonstrated in Figure 2, where we show (in the middle and right plots) the objective values for a series of scalar TGV denoising results and for a variety of parameters (α_0, α_1) for the image depicted on the left. Regarding the choices of $\underline{\sigma}, \bar{\sigma}, w$ we refer to Section 5. Upon inspection of Figure 2 we find that the functional $F(R\cdot)$ is minimized for a pair of scalar parameters (α_0, α_1) that is close to the one maximizing the peak-signal-to-noise-ratio (PSNR). Note, however, that in order to truly optimize the PSNR, one would need the ground truth image u_{true} , which is of course typically not available. In contrast to this, we emphasize that $F(R\cdot)$ does *not* involve any ground truth information. Rather, it only relies on statistical properties of the noise.

For analytical and numerical reasons, rather than having (1.1) as the lower level problem for the bilevel minimization framework (1.4), one can use alternative formulations, as was done for instance in [37, 39] for TV models, where the Fenchel predual problem was used instead, see also [33] for a thorough discussion on the corresponding TGV model. This yields a bilevel problem which is expressed in terms of dual variables and is equivalent to the one stated in terms of the primal variable u . In this way, one has to treat a more amenable variational inequality of the first kind rather than one of second kind in the primal setting in the constraint system of the resulting bilevel optimization problem. Numerically, one may then utilize very efficient and resolution independent,

function space based solution algorithms, like (inexact) semismooth Newton methods [47]. The other option that we will consider in this work, is to minimize the upper level objective subject to the primal-dual optimality conditions, for which Newton methods can also be applied for their solution, see for instance [40] for an inexact semismooth Newton solver which operates on the primal-dual optimality conditions for TV regularization. We should also mention that this approach requires T^*T to be invertible, with T^* being the adjoint of T , which is true when T is injective with closed range. We note that this does not exclude the use of our bilevel scheme to inverse problems whose forward operator does not satisfy this condition. A noninvertibility of T^*T can for instance be treated by adding a small regularization term of the form $\frac{\kappa}{2} \int_{\Omega} u^2 dx$, in combination with a Levenberg-Marquardt algorithmic scheme that sends the parameter κ to zero along the iterations. Alternatively, instead of Newton, one can resort to first-order methods for solving the lower level problem for which invertibility of T^*T is not required. Such approaches are definitely interesting and necessary for many inverse problems however since they would deviate from the main focus of our paper which is the automatic computation of spatially distributed regularization weights, we will keep this invertibility assumption in the following.

The structure of the paper. Basic facts on the TGV functional with spatially varying parameters along with functional analytic foundations needed for (pre)dualization are the subjects of Section 2. Section 2.4 is concerned with the derivation of the predual problem of (1.1) and the corresponding primal-dual optimality conditions. Regularized versions of these conditions are the focus of Section 3. Besides respective primal-dual optimality conditions, we study the asymptotic behavior of these problems and their associated solutions under vanishing regularization. Section 4 introduces the bilevel TGV problem for which the primal-dual optimality conditions serve as constraints. The numerical solution of the proposed bilevel problem is the subject of Section 5. It is also argued that every regularized instance of the lower level problem can be solved efficiently by employing an (inexact) semismooth Newton method. The paper ends by a report on extensive numerical tests along with conclusions drawn from these computational results.

Summarizing, this work provides not only a user-friendly and novel hierarchical variational framework for automatic selection of the TGV regularization parameters, but by making these parameters spatially dependent it leads to an overall performance improvement; compare, e.g., the results in Section 5.

2. THE DUAL FORM OF THE WEIGHTED TGV FUNCTIONAL

2.1. Total Generalized Variation. We recall here some basic facts about the TGV functional (1.2) with constant parameters α_0, α_1 and assume throughout that the reader is familiar with the basic concepts of functions of bounded variation (BV); see [2] for a detailed account. For a function $\phi \in C_c^\infty(\Omega, \mathcal{S}^{d \times d})$ the first- and second-order divergences are respectively given by

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

In [17] it was shown that a function $u \in L^1(\Omega)$ has finite TGV value if and only if it belongs to $\operatorname{BV}(\Omega)$. Here $\operatorname{BV}(\Omega)$ denotes the Banach space of function of bounded variation over Ω with associated norm $\|\cdot\|_{\operatorname{BV}(\Omega)}$. Moreover, the bounded generalized variation norm $\|\cdot\|_{\operatorname{BGV}} := \|\cdot\|_{L^1(\Omega)} + \operatorname{TGV}_\alpha^2(\cdot)$ is equivalent to $\|\cdot\|_{\operatorname{BV}(\Omega)}$. Similarly to TV, TGV is a convex functional which is lower semicontinuous with respect to the strong L^1 convergence. In [11, 17] it is demonstrated that the

TGV functional can be equivalently written as

$$(2.1) \quad \text{TGV}_{\alpha}^2(u) = \min_{w \in \text{BD}(\Omega)} \alpha_1 |Du - w|(\Omega) + \alpha_0 |\mathcal{E}w|(\Omega),$$

where $\text{BD}(\Omega)$ is the Banach space of functions of bounded deformation, with $\mathcal{E}$ denoting the distributional symmetrized gradient [56, 57]. The asymptotical behavior of the TGV model in image restoration with respect to scalars α_0, α_1 was studied in [51]; see also [58]. For instance, when $T = Id$ and either α_0 or α_1 converges to zero, then the corresponding solutions of (1.1) converge (weakly* in $\text{BV}(\Omega)$) to f . When both of the parameters are sent to infinity, then the solutions converge weakly* to the L^2 -linear regression solution for f . We further note that the set of affine functions constitutes the kernel of the TGV functional.

For specific symmetric functions u , there exist combinations of α_0, α_1 such that $\text{TGV}_{\alpha}(u) = \alpha_1 \text{TV}(u)$. In general one can show that there exists a constant $C > 0$ such that if $\alpha_0/\alpha_1 > C$, then the TGV value does not depend on α_0 and, up to an affine correction, it is equivalent to TV. In that case the reconstructed images still suffer from a kind of (affine) staircasing effect [51].

The fine structure of TGV reconstructions has been studied analytically mainly in dimension one in [5, 16, 49, 50, 52]. Under some additional regularity assumptions (compare [58]) it can be shown that for TGV denoising the jump set of the solution is essentially contained in the jump set of the data; see [22] for the TV case.

2.2. The space $W_0^q(\text{div}^2; \Omega)$. Next we introduce several function spaces which will be useful in our subsequent development. For this purpose, let $1 \leq q \leq \infty$ and $p \in L^q(\Omega, \mathbb{R}^d)$. Recall that $\text{div}p \in L^q(\Omega)$ if there exists $w \in L^q(\Omega)$ such that

$$\int_{\Omega} \nabla \phi \cdot p \, dx = - \int_{\Omega} \phi w \, dx, \quad \text{for all } \phi \in C_c^\infty(\Omega).$$

In that case w is unique and we set $\text{div}p = w$. Based on this first-order divergence, we define the Banach space

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

endowed with the norm $\|p\|_{W^q(\text{div}; \Omega)}^q := \|p\|_{L^q(\Omega, \mathbb{R}^d)}^q + \|\text{div}p\|_{L^q(\Omega)}^q$. Similarly one obtains the Banach space $W^q(\text{div}^2; \Omega)$ as the space of all functions $p \in L^q(\Omega, \mathcal{S}^{d \times d})$ whose first- and second-order divergences, $\text{div}p$ and div^2p , respectively, belong to $L^q(\Omega)$. We note that for a $p \in L^q(\Omega, \mathcal{S}^{d \times d})$, we have that $\text{div}p \in L^q(\Omega, \mathbb{R}^d)$ if there exists an $\omega \in L^q(\Omega, \mathbb{R}^d)$ such that

$$\int_{\Omega} E\phi \cdot p \, dx = - \int_{\Omega} \phi \cdot \omega \, dx, \quad \text{for all } \phi \in C_c^\infty(\Omega, \mathbb{R}^d),$$

with $E\phi$ denoting the L^1 function representing the absolutely continuous part of $\mathcal{E}\phi$, with respect to the Lebesgue measure. Note that since ϕ is smooth, we have $\mathcal{E}\phi = E\phi$. As before ω is unique and we set $\text{div}p = \omega$. Finally $\text{div}^2p \in L^q(\Omega)$ if there exists a function $v \in L^q(\Omega)$ such that

$$\int_{\Omega} \nabla \phi \cdot \text{div}p \, dx = - \int_{\Omega} \phi v \, dx, \quad \text{for all } \phi \in C_c^\infty(\Omega).$$

This space is equipped with the norm $\|p\|_{W^q(\text{div}^2; \Omega)}^q := \|p\|_{L^q(\Omega)}^q + \|\text{div}p\|_{L^q(\Omega, \mathbb{R}^d)}^q + \|\text{div}^2p\|_{L^q(\Omega)}^q$. We refer to [12] for a more general definition of these spaces. Note that when $q = 2$ these spaces are Hilbertian and then the standard notation is $H(\text{div}; \Omega)$ and $H(\text{div}^2; \Omega)$; see [30]. The Banach spaces $W_0^q(\text{div}; \Omega)$ and $W_0^q(\text{div}^2; \Omega)$ are defined as

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

with the analogous notation $H_0(\operatorname{div}; \Omega)$ and $H_0(\operatorname{div}^2; \Omega)$ for $q = 2$. Using the definitions above, the following integration by parts formulae hold true:

$$(2.2) \quad \int_{\Omega} \nabla \phi \cdot p \, dx = - \int_{\Omega} \phi \operatorname{div} p \, dx, \quad \text{for all } p \in W_0^q(\operatorname{div}; \Omega), \phi \in C^\infty(\overline{\Omega}, \mathbb{R}),$$

$$(2.3) \quad \int_{\Omega} E \phi \cdot p \, dx = - \int_{\Omega} \phi \cdot \operatorname{div} p \, dx, \quad \text{for all } p \in W_0^q(\operatorname{div}^2; \Omega), \phi \in C^\infty(\overline{\Omega}, \mathbb{R}^d),$$

$$(2.4) \quad \int_{\Omega} \nabla \phi \cdot \operatorname{div} p \, dx = - \int_{\Omega} \phi \operatorname{div}^2 p \, dx, \quad \text{for all } p \in W_0^q(\operatorname{div}^2; \Omega), \phi \in C^\infty(\overline{\Omega}, \mathbb{R}).$$

2.3. Weighted TGV. Throughout the remainder of this work we use the weighted TGV functional (1.2) with $\alpha_0, \alpha_1 \in C(\overline{\Omega})$ and $\alpha_0(x), \alpha_1(x) > \underline{\alpha} > 0$, $\underline{\alpha} \in \mathbb{R}$, $x \in \Omega$. We denote by $|\cdot|$ the finite dimensional Euclidean norm. We note that for an $\mathbb{R}^\ell$ -valued finite Radon measure $\mu = (m_1, \dots, m_\ell)$, we denote by $|\mu|$ its total variation measure, where for every Borel $E \subseteq \Omega$

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

Note also that it can be shown

$$|\mu|(\Omega) = \sup \left\{ \sum_{i=1}^{\ell} \int_{\Omega} \phi_i \, d\mu_i : \phi \in C_c^\infty(\Omega, \mathbb{R}^\ell), |\phi(x)| \leq 1, \text{ for all } x \in \Omega \right\}.$$

We will show that for $u \in L^2(\Omega)$ the space $C_c^\infty(\Omega, \mathcal{S}^{d \times d})$ in (1.2) can be substituted by $H_0(\operatorname{div}^2; \Omega)$ (and by $W_0^d(\operatorname{div}^2; \Omega)$ for $u \in L^{d/d-1}(\Omega)$). This fact will be instrumental when deriving the predual of the TGV minimization problem. For this we need the following result:

Proposition 2.1. *The weighted $\operatorname{TGV}_\alpha^2$ functional (1.2) admits the equivalent expression*

$$(2.5) \quad \operatorname{TGV}_\alpha^2(u) = \min_{w \in \operatorname{BD}(\Omega)} \int_{\Omega} \alpha_1 \, d|Du - w| + \int_{\Omega} \alpha_0 \, d|\mathcal{E}w|.$$

Proof. The proof is analogous to the one for the scalar TGV functional; see for instance [11, Theorem 3.5] or [12, Proposition 2.8]. Here, we highlight only the significant steps. Indeed, given $u \in L^1(\Omega)$, the idea is to define

$$\begin{aligned} U &= C_0^1(\Omega, \mathbb{R}^d) \times C_0^2(\Omega; \mathcal{S}^{d \times d}), \quad V = C_0^1(\Omega, \mathbb{R}^d), \\ \Lambda : U &\rightarrow V, \quad \Lambda(u_1, u_2) = -u_1 - \operatorname{div} u_2, \\ F_1 : U &\rightarrow \overline{\mathbb{R}}, \quad F_1(u_1, u_2) = - \int_{\Omega} u \operatorname{div} u_1 + \mathcal{I}_{\{|\cdot(x)| \leq \alpha_1(x)\}}(u_1) + \mathcal{I}_{\{|\cdot(x)| \leq \alpha_0(x)\}}(u_2), \\ F_2 : V &\rightarrow \overline{\mathbb{R}}, \quad F_2(v) = \mathcal{I}_{\{0\}}(v). \end{aligned}$$

Here, $\mathcal{I}_S(\cdot)$ denotes the indicator function of a set S . Now, after realizing that

$$(2.6) \quad \operatorname{TGV}_\alpha^2(u) = \sup_{(u_1, u_2) \in U} -F_1(u_1, u_2) - F_2(\Lambda(u_1, u_2)),$$

the proof proceeds by showing that the dual problem of (2.6) is equivalent to (2.5) and then applying the Fenchel duality result [29]. We note that in order to achieve zero duality gap between the primal and the dual problem the so-called Attouch-Brezis condition needs to be satisfied see [3], that is

one needs to show that the set $\bigcup_{\lambda \geq 0} \lambda(\text{dom}(F_2) - \Lambda(\text{dom}(F_1)))$ is a closed subspace of V . Indeed one can easily see that

$$\bigcup_{\lambda \geq 0} \lambda(\text{dom}(F_2) - \Lambda(\text{dom}(F_1))) = \{ \lambda(u_1 + \text{div} u_2) : (u_1, u_2) \in U, \\ |u_1(x)| \leq \alpha_1(x), |u_2(x)| \leq \alpha_0(x), \forall x \in \Omega \} = V.$$

Note that for the above equality we crucially used the fact that α_1 is bounded away from zero.

The other subtle point is the following density result which is required in order to show that (2.6) is indeed equal to (1.2):

$$(2.7) \quad \overline{\{ \phi \in C_c^\infty(\Omega, \mathcal{S}^{d \times d}) : |\phi(x)| \leq \alpha_0(x), |\text{div} \phi(x)| \leq \alpha_1(x), \text{ for all } x \in \Omega \}}^{\|\cdot\|_{C_0^2}} \\ = \left\{ \psi \in C_0^2(\Omega, \mathcal{S}^{d \times d}) : |\psi(x)| \leq \alpha_0(x), |\text{div} \psi(x)| \leq \alpha_1(x), \text{ for all } x \in \Omega \right\}.$$

Indeed let ψ belong to the second set in (2.7), and let $\epsilon > 0$. Choose $0 < \lambda_\epsilon < 1$ such that

$$(2.8) \quad \|\psi - \lambda_\epsilon \psi\|_{C_0^2} < \epsilon/2.$$

Since α_0 and α_1 are continuous and bounded away from zero there exists $\alpha_\epsilon > 0$, smaller than the minimum of α_0, α_1 , such that

$$|\lambda_\epsilon \psi(x)| \leq \alpha_0(x) - \alpha_\epsilon, \quad |\text{div} \lambda_\epsilon \psi(x)| \leq \alpha_1(x) - \alpha_\epsilon, \quad \text{for all } x \in \Omega.$$

From standard density properties there exists a function $\phi_\epsilon \in C_c^\infty(\Omega, \mathcal{S}^{d \times d})$ such that the following conditions hold for all $x \in \Omega$:

$$(2.9) \quad \|\phi_\epsilon - \lambda_\epsilon \psi\|_{C_0^2} < \epsilon/2, \quad |\phi_\epsilon(x) - \lambda_\epsilon \psi(x)| \leq \alpha_\epsilon/2, \quad |\text{div} \phi_\epsilon(x) - \text{div} \lambda_\epsilon \psi(x)| \leq \alpha_\epsilon/2,$$

which implies

$$(2.10) \quad |\phi_\epsilon(x)| \leq \alpha_0(x) - \alpha_\epsilon/2, \quad |\text{div} \phi_\epsilon(x)| \leq \alpha_1(x) - \alpha_\epsilon/2, \quad \text{for all } x \in \Omega.$$

Then, from (2.10) it follows that ϕ_ϵ belongs to the first set in (2.7) and from (2.8) and (2.9) we get that $\|\psi - \phi_\epsilon\|_{C_0^2} < \epsilon$. $\square$

Now we are ready to establish the density result needed for dualization. For the sake of the flow of presentation we defer the proof, which parallels the one of [12, Proposition 3.3], to Appendix A. Below “a.e.” stands for “almost every” with respect to the Lebesgue measure.

Proposition 2.2. *For $u \in L^2(\Omega)$, the weighted TGV functional (1.2) can be equivalently written as*

$$(2.11) \quad \text{TGV}_\alpha^2(u) = \sup \left\{ \int_\Omega u \text{div}^2 p \, dx : p \in H_0(\text{div}^2; \Omega), |p(x)| \leq \alpha_0(x), \right. \\ \left. |\text{div} p(x)| \leq \alpha_1(x), \text{ for a.e. } x \in \Omega \right\}.$$

Remark: By slightly amending the proof of Proposition 2.2, one can also show that if $u \in L^{d/d-1}(\Omega) \supset \text{BV}(\Omega)$ then the dual variables p can be taken to belong to $W_0^d(\text{div}^2; \Omega)$ rather than $H_0(\text{div}^2; \Omega)$.

2.4. The predual weighted TGV problem. Now we study the predual problem for the weighted TGV model with continuous weights, i.e., we use the regularization functional (1.2) or equivalently (2.11). For $T \in \mathcal{L}(L^2(\Omega), L^2(\Omega))$ we assume that $\mathcal{B} := T^*T$ is invertible and define $\|v\|_{\mathcal{B}}^2 = \int_{\Omega} v \mathcal{B}^{-1} v$, which induces a norm in $L^2(\Omega)$ equivalent to the standard one; compare [47]. When T^*T is not invertible one can add an extra regularization term of the form $\frac{\kappa}{2} \int_{\Omega} u^2 dx$, for some small $\kappa > 0$, to the objective in (1.1). Then $\mathcal{B} = \kappa \text{Id} + T^*T$ will be invertible, see a similar approach in [37, 39, 47]. We mention that a different set-up, as adopted for instance in [31], assumes $T \in \mathcal{L}(L^{d/d-1}, L^2(\Omega))$, since $\text{BV}(\Omega)$ continuously embeds to $L^{d/d-1}(\Omega)$. Then one readily finds that for $d > 2$, $\mathcal{B}$ cannot be invertible, in general. Even though invertibility of T^*T is not necessary when dealing only with the TGV regularization problem (1.1), additional challenges would arise in our overall bilevel optimization context, such as, e.g., a lack of uniqueness of the solution for the lower level problem, i.e., genuine set-valuedness of the regularization-weight-to-reconstruction map cf. the constraint set in (1.4). In the next proposition we prove zero duality gap for (1.1) and its predual. We note however that the invertibility assumption on T^*T is not needed to establish zero duality gap but rather to get an explicit formulation of the predual problem. An adaptation of the following proposition without this invertibility assumption can be done for instance by simply adjusting the corresponding proof of [31, Theorem 5.4] where analogous dualization results for a class of structural TV regularizers were considered.

