J Math Imaging Vis (2017) 59:515–533
DOI 10.1007/s10851-017-0736-2

Optimal Selection of the Regularization Function in a Weighted Total Variation Model. Part II: Algorithm, Its Analysis and Numerical Tests

Michael Hintermüller[1] · **Carlos N. Rautenberg**[1] · **Tao Wu**[1] · **Andreas Langer**[2]

Received: 26 February 2016 / Accepted: 2 May 2017 / Published online: 9 June 2017

Abstract Based on the weighted total variation model and its analysis pursued in Hintermüller and Rautenberg 2016, in this paper a continuous, i.e., infinite dimensional, projected gradient algorithm and its convergence analysis are presented. The method computes a stationary point of a regularized bilevel optimization problem for simultaneously recovering the image as well as determining a spatially distributed regularization weight. Further, its numerical realization is discussed and results obtained for image denoising and deblurring as well as Fourier and wavelet inpainting are reported on.

Keywords Image restoration · Weighted total variation regularization · Spatially distributed regularization weight · Fenchel predual · Bilevel optimization · Variance corridor · Projected gradient method · Convergence analysis

Mathematics Subject Classification 94A08 · 68U10 · 49K20 · 49K30 · 49K40 · 49M37 · 65K15

This research was carried out in the framework of MATHEON supported by the Einstein Foundation Berlin within the ECMath projects OT1, SE5 and SE15 as well as by the DFG under Grant No. HI 1466/7-1 "Free Boundary Problems and Level Set Methods".

A. Langer is listed as a co-author as he was involved in early numerical tests prior to writing this paper. In particular, he found the discretization of the $\nabla\circ$div-operator of [15] suitable for the present context, performed numerical tests concerning the choice of the upper level objective and the initial choice of $\alpha = 2.5 \times 10^{-3}$ when solving the bilevel problem. He also provided the original source images used in Figs. 6 and 8.

✉ Michael Hintermüller
hint@math.hu-berlin.de

Carlos N. Rautenberg
carlos.rautenberg@math.hu-berlin.de

Tao Wu
wutao@math.hu-berlin.de

[1] Department of Mathematics, Humboldt-University of Berlin, Unter den Linden 6, 10099 Berlin, Germany

[2] Department of Mathematics, University of Stuttgart, Pfaffenwaldring 57/8.345, 70569 Stuttgart, Germany

1 Introduction

The following novel duality-based bilevel optimization framework is proposed in [31] for the development of a monolithic variational, i.e., optimization approach to simultaneously recovering an image $u : \Omega \to \mathbb{R}$ and a spatially varying regularization weight $\alpha : \Omega \to \mathbb{R}_+$ from measurement data $f \in L^2(\Omega)$:

$$\text{minimize} \quad J(\mathbf{p}, \alpha) \quad \text{over } (\mathbf{p}, \alpha) \in H_0(\text{div}) \times \mathcal{A}_{\text{ad}} \qquad (\mathbb{P})$$
$$\text{subject to (s.t.) } \mathbf{p} \text{ solves } \mathrm{D}(\alpha),$$

where $J(\cdot,\cdot)$ is defined in ($\tilde{\mathbb{P}}$) below along with the motivation for its choice. Specifically, it contains a term involving a localized variance estimator and a H^1-regularization term. We define $\mathbf{K}(\alpha)$ as

$$\mathbf{K}(\alpha) := \{\mathbf{q} \in H_0(\text{div}) \ : \ |\mathbf{q}(x)|_\infty \le \alpha(x) \text{ f.a.a. } x \in \Omega\},$$

and the lower level problem $\mathrm{D}(\alpha)$ is given by

$$\text{minimize} \quad J_D(\mathbf{p}) := \frac{1}{2}|\operatorname{div}\mathbf{p} + K^* f|_B^2 \qquad (\mathrm{D}(\alpha))$$
$$\text{s.t. } \mathbf{p} \in \mathbf{K}(\alpha),$$

with $\operatorname{div}(\cdot) = \sum_i \frac{\partial(\cdot)_i}{\partial x_i}$ the divergence operator, and K a linear and continuous transfer operator from $L^2(\Omega)$ to $L^2(\Omega)$,

i.e., $K \in \mathcal{L}(L^2(\Omega))$, and K^* standing for its adjoint. Specific examples for K are the identity (denoising), convolution (deblurring), and Fourier or wavelet transforms. The image domain $\Omega \subset \mathbb{R}^\ell$, where $\ell = 1$ or 2 (unless stated differently), is a bounded connected open set with Lipschitz boundary $\partial\Omega$. The given datum satisfies $f = Ku_{\text{true}} + \eta \in L^2(\Omega)$, where u_{true} denotes the original image and η additive "noise," which has zero mean on Ω and satisfies $|\eta|^2_{L^2(\Omega)} \leq \sigma^2|\Omega|$ with $\sigma^2 > 0$ and $|\cdot|$ the (Lebesgue) measure of Ω. Further, $|w|^2_B := (w, B^{-1}w)_{L^2(\Omega)}$ with $B = K^*K$, which—for simplicity—is assumed invertible, and $|\cdot|_\infty$ denotes the maximum norm on $\mathbb{R}^\ell$. We use $(\cdot,\cdot)_{L^2(\Omega)}$ to denote the $L^2(\Omega)$-inner product, for which we sometimes also write $(\cdot,\cdot)_{L^2}$ or just $(\cdot,\cdot)$. Note also that with inner products and pairings we do not distinguish notationwise between scalar functions and vector fields. The underlying function space is

$$H_0(\text{div}) := \{\mathbf{v} \in L^2(\Omega)^\ell : \text{div}\,\mathbf{v} \in L^2(\Omega) \text{ and } \mathbf{v}\cdot\boldsymbol{n}|_{\partial\Omega}=0\}, \tag{1.1}$$

where $\boldsymbol{n}$ denotes the outer unit normal vector and the boundary condition is taken in the $H^{-1/2}(\partial\Omega)$-sense. Endowed with the inner product

$$(\mathbf{v},\mathbf{w})_{H_0(\text{div})} := (\mathbf{v},\mathbf{w}) + (\text{div}\,\mathbf{v}, \text{div}\,\mathbf{w}),$$

$H_0(\text{div})$ is a Hilbert space. Moreover,

$$\mathcal{A}_{\text{ad}} := \{\alpha \in H^1(\Omega) : \underline{\alpha} \leq \alpha \leq \overline{\alpha}, \text{ a.e. on } \Omega\}, \tag{1.2}$$

with scalars $0 < \underline{\alpha} < \overline{\alpha} < +\infty$, denotes the set of admissible filtering weights. Further, we note already here that throughout this work vector-valued quantities are written in bold font, "s.t." and "f.a.a." stand for "subject to" and "for almost all," respectively. Moreover, we use standard notation for Lebesgue spaces ($L^p(\Omega)$, $p \in [1,+\infty]$) and Sobolev spaces ($W^{s,p}(\Omega)$, $s \in [1,+\infty)$, and $H^s(\Omega) = W^{s,2}(\Omega)$); see, e.g., [1] for more on this. We denote by $\langle\cdot,\cdot\rangle$ the duality pairings $\langle\cdot,\cdot\rangle_{H^{-1},H^1_0}$ and $\langle\cdot,\cdot\rangle_{H^1(\Omega)^*,H^1(\Omega)}$. For the sake of completeness, we also mention that $H^{-1/2}(\partial\Omega)$ denotes the dual space of $H^{1/2}(\partial\Omega)$.

Provided that α is regular enough, in [31] (see also [32]) it is argued that (D(α)) is the Fenchel pre-dual problem of the following weighted total variation problem:

$$\text{minimize} \quad J_P(u,\alpha) \quad \text{over } u \in BV(\Omega), \tag{P}$$

with

$$J_P(u,\alpha) := \frac{1}{2}\int_\Omega |Ku-f|^2\text{d}x + \int_\Omega \alpha(x)|\mathcal{D}u|,$$

where $BV(\Omega) := \{u \in L^1(\Omega) : \mathcal{D}u \in \mathbf{M}(\Omega,\mathbb{R}^\ell)\}$, and with $\mathcal{D}u$ representing the distributional gradient of u. Further, by $\mathbf{M}(\Omega,\mathbb{R}^\ell)$ we denote the space of ℓ-valued Borel measures, which is the dual of $C_c(\Omega;\mathbb{R}^\ell)$, the space of continuous $\mathbb{R}^\ell$-valued functions with compact support in Ω. The quantity $|\mathcal{D}u|$ stands for the smallest nonnegative scalar Borel measure associated with the sum of the total variation norms of the component measures of $\mathcal{D}u$.

The bilevel optimization problem (ℙ) falls into the realm of mathematical programs with equilibrium constraints (MPECs) (in function space); see, e.g., [41,44] for an account of MPECs in $\mathbb{R}^n$, [5,29,34] for infinite dimensional settings, and [35,40,47] for recent applications in mathematical image processing. This problem class suffers from notoriously degenerate constraints ruling out the applications of the celebrated Karush–Kuhn–Tucker theory (compare, e.g., [52]) for deriving first-order optimality or stationarity conditions.

As a remedy, for scalar parameters $\beta, \delta, \epsilon, \gamma, \lambda > 0$ the following regularized version of (ℙ) is studied in [31]:

$$\begin{cases} \text{minimize} \quad J(\mathbf{p},\alpha) := F\circ R(\text{div}\,\mathbf{p}) + \dfrac{\lambda}{2}|\alpha|^2_{H^1(\Omega)} \\ \text{over } (\mathbf{p},\alpha) \in H^1_0(\Omega)^\ell \times \mathcal{A}_{\text{ad}}, \\ \text{s.t. } \mathbf{p} \in \underset{\mathbf{w}\in H^1_0(\Omega)^\ell}{\arg\min} \dfrac{\beta}{2}|\mathbf{w}|^2_{H^1_0(\Omega)^\ell} + \dfrac{\gamma}{2}|\mathbf{w}|^2_{L^2(\Omega)^\ell} \\ \qquad\qquad + J_D(\mathbf{w}) + \dfrac{1}{\epsilon}\mathcal{P}_\delta(\mathbf{w},\alpha), \end{cases} \tag{$\tilde{\mathbb{P}}$}$$

where $F : L^2(\Omega) \to \mathbb{R}^+_0$ with

$$F(v) := \frac{1}{2}\int_\Omega \max(v-\overline{\sigma}^2,0)^2\text{d}x + \frac{1}{2}\int_\Omega \min(v-\underline{\sigma}^2,0)^2\text{d}x,$$

and the max- and min-operations are understood in the pointwise sense. The choice of the bounds $0 < \underline{\sigma} \leq \overline{\sigma} < \infty$ is based on statistical properties related to the noise contained in the measurement f; see Sect. 4.2.1 below for details. Moreover, R

$$R(v)(x) := \int_\Omega w(x,y)\,(KB^{-1}v + (KB^{-1}K^* - I)f)^2(y)\text{d}y \tag{1.3}$$

with a normalized weight $w \in L^\infty(\Omega\times\Omega)$ with $\int_\Omega\int_\Omega w(x,y)\,dxdy = 1$. Note that if $\mathbf{p}$ solves (D(α)), then we have $\text{div}\,\mathbf{p} = Bu - K^*f$, where u is the solution to (P) (see [31, Theorem 3.4]). This implies that

$$R(\text{div}\,\mathbf{p})(x) = \int_\Omega w(x,y)\,(Ku-f)^2(y)\text{d}y,$$

where the right-hand side represents a convolved version of the image residual $Ku - f$.

We now provide the motivation and reasoning behind the definition of $(\mathbf{p},\alpha)\mapsto J(\mathbf{p},\alpha)$. We start with the functional $\mathbf{p}\mapsto F\circ R(\operatorname{div}\mathbf{p})$: Since F penalizes violations above $\overline{\sigma}^2$ and below $\underline{\sigma}^2$, we induce residuals $R(\operatorname{div}\mathbf{p})$ to satisfy $\underline{\sigma}^2\le R(\operatorname{div}\mathbf{p})\le\overline{\sigma}^2$. The map $R(\operatorname{div}\mathbf{p})(x)=\int_\Omega w(x,y)(Ku-f)^2(y)\mathrm{d}y$ for $x\in\Omega$ may be considered a local variance (see [23]) and note that $f=Ku_{\text{true}}+\eta$ where $\int_\Omega|\eta|^2\mathrm{d}x=\sigma^2|\Omega|$. Consequently, if for some α^* we would have $u(\alpha^*)=u_{\text{true}}$, then we expect $R(\operatorname{div}\mathbf{p})\simeq\sigma^2$. Thus, by choosing $\underline{\sigma}<\sigma<\overline{\sigma}$ one would get $F\circ R(\operatorname{div}\mathbf{p}^*)\simeq 0$, where $\operatorname{div}\mathbf{p}^*=Bu(\alpha^*)-K^*f$. Secondly, the H^1-regularity of α induced by the term $\frac{\lambda}{2}|\alpha|^2_{H^1(\Omega)}$ in the objective yields that (P) and (D(α)) are dual to each other. This comes as a consequence of Theorem 3.1 below.

The map $P_\delta:H_0^1(\Omega)^\ell\times L^2(\Omega)\to L^2(\Omega)^\ell$ is defined as

$$P_\delta(\mathbf{p},\alpha):=(\mathbf{p}-\alpha\mathbf{1})^+_\delta-(\mathbf{p}+\alpha\mathbf{1})^-_\delta,\tag{1.4}$$

where, for $\delta>0$, $\mathbb{R}\ni r\mapsto(r)^+_\delta\in\mathbb{R}$ is given by

$$(r)^+_\delta=\begin{cases}r-\delta/2, & r\ge\delta;\\ r^2/2\delta, & r\in(0,\delta);\\ 0, & r\le 0.\end{cases}\tag{1.5}$$

The function $r\mapsto(r)^+_\delta$ is a differentiable approximation of the positive part $r\mapsto(r)^+:=\max(r,0)$ and analogously, for $(r)^-_\delta:=(-r)^+_\delta$ and the negative part $(r)^-:=(-r)^+$. Additionally, for $\delta=0$, $(r)^+_\delta:=(r)^+$ and $(r)^-_\delta:=(r)^-$. For $\mathbf{r}\in\mathbb{R}^\ell$, $(\mathbf{r})^+_\delta$ is defined componentwise, i.e., $(\mathbf{r})^+_\delta=((r_1)^+_\delta,(r_1)^+_\delta,,\dots,(r_l)^+_\delta)$ and $(\mathbf{r})^-_\delta$ analogously. The functional $\mathcal{P}_\delta(\cdot,\alpha):H_0^1(\Omega)^\ell\to\mathbb{R}^+_0$ in ($\tilde{\mathbb{P}}$) penalizes violations of $\mathbf{p}\in\mathbf{K}(\alpha)$ and is defined as

$$\mathcal{P}_\delta(\mathbf{p},\alpha):=\int_\Omega\sum_{i=1}^{\ell}\big(G_\delta(-(p_i+\alpha))+G_\delta(p_i-\alpha)\big)\mathrm{d}x,\tag{1.6}$$

with $\mathbf{p}=(p_1,p_2,\dots,p_l)$ and $G_\delta:\mathbb{R}\to\mathbb{R}$,

$$G_\delta(r)=\begin{cases}\frac{1}{2}r^2-\frac{\delta}{2}r+\frac{\delta^2}{6}, & r\ge\delta;\\ r^3/6\delta, & r\in(0,\delta);\\ 0, & r\le 0,\end{cases}\tag{1.7}$$

for $\delta>0$. The function G_δ is a primitive of $(\cdot)^+_\delta$ defined in (1.5), specifically $G_\delta(r):=\int_{-\infty}^r(s)^+_\delta\mathrm{d}s$ and hence G_δ is twice continuously differentiable. For $\delta=0$, we use $r\mapsto G_0(r):=r^2/2$ for $r\ge 0$ and $G_0(r):=0$ otherwise. Note that the derivative of the map $\mathbf{p}\mapsto\mathcal{P}_\delta(\mathbf{p},\alpha)$ is given by $P_\delta(\mathbf{p},\alpha)$ (see [31] for details).

Utilizing [52], an optimal solution $(\mathbf{p}^*,\alpha^*)\in H_0^1(\Omega)^\ell\times\mathcal{A}_{\text{ad}}$ of ($\tilde{\mathbb{P}}$) can be characterized by an adjoint state (a Lagrange multiplier) $\mathbf{q}^*\in H_0^1(\Omega)^\ell$ such that

$$(J_0'(\operatorname{div}\mathbf{p}^*),\operatorname{div}\mathbf{p})+\Big\langle-\beta\boldsymbol{\Delta}\mathbf{q}^*+\gamma\mathbf{q}^*+A\mathbf{q}^*+\frac{1}{\epsilon}D_1P_\delta(\mathbf{p}^*,\alpha^*)\mathbf{q}^*,\mathbf{p}\Big\rangle=0,\tag{1.8a}$$

$$\Big\langle\lambda(-\Delta+I)\alpha^*+\frac{1}{\epsilon}\big(D_2P_\delta(\mathbf{p}^*,\alpha^*)\big)^\top\mathbf{q}^*,\alpha-\alpha^*\Big\rangle\ge 0,\tag{1.8b}$$

for all $\mathbf{p}\in H_0^1(\Omega)^\ell$ and all $\alpha\in\mathcal{A}_{\text{ad}}$, where $J_0:=F\circ R$ and further

$$-\beta\boldsymbol{\Delta}\mathbf{p}^*+\gamma\mathbf{p}^*+A\mathbf{p}^*+\mathbf{f}+\frac{1}{\epsilon}P_\delta(\mathbf{p}^*,\alpha^*)=0,\quad\text{in } H^{-1}(\Omega)^\ell,\tag{1.8c}$$

where $A:H_0(\operatorname{div})\to H_0(\operatorname{div})^*$ is defined as $A\mathbf{p}:=-\nabla B^{-1}\operatorname{div}\mathbf{p}$, with $\mathbf{p}\in H_0(\operatorname{div})$ and $\mathbf{f}=-\nabla B^{-1}K^*f\in H_0(\operatorname{div})^*$; see [31, Thm. 6.3]. Further, $D_1P_\delta(\mathbf{p},\alpha)$ and $D_2P_\delta(\mathbf{p},\alpha)$ denote the Fréchet derivatives of $\mathbf{p}\mapsto P_\delta(\mathbf{p},\alpha)$ and $\alpha\mapsto P_\delta(\mathbf{p},\alpha)$, respectively. The latter are given by

$$D_1P_\delta(\mathbf{p},\alpha)\mathbf{r}_1:=\big(\boldsymbol{G}''_\delta(\mathbf{p}-\alpha\mathbf{1})+\boldsymbol{G}''_\delta(-\mathbf{p}-\alpha\mathbf{1})\big)\mathbf{r}_1,\tag{1.9a}$$

$$D_2P_\delta(\mathbf{p},\alpha)r_2:=\big(\boldsymbol{G}''_\delta(-\mathbf{p}-\alpha\mathbf{1})-\boldsymbol{G}''_\delta(\mathbf{p}-\alpha\mathbf{1})\big)\mathbf{1}r_2,\tag{1.9b}$$

with $\boldsymbol{G}'_\delta:L^{2+\xi}(\Omega)^\ell\to L^2(\Omega)^\ell$ (with $\xi>0$) given by $\boldsymbol{G}'_\delta(\mathbf{p})=(G'_\delta(p_1),\dots,G'_\delta(p_l))$ and where $G'_\delta:L^{2+\xi}(\Omega)\to L^2(\Omega)$ is the Nemytskii (superposition) operator induced by the real-valued function $r\mapsto(r)^+_\delta$. In order to facilitate navigation through the text, we provide a glossary in Table 1.

Besides characterizing stationarity, another benefit of (1.8) is related to the reduced bilevel problem. In fact, the solution map $\alpha\mapsto\mathbf{p}(\alpha)$ for the regularized lower-level problem allows to reduce ($\tilde{\mathbb{P}}$) to

$$\text{minimize}\quad\hat{J}(\alpha):=J(\mathbf{p}(\alpha),\alpha)\quad\text{over }\alpha\in\mathcal{A}_{\text{ad}}.\tag{$\tilde{\mathbb{P}}_{\text{red}}$}$$

Then, the adjoint state $\mathbf{q}$ allows to compute the derivative of the reduced objective $\hat{J}'$ at some α in an amenable way. In fact, one has

$$\hat{J}'(\alpha)=\lambda(-\Delta+I)\alpha+\frac{1}{\epsilon}\left(D_2P_\delta(\mathbf{p}(\alpha),\alpha)\right)^\top\mathbf{q}(\alpha),\tag{1.10}$$

where $\alpha\mapsto\mathbf{q}(\alpha)$ solves (1.8a) for $\mathbf{p}^*=\mathbf{p}(\alpha)$ and $\alpha^*=\alpha$. The expression for the derivative $\hat{J}'(\alpha)$ follows from the optimality system (1.8), and the adjoint state formalism (see, for example, [36]).

The starting point for the development in this paper is the reduced problem ($\tilde{\mathbb{P}}_{\text{red}}$). It is the basis for developing a projected gradient method for solving the problem algorithmically.

Table 1 Glossary of functions and variables

Glossary Variable	Description	Location
$\mathcal{A}_{\text{ad}}$	Admissible set for regularization functions $\alpha : \Omega \to \mathbb{R}$	(1.2)
$\overline{\alpha}, \underline{\alpha}$	Upper and lower bounds in $\mathcal{A}_{\text{ad}}$	(1.2)
$\mathbf{K}(\alpha)$	Constraint set for the pre-dual variable	(D(α))—p. 1
K	Data forming operator	(D(α))—p. 1
B	K^*K	(D(α))—p. 1
$J_D(\cdot)$	Objective functional of the lower-level problem (pre-dual)	(D(α))—p. 1
$J_P(\cdot,\cdot)$	Objective functional of the lower-level problem (primal)	(P)—p. 2
$J(\cdot,\cdot)$	Objective functional of the upper-level problem	($\tilde{\mathbb{P}}$)—p. 3
λ	H^1-regularization parameter for α	($\tilde{\mathbb{P}}$)—p. 3
β	H^1_0-regularization parameter for pre-dual problem	($\tilde{\mathbb{P}}$)—p. 3
γ	L^2-regularization parameter for pre-dual problem	($\tilde{\mathbb{P}}$)—p. 3
ϵ	Penalty parameter for violations above and below $\overline{\alpha}$ and $\underline{\alpha}$, respectively	($\tilde{\mathbb{P}}$)—p. 3
δ	Smoothing parameter for max and min functions	(1.5)—p. 3
$P_\delta(\mathbf{p},\alpha)$	Derivative of the map $\mathbf{p} \mapsto \mathcal{P}_\delta(\mathbf{p},\alpha)$	(1.4)—p. 3
$(\cdot)^+_\delta$	Smooth version of $r \mapsto \max(0,r)$ function	(1.5)—p. 3
$\mathcal{P}_\delta(\mathbf{p},\alpha)$	Penalty functional for violations of $\mathbf{p} \in \mathbf{K}(\alpha)$	(1.6)—p. 4
$G_\delta(\cdot)$	Primitive of $(\cdot)^+_\delta$, i.e., $G_\delta(r) := \int_{-\infty}^r (s)^+_\delta \mathrm{d}s$	(1.7)—p. 4
$\hat{J}(\cdot)$	Reduced upper-level objective functional	($\tilde{\mathbb{P}}_{\text{red}}$)—p. 4
$\hat{J}'(\alpha)$	Fréchet derivative of the reduced functional at α	(1.10)—p. 4
$P_{\mathcal{A}_{\text{ad}}}$	Minimal distance H^1-projection operator onto $\mathcal{A}_{\text{ad}}$	p. 6
$\nabla \hat{J}(\alpha)$	Gradient of $\hat{J}$ at α, i.e., $\mathcal{R}^{-1}\hat{J}'(\alpha)$ where $\mathcal{R}$ is the Riesz map	p. 9
$\underline{\sigma}, \overline{\sigma}$	Local variance bounds	p. 15

In order to study regularity properties of the solutions of H^1-projections onto $\mathcal{A}_{\text{ad}}$, in the following Sect. 2 we provide higher-order regularity results for solutions of elliptic variational inequality problems. The projected gradient method is defined in Sect. 3, and global convergence results are established. Section 4 is devoted to the discrete version of our algorithm and the proper choice of the variance bounds $\underline{\sigma}$ and $\overline{\sigma}$. Moreover, it contains a report on numerical tests for image denoising, deblurring as well as Fourier and wavelet inpainting.

Before we commence with our analysis, we close this section by mentioning that total variation models of a generalized type can be found in [38] and [3]. Moreover, spatially adapted regularization or data weighting has been studied in [2,6,21,22,24,33,37]. For a brief discussion of these references, we refer to part I of this work; see [31]. The bilevel formulation approach for inverse problems seems to have been pioneered by Haber, Tenorio, and Ghattas (see [11,28]). In the context of image reconstruction, the bilevel approach has also been studied by De Los Reyes, Schönlieb, Valkonen and collaborators (see [12,18,19] and references therein). In addition, splitting methods in image/signal processing involving statistical estimators for parameter selection and deconvolution have been successfully treated in [17,20,45] and the references therein, for instance. It should be noted that the present work does not deal with bilevel "learning" (as in many of the aforementioned references), but tackles image reconstruction via a bilevel optimization approach where the upper-level problem enforces local variances within a certain range and the reconstruction itself is obtained in the lower-level one.

2 An Obstacle Problem and Projection Results

Returning to ($\tilde{\mathbb{P}}_{\text{red}}$) we note that its associated first-order necessary conditions are given by the variational inequality

$$\text{Find } \alpha^* \in \mathcal{A}_{\text{ad}} : \langle \hat{J}'(\alpha^*), \alpha - \alpha^* \rangle \geq 0, \quad \forall \alpha \in \mathcal{A}_{\text{ad}}. \tag{2.1}$$

Given the structure of the derivative $\hat{J}'(\alpha^*)$ in (1.10), (2.1) becomes a so-called double obstacle problem. Hence, the characterization of solutions to ($\tilde{\mathbb{P}}_{\text{red}}$) hinges on the study of (2.1). In addition, using a gradient descent method for solving problem ($\tilde{\mathbb{P}}_{\text{red}}$) yields the sequence $\{\alpha_n\}$ of iterates defined as

$$\alpha_{n+1} = P_{\mathcal{A}_{\text{ad}}}(\alpha_n - \tau_n \nabla \hat{J}(\alpha_n)), \text{ for } n = 0, \ldots, \tag{2.2}$$

with α_0 given . Here, $P_{\mathcal{A}_{\text{ad}}} : H^1(\Omega) \to \mathcal{A}_{\text{ad}} \subset H^1(\Omega)$ is the minimum distance projector onto $\mathcal{A}_{\text{ad}}$ and $\nabla \hat{J}(\alpha_n) \in H^1(\Omega)$ denotes the gradient of $\hat{J}$ at α_n. From (2.2), it follows that α_{n+1} solves: α_{n+1} solves: Find $\alpha^* \in \mathcal{A}_{\text{ad}}$ such that

$$\langle (-\Delta + I)\alpha^* + M(\alpha_n, \tau_n), \alpha - \alpha^* \rangle \geq 0, \quad \forall \alpha \in \mathcal{A}_{\text{ad}};$$

for some $M(\alpha_n, \tau_n)$, yet another double obstacle problem. This motivates the following study of this type of problems.

The subsequent result establishes the $H^2(\Omega) \cap C^{0,r}(\overline{\Omega})$ regularity of the solution to the bilateral obstacle problem with Neumann boundary conditions. The $H^2(\Omega)$-regularity for a single obstacle and with a C^∞-boundary was established by Brézis in [10]. Similar and related partial results can also be found in the classical texts by Rodrigues [46] and Kinderlehrer and Stampacchia [39]. For dimensions $\ell = 1, 2, 3$ (note $\Omega \subset \mathbb{R}^\ell$), the $C^{0,r}(\overline{\Omega})$-regularity is implied by Sobolev embedding results for $H^2(\Omega)$ (see, for example, [1]). For dimensions $\ell \geq 2$, we show that the $C^{0,r}(\overline{\Omega})$-regularity can also be obtained from estimates due to Serrin; see [48].

While this result may be considered of stand-alone importance in the regularity theory for solutions of elliptic variational inequalities, in our generalized total variation context it is of particular relevance to guarantee continuity of iterates α_n of the regularization weight generated by some projection-based descent method.

Theorem 2.1 *Let $\Omega \subset \mathbb{R}^\ell$, with $\ell = 1, 2, 3$, be a bounded convex subset, and let $\mathcal{A} = \{\alpha \in H^1(\Omega) : \underline{\alpha} \leq \alpha \leq \overline{\alpha}$ a.e. on $\Omega\}$ where $\underline{\alpha}, \overline{\alpha} \in H^2(\Omega)$, such that*

$$\underline{\alpha} \leq \overline{\alpha}, \text{ a.e. on } \Omega \quad \text{and} \quad \frac{\partial \underline{\alpha}}{\partial \boldsymbol{\nu}} = \frac{\partial \overline{\alpha}}{\partial \boldsymbol{\nu}} = 0 \text{ in } H^{1/2}(\partial\Omega).$$

Then, for $f \in L^2(\Omega)$, there exists a unique $u^* \in H^2(\Omega) \cap C^{0,r}(\overline{\Omega}) \cap \mathcal{A}$ for some $r \in (0, 1)$ that solves: Find $u \in \mathcal{A}$ such that

$$\int_\Omega \nabla u \cdot \nabla(v - u) + (u - f)(v - u) \mathrm{d}x \geq 0, \quad \forall v \in \mathcal{A}. \tag{2.3}$$

In addition u^* solves uniquely: Find $u \in \mathcal{A}$ and $\frac{\partial u}{\partial \boldsymbol{\nu}} = 0$ on $\partial\Omega$ such that

$$\langle Lu - f, v - u \rangle \geq 0, \quad \forall v \in \mathcal{A}, \tag{2.4}$$

where $L = -\Delta + I$. Furthermore, for some constant $C > 0$ the following estimates hold:

$$\begin{aligned} &\max(|u^*|_{C^{0,r}(\overline{\Omega})}, |u^*|_{H^2(\Omega)}) \\ &\quad \leq C(|f|_{L^2(\Omega)} + |L\underline{\alpha}|_{L^2(\Omega)} + |L\overline{\alpha}|_{L^2(\Omega)}). \end{aligned} \tag{2.5}$$

Proof For $\rho > 0$ consider the approximating problem: Find $u \in H^1(\Omega)$ such that

$$a(u, w) + (F_\rho(u) - f, w) = 0, \quad \forall w \in H^1(\Omega), \tag{2.6}$$

where, for any $v, w \in H^1(\Omega)$, a and F_ρ are defined as

$$\begin{aligned} a(v, w) &:= \int_\Omega \nabla u \cdot \nabla w + uw \mathrm{d}x \\ (F_\rho(v), w) &:= \int_\Omega \frac{1}{\rho}(v - \overline{\alpha})^+ w - \frac{1}{\rho}(v - \underline{\alpha})^- w \mathrm{d}x. \end{aligned}$$

Note that (2.6) is the first-order optimality condition for the problem:

$$\begin{aligned} &\text{minimize} \quad J(u) := \frac{1}{2}|u|^2_{H^1(\Omega)} + \frac{1}{2\rho}G(u) - (f, u) \\ &\quad \text{over } u \in H^1(\Omega), \end{aligned}$$

with $G(u) := |(u - \overline{\alpha})^+|^2_{L^2(\Omega)} + |(\underline{\alpha} - u)^+|^2_{L^2(\Omega)}$. The existence and uniqueness of a solution are guaranteed since $J : H^1(\Omega) \to \mathbb{R}$ is bounded below, coercive, strictly convex and weakly lower semicontinuous (for being convex and continuous).