Proposition 2.3. *Let $f \in L^2(\Omega)$, then there exists a solution to the primal problem*

$$(2.12) \quad \text{minimize} \quad \frac{1}{2} \|Tu - f\|_{L^2(\Omega)}^2 + \text{TGV}_{\alpha}^2(u) \quad \text{over } u \in \text{BV}(\Omega),$$

as well as to its predual problem

$$(2.13) \quad \begin{aligned} &\text{minimize} \quad \frac{1}{2} \|T^*f - \text{div}^2 p\|_{\mathcal{B}}^2 - \frac{1}{2} \|f\|_{L^2}^2 \quad \text{over } p \in H_0(\text{div}^2; \Omega) \\ &\text{subject to} \quad |p(x)| \leq \alpha_0(x), \quad |\text{div} p(x)| \leq \alpha_1(x), \quad \text{for a.e. } x \in \Omega, \end{aligned}$$

and there is no duality gap, i.e., the primal and predual optimal objective values are equal. Moreover, the solutions u and p of these problems satisfy

$$(2.14) \quad \mathcal{B}u = T^*f - \text{div}^2 p.$$

Proof. We set $U = H_0(\text{div}^2; \Omega)$, $V = L^2(\Omega)$, $\Lambda : U \rightarrow V$ with $\Lambda p = \text{div}^2 p$, and also $F_1 : U \rightarrow \overline{\mathbb{R}}$ and $F_2 : V \rightarrow \overline{\mathbb{R}}$ with

$$(2.15) \quad F_1(p) = \mathcal{I}_{\{|\cdot(x)| \leq \alpha_0(x), \text{ for a.e. } x\}}(p) + \mathcal{I}_{\{|\text{div} \cdot(x)| \leq \alpha_1(x), \text{ for a.e. } x\}}(p),$$

$$(2.16) \quad F_2(\psi) = \frac{1}{2} \|T^*f - \psi\|_{\mathcal{B}}^2 - \frac{1}{2} \|f\|_{L^2(\Omega)}^2.$$

Immediately one gets that

$$(2.17) \quad \inf_{p \in U} F_1(p) + F_2(\Lambda p) = \inf_{\substack{p \in H_0(\text{div}^2; \Omega) \\ |p(x)| \leq \alpha_0(x) \\ |\text{div} p(x)| \leq \alpha_1(x)}} \frac{1}{2} \|T^*f - \text{div}^2 p\|_{\mathcal{B}}^2 - \frac{1}{2} \|f\|_{L^2(\Omega)}^2.$$

The problem in (2.17) admits a solution. Indeed, first observe that the objective is bounded from below. Then note that since $\frac{1}{2} \|T \cdot - f\|_{L^2(\Omega)}^2$ is continuous at $0 \in L^2(\Omega)$, its convex conjugate (see [29] for a general definition) which is equal to $\frac{1}{2} \|T^*f + \cdot\|_{\mathcal{B}}^2 - \frac{1}{2} \|f\|_{L^2(\Omega)}^2$ is coercive in $L^2(\Omega)$; see [7, Theorem 4.4.10]. Hence, any infimizing sequence $(p_n)_{n \in \mathbb{N}}$ is bounded in $H_0(\text{div}^2; \Omega)$, and thus

there exist an (unrelabeled) subsequence and $p \in H(\operatorname{div}^2; \Omega)$ such that $p_n \rightharpoonup p$, $\operatorname{div} p_n \rightharpoonup \operatorname{div} p$ and $\operatorname{div}^2 p_n \rightharpoonup \operatorname{div}^2 p$ weakly in L^2 . We also have that p is a feasible point since the set

$$\{(h, \operatorname{div} h, \operatorname{div}^2 h) : h \in H_0(\operatorname{div}^2; \Omega), |h(x)| \leq \alpha_0(x), |\operatorname{div} h(x)| \leq \alpha_1(x), \text{ for a.e. } x \in \Omega\},$$

is weakly closed. Then p is a minimizer of (2.17) as $\frac{1}{2}\|T^*f - \cdot\|_{\mathcal{B}}^2$ is weakly lower semicontinuous in $L^2(\Omega)$.

We now calculate the expression $F_1^*(\Lambda^*u) + F_2^*(-u)$ for $u \in V^* = L^2(\Omega)$. As before one verifies by direct computation that $F_2^*(-u) = \frac{1}{2}\|Tu - f\|_{L^2(\Omega)}^2$. Moreover,

$$\begin{aligned} F_1^*(\Lambda^*u) &= \sup_{p \in U} \{\langle \Lambda^*u, p \rangle_{U^*, U} - F_1(p)\} = \sup_{p \in U} \{\langle u, \Lambda p \rangle_{L^2(\Omega), L^2(\Omega)} - F_1(p)\} \\ &= \sup_{\substack{p \in H_0(\operatorname{div}^2; \Omega) \\ |p(x)| \leq \alpha_0(x) \\ |\operatorname{div} p(x)| \leq \alpha_1(x)}} \int_{\Omega} \operatorname{div}^2 p \, dx = \operatorname{TGV}_{\alpha}^2(u), \end{aligned}$$

where for the last equality we used Proposition 2.2 and its underlying density result. In order to prove that there is no duality gap, it suffices again to verify the Attouch-Brezis condition and show that the set $\bigcup_{\lambda \geq 0} \lambda(\operatorname{dom}(F_2) - \Lambda(\operatorname{dom}(F_1)))$ is a closed subspace of V . It is immediate to see that $\operatorname{dom}(F_2) = L^2(\Omega)$, and hence the condition holds true. Thus, we also get existence of a solution for the primal problem (2.12). Finally (2.14) follows from the optimality condition (Euler-Lagrange system) that corresponds to $\Lambda p \in \partial F_2^*(-u)$. $\square$

The primal-dual optimality conditions for the problems (2.12) and (2.13) read

$$(2.18) \quad p \in \partial F_1^*(\Lambda^*u),$$

$$(2.19) \quad \Lambda p \in \partial F_2^*(-u),$$

and we note once again that (2.14) corresponds to (2.19) with F_2 and Λ as in the proof of Proposition 2.12. Instead of making the optimality condition that corresponds to (2.18) explicit, we are interested in the analogous optimality conditions written in the variables u and w of the equivalent primal weighted TGV problem

$$(2.20) \quad \min_{\substack{u \in \operatorname{BV}(\Omega) \\ w \in \operatorname{BD}(\Omega)}} \frac{1}{2}\|Tu - f\|_{L^2(\Omega)}^2 + \int_{\Omega} \alpha_1 d|Du - w| + \int_{\Omega} \alpha_0 d|\mathcal{E}w|.$$

For this purpose note first that the predual problem (2.13) can be equivalently written as

$$(2.21) \quad \begin{cases} \text{minimize} & \frac{1}{2}\|T^*f + \operatorname{div} q\|_{\mathcal{B}}^2 - \frac{1}{2}\|f\|_{L^2(\Omega)}^2 & \text{over } (p, q) \in H_0(\operatorname{div}^2; \Omega) \times H_0(\operatorname{div}; \Omega), \\ \text{subject to} & -\operatorname{div} p = q, |p(x)| \leq \alpha_0(x), |q(x)| \leq \alpha_1(x), \text{ for a.e. } x \in \Omega. \end{cases}$$

The next proposition characterizes the solutions (w, u) and (p, q) of the problems (2.20) and (2.21) respectively. Before we state it, we note that in its proof we will make use of the following density results:

$$(2.22) \quad \overline{C_{\alpha_0}}^{L^2(\Omega)} = K_{\alpha_0}, \quad \overline{C_{\alpha_1}}^{H_0(\operatorname{div}; \Omega)} = K_{\alpha_1},$$

where

$$(2.23) \quad C_{\alpha_0} := \left\{ (\operatorname{div}^2 \phi, \operatorname{div} \phi) : \phi \in C_c^\infty(\Omega, \mathcal{S}^{d \times d}), |\phi(x)| \leq \alpha_0(x), \text{ for all } x \in \Omega \right\},$$

$$(2.24) \quad K_{\alpha_0} := \left\{ (\operatorname{div}^2 p, \operatorname{div} p) : p \in H_0(\operatorname{div}^2; \Omega), |p(x)| \leq \alpha_0(x), \text{ for a.e. } x \in \Omega \right\},$$

$$(2.25) \quad C_{\alpha_1} := \left\{ \psi : \psi \in C_c^\infty(\Omega, \mathbb{R}^d), |\psi(x)| \leq \alpha_1(x), \text{ for all } x \in \Omega \right\},$$

$$(2.26) \quad K_{\alpha_1} := \left\{ q : q \in H_0(\operatorname{div}; \Omega), |q(x)| \leq \alpha_1(x), \text{ for a.e. } x \in \Omega \right\}.$$

These results can be proven by using the duality arguments of the proof of Proposition 2.2, which originate from [12], or with the use of mollification techniques; see [35, 36, 38].

Proposition 2.4. *The pair $(p, q) \in H_0(\operatorname{div}^2; \Omega) \times H_0(\operatorname{div}; \Omega)$ is a solution to (2.21), and $(w, u) \in \operatorname{BD}(\Omega) \times \operatorname{BV}(\Omega)$ is a solution to (2.20) if and only if the following optimality conditions are satisfied:*

$$(2.27) \quad \mathcal{B}u = T^*f + \operatorname{div} q,$$

$$(2.28) \quad q = -\operatorname{div} p,$$

$$(2.29) \quad |q(x)| \leq \alpha_1(x) \text{ for a.e. } x \in \Omega$$

$$\text{and } \langle Du - w, \tilde{q} - q \rangle \leq 0 \text{ for every } \tilde{q} \in H_0(\operatorname{div}; \Omega) \text{ with } |\tilde{q}(x)| \leq \alpha_1(x) \text{ for a.e. } x \in \Omega,$$

$$(2.30) \quad |p(x)| \leq \alpha_0(x) \text{ for a.e. } x \in \Omega$$

$$\text{and } \langle \mathcal{E}w, \tilde{p} - p \rangle \leq 0 \text{ for every } \tilde{p} \in H_0(\operatorname{div}^2; \Omega) \text{ with } |\tilde{p}(x)| \leq \alpha_0(x) \text{ for a.e. } x \in \Omega.$$

Proof. Define $X = (X_1, X_2) = H_0(\operatorname{div}^2; \Omega) \times H_0(\operatorname{div}; \Omega)$, $Y = (Y_1, Y_2) = H_0(\operatorname{div}; \Omega) \times L^2(\Omega)$, $\Lambda : X \rightarrow Y$ with $\Lambda(p, q) = (q + \operatorname{div} p, \operatorname{div} q)$, and $F_1 : X \rightarrow \mathbb{R}$, $F_2 : Y \rightarrow \mathbb{R}$ with

$$(2.31) \quad F_1(p, q) = \mathcal{I}_{\{|\cdot(x)| \leq \alpha_0(x), \text{ for a.e. } x\}}(p) + \mathcal{I}_{\{|\cdot(x)| \leq \alpha_1(x), \text{ for a.e. } x\}}(q),$$

$$(2.32) \quad F_2(\phi, \psi) = \mathcal{I}_{\{0\}}(\phi) + \frac{1}{2} \|T^*f + \psi\|_{\mathcal{B}}^2 - \frac{1}{2} \|f\|_{L^2(\Omega)}^2.$$

One checks immediately that $\min_{(p,q) \in X} F_1(p, q) + F_2(\Lambda(p, q))$ corresponds to (2.21) with the dual problem reading $\min_{(w,u) \in Y^*} F_1^*(-\Lambda^*(w, u)) + F_2^*(w, u)$. Observe that since

$$-\langle \Lambda^*(w, u), (p, q) \rangle_{X^*, X} = -\langle (w, u), \Lambda(p, q) \rangle_{Y^*, Y} = -\langle w, \operatorname{div} p \rangle_{Y_1^*, Y_1} - \langle w, q \rangle_{Y_1^*, Y_1} - \langle u, \operatorname{div} q \rangle_{Y_2^*, Y_2},$$

we have

$$(2.33) \quad F_1^*(-\Lambda^*(w, u)) = \sup_{\substack{p \in H_0(\operatorname{div}^2; \Omega) \\ |p(x)| \leq \alpha_0(x)}} -\langle w, \operatorname{div} p \rangle_{Y_1^*, Y_1} + \sup_{\substack{q \in H_0(\operatorname{div}; \Omega) \\ |q(x)| \leq \alpha_1(x)}} -\langle w, q \rangle_{Y_1^*, Y_1} - \langle u, \operatorname{div} q \rangle_{Y_2^*, Y_2}.$$

Note that the suprema above are always greater or equal to the corresponding suprema over $C_c^\infty(\Omega, \mathcal{S}^{d \times d}) \subset H_0(\operatorname{div}^2; \Omega)$ and $C_c^\infty(\Omega, \mathbb{R}^d) \subset H_0(\operatorname{div}; \Omega)$. One can easily check that there is no duality gap between these primal and dual problems – the Attouch-Brezis condition is satisfied – and thus since the primal problem is finite so is the dual. Using the fact that α_0 is bounded from below, this implies in particular that w , seen as distribution through its action on $C_c^\infty(\Omega, \mathbb{R}^d)$, has a distributional symmetrized gradient $\mathcal{E}w$ with bounded Radon norm, and hence it is a Radon measure. It follows that $w \in L^1(\Omega, \mathbb{R}^d)$ yielding $w \in \operatorname{BD}(\Omega)$; see [11], which means that for $\psi \in C_c^\infty(\Omega, \mathbb{R}^d)$ we have $\langle w, \psi \rangle_{Y_1^*, Y_1} = \int_\Omega w \cdot \psi \, dx$. Using density of $C_c^\infty(\Omega, \mathbb{R}^d)$ in $H_0(\operatorname{div}; \Omega)$ this also implies $\langle w, q \rangle_{Y_1^*, Y_1} = \int_\Omega w \cdot q \, dx$ and similarly $\langle w, \operatorname{div} p \rangle_{Y_1^*, Y_1} = \int_\Omega w \cdot \operatorname{div} p \, dx$ in (2.33). Using

now also the density results (2.22) we have

$$\begin{aligned}
F_1^*(-\Lambda^*(w, u)) &= \sup_{\substack{p \in H_0(\operatorname{div}^2; \Omega) \\ |p(x)| \leq \alpha_0(x)}} - \int_{\Omega} w \cdot \operatorname{div} p \, dx + \sup_{\substack{q \in H_0(\operatorname{div}; \Omega) \\ |q(x)| \leq \alpha_1(x)}} - \int_{\Omega} w \cdot q \, dx - \int_{\Omega} u \operatorname{div} q \, dx \\
&= \sup_{\substack{\phi \in C_c^\infty(\Omega, \mathcal{S}^{d \times d}) \\ |\phi(x)| \leq \alpha_0(x)}} - \int_{\Omega} w \cdot \operatorname{div} \phi \, dx + \sup_{\substack{\psi \in C_c^\infty(\Omega, \mathbb{R}^d) \\ |\psi(x)| \leq \alpha_1(x)}} - \int_{\Omega} w \cdot \psi \, dx - \int_{\Omega} u \operatorname{div} \psi \, dx \\
&= \sup_{\substack{\phi \in C_c^\infty(\Omega, \mathcal{S}^{d \times d}) \\ |\phi(x)| \leq \alpha_0(x)}} \langle \mathcal{E}w, \phi \rangle + \sup_{\substack{\psi \in C_c^\infty(\Omega, \mathbb{R}^d) \\ |\psi(x)| \leq \alpha_1(x)}} \langle Du - w, \psi \rangle, \\
&= \int_{\Omega} \alpha_0 \, d|\mathcal{E}w| + \int_{\Omega} \alpha_1 \, d|Du - w|.
\end{aligned}$$

Here we used the fact that since the distribution $Du - w$ has a finite Radon norm, due to α_1 being bounded away from zero, it can be represented by an $\mathbb{R}^d$ -valued finite Radon measure and in particular $u \in \operatorname{BV}(\Omega)$. Furthermore, as in the proof of Proposition 2.3 we have $F_2^*(w, u) = \frac{1}{2} \|Tu - f\|_{L^2(\Omega)}^2$.

The fact that there is no duality gap is ensured by Propositions 2.1, 2.2 and 2.3. We now turn our attention to the optimality conditions

$$(2.34) \quad (p, q) \in \partial F_1^*(-\Lambda^*(w, u)),$$

$$(2.35) \quad \Lambda(p, q) \in \partial F_2^*((w, u)).$$

It can be checked again that (2.35) gives (2.27) and (2.28). We now expand on (2.34). We have that $(p, q) \in \partial F_1^*(-\Lambda^*(w, u))$ which is equivalent to $-\Lambda^*(w, u) \in \partial F_1(p, q)$, that is $F_1(p, q) = 0$ and

$$\begin{aligned}
&\langle -\Lambda^*(w, u), (\tilde{p} - p, \tilde{q} - q) \rangle_{X^*, X} \leq F_1(\tilde{p}, \tilde{q}) \\
&\iff -\langle w, \operatorname{div}(\tilde{p} - p) \rangle - \langle w, \tilde{q} - q \rangle - \langle u, \operatorname{div} \tilde{q} - \operatorname{div} q \rangle \leq F_1(\tilde{p}, \tilde{q}) \\
&\iff \begin{aligned} \langle \mathcal{E}w, \tilde{p} - p \rangle &\leq \mathcal{I}_{\{|\cdot(x)| \leq \alpha_0(x), \text{ f.a.e. } x\}}(\tilde{p}) \\ \langle Du - w, \tilde{q} - q \rangle &\leq \mathcal{I}_{\{|\cdot(x)| \leq \alpha_1(x), \text{ f.a.e. } x\}}(\tilde{q}) \end{aligned} \\
&\iff \begin{aligned} \langle \mathcal{E}w, \tilde{p} - p \rangle &\leq 0 \\ \langle Du - w, \tilde{q} - q \rangle &\leq 0, \end{aligned}
\end{aligned}$$

with the last two inequalities holding for any $\tilde{p} \in H_0(\operatorname{div}^2; \Omega)$ with $|\tilde{p}(x)| \leq \alpha_0(x)$ for a.e. $x \in \Omega$ and for any $\tilde{q} \in H_0(\operatorname{div}; \Omega)$ with $|\tilde{q}(x)| \leq \alpha_1(x)$ for a.e. $x \in \Omega$. Hence we obtain (2.29) and (2.30). $\square$

3. A SERIES OF REGULARIZED PROBLEMS

3.1. Regularization of the primal problem. With the aim of lifting the regularity of u and w to avoid measure-valued derivatives, we next consider the following regularized version of the primal weighted TGV problem (2.20):

$$\begin{aligned}
(3.1) \quad \text{minimize} \quad & \frac{1}{2} \|Tu - f\|_{L^2(\Omega)}^2 + \int_{\Omega} \alpha_1 |\nabla u - w| \, dx + \int_{\Omega} \alpha_0 |Ew| \, dx \\
& + \frac{\mu}{2} \|\nabla u\|_{L^2(\Omega)}^2 + \frac{\nu}{2} \|w\|_{H^1(\Omega, \mathbb{R}^d)}^2 \quad \text{over } (w, u) \in H^1(\Omega, \mathbb{R}^d) \times H^1(\Omega),
\end{aligned}$$

for some constants $0 < \mu, \nu \ll 1$. Existence of solutions for (3.1) follows from standard arguments.