Note that (2.6) is the variational form of a semilinear Neumann problem, i.e., the solution u^*_ρ to (2.6) satisfies

$$Lu^*_\rho + F_\rho(u^*_\rho) - f = 0 \text{ in } \Omega, \quad \text{and} \quad \frac{\partial u^*_\rho}{\partial \boldsymbol{\nu}} = 0 \text{ on } \partial\Omega;$$

see [49,50] or [4]. Let $f_\rho := f - F_\rho(u^*_\rho)$. Then $f_\rho \in L^2(\Omega)$ and $Lu^*_\rho = f_\rho$ in Ω with $\partial u^*_\rho / \partial \boldsymbol{\nu} = 0$ on $\partial\Omega$. From Theorem 3.2.1.3 and its proof in [25], it follows that $u^*_\rho \in H^2(\Omega)$ and $|u^*_\rho|_{H^2(\Omega)} \leq \tilde{C}_1 |f_\rho|_{L^2(\Omega)}$ for some $\tilde{C}_1 > 0$ depending only on ℓ. Also, for $\ell \geq 2$ we have $u^*_\rho \in C^{0,r}(\overline{\Omega})$ (see [43,48] or Theorem 3.1.5 in [42]) for some $r \in (0, 1)$ depending only on ℓ such that $|u^*_\rho|_{C^{0,r}(\overline{\Omega})} \leq \tilde{C}_2(|u^*_\rho|_{L^2(\Omega)} + |f_\rho|_{L^2(\Omega)})$ with $\tilde{C}_2$ independent on f_ρ. Therefore, we have

$$\begin{aligned} |u^*_\rho|_{H^2(\Omega)} \leq \tilde{C}_1 \Bigg(&|f|_{L^2(\Omega)} + \left| \frac{1}{\rho}(u^*_\rho - \overline{\alpha})^+ \right|_{L^2(\Omega)} \\ &+ \left| \frac{1}{\rho}(u^*_\rho - \underline{\alpha})^- \right|_{L^2(\Omega)} \Bigg), \end{aligned} \tag{2.7}$$

and

$$\begin{aligned} |u^*_\rho|_{C^{0,r}(\overline{\Omega})} \leq \tilde{C}_2 \Bigg(&|u^*_\rho|_{L^2(\Omega)} + |f|_{L^2(\Omega)} + \left| \frac{1}{\rho}(u^*_\rho - \overline{\alpha})^+ \right|_{L^2(\Omega)} \\ &+ \left| \frac{1}{\rho}(u^*_\rho - \underline{\alpha})^- \right|_{L^2(\Omega)} \Bigg) \end{aligned}$$

$$\leq 2\max(\tilde{C}_2, \tilde{C}_1)\left(|f|_{L^2(\Omega)} + \left|\frac{1}{\rho}(u_\rho^* - \overline{\alpha})^+\right|_{L^2(\Omega)} + \left|\frac{1}{\rho}(u_\rho^* - \underline{\alpha})^-\right|_{L^2(\Omega)}\right). \tag{2.8}$$

Note that by Green's theorem, $a(v, w) = (Lv, w)_{H^1(\Omega)^*, H^1(\Omega)} + \int_{\partial\Omega}(\frac{\partial v}{\partial \boldsymbol{\nu}})(w)\mathrm{d}S$, and also $L\overline{\alpha} \in L^2(\Omega)$, $\partial\overline{\alpha}/\partial\boldsymbol{\nu} = 0$ and $\underline{\alpha} \leq \overline{\alpha}$. Then, by taking $w = \frac{1}{\rho}(u_\rho^* - \overline{\alpha})^+ \in H^1(\Omega)$ in (2.6) together with adding and subtracting $(L\overline{\alpha}, w)$ we observe that

$$\frac{1}{\rho}a(u_\rho^* - \overline{\alpha}, (u_\rho^* - \overline{\alpha})^+) + \left|\frac{1}{\rho}(u_\rho^* - \overline{\alpha})^+\right|^2_{L^2(\Omega)} = (f - L\overline{\alpha}, \frac{1}{\rho}(u_\rho^* - \overline{\alpha})^+), \tag{2.9}$$

where we have used that $(F_\rho(u_\rho^*), w) = |w|^2_{L^2(\Omega)}$. Furthermore,

$$a(u_\rho^* - \overline{\alpha}, (u_\rho^* - \overline{\alpha})^+) = \left|(u_\rho^* - \overline{\alpha})^+\right|^2_{L^2(\Omega)} + \left|\nabla(u_\rho^* - \overline{\alpha})^+\right|^2_{L^2(\Omega)^\ell}.$$

Here we exploit that if $v \in H^1(\Omega)$, then $v^+ \in H^1(\Omega)$, and $\nabla v^+ = \nabla v$ if $v > 0$ and $\nabla v^+ = 0$, otherwise. From this, we infer

$$\left|\frac{1}{\rho}(u_\rho^* - \overline{\alpha})^+\right|_{L^2(\Omega)} \leq |f - L\overline{\alpha}|_{L^2(\Omega)}.$$

Analogously, for $w = -\frac{1}{\rho}(u_\rho^* - \underline{\alpha})^-$ in (2.6), we obtain

$$\left|\frac{1}{\rho}(u_\rho^* - \underline{\alpha})^-\right|_{L^2(\Omega)} \leq |f - L\underline{\alpha}|_{L^2(\Omega)}.$$

Hence, it follows that (2.5) holds for u_ρ^* and $C = 6\max(\tilde{C}_1, \tilde{C}_2)$.

The boundedness of $\{u_\rho^*\}_{\rho>0}$ in $H^2(\Omega)$ implies that $Lu_\rho^* \rightharpoonup L\tilde{u}$, $u_\rho^* \to \tilde{u}$ in $L^2(\Omega)$ and $u_\rho^* \rightharpoonup \tilde{u}$ in $H^2(\Omega)$, along a subsequence that we also denote by $\{u_\rho^*\}$. The above two inequalities imply that $\tilde{u} \in \mathcal{A}$. Furthermore, since $u \mapsto \frac{1}{\rho}(u - \overline{\alpha})^+ - \frac{1}{\rho}(u - \underline{\alpha})^-$ is a monotone mapping, using $w = v - u_\rho^*$ with an arbitrary $v \in \mathcal{A}$ in (2.6) (note that $(v - \overline{\alpha})^+ + (v - \underline{\alpha})^- = 0$) we observe

$$a(u_\rho^*, v - u_\rho^*) \geq (f, v - u_\rho^*).$$

Since $a(v - u_\rho^*, v - u_\rho^*) \geq 0$, it follows from the above inequality that $a(v, v - u_\rho^*) \geq (f, v - u_\rho^*)$. Taking the limit as $\rho \downarrow 0$, we get

$$a(v, v - \tilde{u}) \geq (f, v - \tilde{u}), \qquad \forall v \in \mathcal{A}.$$

Finally, since $\tilde{u} \in \mathcal{A}$, Minty's lemma [16,46] implies that $\tilde{u}$ solves (2.3) and uniqueness follows from standard results.

Additionally, the trace map $H^2(\Omega) \ni u \mapsto \partial u/\partial\boldsymbol{\nu} \in H^{1/2}(\partial\Omega)$ is a continuous linear map, and hence, it is weakly continuous. Moreover, since the norm is weakly lower semi-continuous, $|\partial\tilde{u}/\partial\boldsymbol{\nu}|_{H^{1/2}(\partial\Omega)} \leq \liminf_{\rho\to 0}\left|\partial u_\rho^*/\partial\boldsymbol{\nu}\right|_{H^{1/2}(\partial\Omega)} = 0$. From $a(v, w) = (Lv, w)_{H^1(\Omega)^*, H^1(\Omega)} + \int_{\partial\Omega}(\frac{\partial v}{\partial\boldsymbol{\nu}})(w)\mathrm{d}S$ for all $v, w \in H^1(\Omega)$, it follows that $\tilde{u}$ solves (2.4), as well. □

Remark 2.2 The boundary conditions $\partial\underline{\alpha}/\partial\boldsymbol{\nu} = 0$ and $\partial\overline{\alpha}/\partial\boldsymbol{\nu} = 0$ may be relaxed to $\partial\overline{\alpha}/\partial\boldsymbol{\nu} \geq 0$ and $\partial\underline{\alpha}/\partial\boldsymbol{\nu} \leq 0$, respectively.

An important application of the previous result is related to the preservation of regularity of the minimal distance projection operator in $H^1(\Omega)$ onto $\mathcal{A} = \{\alpha \in H^1(\Omega) : \underline{\alpha} \leq \alpha \leq \overline{\alpha} \text{ a.e. on } \Omega\}$.

Corollary 2.3 *Let Ω and $\mathcal{A}$ be as in Theorem 2.1. Let $P_{\mathcal{A}} : H^1(\Omega) \to \mathcal{A} \subset H^1(\Omega)$ denote the minimal distance projection operator, i.e., for $\omega \in H^1(\Omega)$,*

$$P_{\mathcal{A}}(\omega) := \arg\min_{\alpha\in\mathcal{A}} \frac{1}{2}|\alpha - \omega|^2_{H^1(\Omega)}. \tag{2.10}$$

Let $\omega^* = P_{\mathcal{A}}(\omega)$. Then it holds that

$$\omega \in H^2(\Omega) \text{ and } \frac{\partial\omega}{\partial\boldsymbol{\nu}} = 0 \Longrightarrow \omega^* \in H^2(\Omega) \text{ and } \frac{\partial\omega^*}{\partial\boldsymbol{\nu}} = 0,$$

and furthermore,

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

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

Proof The first-order optimality condition for (2.10) is equivalent to

$$\int_\Omega \nabla(\omega^* - \omega)\cdot\nabla(v - \omega^*) + (\omega^* - \omega)(v - \omega^*)\mathrm{d}x \geq 0, \quad \forall v \in \mathcal{A}.$$

Since $\omega \in H^2(\Omega)$ and $\partial\omega/\partial\boldsymbol{\nu} = 0$, by Green's Theorem, the previous variational inequality is equivalent to

$$\int_\Omega \nabla\omega^*\cdot\nabla(v - \omega^*) + (\omega^* - f_\omega)(v - \omega^*)\mathrm{d}x \geq 0, \quad \forall v \in \mathcal{A},$$

with $f_\omega := (-\Delta + I)\omega \in L^2(\Omega)$. The proof then follows from a direct application of Theorem 2.1. □

3 Descent Algorithm and Its Convergence

In this section, we study a basic projected gradient method for solving the regularized bilevel optimization problem ($\tilde{\mathbb{P}}$). We are in particular interested in its global convergence properties in the underlying function space setting as this suggests an image resolution (or, from a discretization point of view, mesh) independent convergence when solving discrete, finite dimensional instances of the problem. As a consequence of such a property, the number of iterations of the solver for computing an ϵ-approximation of a solution (or stationary point) should be expected to behave stably on all sufficiently fine meshes resp. image resolutions.

One of the main focus points of our analysis is to provide guarantee that the iterates α_n remain in $C(\bar{\Omega})$ for all $n \in \mathbb{N}$. This property keeps the primal/dual relation between (P) and (D(α)) vital. We recall here also that for the study of (D(α)) alone, $\alpha_n \in L^2(\Omega)$ suffices, but does no longer allow to link (D(α)) to (P) through dualization. This refers to the fact that given a dual solution $\mathbf{p}$ one no longer can infer a primal solution (recovered image) u from primal-dual first-order optimality conditions. We also note here that, of course, more elaborate techniques may be employed as long as the aforementioned primal/dual relation remains intact.

We employ the following projected gradient method given in Algorithm 1 where the steps $\{\tau_n\}$, $\tau_n \geq 0$ for all $n \in \mathbb{N}$, are chosen according to the Armijo rule with backtracking; compare step 1 of Algorithm 1 and see, e.g., [7,9] for further details.

Algorithm 1 Projected Gradient Method in Function Space.

Require: $\alpha_0 \in H^2(\Omega)$ with $\frac{\partial \alpha_0}{\partial \boldsymbol{\nu}} = 0$ in $\partial\Omega$, $0 < \underline{\mu} \leq \mu_0 \leq \overline{\mu} < \infty$, $0 < \theta_- < 1 \leq \theta_+$, $0 < c < 1$, and set $n := 0$.

1: Compute m_n as the smallest $m \in \mathbb{N}_0$ for which the following holds:

$$\hat{J}(\alpha_n) - \hat{J}(\alpha_n(\theta_-^m \mu_n)) \geq c(\nabla \hat{J}(\alpha_n), \alpha_n - \alpha_n(\theta_-^m \mu_n))_{H^1(\Omega)},$$

with

$$\alpha_n(\theta_-^m \mu_n) = P_{\mathcal{A}_{\mathrm{ad}}}(\alpha_n - \theta_-^m \mu_n \nabla \hat{J}(\alpha_n)),$$

where $P_{\mathcal{A}_{\mathrm{ad}}} : H^1(\Omega) \to \mathcal{A}_{\mathrm{ad}} \subset H^1(\Omega)$ is the H^1-projection operator onto the closed, convex set $\mathcal{A}_{\mathrm{ad}}$.

2: Set $\tau_n = \theta_-^{m_n} \mu_n$ and compute

$$\alpha_{n+1} = P_{\mathcal{A}_{\mathrm{ad}}}(\alpha_n - \tau_n \nabla \hat{J}(\alpha_n)). \tag{3.1}$$

3: **Check stopping criteria**. Unless suitable stopping criteria are met, set $n := n + 1$, $\mu_n = \min(\max(\theta_+ \tau_{n-1}, \underline{\mu}), \overline{\mu})$ and go to step 1.

Recall that our duality result in [31, Thm. 3.4] requires $C(\bar{\Omega})$-regularity of the regularization weight. Below, α_{n+1} represents a suitable approximation. Since it results from an $H^1(\Omega)$-projection, and $H^1(\Omega) \not\hookrightarrow C(\bar{\Omega})$, unless $\ell = 1$, the required regularity for dualization seems in jeopardy. Under mild assumptions and in view of Theorem 2.1, our next result guarantees $\alpha_{n+1} \in C^{0,r}(\bar{\Omega})$ for some $r \in (0, 1)$, and thus the required regularity property.

Theorem 3.1 *Let $\{\alpha_n\}$ be generated by Algorithm 1. Then, $\alpha_n \in H^2(\Omega) \cap C^{0,r}(\bar{\Omega})$ for all $n \in \mathbb{N}$, every limit point α^* of $\{\alpha_n\}$ is stationary for ($\tilde{\mathbb{P}}_{red}$), i.e., $\alpha^* = P_{\mathcal{A}_{ad}}(\alpha^* - \nabla \hat{J}(\alpha^*))$, and belongs to $H^2(\Omega) \cap C^{0,r}(\bar{\Omega})$. Furthermore, we have*

$$\lim_{n \to \infty} \alpha_n - P_{\mathcal{A}_{ad}}(\alpha_n - \nabla \hat{J}(\alpha_n)) = 0, \quad \text{in } H^1(\Omega). \tag{3.2}$$

Proof We split the proof into several steps. *Step 1: Regularity of α^* and α_n.* Let $(\mathbf{p}^*, \alpha^*) \in H_0^1(\Omega)^\ell \times \mathcal{A}_{\mathrm{ad}}$ be a solution to problem ($\tilde{\mathbb{P}}$). Setting $K(\mathbf{p}^*, \alpha^*) := \frac{1}{\epsilon} D_2 P_\delta(\mathbf{p}^*, \alpha^*)$, by [31, Prop. 6.3] (compare (1.8)) there exists an adjoint state $\mathbf{q}^* \in H_0^1(\Omega)^\ell$ satisfying

$$\int_\Omega \nabla \alpha^* \cdot \nabla(\alpha - \alpha^*) + \left(\alpha^* - \frac{1}{\lambda} K(\mathbf{p}^*, \alpha^*)^\top \mathbf{q}^*\right) \times (\alpha - \alpha^*) \mathrm{d}x \geq 0, \qquad \forall \alpha \in \mathcal{A}_{\mathrm{ad}}.$$

Let $\boldsymbol{G}'_\delta$ be the Nemytskii operator induced (componentwise) by $r \mapsto G'_\delta(r) = (r)_\delta^+$ where G_δ is defined in (1.7). Since $G'_\delta(r) \in C^1(\mathbb{R})$, G''_δ is Lipschitz with $|G''_\delta|_{L^\infty(\mathbb{R})}, |G'''_\delta|_{L^\infty(\mathbb{R})} \leq \max(1, \delta)$, it follows that $K(\mathbf{p}^*, \alpha^*)^{\mathrm{T}} \mathbf{q}^* \in W^{1,1}(\Omega) \cap L^2(\Omega)$ as $(\mathbf{p}^*, \alpha^*) \in H_0^1(\Omega)^\ell \times H^1(\Omega)$. The application of Theorem 2.1 yields $\alpha^* \in H^2(\Omega) \cap C^{0,r}(\bar{\Omega})$. Given that $L^2(\Omega) \ni \alpha \mapsto \mathbf{p}(\alpha) \in H_0^1(\Omega)^\ell$ is Lipschitz continuous, note also that, by composition with Lipschitz functions, the map $H^1(\Omega) \ni \alpha \mapsto K(\mathbf{p}(\alpha), \alpha) \in L^4(\Omega)^\ell$ for $\ell \leq 4$ is Lipschitz continuous too, and $G''_\delta : \mathbb{R} \to \mathbb{R}$ is uniformly bounded and Lipschitz continuous so that $\boldsymbol{G}''_\delta : L^4(\Omega)^\ell \to L^4(\Omega)^\ell$ is Lipschitz continuous (see Lemma 4.1 in [51] and the remark at the end of its proof).

Suppose that $\alpha \in H^2(\Omega)$ and $\frac{\partial \alpha}{\partial \boldsymbol{\nu}} = 0$ in $\partial\Omega$. Then we have

$$\langle \hat{J}'(\alpha), \omega \rangle = \int_\Omega (\lambda(-\Delta \alpha + \alpha) - K(\mathbf{p}(\alpha), \alpha)^\top \mathbf{q}(\alpha)) \omega \mathrm{d}x, \tag{3.3}$$

for $\omega \in H^1(\Omega)$. Hence, $\hat{J}'(\alpha) \in L^2(\Omega)$ and $\nabla \hat{J}(\alpha) \in H^2(\Omega)$ with $\frac{\partial \nabla \hat{J}(\alpha)}{\partial n} = 0$ on $\partial\Omega$. The application of Corollary 2.3 yields $P_{\mathcal{A}_{\mathrm{ad}}}(\alpha - \tau \nabla \hat{J}(\alpha)) \in H^2(\Omega) \cap C^{0,r}(\bar{\Omega})$ and that it satisfies homogeneous Neumann boundary conditions. By induction one shows $\alpha_n \in H^2(\Omega) \cap C^{0,r}(\bar{\Omega})$ and $\partial \alpha_n / \partial \boldsymbol{\nu} = 0$ on $\partial\Omega$ for all $n \in \mathbb{N}$.

Step 2: The limit in (3.2) holds. It is known that every cluster point of $\{\alpha_n\}$ is stationary (see [9]) and that $\alpha_n - P_{\mathcal{A}_{\mathrm{ad}}}(\alpha_n - \tau_n \nabla \hat{J}(\alpha_n)) \to 0$ as $n \to \infty$ provided that $H^1(\Omega) \ni \alpha \mapsto \nabla \hat{J}(\alpha) \in H^1(\Omega)$ is Lipschitz continuous

(see Theorem 2.4 in [36]). We first prove the Lipschitz continuity of the map $\alpha \mapsto \mathbf{q}(\alpha)$. Let $\mathbf{p}_1, \mathbf{q}_1$ and $\mathbf{p}_2, \mathbf{q}_2$ (satisfying the system in (1.8)) denote the states and adjoint states associated with α_1 and α_2 in $\mathcal{A}_{\text{ad}}$, respectively. Given the structure of $J_0 = F \circ R$, we observe that

$$|(J_0'(\operatorname{div}\mathbf{p}_2) - J_0'(\operatorname{div}\mathbf{p}_1), \operatorname{div}(\mathbf{q}_2 - \mathbf{q}_1))| \leq C_1 |\operatorname{div}(\mathbf{p}_2 - \mathbf{p}_1)|_{L^2(\Omega)} |\operatorname{div}(\mathbf{q}_2 - \mathbf{q}_1)|_{L^2(\Omega)},$$

where $C_1 = C_1(\alpha_1, \alpha_2)$ is bounded by

$$C_1 \leq M_1 \left(|\operatorname{div}\mathbf{p}_2|_{L^2(\Omega)} + \int_\Omega |\max(R(\operatorname{div}\mathbf{p}_1) - \sigma_1^2, 0)| + |\min(R(\operatorname{div}\mathbf{p}_1) - \sigma_2^2, 0)| \mathrm{d}x \right),$$

with $M_1 \geq 0$ depending on the filter kernel w and f, so that $C_1(\alpha_1, \alpha_2) \leq M_2 < \infty$ uniformly in α_1, α_2. Additionally, as stated before, the map $H^1(\Omega) \ni \alpha \mapsto \frac{1}{\epsilon} D_2 P(\mathbf{p}(\alpha), \alpha) = K(\mathbf{p}(\alpha), \alpha) \in L^4(\Omega)^\ell$ is Lipschitz continuous, $D_1 P(\mathbf{p}(\alpha), \alpha)$ is a monotone operator (this follows since $H_0^1(\Omega)^\ell \ni \mathbf{p} \mapsto P_\delta(\mathbf{p}, \alpha) \in H^{-1}(\Omega)^\ell$ is monotone and differentiable), and by composition of maps one shows that $H^1(\Omega) \ni \alpha \mapsto \mathbf{q}(\alpha) \in H_0^1(\Omega)^\ell$ is Lipschitz continuous. This implies in turn that the map $H^1(\Omega) \ni \alpha \mapsto K(\mathbf{p}(\alpha), \alpha)^{\mathrm{T}} \mathbf{q}(\alpha) \in L^2(\Omega)$ is Lipschitz, as well. Since $\nabla \hat{J}(\alpha) = (-\Delta + I)^{-1} \hat{J}'(\alpha)$, we have that $H^1(\Omega) \ni \alpha \mapsto \nabla \hat{J}(\alpha) \in H^1(\Omega)$ is Lipschitz continuous. This ends the proof. □

The above convergence result can be strengthened. In fact, the following theorem shows that under suitable assumptions one has $\alpha_n \to \alpha^*$ in $H^1(\Omega)$ at a q-linear rate. In particular, this requires that the sequence of step lengths $\{\tau_n\}$ is non-increasing and bounded from below. We note that the sequence $\{\tau_n\}$ can be made non-increasing by setting $\mu_n := \tau_{n-1}$ for all $n \in \mathbb{N}$ in step 3 of Algorithm 1. Concerning proving the existence of a uniform lower bound on the step lengths the Lipschitz continuity of the map $H^1(\Omega) \ni \alpha \mapsto \nabla \hat{J}(\alpha) \in H^1(\Omega)$, as shown in the proof of Theorem 3.1, suffices. In fact, in finite dimensions and under simple constraints, the result can be found in [8] and the proof there can easily be adapted to a Hilbert space setting with arbitrary nonempty closed convex set. Further, we make use of the following result which can be found in Theorem 5.1 and Remark 5.1 in [31] that we state here as a lemma.

Lemma 3.2 *Given $f \in L^2(\Omega)$ and let $\mathbf{p}(\alpha, f)$ be the solution to the lower-level problem in ($\tilde{\mathbb{P}}$). Then, $\mathbf{p}(\alpha, f) \to 0$ in $H_0^1(\Omega)^\ell$ as $f \downarrow 0$ in $L^2(\Omega)$.*

Theorem 3.3 *Let $\{\alpha_n\}$ be generated by Algorithm 1. If the sequence of step lengths $\{\tau_n\} = \{\theta_-^{m_n} \mu_n\}$ is non-increasing in the sense that $\mu_n = \tau_{n-1}$, then $\alpha_n \to \alpha^*$ q-linearly in $H^1(\Omega)$ provided that $\lambda > 0$ and the data $f \in L^2(\Omega)$ are sufficiently small, respectively.*

Proof We first prove that the Lipschitz constant of the map $H^1(\Omega) \ni \alpha \mapsto K(\mathbf{p}(\alpha), \alpha)^{\mathrm{T}} \mathbf{q}(\alpha) \in L^2(\Omega)$ denoted as $L(f)$ satisfies $L(f) \to 0$ as $f \to 0$ in $L^2(\Omega)$. Let $\mathbf{p}_i := \mathbf{p}(\alpha_i)$ and $\mathbf{q}_i := \mathbf{q}(\alpha_i)$. Then, by the triangle inequality

$$|K(\mathbf{p}_2, \alpha_2)^{\mathrm{T}} \mathbf{q}_2 - K(\mathbf{p}_1, \alpha_1)^{\mathrm{T}} \mathbf{q}_1|_{L^2(\Omega)} \leq |\mathbf{q}_1|_{L^4(\Omega)^\ell} C(|\mathbf{p}_2 - \mathbf{p}_1|_{L^4(\Omega)^\ell} + |\alpha_2 - \alpha_1|_{L^4(\Omega)}) + |K(\mathbf{p}_2, \alpha_2)|_{L^4(\Omega)^\ell} |\mathbf{q}_2 - \mathbf{q}_1|_{L^4(\Omega)^\ell},$$

for some $C > 0$. We know that $H^1(\Omega) \ni \alpha \mapsto \mathbf{q}(\alpha) \in H_0^1(\Omega)^\ell$ and $L^2(\Omega) \ni \alpha \mapsto \mathbf{p}(\alpha) \in H_0^1(\Omega)^\ell$ are Lipschitz continuous. Furthermore, Lemma 3.2 implies $\mathbf{p}(\alpha, f) \to 0$ in $H_0^1(\Omega)^\ell$ as $f \downarrow 0$ in $L^2(\Omega)$ and analogously, one shows that $\mathbf{q}(\alpha, f) \to 0$ in $H_0^1(\Omega)^\ell$ as $f \downarrow 0$ in $L^2(\Omega)$ since $K(\mathbf{p}(\alpha, f), \alpha) \to 0$ in $L^4(\Omega)^\ell$ and $-\nabla J_0'(\operatorname{div}\mathbf{p}(\alpha, f)) \to 0$ in $H^{-1}(\Omega)^\ell$ as $f \downarrow 0$ in $L^2(\Omega)$. Hence, since $H^1(\Omega) \hookrightarrow L^4(\Omega)$ for $\ell \leq 4$, the map under investigation is Lipschitz continuous with constant $L(f)$, and $L(f) \to 0$ as $f \to 0$ in $L^2(\Omega)$.

Since $H^1(\Omega) \ni \alpha \mapsto \nabla \hat{J}(\alpha) \in H^1(\Omega)$ is Lipschitz continuous (see the proof of Theorem 3.1), it follows that step sizes τ_n are bounded from below (see [8]). The sequence $\{\tau_n\}$ is non-increasing by hypothesis and then, since $\tau_n = \theta_-^{m_n} \tau_{n-1}$, and $m_n \in \mathbb{N}_0$, we have $m_n = 0$ for $n \geq \tilde{N}$ for some $\tilde{N} \in \mathbb{N}$ sufficiently large: Suppose there is no such an $\tilde{N}$. Then, there is a subsequence $\{m_{n_j}\}$ such that $m_{n_j} \geq 1$ for $j \in \mathbb{N}$, which implies that $\tau_{n_j} \leq \theta_-^j \tau_0$. Hence, $\tau_{n_j} \to 0$ as $j \to \infty$ and then $\{\tau_n\}$ is not bounded below.

Then, it is enough to consider $\{\alpha_n\}_{n > \tilde{N}}$ and such that $\tau_n = \tilde{\tau}$ for some fixed $\tilde{\tau} > 0$. Define $Q(\alpha) := K(\mathbf{p}(\alpha), \alpha)^{\mathrm{T}} \mathbf{q}(\alpha)$, let $\Psi = P_{\mathcal{A}_{\text{ad}}}(\psi - \tilde{\tau} \nabla \hat{J}(\psi))$ and $\Theta = P_{\mathcal{A}_{\text{ad}}}(\theta - \tilde{\tau} \nabla \hat{J}(\theta))$ for some $\psi, \theta \in \mathcal{A}_{\text{ad}}$. Then, using that the projection map $P_{\mathcal{A}_{\text{ad}}}$ is non-expansive, $\nabla \hat{J}(\alpha) = (-\Delta + I)^{-1} \hat{J}'(\alpha) = \mathcal{R}^{-1} \hat{J}'(\alpha)$ (where $\mathcal{R}$ is the Riesz map for $H^1(\Omega)$) and (3.3), we have

$$|\Psi - \Theta|_{H^1(\Omega)}^2 \leq |(1 - \tilde{\tau}\lambda)(\psi - \theta) + \tilde{\tau} \mathcal{R}^{-1}(Q(\psi) - Q(\theta))|_{H^1(\Omega)}^2$$

The structure of the norm in $H^1(\Omega)$ implies

$$\begin{aligned} |\Psi - \Theta|_{H^1(\Omega)}^2 &\leq (1 - \tilde{\tau}\lambda)^2 |\psi - \theta|_{H^1(\Omega)}^2 + \tilde{\tau}^2 |\mathcal{R}^{-1}(Q(\theta) - Q(\psi))|_{H^1(\Omega)}^2 \\ &\quad + 2(1 - \tilde{\tau}\lambda)\tilde{\tau}(\psi - \theta, \mathcal{R}^{-1}(Q(\theta) - Q(\psi)))_{H^1(\Omega)} \\ &\leq (1 - \tilde{\tau}\lambda)^2 |\psi - \theta|_{H^1(\Omega)}^2 + \tilde{\tau}^2 L(f)^2 |\psi - \theta|_{L^2(\Omega)}^2 \\ &\quad + 2|1 - \tilde{\tau}\lambda| \tilde{\tau} L(f) |\psi - \theta|_{H^1(\Omega)} |\psi - \theta|_{L^2(\Omega)} \\ &\leq \left((1 - \tilde{\tau}\lambda)^2 + \tilde{\tau}^2 L(f)^2 + 2(1 - \tilde{\tau}\lambda)\tilde{\tau} L(f) \right) |(\psi - \theta)|_{H^1(\Omega)}^2. \end{aligned}$$

Here, we have used the Lipschitz properties of the map $\alpha \mapsto Q(\alpha)$ described before. Finally, for $\lambda > 0$ and $f \in L^2(\Omega)$ sufficiently small, the map $H^1(\Omega) \ni \varphi \mapsto P_{\mathcal{A}_{\mathrm{ad}}}(\varphi - \tilde{\tau}\nabla \hat{J}(\varphi)) \in H^1(\Omega)$ is contractive and the iteration (3.1) converges linearly by Banach Fixed Point Theorem.

4 Numerical Experiments

In this section, we provide numerical results for image denoising, deblurring, and Fourier as well as wavelet inpainting.

4.1 Implementation

Utilizing a finite difference discretization of the regularized and penalized lower-level problem in ($\tilde{\mathbb{P}}$), we arrive at the discretized bilevel problem

$$\begin{cases} \text{minimize} \quad J(\mathbf{p}, \alpha) \quad \text{over } \mathbf{p} \in (\mathbb{R}^{|\Omega_h|})^2,\ \alpha \in \mathcal{A}_{\mathrm{ad}}, \\ \text{s.t. } g(\mathbf{p}, \alpha) := -\beta \mathbf{\Delta} \mathbf{p} + \gamma \mathbf{p} + A\mathbf{p} + \mathbf{f} + \dfrac{1}{\epsilon} P_\delta(\mathbf{p}, \alpha) = 0, \end{cases} \tag{4.1}$$

with $A\mathbf{p} := -\nabla B^{-1} \operatorname{div} \mathbf{p}$, and $\mathbf{f} = -\nabla B^{-1} K^* f$, and where we set $\Omega_h := \{1, 2, ..., n_1\} \times \{1, 2, ..., n_2\}$ and define the mesh size $h := \sqrt{1/(n_1 n_2)}$. Assuming constant bounds in $\mathcal{A}_{\mathrm{ad}}$, the discrete admissible set, again denoted by $\mathcal{A}_{\mathrm{ad}}$, is given by

$$\mathcal{A}_{\mathrm{ad}} := \{\alpha \in \mathbb{R}^{|\Omega_h|} : \underline{\alpha} \le \alpha_j \le \overline{\alpha},\ \forall j = (j_1, j_2) \in \Omega_h\}.$$

The discrete objective reads

$$\begin{aligned} J(\mathbf{p}, \alpha) &:= \frac{1}{2} \left| (R(\operatorname{div} \mathbf{p}) - \overline{\sigma}^2)^+ \right|^2_{\ell^2(\Omega_\omega)} \\ &\quad + \frac{1}{2} \left| (\underline{\sigma}^2 - R(\operatorname{div} \mathbf{p}))^+ \right|^2_{\ell^2(\Omega_\omega)} + \frac{\lambda}{2} |\alpha|^2_{H^1(\Omega_h)}, \\ R(\operatorname{div} \mathbf{p}) &:= w * |K(\mu I + K^*K)^{-1}(\operatorname{div} \mathbf{p} + K^* f) - f|^2, \end{aligned}$$

where Ω_ω is the (index) domain for the acquired data f (we use $\Omega_\omega = \Omega_h$ in denoising and deblurring), and define $|f|^2_{\ell^2(\Omega_\omega)} := (\sum_{j \in \Omega_\omega} |f_j|^2)/|\Omega_\omega|$. In our experiments, w is a (spatially invariant) averaging filter of size $n_{(\mathrm{w})}$-by-$n_{(\mathrm{w})}$ (i.e., the local window size is $n^2_{(\mathrm{w})}$ many pixels), and thus the computation of the local variance estimator $R(\operatorname{div} \mathbf{p})$ becomes a discrete convolution denoted by "$*$". The term "μI" in the definition of $R(\operatorname{div} \mathbf{p})$, with $0 < \mu \ll 1$, serves as a regularization of K^*K.