Observe that (3.1) is equivalent to $\min_{(w,u) \in \hat{X}} Q_1(w,u) + Q_2(R(w,u))$ where $\hat{X} = H^1(\Omega, \mathbb{R}^d) \times H^1(\Omega)$, $\hat{Y} = L^2(\Omega, \mathcal{S}^{d \times d}) \times L^2(\Omega, \mathbb{R}^d)$, $R : \hat{X} \rightarrow \hat{Y}$ with $R(w,u) = (Ew, \nabla u - w)$, $Q_1 : \hat{X} \rightarrow \mathbb{R}$, $Q_2 : \hat{Y} \rightarrow \mathbb{R}$ with $Q(w,u) = \frac{1}{2} \|Tu - f\|_{L^2(\Omega)}^2 + \frac{\mu}{2} \|\nabla u\|_{L^2(\Omega, \mathbb{R}^d)}^2 + \frac{\nu}{2} \|w\|_{H^1(\Omega, \mathbb{R}^d)}^2$ and $Q_2(\psi, \phi) = \int_{\Omega} \alpha_1 |\phi| dx + \int_{\Omega} \alpha_0 |\psi| dx$. Note that the Attouch-Brezis condition is satisfied since $\text{dom}(Q_2) = Y$.

Proposition 3.1. *The pairs $(w,u) \in H^1(\Omega, \mathbb{R}^d) \times H^1(\Omega)$ and $(p,q) \in L^2(\Omega, \mathbb{R}^{d \times d}) \times L^2(\Omega, \mathbb{R}^d)$ are solutions to (3.1) and its predual problem, respectively, if and only if the following optimality conditions are satisfied:*

$$(3.2) \quad \mathcal{B}u - \mu \Delta u + \nabla^* q - T^* f = 0 \quad \text{in } H^1(\Omega)^*,$$

$$(3.3) \quad \nu w - \nu \Delta w - q + E^* p = 0 \quad \text{in } H^1(\Omega, \mathbb{R}^d)^*,$$

$$(3.4) \quad \begin{cases} |q| \leq \alpha_1, \\ \alpha_1(\nabla u - w) - q|\nabla u - w| = 0 & \text{if } |q(x)| = \alpha_1(x), \\ \nabla u - w = 0 & \text{if } |q(x)| < \alpha_1(x), \end{cases}$$

$$(3.5) \quad \begin{cases} |p| \leq \alpha_0, \\ \alpha_0 Ew - p|Ew| = 0 & \text{if } |p(x)| = \alpha_0(x), \\ Ew = 0 & \text{if } |p(x)| < \alpha_0(x), \end{cases}$$

where the multiplications are regarded component-wise.

Proof. The proof follows again easily by calculating the corresponding primal-dual optimality conditions. $\square$

Next we study the relationship between the solutions of (2.20) and (3.1) as the parameters μ, ν tend to zero.

Proposition 3.2. *Let $\mu_n, \nu_n \rightarrow 0$ and let $(w_n, u_n)_{n \in \mathbb{N}}$ be a sequence of solution pairs of the problem (3.1). Then $u_n \xrightarrow{*} u^*$ and $w_n \xrightarrow{*} w^*$ in $\text{BV}(\Omega)$ and $\text{BD}(\Omega)$ respectively, where (w^*, u^*) is a solution pair for (2.20). The convergence of w_n is up to a subsequence.*

Proof. For convenience of notation, define the energies

$$E_n(w, u) = \frac{1}{2} \|Tu - f\|_{L^2(\Omega)}^2 + \int_{\Omega} \alpha_1 |\nabla u - w| dx + \int_{\Omega} \alpha_0 |Ew| dx + \frac{\mu_n}{2} \|\nabla u\|_{L^2(\Omega)}^2 + \frac{\nu_n}{2} \|w\|_{H^1(\Omega, \mathbb{R}^d)}^2,$$

$$E(w, u) = \frac{1}{2} \|Tu - f\|_{L^2(\Omega)}^2 + \int_{\Omega} \alpha_1 d|Du - w| + \int_{\Omega} \alpha_0 d|\mathcal{E}w|.$$

We have

$$(3.6) \quad \frac{1}{2} \|Tu_n - f\|_{L^2(\Omega)}^2 + \int_{\Omega} \alpha_1 |\nabla u_n - w_n| dx + \int_{\Omega} \alpha_0 |Ew_n| dx \leq E_n(w_n, u_n) \leq E_n(0, 0) \leq \frac{1}{2} \|f\|_{L^2(\Omega)}^2.$$

Thus, the sequences $(u_n)_{n \in \mathbb{N}}$ and $(w_n)_{n \in \mathbb{N}}$ are bounded in $\text{BV}(\Omega)$ and $\text{BD}(\Omega)$, respectively. In order to see this, note that by setting $\underline{\alpha}_i := \min_{x \in \overline{\Omega}} \alpha_i(x)$, $i = 0, 1$, we get

$$\begin{aligned} \text{TGV}_{\underline{\alpha}_0, \underline{\alpha}_1}^2(u_n) &= \min_{w \in \text{BD}(\Omega)} \underline{\alpha}_1 \|\nabla u_n - w\|_{\mathcal{M}} + \underline{\alpha}_0 \|\mathcal{E}w\|_{\mathcal{M}} \\ &\leq \int_{\Omega} \alpha_1 |\nabla u_n - w_n| dx + \int_{\Omega} \alpha_0 |Ew_n| dx \leq \frac{1}{2} \|f\|_{L^2(\Omega)}^2. \end{aligned}$$

Hence, $(u_n)_{n \in \mathbb{N}}$ is bounded in the sense of second-order TGV. Again using [7, Theorem 4.4.10], we get that $\frac{1}{2} \|T \cdot - f\|_{L^2(\Omega)}^2$ is coercive, since it is the convex conjugate of $\frac{1}{2} \|T^* f + \cdot\|_{\mathcal{B}}^2 - \frac{1}{2} \|f\|_{L^2}^2$ which

is continuous at $0 \in L^2(\Omega)$. This implies further that this sequence is bounded both on $L^2(\Omega)$ and $\text{BV}(\Omega)$. The bound on $(w_n)_{n \in \mathbb{N}}$ in $\text{BD}(\Omega)$ then follows from (3.6).

From compactness theorems in those spaces (for $\text{BD}(\Omega)$ see for instance [57, Remark 2.4]) we have that there exist $u^* \in \text{BV}(\Omega) \cap L^2(\Omega)$ and $w^* \in \text{BD}(\Omega)$ such that $u_{n_k} \xrightarrow{*} u^*$ in $\text{BV}(\Omega)$ and weakly in $L^2(\Omega)$, and $w_{n_k} \xrightarrow{*} w^*$ in $\text{BD}(\Omega)$ along suitable subsequences. Due to the lower semicontinuity of the functional E with respect to these convergences, we have for any pair $(\tilde{w}, \tilde{u}) \in H^1(\Omega, \mathbb{R}^d) \times H^1(\Omega)$

$$(3.7) \quad E(w^*, u^*) \leq \liminf_{k \rightarrow \infty} E(w_{n_k}, u_{n_k}) \leq \liminf_{k \rightarrow \infty} E_{n_k}(w_{n_k}, u_{n_k}) \leq \liminf_{k \rightarrow \infty} E_{n_k}(\tilde{w}, \tilde{u}) = E(\tilde{w}, \tilde{u}).$$

Recall now that $\text{LD}(\Omega) = \{w \in L^1(\Omega, \mathbb{R}^d) : Ew \in L^1(\Omega, \mathbb{R}^{d \times d})\}$ is a Banach space endowed with the norm $\|w\|_{\text{LD}(\Omega)} = \|w\|_{L^1(\Omega, \mathbb{R}^d)} + \|Ew\|_{L^1(\Omega, \mathbb{R}^{d \times d})}$ and that $C^\infty(\overline{\Omega}, \mathbb{R}^d)$ is dense in that space; see [56, Proposition 1.3]. From this, in combination with the fact that $C^\infty(\overline{\Omega})$ is dense in $W^{1,1}(\Omega) \cap L^2(\Omega)$ we have that for every $(\hat{w}, \hat{u}) \in \text{LD}(\Omega) \times (W^{1,1}(\Omega) \cap L^2(\Omega))$ there exists a sequence

$$(\hat{w}_h, \hat{u}_h)_{h \in \mathbb{N}} \in C^\infty(\overline{\Omega}, \mathbb{R}^d) \times C^\infty(\overline{\Omega}) \subseteq H^1(\Omega, \mathbb{R}^d) \times H^1(\Omega),$$

such that $E(\hat{w}_h, \hat{u}_h) \rightarrow E(\hat{w}, \hat{u})$. Hence, since (3.7) holds we have that

$$(3.8) \quad E(w^*, u^*) \leq E(\hat{w}, \hat{u}), \quad \text{for all } (\hat{w}, \hat{u}) \in \text{LD}(\Omega) \times (W^{1,1}(\Omega) \cap L^2(\Omega)).$$

Finally, by following similar steps as in the proof of [58, Theorem 3], we can show that for every $(w, u) \in \text{BD}(\Omega) \times (\text{BV}(\Omega) \cap L^2(\Omega))$ there exists a sequence $(w_h, u_h)_{h \in \mathbb{N}} \in \text{LD}(\Omega) \times (W^{1,1}(\Omega) \cap L^2(\Omega))$ such that

$$\|u_h - u\|_{L^2(\Omega)} \rightarrow 0, \quad \int_{\Omega} \alpha_1 |\nabla u_h - w_h| dx \rightarrow \int_{\Omega} \alpha_1 d|Du - w|, \quad \int_{\Omega} \alpha_0 |Ew_h| dx \rightarrow \int_{\Omega} \alpha_0 d|\mathcal{E}w|,$$

which implies again that $E(w_h, u_h) \rightarrow E(w, u)$. This, together with (3.8) yields

$$E(w^*, u^*) \leq E(w, u), \quad \text{for all } (w, u) \in \text{BD}(\Omega) \times \text{BV}(\Omega).$$

This yields that (w^*, u^*) is a solution pair for (2.20). Finally, from the uniqueness of the solution u^* for (2.20) we get that the whole initial sequence $(u_n)_{n \in \mathbb{N}}$ converges to u^* weakly* in $\text{BV}(\Omega)$. The uniqueness follows from the fact that T is injective and hence the functional $\frac{1}{2} \|T \cdot - f\|_{L^2(\Omega)}^2$ is strictly convex. $\square$

We now proceed to the second level of regularization of the problem (3.1), which, in addition to lifting the regularity of u and w , respectively, also smoothes the non-differentiable constituents. For this purpose, we define the following primal problem, which will also be treated numerically, below:

$$(P_\gamma) \quad \begin{aligned} \text{minimize} \quad & \frac{1}{2} \|Tu - f\|_{L^2(\Omega)}^2 + \int_{\Omega} \alpha_1 \varphi_\gamma(\nabla u - w) dx + \int_{\Omega} \alpha_0 \varphi_\gamma(Ew) dx \\ & + \frac{\mu}{2} \|\nabla u\|_{L^2(\Omega)}^2 + \frac{\nu}{2} \|w\|_{H^1(\Omega, \mathbb{R}^d)}^2 \text{ over } (w, u) \in H^1(\Omega, \mathbb{R}^d) \times H^1(\Omega). \end{aligned}$$

Here φ_γ denotes the Huber-regularized version of the Euclidean norm. That is, for a vector $v \in S$, $S = \mathbb{R}^d$ or $\mathbb{R}^{d \times d}$ and $\gamma > 0$ we use

$$(3.9) \quad \varphi_\gamma(v)(x) = \begin{cases} |v(x)| - \frac{1}{2}\gamma & \text{if } |v(x)| \geq \gamma, \\ \frac{1}{2\gamma} |v(x)|^2 & \text{if } |v(x)| < \gamma, \end{cases}$$

with $|\cdot|$ denoting either the Euclidean norm in $\mathbb{R}^d$ or the Frobenius norm in $\mathbb{R}^{d \times d}$. We mention that this type of Huber regularization of TV-type terms in the primal problem corresponds to an

L^2 regularization of the dual variables in the predual [18, 40]. In order to illustrate this consider the following denoising problem (P_γ) without any H^1 regularization:

$$(3.10) \quad \text{minimize} \quad \frac{1}{2} \|u - f\|_{L^2(\Omega)}^2 + \int_{\Omega} \alpha_1 d\varphi_{\gamma_1}(Du - w) + \int_{\Omega} \alpha_0 d\varphi_{\gamma_2}(\mathcal{E}w) \quad \text{over } (w, u) \in \text{BD}(\Omega) \times \text{BV}(\Omega),$$

where

$$\begin{aligned} \int_{\Omega} \alpha_1 d\varphi_{\gamma_1}(Du - w) &= \int_{\Omega} \alpha_1 \varphi_{\gamma_1}(\nabla u - w) dx + \int_{\Omega} \alpha_1 d|D^s u|, \\ \int_{\Omega} \alpha_0 d\varphi_{\gamma_2}(\mathcal{E}w) &= \int_{\Omega} \alpha_0 \varphi_{\gamma_2}(Ew) dx + \int_{\Omega} \alpha_0 d|\mathcal{E}^s w|. \end{aligned}$$

Here $D^s u$ and $\mathcal{E}^s w$ denote the singular parts with respect to the Lebesgue measure of Du and $\mathcal{E}w$ respectively, following the Lebesgue decomposition $Du = \int_{\Omega} \nabla u dx + D^s u$ and $\mathcal{E}w = \int_{\Omega} Ew dx + \mathcal{E}^s w$. The corresponding predual problem of (3.10) is given by

$$(3.11) \quad \begin{aligned} &\text{maximize} \quad -\frac{1}{2} \|f + \text{div} q\|_{L^2(\Omega)}^2 - \frac{\gamma_0}{2} \int_{\Omega} \frac{1}{\alpha_0} |p|^2 dx - \frac{\gamma_1}{2} \int_{\Omega} \frac{1}{\alpha_1} |q|^2 dx + \frac{1}{2} \|f\|_{L^2(\Omega)}^2, \\ &\text{over } (p, q) \in W_0^d(\text{div}^2; \Omega) \times W_0^d(\text{div}; \Omega), \\ &\text{subject to } q = -\text{div} p, |p(x)| \leq \alpha_0(x), |q(x)| \leq \alpha_1(x). \end{aligned}$$

The proof is similar to the one of Proposition 2.4 with

$$F_1(p, q) = \mathcal{I}_{\{|\cdot(x)| \leq \alpha_0(x)\}}(p) + \mathcal{I}_{\{|\cdot(x)| \leq \alpha_1(x)\}}(q) - \frac{\gamma_0}{2} \int_{\Omega} \frac{1}{\alpha_0} |p|^2 dx - \frac{\gamma_1}{2} \int_{\Omega} \frac{1}{\alpha_1} |q|^2 dx,$$

and in the dualization process we use the fact that for an S -valued measure μ we have,

$$\int_{\Omega} \alpha d\varphi_{\gamma}(\mu) = \sup \left\{ \int_{\Omega} \phi d\mu - \mathcal{I}_{\{|\cdot(x)| \leq \alpha(x)\}}(\phi) - \frac{\gamma}{2} \int_{\Omega} \frac{1}{\alpha} |\phi|^2 dx : \phi \in C_c^\infty(\Omega, S) \right\};$$

see for instance [28].

Returning to the (doubly) regularized primal problem (P_γ) , we are primarily interested in its associated first-order optimality conditions.

Proposition 3.3. *We have that the pairs $(w, u) \in H^1(\Omega, \mathbb{R}^d) \times H^1(\Omega)$ and $(p, q) \in L^2(\Omega, \mathbb{R}^{d \times d}) \times L^2(\Omega, \mathbb{R}^d)$ are solution to (P_γ) and its predual problem, respectively, if and only if the following optimality conditions are satisfied:*

$$(3.12) \quad \mathcal{B}u - \mu \Delta u + \nabla^* q - T^* f = 0 \quad \text{in } H^1(\Omega)^*,$$

$$(3.13) \quad \nu w - \nu \Delta w - q + E^* p = 0 \quad \text{in } H^1(\Omega, \mathbb{R}^d)^*,$$

$$(3.14) \quad \max(|\nabla u - w|, \gamma_1) q - \alpha_1(\nabla u - w) = 0 \quad \text{in } L^2(\Omega, \mathbb{R}^d),$$

$$(3.15) \quad \max(|Ew|, \gamma_0) p - \alpha_0 Ew = 0 \quad \text{in } L^2(\Omega, \mathcal{S}^{d \times d}).$$

$$(3.16)$$

The proof of Proposition 3.3 follows from calculating the corresponding primal-dual optimality conditions as in Proposition 3.1. Finally, in order to avoid constraint degeneracy and for the sake of differentiability for the bilevel scheme in the next section, we also employ here a smoothed version $\max_\delta(\cdot, \gamma)$ of $\max$ and its derivative, denoted by $\mathcal{X}_\delta$, defined as follows for $r \geq 0$ and for $0 < \frac{\delta}{2} < \gamma$:

$$\max_\delta(r, \gamma) = \begin{cases} \gamma & \text{if } r \leq \gamma - \frac{\delta}{2}, \\ \frac{1}{2\delta}(r + \frac{\delta}{2} - \gamma)^2 + \gamma, & \text{if } \gamma - \frac{\delta}{2} < r < \gamma + \frac{\delta}{2}, \\ r & \text{if } r > \gamma + \frac{\delta}{2}. \end{cases} \quad \mathcal{X}_\delta(r, \gamma) = \begin{cases} 0 & \text{if } r \leq \gamma - \frac{\delta}{2}, \\ \frac{1}{\delta}(r + \frac{\delta}{2} - \gamma) & \text{if } \gamma - \frac{\delta}{2} < r < \gamma + \frac{\delta}{2}, \\ 1 & \text{if } r > \gamma + \frac{\delta}{2}. \end{cases}$$

Thus we arrive at the following regularized optimality conditions:

$$\begin{aligned} (\text{Opt}_1) \quad & \mathcal{B}u - \mu\Delta u + \nabla^* q - T^* f = 0, \\ (\text{Opt}_2) \quad & \nu w - \nu\Delta w - q + E^* p = 0, \\ (\text{Opt}_3) \quad & \max_\delta(|\nabla u - w|, \gamma_1) q - \alpha_1(\nabla u - w) = 0, \\ (\text{Opt}_4) \quad & \max_\delta(|Ew|, \gamma_0) p - \alpha_0 Ew = 0. \end{aligned}$$

Note that these would correspond to the optimality conditions of an analogue problem to (P_γ) where the Huber function φ_γ is substituted by an analogous C^2 version $\varphi_{\gamma,\delta}$, having the property $\nabla\varphi_{\gamma,\delta}(x) = x/\max_\delta(|x|, \gamma)$ pointwise, and which do not write down explicitly here. The relevant approximation result now follows, where we have set $\gamma_0 = \gamma_1 = \gamma$ for simplicity.

Proposition 3.4. *Let (w, u, q, p) and $(w_{\gamma,\delta}, u_{\gamma,\delta}, p_{\gamma,\delta}, q_{\gamma,\delta})$ satisfy the optimality conditions (3.2)–(3.5) and (Opt₁)–(Opt₄), respectively. Then, as $\gamma, \delta \rightarrow 0$, we have $u_{\gamma,\delta} \rightarrow u$ strongly in $H^1(\Omega)$, $w_{\gamma,\delta} \rightarrow w$ strongly in $H^1(\Omega, \mathbb{R}^d)$ as well as $\text{div} q_{\gamma,\delta} \rightarrow \text{div} q$ and $q_{\gamma,\delta} + \text{div} p_{\gamma,\delta} \rightarrow q + \text{div} p$ weakly* in $H^1(\Omega)^*$ and $H^1(\Omega, \mathbb{R}^d)^*$, respectively.*

Proof. By subtracting first two equations of the optimality system of Proposition 3.1 and 3.3, respectively, we get for all $v \in H^1(\Omega)$, $\omega \in H^1(\Omega, \mathbb{R}^d)$

$$(3.17) \quad \int_\Omega \mathcal{B}(u - u_{\gamma,\delta})v \, dx + \mu \int_\Omega \nabla(u - u_{\gamma,\delta}) \nabla v \, dx = \int_\Omega (q_{\gamma,\delta} - q) \nabla v \, dx,$$

$$(3.18) \quad \nu \int_\Omega (w - w_{\gamma,\delta})\omega \, dx + \nu \int_\Omega \nabla(w - w_{\gamma,\delta}) \nabla \omega \, dx = \int_\Omega (q - q_{\gamma,\delta})\omega \, dx + \int_\Omega (p_{\gamma,\delta} - p) E\omega \, dx.$$

When using $v = u - u_{\gamma,\delta}$ and $\omega = w - w_{\gamma,\delta}$ in the equations above and adding them up we get

$$(3.19) \quad \|u - u_{\gamma,\delta}\|_{\mathcal{B}^{-1}}^2 + \mu \|\nabla u - \nabla u_{\gamma,\delta}\|_{L^2(\Omega, \mathbb{R}^d)}^2 + \nu \|w - w_{\gamma,\delta}\|_{H^1(\Omega, \mathbb{R}^d)}^2 = R_1 + R_2,$$

where

$$R_1 := \int_\Omega (q_{\gamma,\delta} - q)^\top [\nabla u - w - (\nabla u_{\gamma,\delta} - w_{\gamma,\delta})] \, dx, \quad R_2 := \int_\Omega (p_{\gamma,\delta} - p)^\top E(w - w_{\gamma,\delta}) \, dx.$$

We now estimate R_1 and R_2 . Consider the partitions of Ω into disjoint sets (up to sets of measure zero) $\Omega = \mathcal{A} \cup \mathcal{I} = \mathcal{A}_{\gamma,\delta} \cup \mathcal{I}_{\gamma,\delta}$, where

$$\begin{aligned} \mathcal{A} &= \{x \in \Omega : |\nabla u - w| > 0\}, & \mathcal{I} &= \Omega \setminus \mathcal{A}, \\ \mathcal{A}_{\gamma,\delta} &= \left\{x \in \Omega : |\nabla u_{\gamma,\delta} - w_{\gamma,\delta}| > \gamma + \frac{\delta}{2}\right\}, & \mathcal{I}_{\gamma,\delta} &= \Omega \setminus \mathcal{A}_{\gamma,\delta}. \end{aligned}$$

We estimate R_1 separately on the disjoint sets $\mathcal{A}_{\gamma,\delta} \cap \mathcal{A}$, $\mathcal{A}_{\gamma,\delta} \cap \mathcal{I}$, $\mathcal{I}_{\gamma,\delta} \cap \mathcal{A}$ and $\mathcal{I}_{\gamma,\delta} \cap \mathcal{I}$. Recall that $|q(x)| \leq \alpha_1(x)$, as well as $|q_{\gamma,\delta}(x)| \leq \alpha_1(x)$ for almost every $x \in \Omega$. Starting from $\mathcal{A}_{\gamma,\delta} \cap \mathcal{A}$ and noticing that

$$q = \alpha_1 \frac{\nabla u - w}{|\nabla u - w|}, \quad q_{\gamma,\delta} = \alpha_1 \frac{\nabla u_{\gamma,\delta} - w_{\gamma,\delta}}{|\nabla u_{\gamma,\delta} - w_{\gamma,\delta}|},$$

it follows that pointwise on $\mathcal{A}_{\gamma,\delta} \cap \mathcal{A}$ (with argument x left off for ease of notation) we have

$$\begin{aligned} (q_{\gamma,\delta} - q)^\top [\nabla u - w - (\nabla u_{\gamma,\delta} - w_{\gamma,\delta})] &= q_{\gamma,\delta}(\nabla u - w) - \alpha_1 |\nabla u_{\gamma,\delta} - w_{\gamma,\delta}| - \alpha_1 |\nabla u - w| + q(\nabla u_{\gamma,\delta} - w_{\gamma,\delta}) \\ &\leq \alpha_1 |\nabla u - w| - \alpha_1 |\nabla u_{\gamma,\delta} - w_{\gamma,\delta}| - \alpha_1 |\nabla u - w| + \alpha_1 |\nabla u_{\gamma,\delta} - w_{\gamma,\delta}| \\ &= 0. \end{aligned}$$

Turning now to the set $\mathcal{A}_{\gamma,\delta} \cap \mathcal{I}$ and recalling $\nabla u - w = 0$ we have

$$(q_{\gamma,\delta} - q)^\top [\nabla u - w - (\nabla u_{\gamma,\delta} - w_{\gamma,\delta})] \leq -\alpha_1 |\nabla u_{\gamma,\delta} - w_{\gamma,\delta}| + |q| |\nabla u_{\gamma,\delta} - w_{\gamma,\delta}| \leq 0.$$

For the set $\mathcal{I}_{\gamma,\delta} \cap \mathcal{A}$, note that

$$|\nabla u_{\gamma,\delta} - w_{\gamma,\delta}| \leq \gamma + \frac{\delta}{2}, \quad \nabla u_{\gamma,\delta} - w_{\gamma,\delta} = \frac{\gamma}{\alpha_1} q_{\gamma,\delta}.$$

Thus, we can estimate

$$\begin{aligned} (q_{\gamma,\delta} - q)^\top [\nabla u - w - (\nabla u_{\gamma,\delta} - w_{\gamma,\delta})] &\leq q_{\gamma,\delta}(\nabla u - w) - \alpha_1 |\nabla u - w| - q_{\gamma,\delta}(\nabla u_{\gamma,\delta} - w_{\gamma,\delta}) + q(\nabla u_{\gamma,\delta} - w_{\gamma,\delta}) \\ &\leq \alpha_1 |\nabla u - w| - \alpha_1 |\nabla u - w| - q_{\gamma,\delta}(\nabla u_{\gamma,\delta} - w_{\gamma,\delta}) + q(\nabla u_{\gamma,\delta} - w_{\gamma,\delta}) \\ &\leq (2\gamma + \delta)\alpha_1 \end{aligned}$$

Similarly, for the set $\mathcal{I}_{\gamma,\delta} \cap \mathcal{I}$ we get

$$(q_{\gamma,\delta} - q)^\top [\nabla u - w - (\nabla u_{\gamma,\delta} - w_{\gamma,\delta})] \leq 2(\gamma + \delta)\alpha_1.$$

Combining the above estimates we have

$$R_1 \leq \int_{\Omega} (2\gamma + \delta)\alpha_1 dx \rightarrow 0$$

and for R_2 we similarly get

$$R_2 \leq \int_{\Omega} (2\gamma + \delta)\alpha_0 dx \rightarrow 0.$$

Hence, from (3.19) and the fact that $\|\cdot\|_{\mathcal{B}^{-1}}$ is equivalent to $\|\cdot\|_{L^2(\Omega)}$, we obtain the desired convergences for $u_{\gamma,\delta}$ and $w_{\gamma,\delta}$. From this result and using (3.17) and (3.18) we get that for every $v \in H^1(\Omega)$ and for every $\omega \in H^1(\Omega, \mathbb{R}^d)$ we have

$$\int_{\Omega} v \operatorname{div} q_{\gamma,\delta} dx \rightarrow \int_{\Omega} v \operatorname{div} q dx \quad \text{and} \quad \int_{\Omega} \omega(q_{\gamma,\delta} + \operatorname{div} p_{\gamma,\delta}) dx \rightarrow \int_{\Omega} \omega(q + \operatorname{div} p) dx,$$

as $\gamma, \delta \rightarrow 0$. This completes the proof. $\square$

Finally, the following approximation result holds true, when all the parameters μ, ν, γ, δ tend to zero.

Proposition 3.5. *Let $\mu_n, \nu_n, \gamma_n, \delta_n \rightarrow 0$, and denote by $u_{\mu_n, \nu_n, \gamma_n, \delta_n} \in H^1(\Omega)$ the solution of (Opt_1) – (Opt_1) with $(\mu, \nu, \gamma, \delta) = (\mu_n, \nu_n, \gamma_n, \delta_n)$. Then $u_{\mu_n, \nu_n, \gamma_n, \delta_n} \xrightarrow{*} u^*$ in $\operatorname{BV}(\Omega)$, where u^* solves (2.20).*

Proof. It is easy to show that $u_{\mu_n, \nu_n, \gamma_n, \delta_n} \rightarrow u^*$ in $L^1(\Omega)$. Indeed, we have

$$\|u_{\mu_n, \nu_n, \gamma_n, \delta_n} - u^*\|_{L^1(\Omega)} \leq \|u_{\mu_n, \nu_n, 0, 0} - u^*\|_{L^1(\Omega)} + \|u_{\mu_n, \nu_n, \gamma_n, \delta_n} - u_{\mu_n, \nu_n, 0, 0}\|_{L^1(\Omega)}.$$

According to Proposition 3.2 it holds that $\|u_{\mu_n, \nu_n, 0, 0} - u^*\|_{L^1(\Omega)} \rightarrow 0$. The other term tends to zero according to equation (3.19) of Proposition 3.4. There, the estimates for R_1, R_2 are not affected if we substitute u and $u_{\gamma,\delta}$ by $u_{\mu_n, \nu_n, 0, 0}$ and $u_{\mu_n, \nu_n, \gamma_n, \delta_n}$, respectively. In other words, the estimate

$$\|u_{\mu_n, \nu_n, 0, 0} - u_{\mu_n, \nu_n, \gamma_n, \delta_n}\|_{L^2(\Omega)}^2 \leq C(2\gamma_n + \delta_n)|\Omega|(\alpha_0 + \alpha_1)\|_\infty$$

holds for some $C > 0$ and hence $\|u_{\mu_n, \nu_n, \gamma_n, \delta_n} - u_{\mu_n, \nu_n, 0, 0}\|_{L^1(\Omega)} \rightarrow 0$.

To finish the proof and show that the convergence is weak* in $BV(\Omega)$, it suffices to establish that $\int_{\Omega} |\nabla u_{\mu_n, \nu_n, \gamma_n, \delta_n}| dx$ is uniformly bounded in n ; see [2, Proposition 3.13]. Observe first that, if $\varphi_{\gamma, \delta}$ is the C^2 regularized Huber function, that corresponds to the δ -smoothing of $\max$, then as in the proof of Proposition 3.2 we get

$$(3.20) \quad \int_{\Omega} \alpha_1 \varphi_{\gamma, \delta}(\nabla u_{\mu_n, \nu_n, \gamma_n, \delta_n} - w_{\mu_n, \nu_n, \gamma_n, \delta_n}) dx + \int_{\Omega} \alpha_0 \varphi_{\gamma, \delta}(E w_{\mu_n, \nu_n, \gamma_n, \delta_n}) dx \leq \frac{1}{2} \|f\|_{L^2(\Omega)}^2.$$

As in (3.9) we also have that $\varphi_{\gamma, \delta}(\cdot) \geq |\cdot| - \frac{1}{2}\gamma$, and hence we obtain

$$(3.21) \quad \int_{\Omega} \alpha_1 |\nabla u_{\mu_n, \nu_n, \gamma_n, \delta_n} - w_{\mu_n, \nu_n, \gamma_n, \delta_n}| dx + \int_{\Omega} \alpha_0 |E w_{\mu_n, \nu_n, \gamma_n, \delta_n}| dx \leq \frac{1}{2} \|f\|_{L^2(\Omega)}^2 + \frac{(\|\alpha_1\|_{\infty} + \|\alpha_0\|_{\infty})|\Omega|\gamma_n}{2} \leq K,$$

for some constant $K > 0$. Then, as in the proof of Proposition 3.2, we get that $(u_{\mu_n, \nu_n, \gamma_n, \delta_n})_{n \in \mathbb{N}}$ is bounded in TGV which, together with the L^1 bound, gives the desired bound in TV. $\square$

4. THE BILEVEL OPTIMIZATION SCHEME

In this section we will adapt the bilevel optimization framework developed in [37, 39] in order to automatically select the regularization functions α_0 and α_1 . The main idea is to minimize a suitable upper level objective over both the image u and the regularization parameters α_0, α_1 subject to u being a solution to a (regularized) TGV-based reconstruction problem with these regularization weights.

It is useful to recall the definitions of the localized residual R and the function F as stated in the introduction:

$$(4.1) \quad Ru(x) = \int_{\Omega} w(x, y)(Tu - f)^2(y) dy,$$

where $w \in L^{\infty}(\Omega \times \Omega)$ with $\int_{\Omega} \int_{\Omega} w(x, y) dx dy = 1$ and

$$(4.2) \quad F(v) := \frac{1}{2} \int_{\Omega} \max(v - \bar{\sigma}^2, 0)^2 dx + \frac{1}{2} \int_{\Omega} \min(v - \underline{\sigma}^2, 0)^2 dx,$$

for some appropriately chosen $\underline{\sigma}^2, \bar{\sigma}^2$.

We thus end up to the following bilevel minimization problem:

$$(4.3) \quad \begin{cases} \min J(u, \alpha_0, \alpha_1) := F(R(u)) + \frac{\lambda_0}{2} \|\alpha_0\|_{H^1(\Omega)}^2 + \frac{\lambda_1}{2} \|\alpha_1\|_{H^1(\Omega)}^2, \\ \text{over } (u, \alpha_0, \alpha_1) \in H^1(\Omega) \times \mathcal{A}_{ad}^0 \times \mathcal{A}_{ad}^1, \\ \text{subject to } (u, w) = \underset{(\tilde{u}, \tilde{w})}{\operatorname{argmin}} \frac{1}{2} \|T\tilde{u} - f\|_{L^2(\Omega)}^2 + \int_{\Omega} \alpha_1 \varphi_{\gamma, \delta}(\nabla \tilde{u} - \tilde{w}) dx + \int_{\Omega} \alpha_0 \varphi_{\gamma, \delta}(E\tilde{w}) dx \\ \quad + \frac{\mu}{2} \|\nabla \tilde{u}\|_{L^2(\Omega)}^2 + \frac{\nu}{2} \|\tilde{w}\|_{H^1(\Omega, \mathbb{R}^d)}^2. \end{cases}$$

Utilizing the regularized primal-dual first-order optimality characterization (Opt_1)–(Opt_4) of the solution to the lower level problem of (4.3), we arrive at the following *mathematical program with equilibrium constraints* (MPEC, for short) which is equivalent to (4.3):

$$(\mathbb{P}_{\text{TGV}}) \quad \left\{ \begin{array}{l} \min J(u, \alpha_0, \alpha_1) := F(R(u)) + \frac{\lambda_0}{2} \|\alpha_0\|_{H^1(\Omega)}^2 + \frac{\lambda_1}{2} \|\alpha_1\|_{H^1(\Omega)}^2, \\ \text{over } (u, \alpha_0, \alpha_1) \in H^1(\Omega) \times \mathcal{A}_{ad}^0 \times \mathcal{A}_{ad}^1, \\ \text{subject to} \quad \mathcal{B}u - \mu \Delta u + \nabla^* q - T^* f = 0, \\ \quad \quad \quad \nu w - \nu \Delta w - q + E^* p = 0, \\ \quad \quad \quad \max_{\delta} (|\nabla u - w|, \gamma_1) q - \alpha_1 (\nabla u - w) = 0, \\ \quad \quad \quad \max_{\delta} (|Ew|, \gamma_0) p - \alpha_0 Ew = 0. \end{array} \right.$$

Note that in view of the equivalence of (4.3) and $(\mathbb{P}_{\text{TGV}})$, we will still refer the latter as the bilevel TGV problem. A few words about $(\mathbb{P}_{\text{TGV}})$ are in order. Here, α_i are forced to be contained in the box constraint sets

$$(4.4) \quad \mathcal{A}_{ad}^i := \{\alpha_i \in H^1(\Omega) : \underline{\alpha}_i \leq \alpha_i \leq \bar{\alpha}_i\}, \quad i = 0, 1,$$

with $\underline{\alpha}_i, \bar{\alpha}_i \in L^2(\Omega)$ and $0 < \underline{\epsilon} \leq \underline{\alpha}_i(x) < \bar{\alpha}_i(x) - \bar{\epsilon}$ in Ω for some $\underline{\epsilon}, \bar{\epsilon} > 0$, $i = 0, 1$. Note that the H^1 regularity on the parameter functions α_0, α_1 facilitates the existence and differential sensitivity analysis as established in [37, 39] for the TV case. Note, however, that this setting does not guarantee a priori that these functions belong to $C(\bar{\Omega})$, the regularity required for applying the dualization results of the previous sections. Nevertheless, under mild data assumptions, one can make use of the following regularity result of the H^1 -projection onto the sets $\mathcal{A}_{ad}^0$ and $\mathcal{A}_{ad}^1$; see [39, Corollary 2.3] for a proof.