We discretize the divergence operator as

$$\begin{aligned} (\operatorname{div} \mathbf{p})_{(j_1, j_2)} = \frac{1}{h} \Big(& \mathbf{p}^1_{(j_1, j_2)} - \mathbf{p}^1_{(j_1 - 1, j_2)} \\ & + \mathbf{p}^2_{(j_1, j_2)} - \mathbf{p}^2_{(j_1, j_2 - 1)} \Big), \quad \forall (j_1, j_2) \in \Omega_h, \end{aligned}$$

with $\mathbf{p}^1_{(\tilde{j}_1, \tilde{j}_2)} = \mathbf{p}^2_{(\tilde{j}_1, \tilde{j}_2)} = 0$ whenever $(\tilde{j}_1, \tilde{j}_2) \notin \Omega_h$ in the above formula. Accordingly, the discrete gradient operator ∇ is defined by the adjoint relation, i.e., $\nabla := -\operatorname{div}^\top$. The discrete vectorial Laplacian Δ is defined by $\Delta \mathbf{p} = (\Delta_{(\mathrm{D})} \mathbf{p}^1, \Delta_{(\mathrm{D})} \mathbf{p}^2)$ for each $\mathbf{p} \in (\mathbb{R}^{|\Omega_h|})^2$, and $\Delta_{(\mathrm{D})}, \Delta_{(\mathrm{N})} \in \mathbb{R}^{|\Omega_h| \times |\Omega_h|}$ denote the discrete five-point-stencil Laplacians with homogenous Dirichlet and Neumann boundary conditions, respectively. For generating $\Delta_{(\mathrm{N})}$, the function value on a ghost grid point (outside the domain) is always set to the function value at the nearest grid point within the domain. For the discrete H^1-norm of $\alpha \in \mathbb{R}^{|\Omega_h|}$ (satisfying homogeneous Neumann conditions) we use

$$|\alpha|_{H^1(\Omega_h)} := h \sqrt{\alpha^\top (I - \Delta_{(\mathrm{N})}) \alpha}.$$

By considering the discrete $H^1(\Omega)$-to-$H^1(\Omega)^*$ Riesz map as $\alpha \mapsto r = (I - \Delta_{(\mathrm{N})})\alpha$, we define the discrete dual H^1-norm as

$$|r|_{H^1(\Omega_h)^*} := \left| (I - \Delta_{(\mathrm{N})})^{-1} r \right|_{H^1(\Omega_h)} = h \sqrt{r^\top (I - \Delta_{(\mathrm{N})})^{-1} r}.$$

The denoising problem is treated specially. In fact, we set $\mu = 0$ and discretize the operator $\nabla \circ \operatorname{div}$ jointly by

$$\begin{aligned} &(\nabla \operatorname{div} \mathbf{p})_{(j_1, j_2)} \\ &= \frac{1}{h^2} \Big(\mathbf{p}^1_{(j_1+1, j_2)} - 2\mathbf{p}^1_{(j_1, j_2)} + \mathbf{p}^1_{(j_1-1, j_2)} + \mathbf{p}^2_{(j_1+1, j_2)} \\ &\quad - \mathbf{p}^2_{(j_1+1, j_2-1)} - \mathbf{p}^2_{(j_1, j_2)} + \mathbf{p}^2_{(j_1, j_2-1)}, \mathbf{p}^2_{(j_1, j_2+1)} \\ &\quad - 2\mathbf{p}^2_{(j_1, j_2)} + \mathbf{p}^2_{(j_1, j_2-1)} + \mathbf{p}^1_{(j_1, j_2+1)} - \mathbf{p}^1_{(j_1-1, j_2+1)} \\ &\quad - \mathbf{p}^1_{(j_1, j_2)} + \mathbf{p}^1_{(j_1-1, j_2)} \Big) \end{aligned}$$

for all $(j_1, j_2) \in \Omega_h$, and $\mathbf{p}^1_{(\tilde{j}_1, \tilde{j}_2)} = \mathbf{p}^2_{(\tilde{j}_1, \tilde{j}_2)} = 0$ whenever $(\tilde{j}_1, \tilde{j}_2) \notin \Omega_h$ in the above formula. Further, this is used to compute the discrete dual $H_0(\mathrm{div})$-norm as

$$|\mathbf{v}|_{H_0(\mathrm{div})^*} := h \sqrt{\mathbf{v}^\top (I - \nabla \circ \operatorname{div})^{-1} \mathbf{v}}, \qquad \text{for } \mathbf{v} \in (\mathbb{R}^{|\Omega_h|})^2.$$

In our numerical tests, we use the discrete version of Algorithm 1 as shown in Algorithm 2 below. For a given α, the solution of the lower-level problem $g(\mathbf{p}, \alpha) = 0$ (compare step 4 of Algorithm 2) is computed by a path-following Newton technique. Its numerical realization can be found in Algorithm 3. Besides, each projection onto $\mathcal{A}_{\mathrm{ad}}$ requires solving an obstacle problem in $H^1(\Omega)$, which is carried out by the semismooth Newton method [30]. For convenience of

the reader, in Algorithm 4 we tailor this semismooth Newton method to the requirements in this paper. The overall algorithm is terminated once $\kappa^n/\kappa^0 < \mathrm{tol}_{(\mathrm{b})}$, where

$$\kappa^n := \left| P_{\mathcal{A}_{\mathrm{ad}}}(\alpha^n - \nabla \hat{J}(\alpha^n)) - \alpha^n \right|_{H^1(\Omega_h)}$$

is our proximity measure and $\mathrm{tol}_{(\mathrm{b})} > 0$ is the user-set tolerance parameter.

Algorithm 2 Discretized projected gradient method.

Require: $\underline{\alpha}, \overline{\alpha}, \overline{\sigma}, \underline{\sigma}, \lambda, \beta, \gamma, \mu, \epsilon, \delta, \tau^0, \mathrm{tol}_{(\mathrm{b})} > 0$, $0 < c < 1$, $0 < \theta_- < 1 \le \theta_+$, $n_{(\mathrm{w})} \in \mathbb{N}$.

Generate the averaging filter w of size $n_{(\mathrm{w})}^2$.
Initialize $\alpha^0 \in \mathcal{A}_{\mathrm{ad}}$ and $k := 0$.
repeat
  Compute $\mathbf{p}^k \in (\mathbb{R}^{|\Omega_h|})^2$ as the solution of $g(\mathbf{p}^k, \alpha^k) = 0$.
  Compute $u^k := (\mu I + K^*K)^{-1}(\operatorname{div} \mathbf{p}^k + K^* f)$.
  Solve the following adjoint equation for $\mathbf{q}^k$:

$$\begin{aligned} &-\nabla(\mu I + K^*K)^{-1} \operatorname{div} \mathbf{q}^k - \beta \Delta \mathbf{q}^k + \gamma \mathbf{q}^k \\ &+ \frac{1}{\epsilon} \operatorname{diag}\Big(G_\delta''(\mathbf{p}^k - \alpha^k \mathbf{1}) + G_\delta''(-\mathbf{p}^k - \alpha^k \mathbf{1}) \Big) \mathbf{q}^k \\ &= \nabla(\mu I + K^*K)^{-1} K^* \operatorname{diag}(K u^k - f) \\ &\Big(w * \big((R(\operatorname{div} \mathbf{p}^k) - \overline{\sigma}^2)^+ - (\underline{\sigma}^2 - R(\operatorname{div} \mathbf{p}^k))^+ \big) \Big). \end{aligned}$$

  Compute the reduced derivative $\hat{J}'(\alpha^k) := \Big(\operatorname{diag}\big(-G_\delta''(\mathbf{p}^k - \alpha^k \mathbf{1}) + G_\delta''(-\mathbf{p}^k - \alpha^k \mathbf{1}) \big) \mathbf{q}^k \Big) \mathbf{1} + \lambda (I - \Delta_{(\mathrm{N})}) \alpha^k$ as well as the reduced gradient $\nabla \hat{J}(\alpha^k) := (I - \Delta_{(\mathrm{N})})^{-1} \hat{J}'(\alpha^k)$.
  Evaluate the proximity measure $\kappa^k := \left| P_{\mathcal{A}_{\mathrm{ad}}}(\alpha^k - \nabla \hat{J}(\alpha^k)) - \alpha^k \right|_{H^1(\Omega)}$.
  if $\kappa^k/\kappa^0 < \mathrm{tol}_{(\mathrm{b})}$ **then**
    return $\alpha^k, \mathbf{p}^k, u^k$.
  end if
  Compute the trial point $\alpha^{k+1} := P_{\mathcal{A}_{\mathrm{ad}}}(\alpha^k - \tau^k \nabla \hat{J}(\alpha^k))$.
  while $\hat{J}(\alpha^{k+1}) > \hat{J}(\alpha^k) + c \hat{J}'(\alpha^k)^\top (\alpha^{k+1} - \alpha^k)$ **do** {Armijo line search}
    Set $\tau^k := \theta_- \tau^k$, and then re-compute $\alpha^{k+1} := P_{\mathcal{A}_{\mathrm{ad}}}(\alpha^k - \tau^k \nabla \hat{J}(\alpha^k))$.
  end while
  Update $\tau^{k+1} := \theta_+ \tau^k$ and $k := k + 1$.
until some stopping criterion is satisfied.

4.2 Parameter Settings

Unless otherwise specified, the following parameters are used throughout our numerical experiments: $\lambda = 10^{-6}$, $\beta = \gamma = 10^{-4}$, $\epsilon = c = 10^{-8}$, $\delta = \tau^0 = 10^{-3}$, $\theta_- = 0.25$, $\theta_+ = 2$, $n_{(\mathrm{w})} = 7$, $\mathrm{tol}_{(\mathrm{b})} = 0.005$. The choice of λ is taken from the range $[10^{-7}, 10^{-5}]$ in which final results seem invariant. The parameters ϵ and δ are chosen to be sufficiently small so that improvements are not significant upon further reduction. Note that the sensitivity of parameters is studied in Sect. 4.5 and results shown in Fig. 8 . The bounds $\underline{\alpha} = 10^{-8}$ and $\overline{\alpha} = 10^{-2}$ are chosen so that the interval $[\underline{\alpha}, \overline{\alpha}]$ is sufficiently large for proper selection of the spatially variant α. The parameter μ is set to be zero for denoising and deblurring, while $\mu = 10^{-4}$ for Fourier- and wavelet inpainting.

Algorithm 3 Path-following Newton method for the lower-level problem in step 4 of Algorithm 2 .

Require: inputs $\mathrm{tol}_{(\mathrm{l})} > 0$, $0 < \theta_\epsilon < 1$, $\alpha \in \mathbb{R}^{|\Omega_h|}$.
Initialize $\mathbf{p}^0 \in (\mathbb{R}^{|\Omega_h|})^2$, $\epsilon^0 := 1$, $\tilde{l} := 0$, and $l := 0$.
while $\epsilon^l > \epsilon$ **or** $\left| g(\mathbf{p}^l, \alpha; \epsilon^l) \right|_{H_0(\mathrm{div})^*} \ge \mathrm{tol}_{(\mathrm{l})} \left| g(\mathbf{p}^{\tilde{l}}, \alpha; \epsilon^l) \right|_{H_0(\mathrm{div})^*}$ **do**
  Compute the Newton step $\delta \mathbf{p}^l$ by solving

$$\begin{aligned} &-\nabla(\mu I + K^*K)^{-1} \operatorname{div} \delta \mathbf{p}^l - \beta \Delta \delta \mathbf{p}^l + \gamma \delta \mathbf{p}^l \\ &+ \frac{1}{\epsilon^l} \operatorname{diag}\Big(G_\delta''(\mathbf{p}^l - \alpha \mathbf{1}) + G_\delta''(-\mathbf{p}^l - \alpha \mathbf{1}) \Big) \delta \mathbf{p}^l \\ &= -g(\mathbf{p}^l, \alpha; \epsilon^l). \end{aligned}$$

  Update $\mathbf{p}^{l+1} := \mathbf{p}^l + \delta \mathbf{p}^l$.
  if $\left| g(\mathbf{p}^{l+1}, \alpha; \epsilon^l) \right|_{H_0(\mathrm{div})^*} < \mathrm{tol}_{(\mathrm{l})} \left| g(\mathbf{p}^{\tilde{l}}, \alpha; \epsilon^l) \right|_{H_0(\mathrm{div})^*}$ **then**
    Set $\epsilon^{l+1} := \max(\theta_\epsilon \epsilon^l, \epsilon)$ and $\tilde{l} := l + 1$.
  else
    Set $\epsilon^{l+1} := \epsilon^l$.
  end if
  Update $l := l + 1$.
end while
Return $\mathbf{p}^l$.

Algorithm 4 α-projection.

Require: Inputs $\epsilon_\alpha, \mathrm{tol}_{(\mathrm{p})} > 0$, $\tilde{\alpha} \in \mathbb{R}^{|\Omega_h|}$.
Initialize $\alpha^0 \in \mathbb{R}^{|\Omega_h|}$ and $l := 0$.
Compute the residual $r^0 := (I - \Delta_{(\mathrm{N})})(\alpha^0 - \tilde{\alpha}) + \frac{1}{\epsilon_\alpha} \left((\alpha^0 - \overline{\alpha})^+ - (\alpha^0 - \underline{\alpha})^+ \right)$.
repeat
  Compute the Newton step $\delta \alpha^l$ by solving

$$\left(I - \Delta_{(\mathrm{N})} + \frac{1}{\epsilon_\alpha} \operatorname{diag}(\xi^l) \right) \delta \alpha^l = -r^l,$$

  where $\xi^l \in \mathbb{R}^{|\Omega_h|}$ is given by

$$\xi_j^l = \begin{cases} 1 & \text{if } \alpha_j^l > \overline{\alpha} \text{ or } \alpha_j^l < \underline{\alpha}, \\ 0 & \text{otherwise.} \end{cases}$$

  Update

$$\begin{aligned} \alpha^{l+1} &:= \alpha^l + \delta \alpha^l, \\ r^{l+1} &:= (I - \Delta_{(\mathrm{N})})(\alpha^{l+1} - \tilde{\alpha}) \\ &\quad + \frac{1}{\epsilon_\alpha} \left((\alpha^{l+1} - \overline{\alpha})^+ - (\alpha^{l+1} - \underline{\alpha})^+ \right). \end{aligned}$$

  Set $l := l + 1$.
until $\left| r^l \right|_{H^1(\Omega)^*} < \mathrm{tol}_{(\mathrm{p})} \left| r^0 \right|_{H^1(\Omega)^*}$.
Return α^l.