Proposition 4.1. *Let $\Omega \subset \mathbb{R}^\ell$ with $\ell = 1, 2, 3$ be a bounded convex set and let $\mathcal{A}_{ad} := \{\alpha \in H^1(\Omega) : \underline{\alpha} \leq \alpha \leq \bar{\alpha}\}$, where $\underline{\alpha}, \bar{\alpha} \in H^2(\Omega)$ such that $\underline{\alpha} \leq \bar{\alpha}$ and $\frac{\partial \underline{\alpha}}{\partial \nu} = \frac{\partial \bar{\alpha}}{\partial \nu} = 0$ in $H^{1/2}(\partial\Omega)$. Then if*

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

it holds

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

In particular, if $\underline{\alpha}_0, \bar{\alpha}_0, \underline{\alpha}_1, \bar{\alpha}_1$ as well as the initializations for α_1 and α_0 are constant functions, then along the projected gradient iterations of Algorithm 2, the weights are guaranteed to belong to $H^2(\Omega)$ which (for dimension $d \leq 2$) embeds into $C(\bar{\Omega})$.

We briefly note that in the TV case it can be shown [31, 33] that $W^{1,1}$ regularity for the regularization parameter α suffices to establish a dualization framework. A corresponding result is not yet known for TGV, even though one expects that it could be shown by similar arguments. Hence, here we will also make use of the H^1 -projection regularity result as described above.

Regarding the box constraints (4.4) in [26] it was shown that for a PSNR-optimizing upper level objective $\tilde{J}(u, \alpha) = \|u(\alpha) - f\|_{L^2(\Omega)}^2$ subject to H^1 and Huber regularized TV and TGV denoising problems, under some mild conditions on the data f , the optimal scalar solutions α and (α_0, α_1) are strictly positive. As depicted in Figure 2 the upper level objective discussed here appears close to optimizing the PSNR, keeping the parameters strictly positive via (4.4) seems, however, necessary for the time being.

We now briefly discuss how to treat the bilevel problem $(\mathbb{P}_{\text{TGV}})$. Let $(\alpha_0, \alpha_1) \mapsto u(\alpha_0, \alpha_1)$ denote the solution map for the lower level problem, equivalently of the optimality conditions (Opt_1) – (Opt_4) . Then the problem $(\mathbb{P}_{\text{TGV}})$ admits the following reduced version

$$(4.5) \quad \min \hat{J}(\alpha_0, \alpha_1) := J(u(\alpha_0, \alpha_1), \alpha_0, \alpha_1) \quad \text{over } \alpha_0 \in \mathcal{A}_{ad}^0, \alpha_1 \in \mathcal{A}_{ad}^1.$$

Similarly to the TV case [37], one can show that the reduced functional $\hat{J} : H^1(\Omega) \times H^1(\Omega) \rightarrow \mathbb{R}$ is differentiable. We can then apply the KKT framework in Banach spaces [60]:

$$(4.6) \quad \begin{cases} \text{minimize} & \mathcal{J}(x) \quad \text{over } x \in X, \\ \text{subject to} & x \in \mathcal{C} \text{ and } g(x) = 0, \end{cases}$$

where $V, \mathcal{A}, Z$ are Banach spaces, $X = V \times \mathcal{A}$, $\mathcal{J} : X \rightarrow \mathbb{R}$ and $g : X \rightarrow Z$ are Fréchet differentiable and continuous differentiable functions, respectively, and $\mathcal{C} \subset X$ is a non-empty, closed convex set. In the bilevel TGV problem ($\mathbb{P}_{\text{TGV}}$) we have $V = H^1(\Omega) \times H^1(\Omega, \mathbb{R}^d) \times L^2(\Omega, \mathbb{R}^d)$, $L^2(\Omega, \mathcal{S}^{d \times d})$, $\mathcal{A} = H^1(\Omega) \times H^1(\Omega)$, $\mathcal{C} = V \times \mathcal{A}_{ad}^0 \times \mathcal{A}_{ad}^1$, and $Z = H^1(\Omega)^* \times H^1(\Omega, \mathbb{R}^d)^* \times L^2(\Omega, \mathbb{R}^d)$, $L^2(\Omega, \mathcal{S}^{d \times d})$. Here $g : X \rightarrow Z$ is defined by the optimality conditions (Opt_1)–(Opt_4). Finally, for $x = (u, w, q, p, \alpha_0, \alpha_1) := (x, \alpha_0, \alpha_1)$, we have $\mathcal{J}(x) = J(u, \alpha_0, \alpha_1)$. Note that the framework of (4.6) guarantees the existence of an adjoint variable $x^* \in V$ with the help of which an optimal solution of ($\mathbb{P}_{\text{TGV}}$) can be characterised. This adjoint variable also allows the computation of the derivative of the reduced objective $\hat{J}'$ in an amenable way, see next section.

We will skip here the proofs for the differentiability of the functions g and the reduced objective J as well as the existence proofs for ($\mathbb{P}_{\text{TGV}}$). These results can be shown similarly to the corresponding assertions for TV; see [37, Thm. 6.1, Prop. 6.2, Prop. 6.3].

5. NUMERICAL IMPLEMENTATION

In this section we will describe a Newton method for the lower level problem, a projected gradient algorithm for the solution of the discretized version of the bilevel problem ($\mathbb{P}_{\text{TGV}}$), as well as provide corresponding numerical examples in denoising.

5.1. Newton solver for the lower level problem. Before we proceed to devising of a projected gradient algorithm for the solution of the bilevel problem, we discuss first a primal-dual Newton algorithm for the solution of the first-order optimality conditions (Opt_1)–(Opt_4) re-written here for the sake of readability:

$$(5.1) \quad \mathcal{B}u - \mu \Delta u - \text{div} q - f = 0,$$

$$(5.2) \quad \nu w - \nu \Delta w - q - \text{div} p = 0,$$

$$(5.3) \quad \max_{\delta} (|\nabla u - w|, \gamma_1) q - \alpha_1 (\nabla u - w) = 0,$$

$$(5.4) \quad \max_{\delta} (|Ew|, \gamma_0) p - \alpha_0 Ew = 0.$$

A few words on the discrete involved quantities are in order. Images ($d = 2$) are considered as elements of $U_h := \{u \mid u : \Omega_h \rightarrow \mathbb{R}\}$ where $\Omega_h = \{1, 2, \dots, n\} \times \{1, 2, \dots, m\}$ is a discrete cartesian grid that corresponds to the image pixels. The mesh size, defined as the distance between the grid points, is set to $h = 1/\sqrt{nm}$. We define the associated discrete function spaces $W_h = U_h \times U_h$, $V_h = U_h \times U_h \times U_h$, so that $p \in V_h$ with $p = (p^{11}, p^{12}, p^{22})$. For the discrete gradient and divergence we have, $\nabla : W_h \rightarrow V_h$ and $\text{div} : V_h \rightarrow W_h$ satisfying the adjoint relation $\nabla = -\text{div}^\top$, setting zero values at the ghost points. The discretized symmetrized gradient Ew is defined as $\frac{1}{2}(\nabla w + (\nabla w)^\top)$. For the discretized versions of the Laplacian, we use the standard five-point stencils with zero Neumann boundary conditions, by setting the function values of ghost grid points to be the same with the function value of the nearest grid point in Ω_h . Note that these act on the primal variables u and w , which satisfy natural boundary conditions in contrast to the dual variable.

The system of equations (5.1)–(5.4) can be shortly written as $g_{\text{pd}}(x) = 0$, where $x = (u, w, q, p)$. We compute the derivative of g_{pd} at a point $x = (u, w, q, p)$ as the following block-matrix:

$$Dg_{\text{pd}}(\mathbf{x}) = Dg_{\text{pd}}(u, w, q, p) = \begin{bmatrix} A & B \\ C & D \end{bmatrix},$$

where

$$(5.5) \quad A = \begin{bmatrix} \mathcal{B} - \mu\Delta & 0 \\ 0 & \nu(I - \Delta) \end{bmatrix}, \quad B = \begin{bmatrix} -\text{div} & 0 \\ -I & -\text{div} \end{bmatrix}, \quad D = \begin{bmatrix} \max_{\delta}(|\nabla u - w|, \gamma_1) & 0 \\ 0 & \max_{\delta}(|Ew|, \gamma_0) \end{bmatrix},$$

$$(5.6) \quad C = \begin{bmatrix} -\alpha_1 \nabla + q\mathcal{X}_{\delta}(|\nabla u - w|, \gamma_1) \frac{\nabla u - w}{|\nabla u - w|} \cdot \nabla & \alpha_1 I + q\mathcal{X}_{\delta}(|\nabla u - w|, \gamma_1) \frac{\nabla u - w}{|\nabla u - w|} \cdot (-I) \\ 0 & -\alpha_0 E + p\mathcal{X}_{\delta}(|Ew|, \gamma_0) \frac{Ew}{|Ew|} \cdot E \end{bmatrix}.$$

Given $\mathbf{x}^k$, the Newton iteration for solving the system of equations (5.1)–(5.4), or $g_{\text{pd}}(\mathbf{x}) = 0$ for short, reads

$$\mathbf{x}^{k+1} = \mathbf{x}^k - D\mathcal{F}(\mathbf{x}^k)^{-1}\mathcal{F}(\mathbf{x}^k),$$

which can also be written as

$$(5.7) \quad Dg_{\text{pd}}(\mathbf{x}^k)\mathbf{x}^{k+1} = Dg_{\text{pd}}(\mathbf{x}^k)\mathbf{x}^k - g_{\text{pd}}(\mathbf{x}^k).$$

Here it is convenient to introduce the notation

$$Dg_{\text{pd}}(\mathbf{x}^k) = Dg_{\text{pd}}(u^k, w^k, q^k, p^k) = \begin{bmatrix} A & B \\ C_k & D_k \end{bmatrix}$$

since only the submatrices C and D depend on k . Note that the right-hand side $Dg_{\text{pd}}(\mathbf{x}^k)\mathbf{x}^k - g_{\text{pd}}(\mathbf{x}^k)$ of the linear system (5.7) can be written as

$$Dg_{\text{pd}}(\mathbf{x}^k)\mathbf{x}^k - \mathcal{F}(\mathbf{x}^k) = \begin{pmatrix} b_1^k \\ b_2^k \end{pmatrix},$$

where

$$b_1^k = (f, 0)^{\top}, \quad \text{and} \quad b_2^k = \left(q^k \mathcal{X}_{\delta}(|\nabla u^k - w^k|, \gamma_1) |\nabla u^k - w^k|, p^k \mathcal{X}_{\delta}(|Ew^k|, \gamma_0) |Ew^k| \right)^{\top}.$$

Notation-wise, the components that appear in b_2^k should be regarded as the diagonals of the corresponding diagonal matrices that we mentioned before, multiplied component-wise. By introducing the notation $\mathbf{x}_1^k = (u^k, w^k)^{\top}$, $\mathbf{x}_2^k = (q^k, p^k)^{\top}$, the Newton system (5.7) can be written as

$$(5.8) \quad \begin{bmatrix} A & B \\ C_k & D_k \end{bmatrix} \begin{pmatrix} \mathbf{x}_1^{k+1} \\ \mathbf{x}_2^{k+1} \end{pmatrix} = \begin{pmatrix} b_1^k \\ b_2^k \end{pmatrix}.$$

The above system can be simplified utilizing the Schur complement: First solve for the primal variables $\mathbf{x}_1^{k+1} = (u^{k+1}, w^{k+1})$ and then recover the dual ones $\mathbf{x}_2^{k+1} = (q^{k+1}, p^{k+1})$. This yields

$$(A - BD_k^{-1}C_k)\mathbf{x}_1^{k+1} = b_1^k - BD_k^{-1}b_2^k,$$

$$\mathbf{x}_2^{k+1} = D_k^{-1}(b_2^k - C_k\mathbf{x}_1^{k+1}).$$

The following result then holds.

Lemma 5.1. *If (q^k, p^k) belong to the feasible set, i.e., $|q^k| \leq \alpha_1$ and $|p^k| \leq \alpha_0$ component-wise, then the matrix $S_k := (A - BD_k^{-1}C_k)$ is positive definite and for the minimum eigenvalues we have $\lambda_{\min}(S_k) \geq \lambda_{\min}(A) > 0$. Furthermore, S_k^{-1} is bounded independently of k .*

Algorithm 1Newton algorithm for the solution of the regularized TGV primal problem (P_γ)

while some stopping criterion is not satisfied **do**Solve the linear system for $\mathbf{x}_1^{k+1} = (u^{k+1}, w^{k+1})$

$$(A - BD_k^{-1}C_k)\mathbf{x}_1^{k+1} = b_1^k - BD_k^{-1}b_2^k$$

Update $\tilde{\mathbf{x}}_2^{k+1} = (\tilde{q}^{k+1}, \tilde{p}^{k+1})$ as follows

$$\tilde{\mathbf{x}}_2^{k+1} = D_k^{-1}(b_2^k - C_k\mathbf{x}_1^{k+1})$$

Compute q^{k+1}, p^{k+1} as projections of $\tilde{q}^{k+1}, \tilde{p}^{k+1}$ onto the feasible sets $\{q : |q| \leq \alpha_1\}, \{p : |p| \leq \alpha_0\}$ **end while**

The proof of Lemma 5.1 follows the steps of the analogous proof in [40] and is hence omitted. Summarizing, the Newton method for the solution of the (5.1)–(5.4) is outlined in Algorithm 1. Here we have followed [40] and project in every iteration the variables q, p onto the feasible sets such that the result of Lemma 5.1 holds.

The projections onto the feasible sets are defined respectively as

$$(5.9) \quad q = \frac{\tilde{q}}{\max\left\{1, \frac{|\tilde{q}|}{\alpha_1}\right\}}, \quad p = \frac{\tilde{p}}{\max\left\{1, \frac{|\tilde{p}|}{\alpha_0}\right\}},$$

with the equalities above to be considered component-wise.

5.2. The numerical algorithm for ($\mathbb{P}_{\text{TGV}}$). We now describe our strategy for solving the discretized version of the bilevel TGV problem ($\mathbb{P}_{\text{TGV}}$). We note that in most of the experiments we will keep α_0 a scalar – this is justified by the numerical results; see the relevant discussion later on. We will always mention the small modifications on the algorithm when α_0 is a scalar. We will also make use here of the discrete Laplacian with zero Neumann boundary conditions $\Delta : U_h \rightarrow U_h$ which is used to act on the weight function α_1 . These are the desired boundary conditions for α_0, α_1 as dictated by the regularity result for the H^1 -projection in [39, Corollary 2.3]. For a function $u \in U_h$ we define the discrete ℓ^2 norm as

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

For the discrete H^1 norm applied to the weight function α_1 we use

$$\|\alpha_1\|_{H^1(\Omega_h)} = h \sqrt{\alpha_1^\top (I - \Delta) \alpha_1},$$

while the dual norm is defined as

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

based on the $H^1 \rightarrow H^1(\Omega)^*$ Riesz map $\alpha \mapsto r = (I - \Delta)\alpha$. For the discrete version of the averaging filter in the definition of the localized residuals (4.1) we use a filter of size $n_w \times n_w$, with entries of equal value whose sum is equal to one. With these definitions the discrete version of the bilevel TGV ($\mathbb{P}_{\text{TGV}}$), is the following:

$$(\mathbb{P}_{\text{TGV}}^h) \left\{ \begin{array}{l} \text{minimize } \frac{1}{2} \| (R(u) - \bar{\sigma}^2)^+ \|_{\ell^2(\Omega_h)}^2 + \frac{1}{2} \| (\underline{\sigma}^2 - R(u))^+ \|_{\ell^2(\Omega_h)}^2 + \frac{\lambda_0}{2} \|\alpha_0\|_{H^1(\Omega_h)}^2 + \frac{\lambda_0}{2} \|\alpha_1\|_{H^1(\Omega_h)}^2, \\ \text{over } (u, \alpha_0, \alpha_1) \in U_h \times (\mathcal{A}_{ad}^0)_h \times (\mathcal{A}_{ad}^1)_h, \\ \text{subject to } \begin{array}{l} \mathcal{B}u - \mu \Delta u - \text{div} q - f = 0, \\ \nu w - \nu \Delta w - q - \text{div} p = 0, \\ \max_{\delta} (|\nabla u - w|, \gamma_1) q - \alpha_1 (\nabla u - w) = 0, \\ \max_{\delta} (|Ew|, \gamma_0) p - \alpha_0 Ew = 0. \end{array} \end{array} \right.$$

Here the box constraint sets are defined as

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

The discretized versions of (5.1)–(5.4) and the upper level objective are still denoted by $g_{\text{pd}}(\mathbf{x}) = 0$ and J respectively.

Regarding the choice of the lower and upper bounds for the local variance $\underline{\sigma}^2$ and $\bar{\sigma}^2$, respectively, we follow here the following rules, where σ^2 is the variance of the “Gaussian” noise contaminating the data:

$$(5.10) \quad \bar{\sigma}^2 = \sigma^2 \left(1 + \frac{\sqrt{2}}{n_w} \right), \quad \underline{\sigma}^2 = \sigma^2 \left(1 - \frac{\sqrt{2}}{n_w} \right).$$

The formulae (5.10) are based on the statistics of the extremes; see [39, Section 4.2.1].

We now proceed by describing the algorithm for the numerical solution of $(\mathbb{P}_{\text{TGV}}^h)$. In essence, we employ a discretized projected gradient method with Armijo line search. We briefly describe how the discrete gradient of the reduced objective functional is computed with the help of the adjoint equation. The corresponding discretized version of the latter

$$D_{\mathbf{x}} g_{\text{pd}}(\mathbf{x}^*)^\top = -D_{\mathbf{x}} J(\mathbf{x}),$$

where $\mathbf{x}^* := (u^*, w^*, q^*, p^*)$ is the adjoint variable, reads

$$(5.11) \quad \begin{bmatrix} A^\top & C^\top \\ B^\top & D^\top \end{bmatrix} \begin{pmatrix} u^* \\ w^* \\ q^* \\ p^* \end{pmatrix} = \begin{pmatrix} -2(u - f) (w * ((R(u) - \bar{\sigma}^2)^+ - (\underline{\sigma}^2 - R(u))^+)) \\ 0 \\ 0 \\ 0 \end{pmatrix} := \begin{pmatrix} b_1^* \\ b_2^* \end{pmatrix},$$

where the matrices above were defined in (5.5) and (5.6). The equation can be solved again for $\mathbf{x}_1^* := (u^*, w^*)$ first and then subsequently for $\mathbf{x}_2^* := (q^*, p^*)$ as follows

$$\begin{aligned} (A^\top - C^\top (D^\top)^{-1} B^\top) \mathbf{x}_1^* &= b_1^*, \\ \mathbf{x}_2^* &= (D^\top)^{-1} (b_2^* - B^\top \mathbf{x}_1^*). \end{aligned}$$

The derivatives of the reduced objective with respect to α_0 and α_1 , respectively, are

$$(5.12) \quad \hat{J}'_{\alpha_0}(\alpha_0, \alpha_1) = (D_{\alpha_0} g_{\text{pd}})^\top x^* + D_{\alpha_0} J(\alpha_0, \alpha_1)$$

$$(5.13) \quad \begin{aligned} &= [Id \quad Id \quad 2Id] \begin{pmatrix} 0 & & & \\ & 0 & & \\ & & 0 & \\ & & & -diag(Ew) \end{pmatrix} \begin{pmatrix} u^* \\ w^* \\ q^* \\ p^* \end{pmatrix} + \lambda_0(Id - \Delta)\alpha_0 \\ &= -[Id \quad Id \quad 2Id] diag(Ew)p^* + \lambda_0(Id - \Delta)\alpha_0, \end{aligned}$$

$$(5.14) \quad \hat{J}'_{\alpha_1}(\alpha_0, \alpha_1) = (D_{\alpha_1} g_{\text{pd}})^\top x^* + D_{\alpha_1} J(\alpha_0, \alpha_1)$$

$$\begin{aligned} &= [Id \quad Id] \begin{pmatrix} 0 & & & \\ & 0 & & \\ & & -diag(Du - w) & \\ & & & 0 \end{pmatrix} \begin{pmatrix} u^* \\ w^* \\ q^* \\ p^* \end{pmatrix} + \lambda_1(Id - \Delta)\alpha_1, \\ &= -[Id \quad Id] diag(Du - w)q^* + \lambda_1(Id - \Delta)\alpha_1, \end{aligned}$$

where $x = (u, w, q, p)$ solves $g_{\text{pd}}(x) = 0$ for α_0, α_1 . The corresponding reduced gradients are

$$(5.15) \quad \nabla_{\alpha_i} \hat{J}(\alpha_0, \alpha_1) = (I - \Delta)^{-1} \hat{J}'_{\alpha_i}(\alpha_0, \alpha_1), \quad i = 0, 1.$$

We note that in the case of a scalar α_0 , we set $\lambda_0 = 0$. Then, $\hat{J}'_{\alpha_0}(\alpha_0, \alpha_1) = -[\mathbb{1} \quad \mathbb{1} \quad 2\mathbb{1}]diag(Ew)p^*$, and $\nabla_{\alpha_0} \hat{J}(\alpha_0, \alpha_1) = \hat{J}'_{\alpha_0}(\alpha_0, \alpha_1)$. Here $\mathbb{1}$ denotes a matrix of size $1 \times nm$ with all entries equal to one. In summary, the projected gradient algorithm for the solutions of $(\mathbb{P}_{\text{TGV}}^h)$ is described in Algorithm 2

We lastly note that the projections $P_{(\mathcal{A}_{ad}^0)_h}$, $P_{(\mathcal{A}_{ad}^1)_h}$ are computed as described in [39, Algorithm 4], that is via the semismooth Newton method developed in [32]. We only mention that the original discretized H^1 -projection problem $P_{(\mathcal{A}_{ad})_h}(\tilde{\alpha})$ given by

$$(5.16) \quad \begin{cases} \min & \frac{1}{2} \|\alpha - \tilde{\alpha}\|_{H^1(\Omega_h)}^2 := \frac{h^2}{2} (\alpha - \tilde{\alpha})^\top (I - \Delta) (\alpha - \tilde{\alpha}), \\ \text{over} & \alpha \in (\mathcal{A}_{ad})_h = \{\alpha \in U_h : \underline{\alpha} \leq \alpha_{i,j} \leq \bar{\alpha}\}, \end{cases}$$

is approximated by the following penalty version:

$$(5.17) \quad \min_{\alpha \in U_h} \frac{1}{2} \|\alpha - \tilde{\alpha}\|_{H^1(\Omega_h)}^2 + \frac{1}{\epsilon_\alpha} \left(\frac{1}{2} \|(\alpha - \bar{\alpha})^+\|_{\ell^2(\Omega_h)}^2 + \frac{1}{2} \|(\underline{\alpha} - \alpha)^+\|_{\ell^2(\Omega_h)}^2 \right),$$

with some small $\epsilon_\alpha > 0$. For the projection regarding a scalar α_0 , we simply set $P_{(\mathcal{A}_{ad}^0)_h}(\alpha_0) = \max(\min(\alpha_0, \bar{\alpha}_0), \underline{\alpha}_0)$.

5.3. Numerical examples in denoising. We now discuss some weighted TGV numerical examples, with regularization weights produced automatically by Algorithm 2. We are particularly interested in the degree of improvement over the scalar TGV examples. We are also interested in whether the statistics-based upper level objective enforces an automatic choice of regularization parameters that ultimately leads to a reduction of the staircasing effect. Our TGV results are also compared with the bilevel weighted TV method of [37, 39]. In order to have a fair comparison, we use a Huber TV regularization for the corresponding lower level problem, with the latter substituted

Algorithm 2Discretized projected gradient method for the bilevel TGV problem ($\mathbb{P}_{\text{TGV}}^h$)

Input: $f, \underline{\alpha}_0, \bar{\alpha}_0, \underline{\alpha}_1, \bar{\alpha}_1, \bar{\sigma}, \underline{\sigma}, \lambda_0, \lambda_1, \mu, \nu, \gamma_0, \gamma_1, \delta, n_w, \tau_0^0, \tau_1^0, 0 < c < 1, 0 < \theta_- < 1 \leq \theta_+$ **Initialize:** $\alpha_0^0 \in (\mathcal{A}_{ad}^0)_h, \alpha_1^0 \in (\mathcal{A}_{ad}^1)_h$ and set $k = 0$.**repeat**Use the Algorithm 1 to compute the solution $\mathbf{x}^k = (u^k, w^k, q^k, p^k)$ of the lower level problem

$$g_{\text{pd}}(u^k, w^k, q^k, p^k) = 0$$

Solve the adjoint equation (5.11) for (u^*, w^*, q^*, p^*) Compute the derivative of the reduced objective with respect to α_0 and α_1 as in (5.13) and (5.14)

Compute the reduced gradients

$$\nabla_{\alpha_i} \hat{J}(\alpha_0^k, \alpha_1^k) = (I - \Delta)^{-1} \hat{J}'_{\alpha_i}(\alpha_0^k, \alpha_1^k), \quad i = 0, 1$$

Compute the trial points

$$\alpha_i^{k+1} = P_{(\mathcal{A}_{ad}^i)_h}(\alpha_i^k - \tau_i^k \nabla_{\alpha_i} \hat{J}(\alpha_0^k, \alpha_1^k)), \quad i = 0, 1$$

while

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

do (Armijo line search)Set $\tau_0^k := \theta_- \tau_0^k, \tau_1^k := \theta_- \tau_1^k$ and re-compute

$$\alpha_i^{k+1} = P_{(\mathcal{A}_{ad}^i)_h}(\alpha_i^k - \tau_i^k \nabla_{\alpha_i} \hat{J}(\alpha_0^k, \alpha_1^k)), \quad i = 0, 1$$

end whileUpdate $\tau_0^{k+1} = \theta_+ \tau_0^k, \tau_1^{k+1} = \theta_+ \tau_1^k$ and $k := k + 1$ **until** some stopping condition is satisfied

by the TV primal-dual optimality conditions,

$$\begin{aligned} u - \operatorname{div} p - f &= 0 \\ \max(|\nabla u|, \gamma) p - \alpha \nabla u &= 0, \end{aligned}$$

where α is now the spatially dependent regularization parameter for TV. We use the same values for the Huber parameter γ both in bilevel TV and bilevel TGV. The associated test images are depicted in Figure 3 with resolution $n = m = 256$. The first one is the well-known ‘‘Cameraman’’ image which essentially consists of a combination of piecewise constant parts and texture. The next two images, ‘‘Parrot’’ and ‘‘Turtle’’ contain large piecewise affine type areas, thus they are more suitable for the TGV prior. The final image ‘‘hatchling’’ is characterized by highly oscillatory patterns of various kinds, depicting sand in various degrees of focus.

Parameter values: For the lower level primal-dual TGV problem we used $\mu = 0, \nu = 1, \delta = 10^{-5}, \gamma_0 = \gamma_1 = 10^{-3}$, and a mesh size $h = 1$. For the H^1 -projection, we set $\epsilon_\alpha = 10^{-6}$, and we also weighted the discrete Laplacian Δ with 6×10^4 . For the lower and upper bounds of α_0 and α_1 we set here $\underline{\alpha}_0 = 10^{-2}, \bar{\alpha}_0 = 10$ and $\underline{\alpha}_1 = 10^{-4}, \bar{\alpha}_1 = 10$. We also set $\lambda_1 = 10^{-11}$ and when we spatially varied α_0 we also set $\lambda_0 = 10^{-11}$. We used a normalized $n_w \times n_w$ filter for w (i.e., with entries $1/n_w^2$), with $n_w = 7$. The local variance barriers $\underline{\sigma}^2$ and $\bar{\sigma}^2$ were set according to (5.10). For our noisy images we have $\sigma^2 = 10^{-2}$, and thus the corresponding values for $(\underline{\sigma}, \bar{\sigma})$ are $(0.00798, 0.01202)$. For

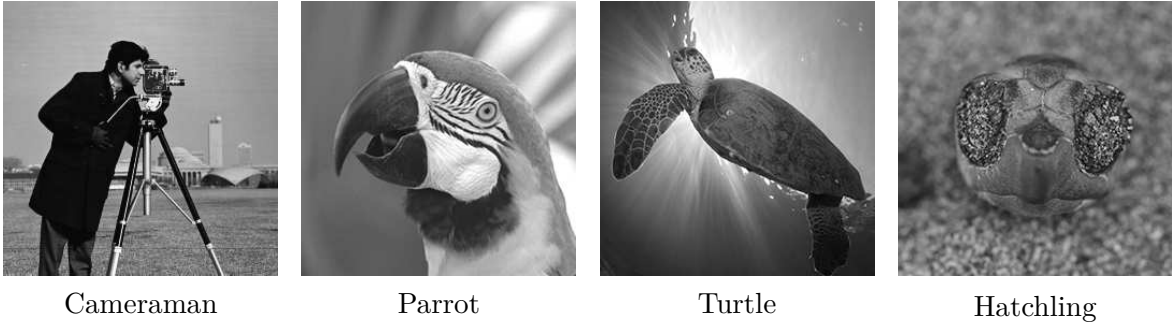


FIGURE 3. Test images, resolution 256×256

the Armijo line search the parameters were $\tau_0^0 = 0.05$, $\tau_1^0 = 100$, $c = 10^{-9}$, $\theta_- = 0.25$, $\theta_+ = 2$. We solved each lower level problem until the residual of each of the optimality conditions (5.1)–(5.4) had Euclidean norm less than 10^{-4} . MATLAB’s backslash was used for the solution of the linear systems.

We note that the initialization of the algorithm needs some attention. As it was done in [39] for 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 falling into the regime, 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 . In that case the derivative of the reduced functional with respect to α_0 will be close to zero, thus making no or little progress with respect to its optimal choice. Indeed this was confirmed after some numerical experimentation. Note that an analogous phenomenon can occur also in the case where α_0 is much smaller than α_1 . In that case it is the effect of α_1 which vanishes. This has been shown theoretically in [51, Proposition 2] for dimension one, but numerical experiments indicate that this phenomenon persists also in higher dimensions. In our examples we used $\alpha_1^0 = 0.25$ and $\alpha_0^0 = 0.2$. Regarding the termination of the projected gradient algorithm, we used a fixed number of iterations $n = 40$. Neither the upper level objective nor the argument changed significantly after running the algorithm for more iterations; see for instance Figure 4. The same holds true for the corresponding PSNR and SSIM values. We also note that a termination criterion as in [39] based on the proximity measures $\|P_{(\mathcal{A}_{ad}^i)_h}(\alpha_i^k - \nabla_{\alpha_i} \hat{J}(\alpha_0^k, \alpha_1^k)) - \alpha_i^k\|_{H^1(\Omega_h)}$, $i = 0, 1$, is also possible here.

We note that due to the line search, the number of times that the lower level problem has to be solved is more than the number of projected gradient iterations. For instance for the four examples Figure 4 the lower level problem had to be solved 58, 57, 57, and 57 times respectively, with typically 8-12 Newton iterations needed per each lower level problem. This resulted in each bilevel problem (40 projected gradient iterations) requiring approximately 45-50 times the CPU time needed to solve one instance of the lower level problem. We note however that in general after 8-9 projected gradient iterations the PSNR and SSIM values of the reconstruction were essentially reaching their top values, requiring about 6-8 times the CPU time needed for the lower level problem. Further improvement with respect to computational times can be achieved using faster descent methods than the one we used here, like for instance [4, 48].

For the first series of examples we keep the parameter α_0 scalar, whose value nevertheless is determined by the bilevel algorithms. We depict the examples in Figure 5. The first row shows the noisy images, while the second contains the bilevel TV results. The third row depicts the best scalar TGV results with respect to SSIM, where we have computed the optimal scalars α_0, α_1 with a manual grid method. The fourth row shows the results of $(\mathbb{P}_{\text{TGV}}^h)$. Detailed sections of all the

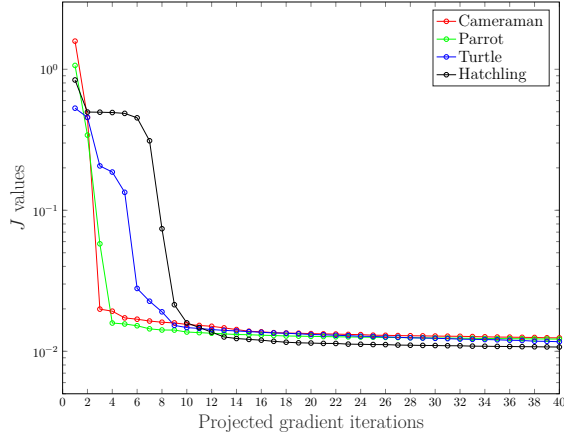


FIGURE 4. Upper level objective values vs projected gradient iterations for the problem $(\mathbb{P}_{\text{TGV}}^h)$ (right) of Figure 5 (scalar α_0 , spatial α_1)

$\sigma^2 = 0.01$	Cameraman	Parrot	Turtle	Hatchling
scalar TV (best PSNR)	27.69, 0.7962	29.25, 0.8354	29.49, 0.8032,	27.75, 0.7701
scalar TV (best SSIM)	27.44, 0.8118	28.93, 0.8508	29.38, 0.8112	27.62, 0.7759
bilevel TV spatial α	27.78, 0.8264	29.18, 0.8550	29.74, 0.8237	27.52, 0.7722
scalar TGV (best PSNR)	27.68, 0.7813	29.47, 0.8504	29.63, 0.8223	28.05 , 0.7950
scalar TGV (best SSIM)	27.44, 0.8135	29.09, 0.8611	29.50, 0.8303	27.88, 0.8066
bilevel TGV scalar α_0 , scalar α_1	27.63, 0.8063	29.41, 0.8560	29.63, 0.8200	27.70, 0.7851
bilevel TGV scalar α_0 , spatial α_1	27.67, 0.8135	29.56, 0.8629	29.81, 0.8328	27.98, 0.8012
bilevel TGV spatial α_0 , spatial α_1	27.68, 0.8151	29.58, 0.8643	29.76, 0.8328	27.97, 0.8009

TABLE 1. PSNR and SSIM comparisons for the images of Figure 5. Every cell contains the corresponding PSNR and SSIM value.

images of Figure 5 are highlighted in Figure 6. The weight functions α_1 for the bilevel TV and the bilevel TGV algorithms are shown in Figure 7. In Table 1 we report all PSNR and SSIM values of the best scalar methods (scalar TV, scalar TGV) with respect to both quality measures, the corresponding values of the bilevel TV and TGV algorithms, as well as the ones that correspond to solving the bilevel TGV with the statistics-based upper level objective (4.2) but using scalar parameters only (third row from the end in Table 1). For the latter case, we used no additional regularization for the weights in the upper level objective. We also report the PSNR and SSIM values of the computed results when both α_0 , α_1 are spatially varying (last row), which we discuss later on in this section. We next comment on the results for each image.

Cameraman: Here both the best PSNR and SSIM are obtained by the bilevel TV algorithm. This is probably not surprising due to the piecewise constant nature of this image. However, the bilevel TGV algorithm improves upon its scalar version with respect to both measures.

Parrot: Here the best results with respect to both PSNR and SSIM are achieved by the bilevel TGV algorithm, $(\mathbb{P}_{\text{TGV}}^h)$. There is significant improvement over all TV methods, which is due to the parameters being chosen in a way such that the staircasing effect diminishes. Furthermore, we observe improvement over the scalar TGV result especially around the parrot’s eye, where the

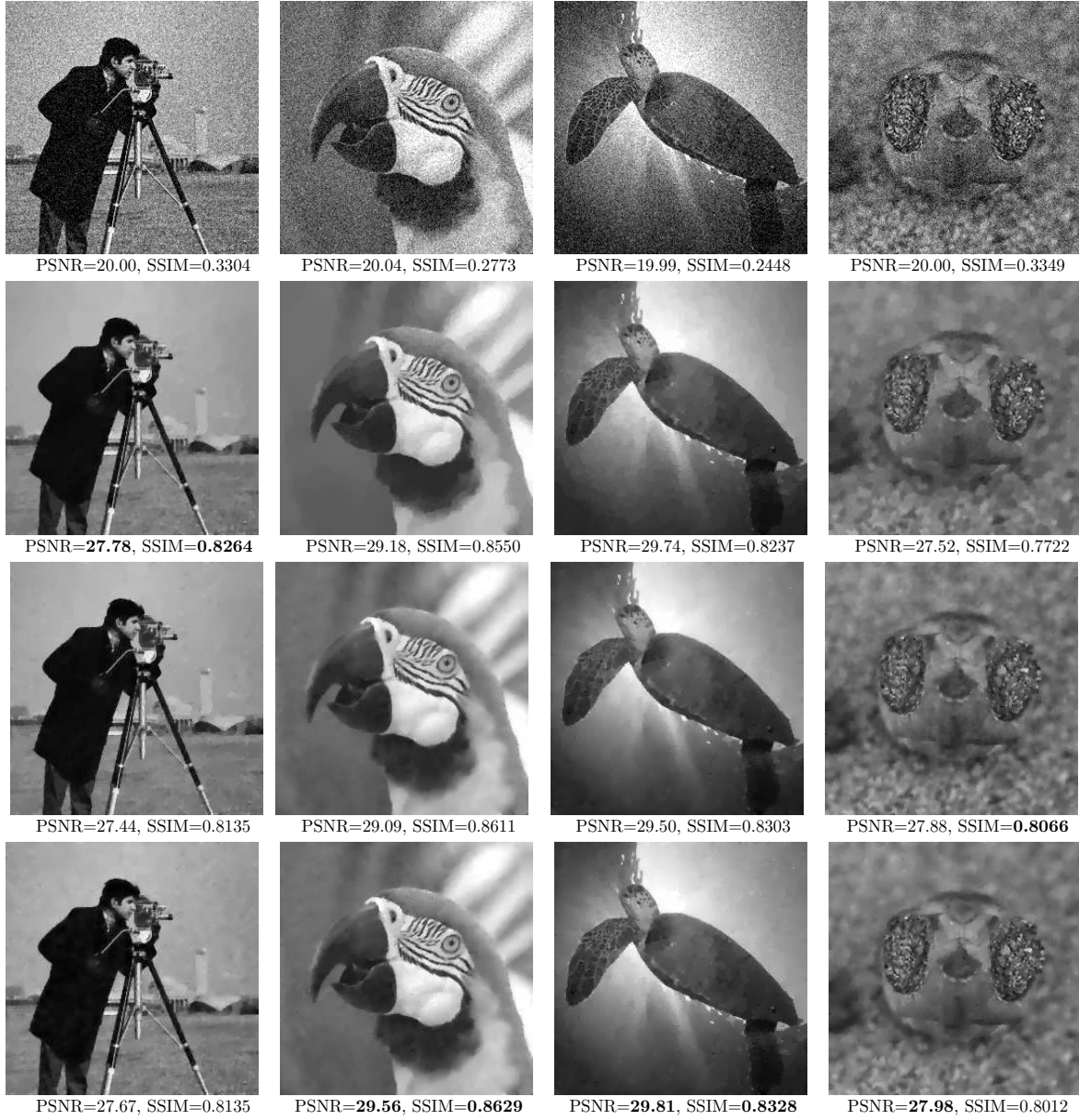


FIGURE 5. First row: noisy images. Second row: bilevel TV. Third row: Best scalar TGV (SSIM). Fourth row: bilevel TGV

weights α_1 drop significantly; see the second column of Figure 7.