Table 2 Comparison with respect to PSNR and SSIM

$\begin{pmatrix}\text{PSNR}\\ \text{SSIM}\end{pmatrix}$	Denoise		Deblur	Fourier		Wavelet
	$\sigma = 0.1$	$\sigma = 0.2$		Teeth	Chest	
Best scalar $\hat{\alpha}$	27.1172	23.9003	25.5452	28.3300	29.1656	27.3100
	0.7937	0.7112	0.7913	0.8136	0.8357	0.8566
SATV	27.9817	24.5544	25.8144	–	–	–
	0.8042	0.6803	0.8004	–	–	–
Bilevel-(#1)	27.4184	23.5480	25.5760	28.3529	28.4044	27.5024
	0.8154	0.7128	0.7916	0.8134	0.8210	0.8533
Bilevel-(#2)	27.5783	24.3556	26.0976	28.5605	28.8902	27.6311
	0.8159	0.7031	0.8092	0.8258	0.8403	0.8554

Finally, concerning the initialization of α, the general guideline is to choose α^0 sufficiently large, depending on the underlying problem, so that it yields a cartoon-like restoration u^0. This is analogous to the spatially adaptive total variation method in [23]. The rationale behind this guideline lies in that a cartoon-like restoration typically injects meaningful information into the local variance estimator, which finally transfers into the spatial adaption of the regularization parameter. In our experiments, $\alpha^0 = 2.5 \times 10^{-3}$ seems universally good for all examples. In particular, our choice of α^0 will be illustrated for the denoising example in Fig. 3.

All experiments reported in this section were performed under MATLAB R2013b. The image intensity is scaled to the interval [0, 1] in our computation. The displayed images will be quantitatively compared with respect to their peak signal-to-noise ratios (PSNR) and the structural similarity measures (SSIM); see Table 2. In all examples, the "best" scalar regularization parameter $\hat{\alpha}$ is selected via a bisection procedure, up to a relative error of 0.02, to maximize the following weighted sum of the PSNR- and SSIM values of the resulting scalar-α restoration

$$\frac{\text{PSNR}(\alpha)}{\max\{\text{PSNR}(\tilde{\alpha}) : \tilde{\alpha} \in I\}} + \frac{\text{SSIM}(\alpha)}{\max\{\text{SSIM}(\tilde{\alpha}) : \tilde{\alpha} \in I\}}$$

over the interval $I = [10^{-5}, 10^{-3}]$. The maximal PSNR and SSIM in the above formula are pre-computed up to a relative error of 0.001.

4.2.1 Choices of $\overline{\sigma}$ and $\underline{\sigma}$

Assuming that the noise level σ is known or estimated beforehand, the local variance bounds $\overline{\sigma}$ and $\underline{\sigma}$ can be chosen as follows. Let $\chi^2(n^2_{(\mathrm{w})})$ denote the Chi-squared distribution with $n^2_{(\mathrm{w})}$ degrees of freedom. Ideally, if $u = (\mu I + K^*K)^{-1}(\operatorname{div}\mathbf{p} + K^*f)$ is equal to the true image, then the local variance estimator $R(\operatorname{div}\mathbf{p}) = w * |Ku - f|^2$ follows the (scaled) Chi-squared distribution componentwise (see [23]), i.e., for each $(i, j) \in \Omega_h$ we have

$$R(\operatorname{div}\mathbf{p})_{(i,j)} \sim \frac{\sigma^2}{n^2_{(\mathrm{w})}} \chi^2(n^2_{(\mathrm{w})}). \tag{4.2}$$

This motivates our selection of the local variance bounds. In the following, we describe two variants of the local variance bounds based on Chi-squared statistics. Both of them will be tested through our numerical experiments.

First choice of $\overline{\sigma}$ and $\underline{\sigma}$ Ignoring certain dependencies of the random variables, our first local variance bounds are based on extreme value estimation (in the sense of Gumbel, see[26]). In fact, Gumbel's theory allows to describe the statistical distribution of the maximum and minimum values of a finite number of random variables. Within the theory, the asymptotic distribution of the maximal and minimal values is determined, and hence, the larger the sample, the more accurate the description becomes. In light of this approach, the upper bound $\overline{\sigma}$ was previously established in [23]. Under conditions analogous to the ones in [23], here we derive the value of the lower bound $\underline{\sigma}$ and argue that the choice of $\overline{\sigma}$ is also proper in the setting where the localized residual is enforced to the interval $[\underline{\sigma}, \overline{\sigma}]$.

Let $\mathfrak{f}$ be the probability density function of $\chi^2(n^2_{(\mathrm{w})})$ and $\mathfrak{F}$ denote its cumulative distribution function, i.e., $\mathfrak{F}(T) := \int_{-\infty}^{T} \mathfrak{f}(z)\mathrm{d}z$. The maximum and minimum values of $N := n_1 n_2$ observations of independent and identically distributed $\chi^2(n^2_{(\mathrm{w})})$-random variables are, respectively, denoted by $T_{\max}$ and $T_{\min}$. Following Gumbel (see [26], eq. 31' on p. 133 and eq. '31 on p. 135 or [27]), the limiting distributions of the maximum and minimum value $\mathfrak{f}_{\max}$ and $\mathfrak{f}_{\min}$ are given by

$$\mathfrak{f}_{\max}(y_{\max}(T_{\max})) = N\mathfrak{f}(\tilde{T}_{\max})e^{-y_{\max}(T_{\max})-e^{-y_{\max}(T_{\max})}},$$
$$\mathfrak{f}_{\min}(y_{\min}(T_{\min})) = N\mathfrak{f}(\tilde{T}_{\min})e^{y_{\min}(T_{\min})-e^{y_{\min}(T_{\min})}},$$

where $\tilde{T}_{\min}$ and $\tilde{T}_{\max}$ are the "dominant values" defined as $\mathfrak{F}(\tilde{T}_{\min}) := 1/N$ and $\mathfrak{F}(\tilde{T}_{\max}) := 1 - 1/N$. Further, $y_{\max}(\cdot)$

and $y_{\min}(\cdot)$ represent the standardizations (of $T_{\max}$ and $T_{\min}$) defined by

$$y_{\max}(T) := N\mathfrak{f}(\tilde{T}_{\max})(T - \tilde{T}_{\max}),$$
$$y_{\min}(T) := N\mathfrak{f}(\tilde{T}_{\min})(T - \tilde{T}_{\min}).$$

The cumulative distributions $\mathfrak{F}_{\max}(T) := \mathrm{P}(T_{\max} \le T)$ and $\mathfrak{F}_{\min}(T) := \mathrm{P}(T_{\min} \le T))$ satisfy

$$\mathrm{P}(T_{\max} \le T) = e^{-e^{-y_{\max}(T)}}, \mathrm{P}(T_{\min} \le T) = 1 - e^{-e^{y_{\min}(T)}},$$

see eq. 32' on p. 133 and eq. '32 on p. 135 in [26] or [27]. The corresponding expectations ($\mathfrak{E}$) and standard deviations ($\mathfrak{d}$) for $y_{\max}(T_{\max})$ and $y_{\min}(T_{\min})$ are given by

$$\mathfrak{E}(y_{\max}(T_{\max})) = \kappa, \qquad \mathfrak{d}(y_{\max}(T_{\max})) = \frac{\pi}{\sqrt{6}},$$
$$\mathfrak{E}(y_{\min}(T_{\min})) = -\kappa, \qquad \mathfrak{d}(y_{\min}(T_{\min})) = \frac{\pi}{\sqrt{6}},$$

where $\kappa \simeq 0.577215$ is the Euler–Mascheroni constant (see [26], p. 141). It follows from the standardizations of $T_{\max}$ and $T_{\min}$ that

$$\mathfrak{E}(T_{\max}) = \tilde{T}_{\max} + \frac{\kappa}{N\mathfrak{f}_{\max}(\tilde{T}_{\max})}, \quad \mathfrak{d}(T_{\max}) = \frac{\pi}{\sqrt{6}N\mathfrak{f}_{\max}(\tilde{T}_{\max})},$$
$$\mathfrak{E}(T_{\min}) = \tilde{T}_{\min} + \frac{\kappa}{N\mathfrak{f}_{\min}(\tilde{T}_{\min})}, \quad \mathfrak{d}(T_{\min}) = \frac{\pi}{\sqrt{6}N\mathfrak{f}_{\min}(\tilde{T}_{\min})}.$$

It can be straightforwardly proven (see [23]) that

$$\mathrm{P}(T_{\max} \le \mathfrak{E}(T_{\max}) + \mathfrak{d}(T_{\max})) = e^{-e^{-k-\frac{\pi}{\sqrt{6}}}} \simeq 0.86,$$

and analogously, since $y_{\min}(\mathfrak{E}(T_{\min}) - \mathfrak{d}(T_{\min})) = -\kappa - \pi/\sqrt{6}$, we have that

$$\mathrm{P}(T_{\min} \ge \mathfrak{E}(T_{\min}) - \mathfrak{d}(T_{\min}))$$
$$= 1 - \mathrm{P}(T_{\min} \le \mathfrak{E}(T_{\min}) - \mathfrak{d}(T_{\min}))$$
$$= 1 - (1 - e^{-e^{-k-\frac{\pi}{\sqrt{6}}}}) \simeq 0.86.$$

Furthermore, although it is not possible to obtain closed-form expressions for $\mathrm{P}(T_{\max} \le \mathfrak{E}(T_{\min}) - \mathfrak{d}(T_{\min}))$ and $P(T_{\min} \ge \mathfrak{E}(T_{\max}) + \mathfrak{d}(T_{\max}))$, it is obtained computationally that these two quantities are almost zero in the range given by $N = 16^2, 32^2, \ldots, 1024^2$ and $n_{(\mathrm{w})} = 3, 4, \ldots, 11$. This implies that

$$\mathrm{P}(\mathfrak{E}(T_{\min}) - \mathfrak{d}(T_{\min}) \le T \le \mathfrak{E}(T_{\max}) + \mathfrak{d}(T_{\max})) \simeq 0.86,$$

for $T = T_{\min}$ or $T = T_{\max}$.

Based on the above derivation and (4.2), our first selection of the local variance bounds is given as follows

$$\overline{\sigma}^2_{(\mathrm{l})} := \frac{\sigma^2}{n^2_{(\mathrm{w})}}(\mathfrak{E}(T_{\max}) + \mathfrak{d}(T_{\max})), \quad \underline{\sigma}^2_{(\mathrm{l})} := \frac{\sigma^2}{n^2_{(\mathrm{w})}}(\mathfrak{E}(T_{\min}) - \mathfrak{d}(T_{\min})). \tag{\#1}$$

Second choice of $\overline{\sigma}$ and $\underline{\sigma}$ Our second choice of the local variance bounds is based on mean and variance estimation. It is known that the mean and the standard deviation of $\chi^2(n^2_{(\mathrm{w})})$ can be, respectively, calculated as

$$\mathfrak{E}(\chi^2(n^2_{(\mathrm{w})})) = n^2_{(\mathrm{w})}, \qquad \mathfrak{d}(\chi^2(n^2_{(\mathrm{w})})) = \sqrt{2}n_{(\mathrm{w})}.$$

Based on this information, one can choose the local variance bounds as

$$\begin{cases} \overline{\sigma}^2_{(\mathrm{t})} := \mathfrak{E}\left(\frac{\sigma^2}{n^2_{(\mathrm{w})}}\chi^2(n^2_{(\mathrm{w})})\right) + \mathfrak{d}\left(\frac{\sigma^2}{n^2_{(\mathrm{w})}}\chi^2(n^2_{(\mathrm{w})})\right) = \sigma^2\left(1 + \frac{\sqrt{2}}{n_{(\mathrm{w})}}\right), \\ \underline{\sigma}^2_{(\mathrm{t})} := \mathfrak{E}\left(\frac{\sigma^2}{n^2_{(\mathrm{w})}}\chi^2(n^2_{(\mathrm{w})})\right) - \mathfrak{d}\left(\frac{\sigma^2}{n^2_{(\mathrm{w})}}\chi^2(n^2_{(\mathrm{w})})\right) = \sigma^2\left(1 - \frac{\sqrt{2}}{n_{(\mathrm{w})}}\right). \end{cases} \tag{\#2}$$

4.3 Experiments on Denoising

We first test our method on a denoising problem. The observed image is generated by adding Gaussian white noise

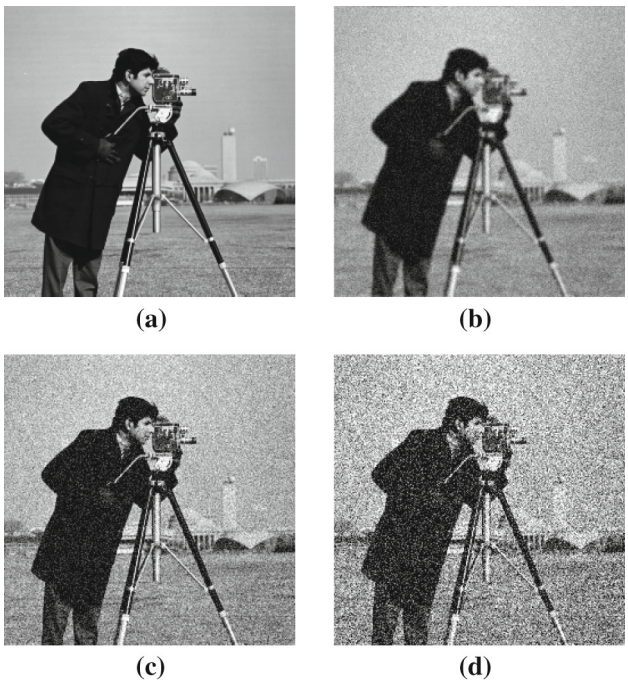

Fig. 1 "Cameraman" image. **a** True image. **b** Noisy blurry image. **c** Noisy image ($\sigma = 0.1$). **d** Noisy image ($\sigma = 0.2$)

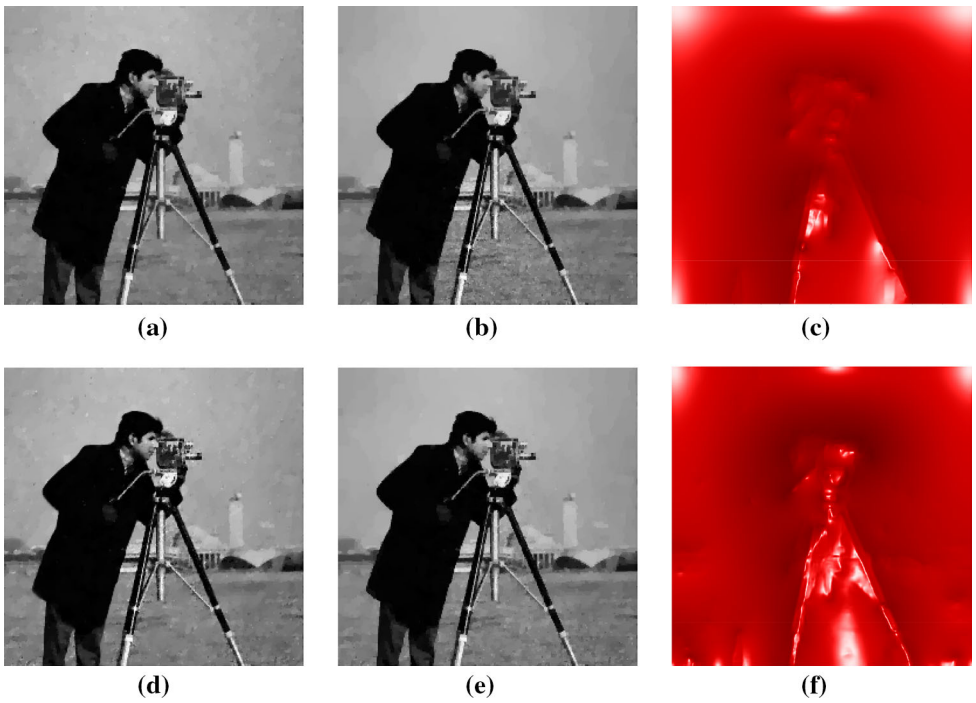

Fig. 2 Denoising: $\sigma = 0.1$. **a** Restor. via $\hat{\alpha} = 2.641\text{e}{-4}$. **b** Restor. via bilevel-(#1). **c** α via bilevel-(#1). **d** Restor. via SATV. **e** Restor. via bilevel-(#2). **f** α via bilevel-(#2)

of standard deviation 0.1 to the test image "Cameraman"; see subplots a and c in Fig. 1. We test our bilevel method with two different local variance bounds in (#1), i.e., $\underline{\sigma}^2_{(\mathrm{l})} = 0.00325$ and $\overline{\sigma}^2_{(\mathrm{l})} = 0.02211$, and in (#2), i.e., $\underline{\sigma}^2_{(\mathrm{t})} = 0.00798$ and $\overline{\sigma}^2_{(\mathrm{t})} = 0.01202$, which are, respectively, referred to as "bilevel-(#1)" and "bilevel-(#2)" in what follows. In Fig. 2, the corresponding restored images and the spatially variant regularization parameters are displayed. These results are compared with the restoration via the best scalar $\hat{\alpha} = 2.641 \times 10^{-4}$, as well as the restoration via the spatially adaptive total variation approach (SATV) [23].