Turtle: We get analogous results here as well, with the bilevel TGV ($\mathbb{P}_{\text{TGV}}^h$) producing the best results both with respect to PSNR and SSIM. There is a significant reduction of the staircasing effect, while the weight α_1 drops in the detailed areas of the image (head and flipper of the turtle).

Hatchling: In this image, the best PSNR and best SSIM are achieved by the scalar TGV when its parameters are manually optimized with respect to each measure (using the ground truth). However, the (ground truth-free) bilevel TGV achieves a better SSIM and PSNR value compared to these two results respectively, achieving a better balance with respect to these measures. Similarly

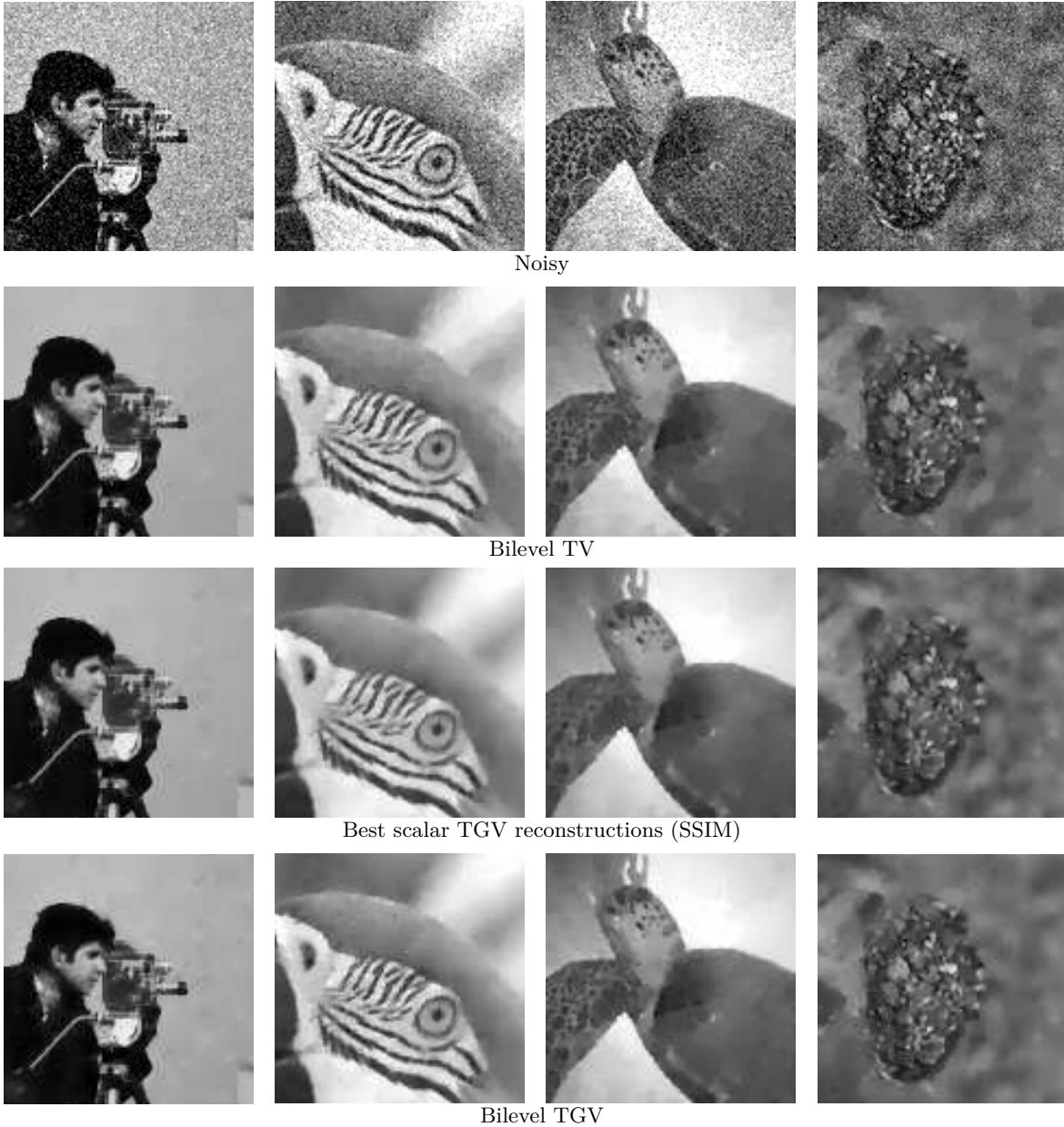


FIGURE 6. Details of the reconstructions shown in Figure 5

the scalar TV results are generally better than the ones of bilevel TV. We attribute this to the fact that the natural oscillatory features of the image are interpreted as noise by the upper level objective. Nevertheless, all the bilevel methods are able to locate and preserve better the eyes area, i.e., sand in focus, with the weight α_1 dropping there significantly.

We remark that in all four examples, the ground truth-free bilevel TGV with at least one spatially varying parameter always produces better results with respect to both PSNR and SSIM, than the corresponding ground truth-free bilevel TGV with both parameters being scalar, compare the third last with the second last row of Table 1.

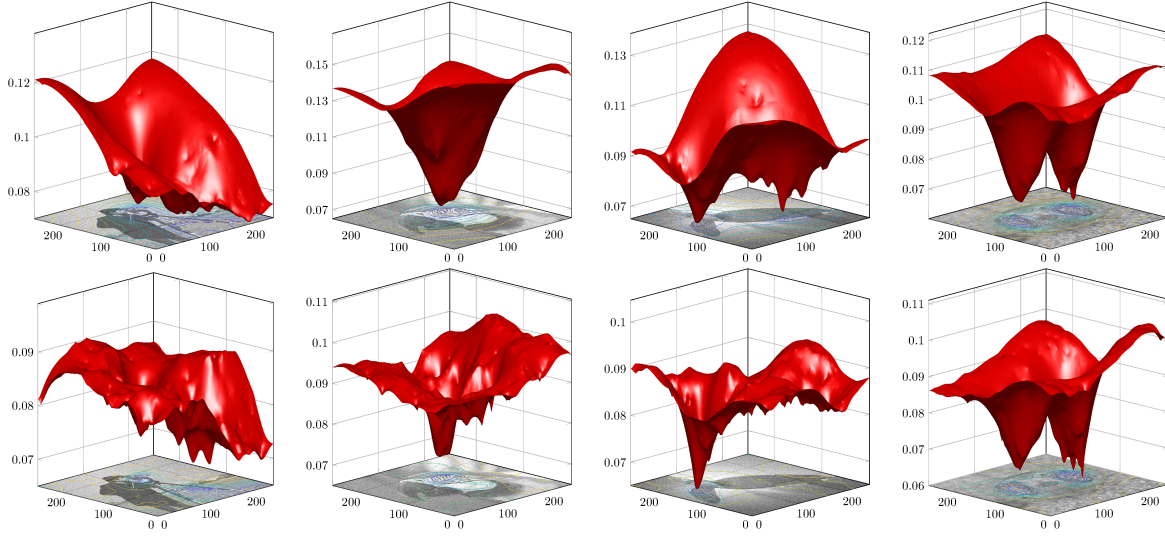


FIGURE 7. First row: the computed regularization functions α for bilevel TV.
Second row: the computed regularization functions α_1 for bilevel TGV

Finally, we would like to experiment with the case where also the weight α_0 varies spatially. We note that by spatially varying both TGV parameters, the reduced problem becomes highly non-convex with many combinations of these parameters leading to similar values for the upper level objective. In order to deal with this, we solve a slightly different problem by performing one step of alternating optimization. Here we optimize only with respect to a spatially varying α_0 having fixed the spatial weight α_1 as it has been computed by the previous experiments, see the last row of Figure 7. According to our numerical experiments, this strategy produces satisfactory results. As initialization for α_0 , we set it constant, equal to 5.

In Figure 8 we depict the computed spatially varying parameters α_0 as well as the corresponding PSNR and SSIM values. Observe that the shape of α_0 is different to the one of α_1 , compare the last row of Figure 7 to the second row of Figure 8. This implies that a non-constant ratio of α_0/α_1 is preferred throughout the image domain. Secondly, by spatially varying α_0 we only get a slight improvement with respect to PSNR and SSIM in the first two images, compare the second last and the last row of Table 1. However, it is interesting to observe the spatial adaptation of α_0 with respect to piecewise constant versus piecewise smooth areas. The values of α_0 are high in large piecewise constant areas, like the background of cameraman, the left area of the parrot image, as well as the top-right and the bottom-left corner of the turtle image. This is not so surprising as large values of α_0 imply a large ratio α_0/α_1 and a promotion of TV like behaviour in those areas. We can observe this in more detail at the parrot image, see last row of Figure 8. On the contrary, the values of α_0 are kept small in piecewise smooth areas like the right part of the parrot image and the sun rays around the turtle's body. This results in low ratio α_0/α_1 and thus to a more TGV like behaviour, reducing the staircasing effect. This is another indication of the fact that by minimizing the statistics-based upper level objective one is able not only to better preserve detailed areas but also to finely adjust the TGV parameters such that the staircasing is reduced.

Lastly, we would like to see the degree of improvement of bilevel TGV with spatial parameters over the other ground truth-free approaches, i.e. using the statistics-based upper level objective, over a larger set of images. We thus used the 24 images of the Kodak dataset <http://r0k.us/graphics/kodak/>, and selected two middle square parts from each photograph, resized to 256×256

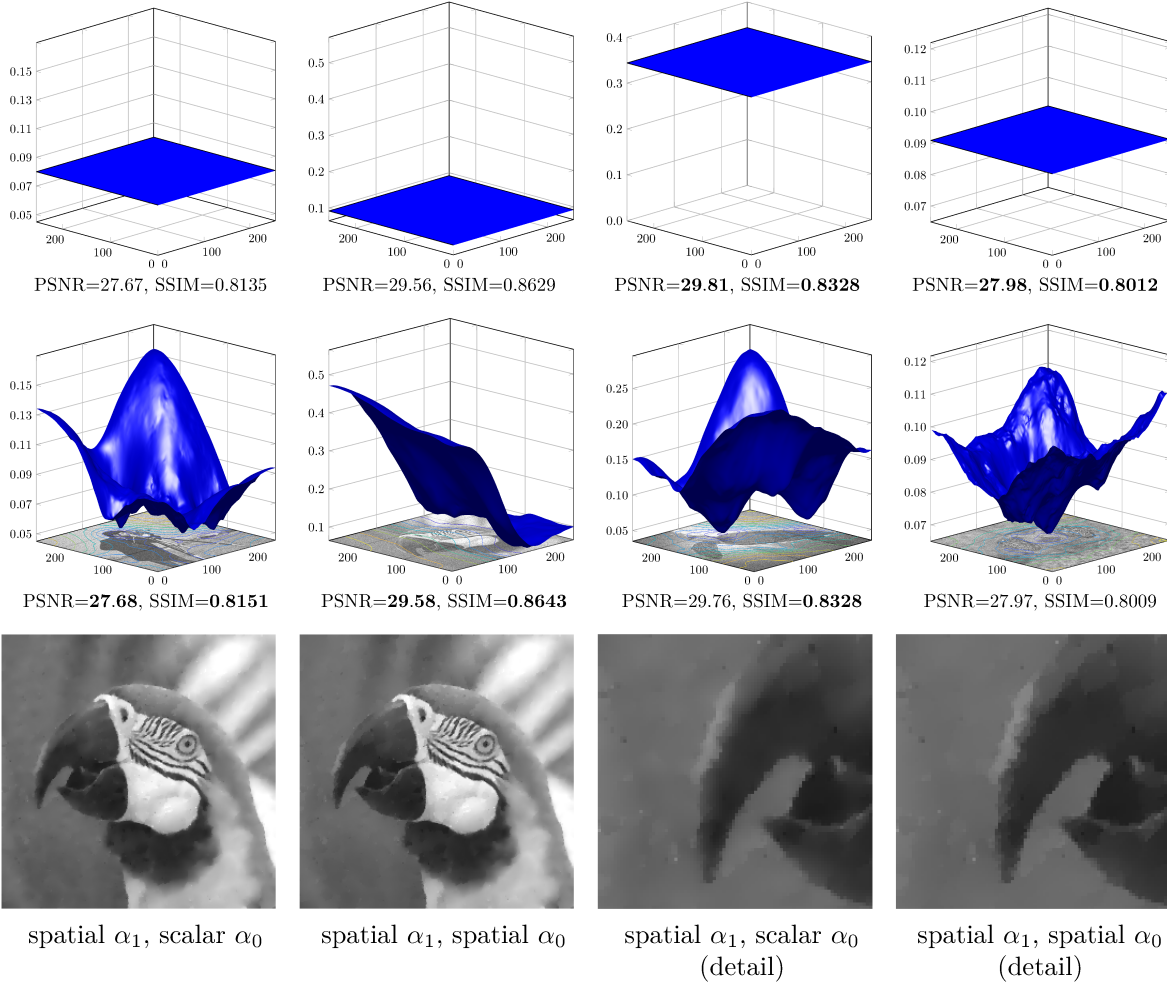


FIGURE 8. Experiments with optimizing over a spatially varying α_0 . Top row: the automatically computed scalar parameters α_0 , that correspond to the images of the last row of Figure 5. Middle row: the automatically computed spatially varying parameters α_0 , where α_1 has been kept fixed (last row of Figure 7). The weight α_0 is adapted to piecewise constant parts having there large values and hence promoting TV like behaviour, see for instance the parrot image at the last row. On the contrary α_0 has low values in piecewise smooth parts promoting a TGV like behaviour reducing the staircasing.

pixels, see left part of Figure 9. We added two more photographs (see bottom-right part in the depicted collection) to reach 50 photographs in total. In each photograph, we run the bilevel TV and bilevel TGV experiments after adding the same type of Gaussian noise (with only α_1 being spatially varying), as well as the bilevel TGV with both parameters to be scalar and we compared the corresponding PSNR differences, $\text{PSNR}_{\text{spatial TGV}} - \text{PSNR}_{\text{spatial TV}}$, $\text{PSNR}_{\text{spatial TGV}} - \text{PSNR}_{\text{scalar TGV}}$, as well as the corresponding SSIM ones, for every one of the 50 images. The PSNR improvement for the bilevel TGV over bilevel TV (both spatial parameters) was 0.05 ± 0.14 (mean $\pm$ standard deviation) while the improvement with respect to SSIM was more moderate, 0.003 ± 0.008 . In Figure 9 (top) we also show the corresponding histograms of this difference. The fact that for

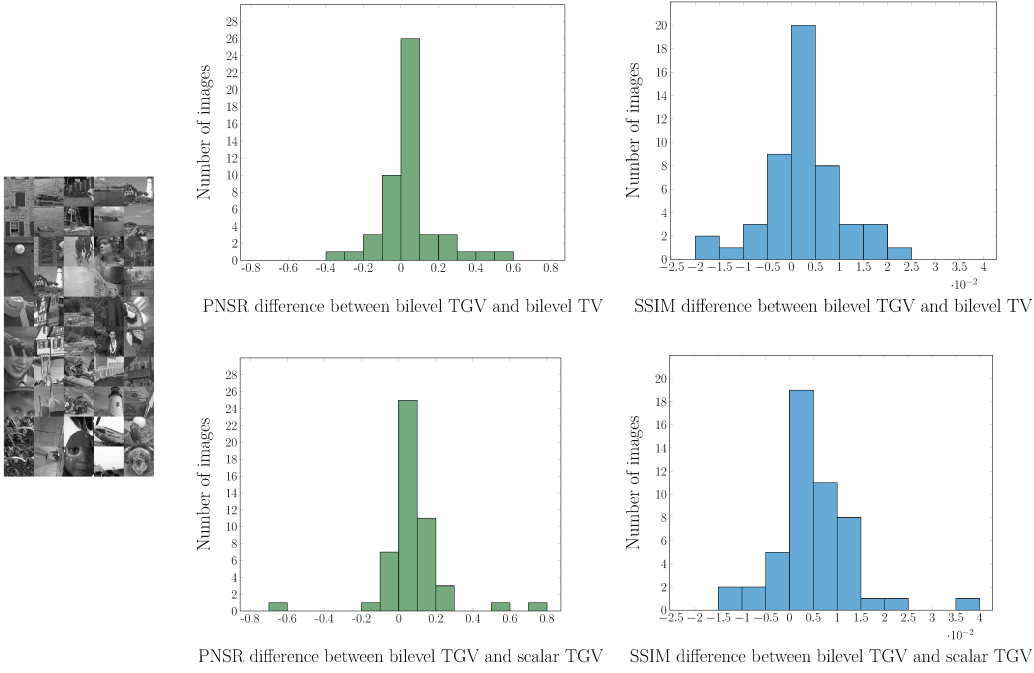


FIGURE 9. Histograms of the PSNR and SSIM differences between the bilevel spatial TGV and bilevel spatial TV (top) and between the bilevel spatial TGV and bilevel scalar TGV (bottom), for reconstructions of noisy versions of the 50 images on the left. In all bilevel algorithms the ground truth-free statistics-based upper level objective was used.

some images the bilevel TV algorithm produces images of higher PSNR and/or SSIM, as it was the case for the cameraman, is perhaps not so surprising as the same has been observed in similar tests in [27] in the case of scalar parameters only and for a bilevel scheme that maximizes the PSNR of the reconstructed image, using a ground truth-based upper level objective. We stress that for large ratio α_0/α_1 TGV is equivalent to TV only up an affine correction [51] and for images that have a piecewise constant structure, setting α_0 in TGV to be large enough might not be enough to achieve the same reconstruction as TV. On the other hand the improvement of bilevel TGV with spatially varying α_1 versus its scalar analogue was 0.08 ± 0.17 for the PSNR and 0.005 ± 0.008 for the SSIM, see bottom histograms of Figure 9. We note that the last two methods require a similar computational effort since the additional H^1 -projection for the spatially varying weight is typically quite fast and does not significantly contribute to the overall computational time.

6. CONCLUSION

In this work we have adapted the bilevel optimization framework of [37, 39] for automatically computing spatially dependent regularization parameters for the TGV regularizer. We established a rigorous dualization framework for the lower level TGV minimization problem that formed the basis for its algorithmic treatment via a Newton method. We showed that the bilevel optimization framework with the statistics/localized residual based upper level objective is able to automatically produce spatially varying parameters that not only adapt to the level of detail in the image but also reduce the staircasing effect.

Future continuation of this work includes adaptation of the bilevel TGV framework for advanced inverse problems tasks, i.e., Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET) reconstruction as well as in multimodal medical imaging problems where structural TV based regularizers (edge aligning) have been suggested. That will also require devising a bilevel scheme where the invertibility assumption on T is dropped, since the linear operators involved in these inverse problems do not satisfy this assumption. Adaptation of the framework for different noise distributions e.g. Poisson, Salt & Pepper as well as combination of those [20, 21], should also be investigated. Finally, a fine structural analysis of the weighted TGV regularized solutions in the spirit of [34, 43] would be also of interest.

Acknowledgements. This work is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – The Berlin Mathematics Research Center MATH+ (EXC-2046/1, project ID: 390685689). It was further supported by the MATHEON Research Center project CH12 funded by the Einstein Center for Mathematics (EC-Math) Berlin. M.H and K.P. acknowledge support of Institut Henri Poincaré (UMS 839 CNRS-Sorbonne Université), and LabEx CARMIN (ANR-10-LABX-59-01). C.N.R. was supported by NSF grant DMS-2012391. H.S. acknowledges the financial support of Alexander von Humboldt Foundation.

APPENDIX A.

We provide here the proof of Proposition 2.2:

Proof of Proposition 2.2. The proof follows [12, Proposition 3.3]. Denote by C_α , K_α the following convex sets

$$(A.1) \quad C_\alpha = \left\{ \operatorname{div}^2 \phi : \phi \in C_c^\infty(\Omega, \mathcal{S}^{d \times d}), |\phi(x)| \leq \alpha_0(x), |\operatorname{div} \phi(x)| \leq \alpha_1(x), \text{ for all } x \in \Omega \right\},$$

$$(A.2) \quad K_\alpha = \left\{ \operatorname{div}^2 p : p \in H_0(\operatorname{div}^2; \Omega), |p(x)| \leq \alpha_0(x), |\operatorname{div} p(x)| \leq \alpha_1(x), \text{ for a.e. } x \in \Omega \right\}.$$

It suffices to show that

$$(A.3) \quad \overline{C_\alpha}^{L^2(\Omega)} = K_\alpha.$$

We first show that K_α is closed in $L^2(\Omega)$. Let $g \in L^2(\Omega)$ and assume that there exists $(p_n)_{n \in \mathbb{N}} \subset H_0(\operatorname{div}^2; \Omega)$ where every p_n satisfies the convex constraints and $\operatorname{div}^2 p_n \rightarrow g$ in $L^2(\Omega)$. By boundedness of α_0, α_1 we have that there exist $h_0 \in L^2(\Omega, \mathcal{S}^{d \times d})$, $h_1 \in L^2(\Omega, \mathbb{R}^d)$ and a subsequence of $(p_{n_k})_{k \in \mathbb{N}}$ such that

$$p_{n_k} \rightharpoonup h_0 \quad \text{and} \quad \operatorname{div} p_{n_k} \rightharpoonup h_1,$$

in $L^2(\Omega)$ and $L^2(\Omega, \mathbb{R}^d)$ respectively. Using that, we have for every $\phi \in C_c^\infty(\Omega, \mathbb{R}^d)$

$$(A.4) \quad \int_\Omega \nabla \phi \cdot h_0 \, dx = \lim_{k \rightarrow \infty} \int_\Omega \nabla \phi \cdot p_{n_k} \, dx = - \lim_{k \rightarrow \infty} \int_\Omega \phi \cdot \operatorname{div} p_{n_k} \, dx = - \int_\Omega \phi \cdot h_1 \, dx,$$

thus $h_1 = \operatorname{div} h_0$. Similarly we derive that $g = \operatorname{div} h_1 = \operatorname{div}^2 h_0$ and hence $h_0 \in H(\operatorname{div}^2; \Omega)$. Finally note that the set

$$\left\{ (h, \operatorname{div} h, \operatorname{div}^2 h) : h \in H_0(\operatorname{div}^2; \Omega), |h(x)| \leq \alpha_0(x), |\operatorname{div} h(x)| \leq \alpha_1(x), \text{ for a.e. } x \in \Omega \right\},$$

is a norm-closed and convex subset of $L^2(\Omega, (\mathcal{S}^{d \times d} \times \mathbb{R}^d \times \mathbb{R}))$ and hence weakly closed. Since $(p_{n_k}, \operatorname{div} p_{n_k}, \operatorname{div}^2 p_{n_k})_{k \in \mathbb{N}}$ belongs to that set, converging weakly to $(h_0, \operatorname{div} h_0, \operatorname{div}^2 h_0)$ we get that the latter also belongs there. Thus, K_α is closed in $L^2(\Omega)$ and since $C_\alpha \subset K_\alpha$, we get $\overline{C_\alpha}^{L^2(\Omega)} \subset K_\alpha$.

It remains to show the other direction, i.e., $K_\alpha \subset \overline{C_\alpha}^{L^2(\Omega)}$. Towards that, note first that the functional $\text{TGV}_\alpha^2(\Omega) : L^2(\Omega) \rightarrow \overline{\mathbb{R}}$, can also be written as

$$\text{TGV}_\alpha^2(u) = \mathcal{I}_{C_\alpha}^*(u).$$

Using the convexity of C_α one gets

$$\text{TGV}_\alpha^{2*}(v) = \mathcal{I}_{C_\alpha}^{**}(v) = \mathcal{I}_{\overline{C_\alpha}^{L^2(\Omega)}}(v).$$

Secondly, note that due to the lower bounds on α_0, α_1 , for $u \in L^2(\Omega)$, we have that $\text{TGV}_\alpha^2(u) < \infty$ if and only if $u \in \text{BV}(\Omega)$. Indeed this holds from the equivalence of the (scalar) $\|\cdot\|_{\text{BGV}}$ with $\|\cdot\|_{\text{BV}(\Omega)}$ and from the estimate

$$\text{TGV}_{\underline{\alpha}, \underline{\alpha}}^2(u) \leq \text{TGV}_\alpha^2(u) \leq \|\alpha_1\|_\infty \text{TV}(u),$$

for every $u \in L^2(\Omega)$. This means that if for $\text{div}^2 p \in K_\alpha$ it holds

$$(A.5) \quad \int_\Omega u \text{div}^2 p \, dx \leq \text{TGV}_\alpha^2(u), \quad \text{for all } u \in \text{BV}(\Omega) \cap L^2(\Omega),$$

then in fact the inequality (A.5) will hold for every $u \in L^2(\Omega)$ and thus $\text{TGV}_\alpha^{2*}(\text{div}^2 p) = 0$ which implies $\text{div}^2 p \in \overline{C_\alpha}^{L^2(\Omega)}$. Thus in order to finish the proof it suffices to show (A.5) for every $\text{div}^2 p \in K_\alpha$. In view of Proposition 2.1 it suffices to show

$$(A.6) \quad \int_\Omega u \text{div}^2 p \, dx \leq \min_{w \in \text{BD}(\Omega)} \int_\Omega \alpha_1 d|Du - w| + \int_\Omega \alpha_0 d|\mathcal{E}w|$$

for all $u \in \text{BV}(\Omega) \cap L^2(\Omega)$. The first step towards that is to show that for every $w \in \text{BD}(\Omega)$ and for every $p \in H_0(\text{div}^2; \Omega)$ with $|p(x)| \leq \alpha_0(x)$ and $|\text{div} p(x)| \leq \alpha_1(x)$ for a.e. $x \in \Omega$, it holds

$$(A.7) \quad \int_\Omega w \cdot \text{div} p \, dx \leq \int_\Omega \alpha_0 d|\mathcal{E}w|.$$

Indeed, note first that from (2.3), we get for every $\phi \in C^\infty(\overline{\Omega}, \mathbb{R}^d)$

$$(A.8) \quad \left| \int_\Omega \phi \cdot \text{div} p \, dx \right| = \left| \int_\Omega p \cdot E\phi \, dx \right| \leq \int_\Omega |p| |E\phi| \, dx \leq \int_\Omega \alpha_0 d|E\phi|.$$

Recall now that every $w \in \text{BD}(\Omega)$ can be strictly approximated by a sequence $(\phi_n)_{n \in \mathbb{N}} \subset C^\infty(\overline{\Omega}, \mathbb{R}^d)$, that is $\phi_n \rightarrow w$ in $L^1(\Omega, \mathbb{R}^d)$ and $|E\phi|(\Omega) \rightarrow |\mathcal{E}w|(\Omega)$, see [12, Proposition 2.10]. Furthermore, using that, along with Reshetnyak's continuity theorem [2, Theorem 2.39] we also get that

$$\int_\Omega \alpha_0 d|E\phi_n| \rightarrow \int_\Omega \alpha_0 d|\mathcal{E}w|, \quad \text{as } n \rightarrow \infty.$$

Using the above and the fact that $\text{div} p \in L^\infty(\Omega, \mathbb{R}^d)$, by taking limits in (A.8) we obtain (A.7). Finally in order to obtain (A.5) let again $p \in H_0(\text{div}^2; \Omega)$ with $|p(x)| \leq \alpha_0(x)$ and $|\text{div} p(x)| \leq \alpha_1(x)$ for a.e. $x \in \Omega$, and let $\phi \in C^\infty(\overline{\Omega}, \mathbb{R})$. Then by using (2.4) and (A.7), we have for every $w \in \text{BD}(\Omega)$

$$\begin{aligned} \int_\Omega \phi \text{div}^2 p \, dx &\leq \left| \int_\Omega \nabla \phi \cdot \text{div} p \, dx \right| = \left| \int_\Omega (\nabla \phi - w) \cdot \text{div} p \, dx + \int_\Omega w \text{div} p \, dx \right| \\ &\leq \int_\Omega |\nabla \phi - w| |\text{div} p| \, dx + \int_\Omega \alpha_0 d|\mathcal{E}w| \leq \int_\Omega \alpha_1 |\nabla \phi - w| \, dx + \int_\Omega \alpha_0 d|\mathcal{E}w|. \end{aligned}$$

Similarly as before given $u \in \text{BV}(\Omega) \cap L^2(\Omega)$ and $w \in \text{BD}(\Omega)$, there exists a sequence $(\phi_n)_{n \in \mathbb{N}} \subset C^\infty(\overline{\Omega}, \mathbb{R})$ such that $\phi_n \rightarrow u$ in $L^2(\Omega)$ and $|\nabla \phi_n - w|(\Omega) \rightarrow |Du - w|(\Omega)$. This follows from a modification of the proof of [12, Proposition 2.10] where one takes advantage of the L^2 integrability

of u . Using again the Reshetnyak's continuity theorem and taking limits we get that for every $w \in \text{BD}(\Omega)$

$$\int_{\Omega} u \operatorname{div}^2 p \, dx \leq \int_{\Omega} \alpha_1 d|Du - w| + \int_{\Omega} \alpha_0 d|\mathcal{E}w|.$$

By taking the minimum over $w \in \text{BD}(\Omega)$, we obtain (A.5). □

REFERENCES

1. R. A. Adams and J.J.F. Fournier, *Sobolev spaces*, 2nd ed., Academic Press, 2003.
2. L. Ambrosio, N. Fusco, and D. Pallara, *Functions of bounded variation and free discontinuity problems*, Oxford University Press, USA, 2000.
3. H. Attouch and H. Brezis, *Duality for the sum of convex functions in general Banach spaces*, North-Holland Mathematical Library **34** (1986), 125–133.
4. J. Barzilai and J.M. Borwein, *Two-point step size gradient methods*, IMA Journal of Numerical Analysis **8** (1988), no. 1, 141–148, <https://doi.org/10.1093/imanum/8.1.141>.
5. M. Benning, C. Brune, M. Burger, and J. Müller, *Higher-order TV methods – Enhancement via Bregman iteration*, Journal of Scientific Computing **54** (2013), no. 2-3, 269–310, <http://dx.doi.org/10.1007/s10915-012-9650-3>.
6. M. Bergounioux, E. Papoutsellis, S. Stute, and C. Tauber, *Infimal convolution spatiotemporal PET reconstruction using total variation based priors*, HAL preprint (2018), <https://hal.archives-ouvertes.fr/hal-01694064>.
7. J.M. Borwein and J.D. Vanderwerff, *Convex Functions*, Cambridge University Press, 2010.
8. 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>.
9. K. Bredies and M. Holler, *Artifact-free JPEG decomposition with total generalized variation*, VISAP 2012: Proceedings of the International Conference on Computer Vision and Applications, 2012.
10. ———, *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.
11. ———, *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>.
12. ———, *A TGV-based framework for variational image decomposition, zooming, and reconstruction. Part I: Analytics*, SIAM Journal on Imaging Sciences **8** (2015), no. 4, 2814–2850, <https://doi.org/10.1137/15M1023865>.
13. ———, *A TGV-based framework for variational image decomposition, zooming, and reconstruction. Part II: Numerics*, SIAM Journal on Imaging Sciences **8** (2015), no. 4, 2851–2886, <https://doi.org/10.1137/15M1023877>.
14. K. Bredies, M. Holler, M. Storath, and A. Weinmann, *Total generalized variation for manifold-valued data*, SIAM Journal on Imaging Sciences **11** (2018), no. 3, 1785–1848, <https://doi.org/10.1137/17M1147597>.
15. 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>.
16. K. Bredies, K. Kunisch, and T. Valkonen, *Properties of L^1 -TGV²: The one-dimensional case*, Journal of Mathematical Analysis and Applications **398** (2013), no. 1, 438–454, <http://dx.doi.org/10.1016/j.jmaa.2012.08.053>.
17. 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.
18. 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>.
19. 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>.
20. L. Calatroni, J.C. De Los Reyes, and C.B. 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>.
21. L. Calatroni and K. Papafitsoros, *Analysis and automatic parameter selection of a variational model for mixed gaussian and salt-and-pepper noise removal*, Inverse Problems **35** (2019), no. 11, 114001, <https://doi.org/10.1088/1361-6420/ab291a>.
22. 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>.

23. 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>.
24. 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>.
25. L. 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), 074005, <https://doi.org/10.1088/1361-6420/33/7/074005>.
26. 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>.
27. 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>.
28. F. Demengel and R. Temam, *Convex functions of a measure and applications*, Indiana University Mathematics Journal **33** (1984), 673–709.
29. I. Ekeland and R. Temam, *Convex analysis and variational problems*, Classics in Applied Mathematics, Society for Industrial and Applied Mathematics, 1999.
30. V. Girault and P.A. Raviart, *Finite Element Method for Navier-Stokes Equation*, Springer, 1986.
31. 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>.
32. 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>.
33. M. Hintermüller and K. Papafitsoros, *Generating structured nonsmooth priors and associated primal-dual methods*, Processing, Analyzing and Learning of Images, Shapes, and Forms: Part 2 (Ron Kimmel and Xue-Cheng Tai, eds.), Handbook of Numerical Analysis, vol. 20, 2019, <https://doi.org/10.1016/bs.hna.2019.08.001>, pp. 437–502.
34. 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>.
35. ———, *Variable step mollifiers and applications*, Integral equations and operator theory **92** (2020), no. 53, <https://doi.org/10.1007/s00020-020-02608-2>.
36. 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>.
37. 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>.
38. 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>.
39. 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>.
40. M. Hintermüller and G. Stadler, *An infeasible primal-dual algorithm for total bounded variation-based inf-convolution-type image restoration*, SIAM Journal on Scientific Computing **28** (2006), no. 1, 1–23, <http://dx.doi.org/10.1137/040613263>.
41. 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>.
42. R. Huber, G. Haberfehlner, M. Holler, G. Kothleitner, and K. Bredies, *Total generalized variation regularization for multi-modal electron tomography*, Nanoscale **11** (2019), 5617–5632, <http://dx.doi.org/10.1039/C8NR09058K>.
43. K. Jalalzai, *Discontinuities of the minimizers of the weighted or anisotropic total variation for image reconstruction*, arXiv preprint 1402.0026 (2014), <http://arxiv.org/abs/1402.0026>.
44. ———, *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>.

45. 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>.
46. F. Knoll, M. Holler, T. Koesters, R. Otazo, K. Bredies, and D. K. Sodickson, *Joint MR-PET reconstruction using a multi-channel image regularizer*, IEEE Transactions on Medical Imaging **36** (2017), no. 1, 1–16, <http://dx.doi.org/10.1109/TMI.2016.2564989>.
47. 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>.
48. Y.E. Nesterov, *A method for solving the convex programming problem with convergence rate $O(1/k^2)$* , **27** (1983), 372–367.
49. 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>.
50. 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>.
51. K. Papafitsoros and T. Valkonen, *Asymptotic behaviour of total generalised variation*, Scale Space and Variational Methods in Computer Vision: 5th International Conference, SSVN 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.
52. 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>.
53. 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>.
54. L.I. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Physica D: Nonlinear Phenomena **60** (1992), no. 1–4, 259–268, [http://dx.doi.org/10.1016/0167-2789\(92\)90242-F](http://dx.doi.org/10.1016/0167-2789(92)90242-F).
55. M. Schloegl, M. Holler, A. Schwarzl, K. Bredies, and R. Stollberger, *Infimal convolution of total generalized variation functionals for dynamic MRI*, Magnetic Resonance in Medicine **78** (2017), no. 1, 142–155, <http://dx.doi.org/10.1002/mrm.26352>.
56. R. Temam, *Mathematical problems in plasticity*, vol. 15, Gauthier-Villars Paris, 1985.
57. R. Temam and G. Strang, *Functions of bounded deformation*, Archive for Rational Mechanics and Analysis **75** (1980), no. 1, 7–21, <http://dx.doi.org/10.1007/BF00284617>.
58. 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>.
59. 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>.
60. 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>.