Subplot a in Fig. 2 indicates that the scalar $\hat{\alpha}$ can not simultaneously recover, to visual satisfaction, the detail regions (e.g., where the camera and the tripod are placed) and the homogenous regions (e.g., the background sky). The SATV restoration yields significant improvement in this respect. Our bilevel restorations in subplots b and e are visually even better, especially in the homogenous regions. Comparing b and e, we observe that the tighter bounds given by (#2) tend to capture more information from the image and yield a slightly better restored image. According to a quantitative comparison in Table 2, the bilevel approaches are always superior to the best scalar $\hat{\alpha}$ with respect to PSNR and SSIM. Compared with SATV, the bilevel approaches lose in PSNR but are better in SSIM.

We note that the α-plots in c and f are reversely scaled for visualization purposes (i.e., a peak in the α-plot indicates small value of α at the point), and similarly for all forthcoming α-plots in Sect. 4. Notably, one can observe patterns in the spatial distribution of α from our bilevel approach. In both subplots c and f, α tends to be small in the detailed regions while being large in the homogenous regions. This explains why the restorations in b and e are superior to the one via the best scalar-valued $\hat{\alpha}$.

We also illustrate the evolution of α^k and u^k along the iterations of the projected gradient algorithm in Fig. 3. As instructed by the guideline at the end of Sect. 4.1, the initial guess α^0 produces a cartoon-like image u^0. As the iterations proceed, it is observed that α^k reveals more and more apparent spatial pattern, and correspondingly the restoration becomes sharper and sharper. The final α^k and u^k after 21 iterations are, respectively, given by subplots f and e in Fig. 2.

To conclude the denoising example, we increase the noise level, i.e., $\sigma = 0.2$, and repeat the above experiment. In this case, the local variance bounds from (#1) and (#2) are given by $\underline{\sigma}^2_{(\mathrm{l})} = 0.01302$, $\overline{\sigma}^2_{(\mathrm{l})} = 0.08843$, $\underline{\sigma}^2_{(\mathrm{t})} = 0.03192$, $\overline{\sigma}^2_{(\mathrm{t})} = 0.04808$. The corresponding results are shown in Fig. 4. From these results, a general observation is that detection of spatial patterns in α becomes more challenging as the noise level increases. For relatively loose bounds such as $\underline{\sigma}^2_{(\mathrm{l})}$ and $\overline{\sigma}^2_{(\mathrm{l})}$, the pattern in the spatially variant α becomes less significant. On the other hand, artifacts due to strong noise tend to appear in α via relatively tight bounds such as $\underline{\sigma}^2_{(\mathrm{t})}$ and $\overline{\sigma}^2_{(\mathrm{t})}$. Nevertheless, the restorations via the bilevel approaches seem never worse off than the restorations via scalar $\hat{\alpha}$ or SATV, both visually and quantitatively.

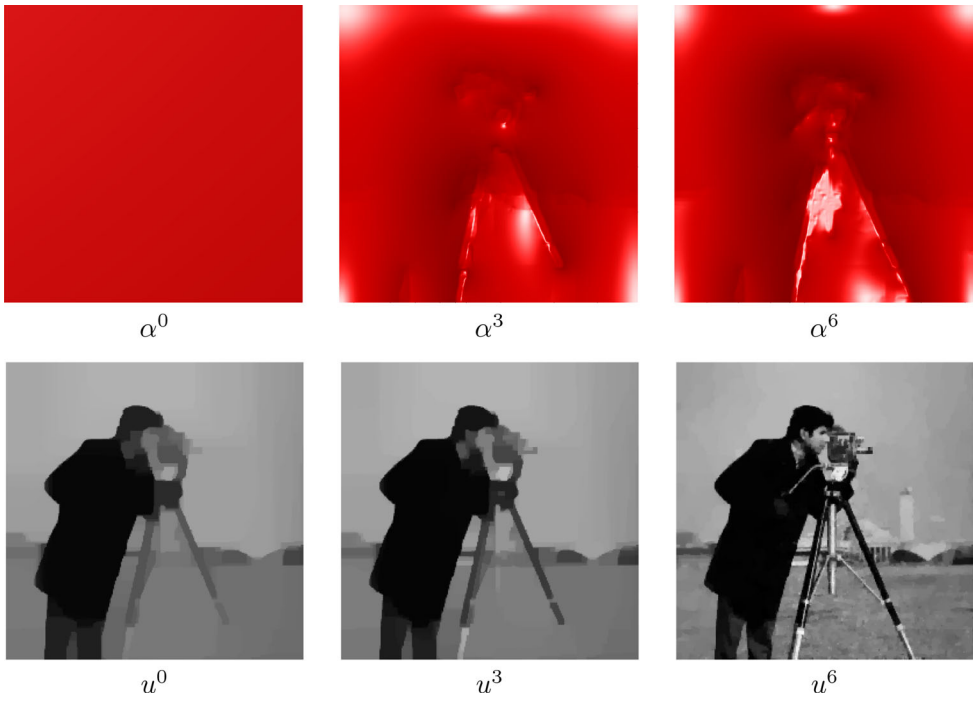

Fig. 3 Evolution of α^k and u^k in bilevel-(#2)

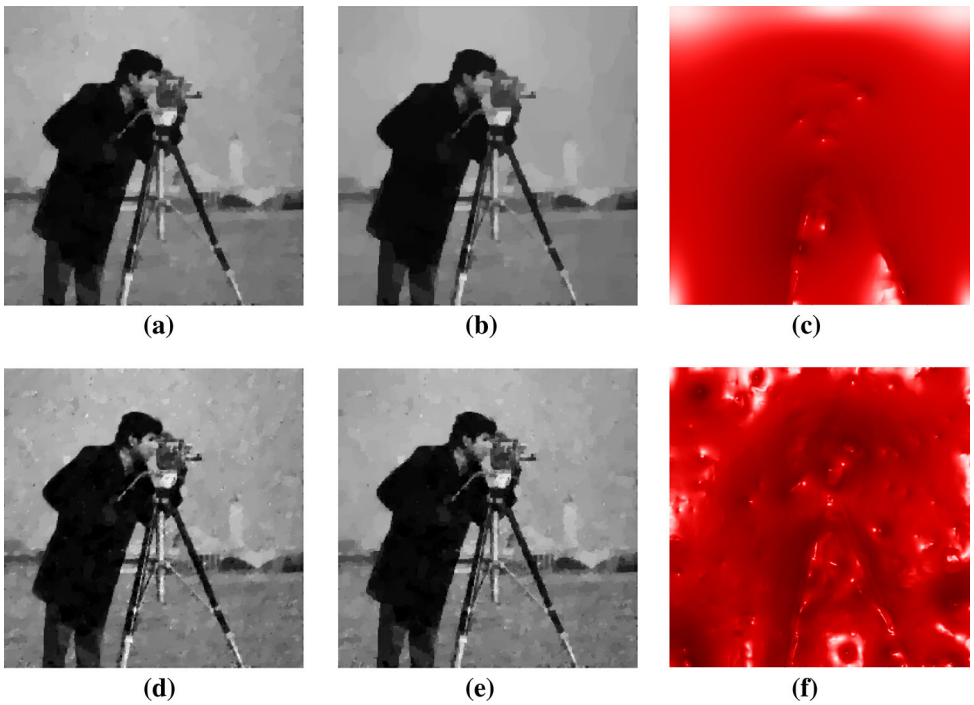

Fig. 4 Denoising: $\sigma = 0.2$. **a** Restor. via $\hat{\alpha} = 6.493\text{e}{-4}$. **b** Restor. via bilevel-(#1). **c** α via bilevel-(#1) **d** Restor. via SATV. **e** Restor. via bilevel-(#2). **f** α via bilevel-(#2)

4.4 Experiments on Deblurring

We continue our experiments by deblurring the "Cameraman" image. Here the image is blurred by Gaussian blur of standard deviation 1 and then degraded by Gaussian white noise of standard deviation 0.05; see Fig. 1b. Again, we have implemented both bilevel-(#1) and bilevel-(#2), where the local variance bounds are given by $\underline{\sigma}^2_{(\mathrm{l})} = 0.000814$, $\overline{\sigma}^2_{(\mathrm{l})} = 0.005527$, $\underline{\sigma}^2_{(\mathrm{t})} = 0.001995$, and $\overline{\sigma}^2_{(\mathrm{t})} = 0.003005$. In Fig. 5, the resulting images and α's are displayed. These results are compared with the restorations via the best scalar $\hat{\alpha} = 4.698 \times 10^{-5}$ and via SATV. In view of subplots c and f, the spatially variant regularization parameters obtained in deblurring share similar patterns to the ones in denoising, particularly in the regions of the camera and the tripod. Both bilevel-(#1) and bilevel-(#2) seem to outperform the

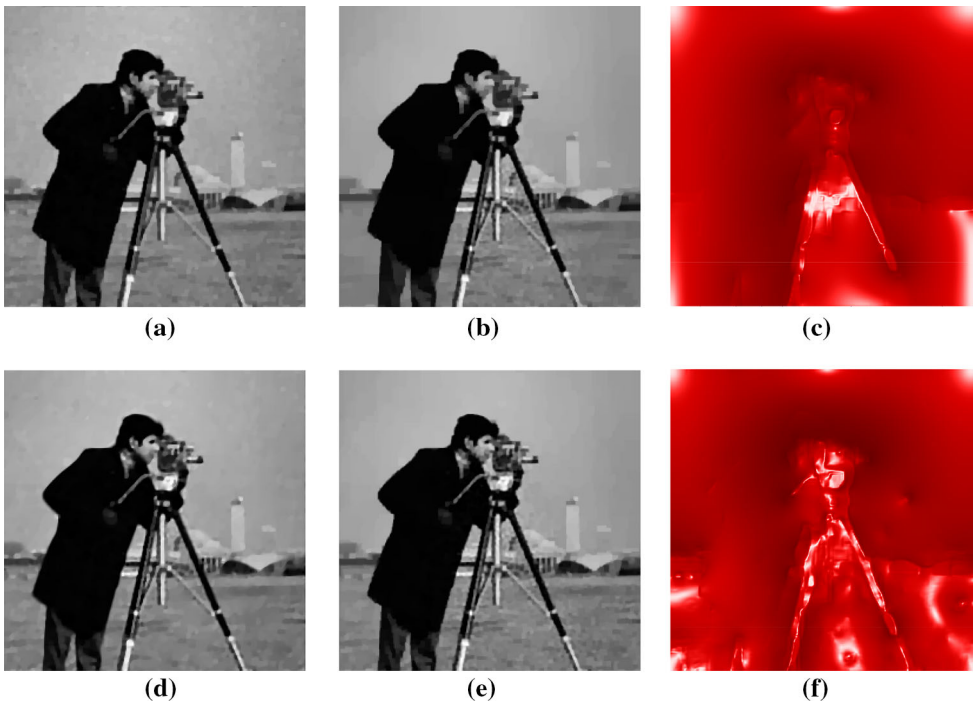

Fig. 5 Deblurring. **a** Restor. via $\hat{\alpha} = 4.698\text{e}{-5}$. **b** Restor. via bilevel-(#1). **c** α via bilevel-(#1). **d** Restor. via SATV. **e** Restor. via bilevel-(#2). **f** α via bilevel-(#2)

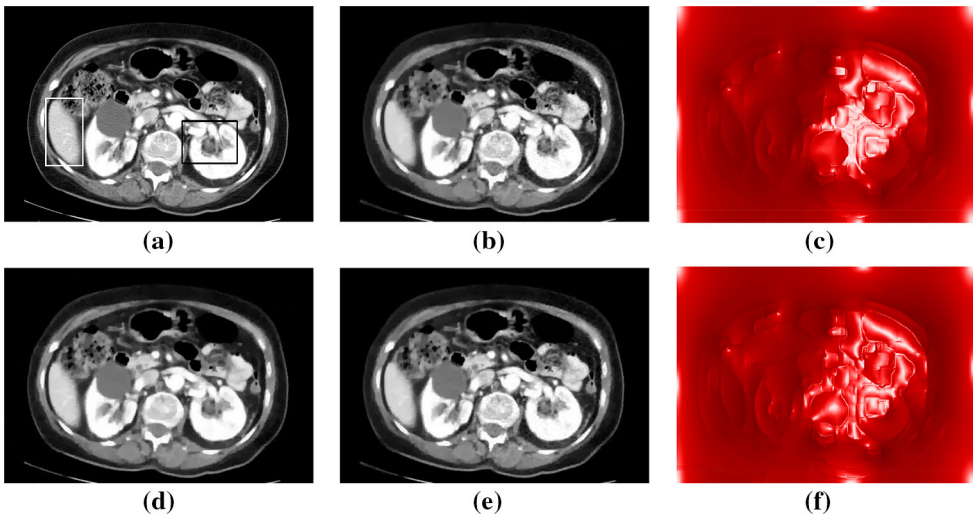

Fig. 6 Fourier inpainting: "Chest." **a** "Chest" image. **b** Restor. via bilevel-(#1). **c** α via bilevel-(#1). **d** Restor. via $\hat{\alpha} = 6.978\text{e}{-5}$. **e** Restor. via bilevel-(#2). **f** α via bilevel-(#2)

best scalar $\hat{\alpha}$ in PSNR and SSIM; see Table 2. Note that the blurring operator has a dampening effect on the artifacts contained in the image. In this circumstance, bilevel-(#2) with tighter local variance bounds is typically more favorable than bilevel-(#1).

4.5 Experiments on Fourier Inpainting

Now we consider Fourier inpainting (restoration with missing samples in the Fourier domain), which is typically encountered in parallel magnetic resonance imaging. For the test image "Chest" in Fig. 6(a), the corresponding data f are generated as $f = K(u+\eta)$. Here K is defined by $K = S \circ F$, where F is the 2D discrete Fourier transform and S is a subsampling operator which collects Fourier coefficients along 120 radial lines centered at zero frequency. Since the subsampled Fourier data are typically *non-uniformly* distributed, the local variance estimator $R(\operatorname{div} \mathbf{p})$ is computed as a 1D convolution, i.e., w is a 1D averaging filter of size $n^2_{(\mathrm{w})}$, and $|Ku - f|^2 \in \mathbb{R}^{|\Omega_\omega|}$ is aligned *lexicographically* as a 1D vector and then convolved with w. Besides, $\eta \in \mathbb{R}^{|\Omega_h|}$ is Gaussian white noise of standard deviation 0.05. In contrast to denoising and deblurring, here the acquired data f are coded in the frequency domain rather than the image domain. This renders the SATV method [23] inapplicable to Fourier inpainting.

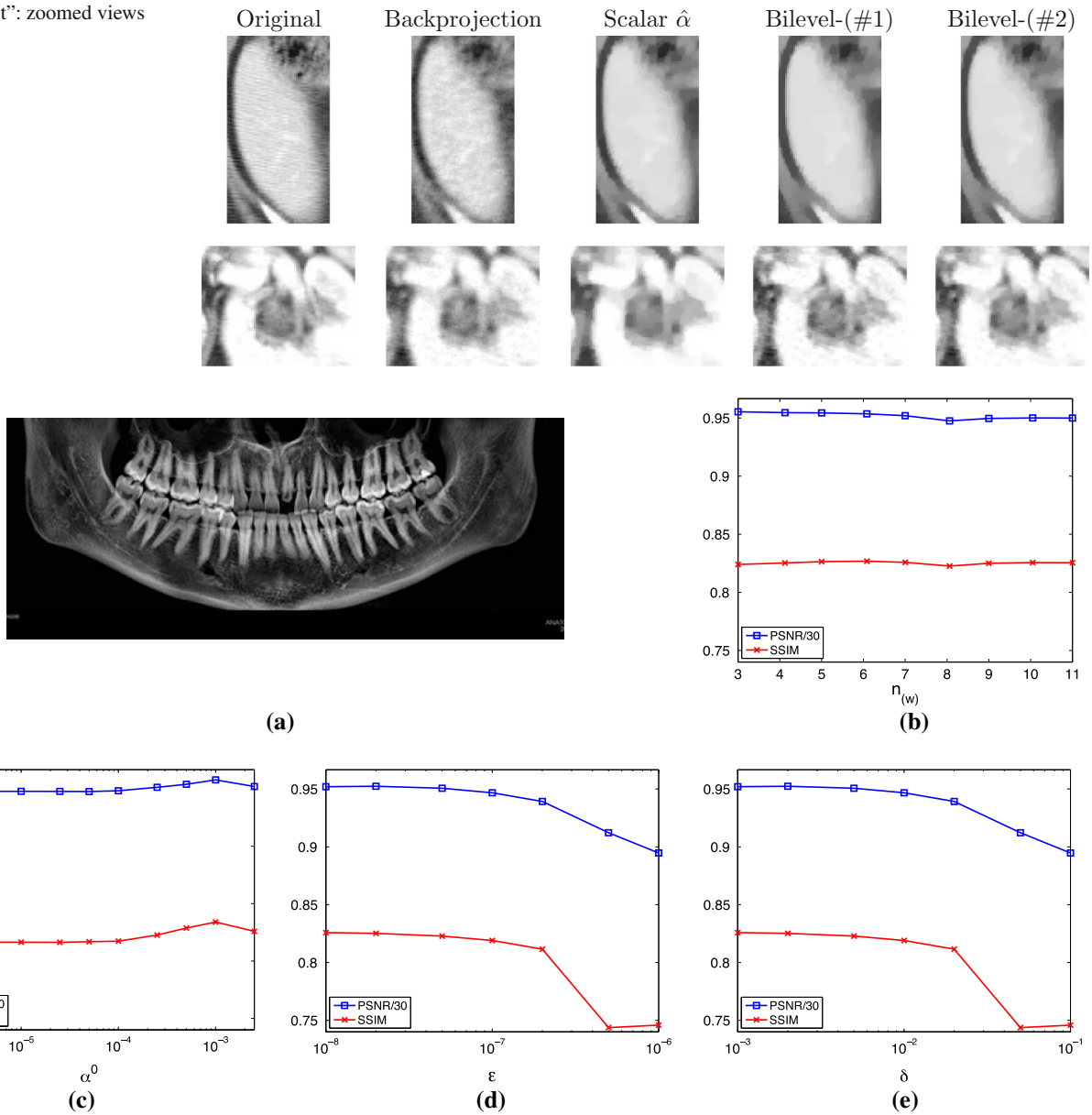

Fig. 7 "Chest": zoomed views

Fig. 8 Sensitivity tests on "Teeth." **a** "Teeth" image. **b** Sensitivity on $n_{(w)}$. **c** Sensitivity on α^0. **d** Sensitivity on ϵ. **e** Sensitivity on δ

The results via bilevel-(#1) and bilevel-(#2) are displayed in Fig. 6, where the corresponding local variance bounds are given by $\underline{\sigma}^2_{(l)} = 0.00077$, $\overline{\sigma}^2_{(l)} = 0.00570$, $\underline{\sigma}^2_{(t)} = 0.00199$, $\overline{\sigma}^2_{(t)} = 0.00301$. It is observed that the spatially distributed α's tend to be small in the regions of interest and large in the backgrounds. For comparison, we also display the restorations via scalar $\hat{\alpha}$; see subplot (c). To highlight the differences among various restorations, we take zoomed views on two framed regions in the "Chest" image; see Fig. 7 for visual comparison. Favorably, the spatial distribution of α allows to handle both local features properly, i.e., homogenize the flat region while preserving the detailed region, which is not attainable by either backprojection or scalar-valued $\hat{\alpha}$.

We also test on another medical image "Teeth"; see Fig. 8a, under the same settings as in "Chest." Similar conclusions can be drawn as before. In addition, we perform sensitivity tests on various parameters in bilevel-(#2) for the "Teeth" example, namely $n_{(w)}$, α^0, ϵ, δ, and λ. Here the parameter $n_{(w)}$ determines the window size in the local variance estimator, α^0 is a scalar which initializes the search for a spatially distributed α, ϵ controls the penalty term in the lower-level problem, δ contributes to the smoothing of the max-function, and λ weights the H^1-regularization on α.

Figure 8 reports the sensitivity measured by PSNR and SSIM. We remark that in general the choice of the window size represents a tradeoff: Small windows typically reduce the

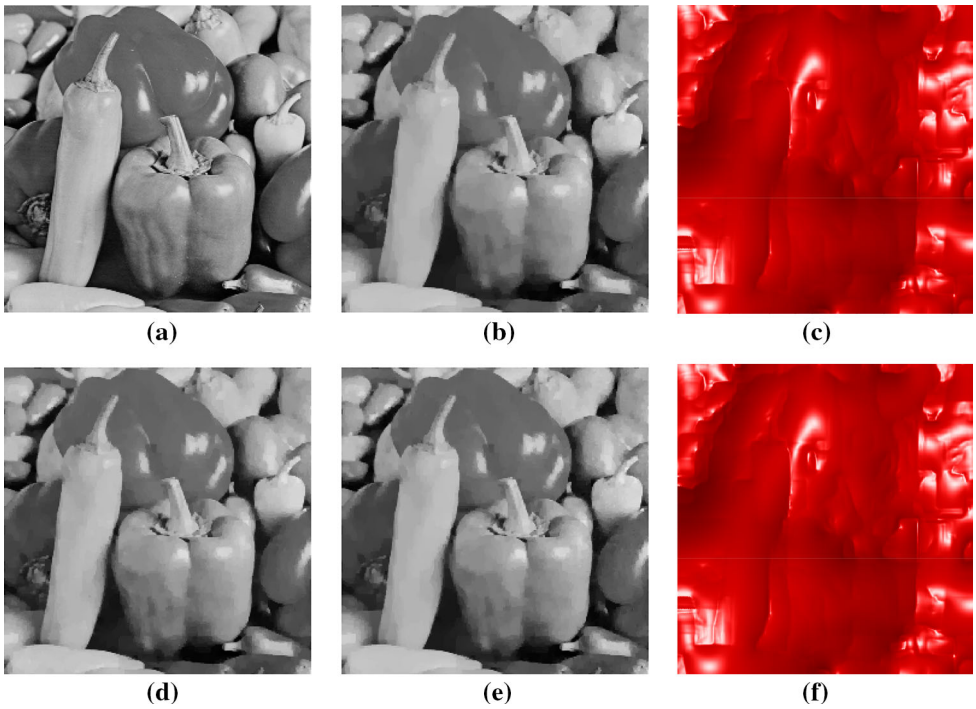

Fig. 9 Wavelet inpainting: "Pepper." **a** "Pepper" image. **b** Restor. via bilevel-(#1). **c** α via bilevel-(#1). **d** Restor. via $\hat{\alpha} = 1.334\text{e}{-4}$. **e** Restor. via bilevel-(#2). **f** α via bilevel-(#2)

reliability of the local variance statistics, while large windows render the local variance less "localized." Observed from subplot b, however, our bilevel approach appears quite stable with respect to the window size in view of PSNR and SSIM. Concerning the initialization of α, as remarked at the end of Sect. 4.2, the bilevel approach benefits from relatively large initial α which yields a blocky initial restoration. This identifies with the test results reported in subplot c. Besides, we observe from the numerical tests that the bilevel approach is almost invariant, in terms of PSNR and SSIM, to λ in the range $[10^{-7}, 10^{-5}]$. In contrast, the parameters ϵ and δ may significantly affect the restoration in case they are too large; see subplots d and e. The present parameters $\epsilon = 10^{-8}$ and $\delta = 10^{-3}$ are chosen to be sufficiently small so that there would be little marginal gain from any further reduction of ϵ or δ.

4.6 Experiments on Wavelet Inpainting

We conclude this section by a wavelet inpainting (restoration with missing samples) problem on the "Pepper" image; see Fig. 9. Our task is to "inpaint" the missing Haar wavelet coefficients due to lossy image transmission or communication; see [13,14] for more background information. The given data are generated by $f = K(u + \eta)$. Here η is Gaussian white noise of standard deviation 0.05, and K is defined by $K = S \circ W$ with the Haar wavelet transform W and the operator S which randomly collects 80% of the wavelet coefficients. Note that the data f are coded in the (wavelet) transform domain rather than the original image domain. Thus, analogous to Fourier inpainting, the local variance estimator $R(\operatorname{div} \mathbf{p})$ is computed as a 1D convolution.

In this example, we set $\tau^0 = 10^{-5}$ for bilevel-(#1) and bilevel-(#2). The local variance bounds in (#1) and (#1) are given by $\underline{\sigma}^2_{(\mathrm{l})} = 0.00081$, $\overline{\sigma}^2_{(\mathrm{l})} = 0.00553$, $\underline{\sigma}^2_{(\mathrm{t})} = 0.00199$, $\overline{\sigma}^2_{(\mathrm{t})} = 0.00301$. Their restorations, together with the restoration from scalar $\hat{\alpha}$, are reported in Fig. 9. The spatially adapted α's via bilevel-(#1) and bilevel-(#2) are also shown in subplots c and f, respectively. Although the three restorations in b, d, e are visually close to each other, the bilevel restorations are superior in PSNR but less good in SSIM according to Table 2.

5 Conclusion

The choice of the regularization parameter for total variation-based image restoration remains a challenging task. At the expense of solving a bilevel optimization problem, this paper generalizes and "robustifies" the classical TV-model by considering a spatially variant regularization parameter α. In particular, an upper-level objective based on local variance estimators is proposed. The overall bilevel model is solved by a projected gradient-type algorithm and yields competitive numerical results in comparison with existing methods. In fact, the reconstructions are almost always better in PSNR or SSIM than those obtained from scalar regularization. Moreover, visually, image details get better preserved and homogeneous regions better denoised for distributed regularization than for scalar one.

Potential future research may include alternative choices for the upper-level objectives, although the statistics-based variance corridors proposed in this work operate satisfactorily. From an analytical point of view, either passage to the limit with the lower-level regularization parameter or employing set-valued analysis tools would be of interest in order to obtain sharp stationarity conditions for the original bilevel formulation. Moreover, the framework may be generalized to other types of priors (such as Total Generalized Variation (TGV)) or alternative noise types (such as random-valued impulse noise). Also, the local adaptation of the filter (e.g., by adjusting the window size according to some confidence criterion) is of interest.

References

1. Adams, R.A., Fournier, J.J.F.: Sobolev Spaces. Volume 140 of Pure and Applied Mathematics. Elsevier/Academic Press, Amsterdam (2003)
2. Almansa, A., Ballester, C., Caselles, V., Haro, G.: A TV based restoration model with local constraints. J. Sci. Comput. **34**(3), 209–236 (2008)
3. Athavale, P., Jerrard, R., Novaga, M., Orlandi, G.: Weighted TV minimization and applications to vortex density models. Technical report, University of Pisa, Department of Mathematics, (2015)
4. Attouch, H., Buttazzo, G., Michaille, G.: Variational analysis in Sobolev and BV spaces. MPS-SIAM, (2006)
5. Barbu, V.: Optimal control of variational inequalities. Res, vol. 100. Notes Math. Pitman, London, United Kingdom (1984)
6. Bertalmio, M., Caselles, V., Rougé, B., Solé, A.: TV based image restoration with local constraints. J. Sci. Comput. **19**, 95–122 (2003)
7. Bertsekas, D.P.: On the Goldstein-Levitin-Polyak gradient projection method. IEEE Trans. Autom. Control **AC–21**(2), 174–184 (1976)
8. Bertsekas, D.P.: Projected Newton methods for optimization problems with simple constraints. SIAM J. Control Optim. **20**(2), 221–246 (1982)
9. Bertsekas, D.P. Gafni, E.M.: Convergence of a gradient projection method. Report P-121, Laboratory for Information and Decision Systems Report, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA, (1982)
10. Brézis, H.: *Problèmes Unilatéraux*. PhD thesis, Sc. math. Paris VI. 1971., (1972)
11. Bui-Thanh, T., Willcox, K., Ghattas, O.: Model reduction for large-scale systems with high-dimensional parametric input space. SIAM J. Sci. Comput. **30**(6), 3270–3288 (2008)
12. Cao, V.C., De los Reyes, J. C., Schoenlieb, C.B.: Learning optimal spatially-dependent regularization parameters in total variation image restoration. *ArXiv e-prints*, Mar. (2016)
13. Chan, R.H., Yang, J., Yuan, X.: Alternating direction method for image inpainting in wavelet domain. SIAM J. Imaging Sci. **4**, 807–826 (2011)
14. Chan, T.F., Shen, J., Zhou, H.-M.: Total variation wavelet inpainting. J. Math. Imaging Vis. **25**, 107–125 (2006)
15. Chen, K., Dong, Y., Hintermüller, M.: A nonlinear multigrid solver with line Gauss-Seidel-semismooth-Newton smoother for the Fenchel pre-dual in total variation based image restoration. Inverse Probl. Imaging **5**(2), 323–339 (2011)
16. Chipot, M.: Variational Inequalities and Flow in Porous Media. Springer, New York (1984)
17. Combettes, P.L., Pesquet, J.-C.: Proximal splitting methods in signal processing. In: *Fixed-point algorithms for inverse problems in science and engineering*, volume 49 of *Springer Optim. Appl.*, pp.185–212. Springer, New York, (2011)
18. De los Reyes, J.C., Schönlieb, C.-B., Valkonen, T.: Bilevel parameter learning for higher-order total variation regularisation models. *Journal of Mathematical Imaging and Vision*, pages 1–25, (2016)
19. De Los Reyes, J.C., Schönlieb, C.-B., Valkonen, T.: The structure of optimal parameters for image restoration problems. J. Math. Anal. Appl. **434**(1), 464–500 (2016)
20. Deledalle, C.-A., Vaiter, S., Fadili, J., Peyré, G.: Stein Unbiased GrAdient estimator of the Risk (SUGAR) for multiple parameter selection. SIAM J. Imaging Sci. **7**(4), 2448–2487 (2014)
21. Dong, Y., Hintermüller, M., Rincon-Camacho, M.: Automated regularization parameter selection in multi-scale total variation models for image restoration. J. Math. Imaging Vis. **40**(1), 82–104 (2011)
22. Dong, Y., Hintermüller, M., Rincon-Camacho, M.: A multi-scale vectorial l^r-TV framework for color image restoration. Int. J. Comput. Vis. **92**(3), 296–307 (2011)
23. Dong, Y., Hintermüller, M., Rincon-Camacho, M.M.: Automated regularization parameter selection in multi-scale total variation models for image restoration. J. Math. Imaging Vis. **40**(1), 82–104 (2011)
24. Frick, K., Marnitz, P., Munk, A.: Statistical multiresolution Dantzig estimation in imaging: fundamental concepts and algorithmic framework. Electron. J. Stat. **6**, 231–268 (2012)
25. Grisvard, P.: Elliptic Problems in Nonsmooth Domains. Volume 24 of Monographs and Studies in Mathematics. Pitman. Advanced Publishing Program, Boston, MA (1985)
26. Gumbel, E.: Les valeurs extrêmes des distributions statistiques. Ann. Inst. H. Poincaré **5**(2), 115–158 (1935)
27. Gumbel, E.J.: *Statistics of extremes*. Dover Publications, Inc., Mineola, NY, 2004. Reprint of the 1958 original [Columbia University Press, New York; MR0096342]
28. Haber, E., Tenorio, L.: Learning regularization functionals—a supervised training approach. Inverse Probl. **19**(3), 611–626 (2003)
29. Hintermüller, M., Kopacka, I.: Mathematical programs with complementarity constraints in function space: C- and strong stationarity and a path-following algorithm. SIAM J. Optim. **20**(2), 868–902 (2009)
30. Hintermüller, M., Kunisch, K.: Path-following methods for a class of constrained minimization problems in function space. SIAM J. Optim. **17**(1), 159–187 (2006). (electronic)
31. Hintermüller, M., Rautenberg, C.N.: Optimal selection of the regularization function in a generalized total variation model. Part I: Modelling and theory. WIAS Preprint No. 2235, (2016)
32. Hintermüller, M., Rautenberg, C.N.: On the density of classes of closed convex sets with pointwise constraints in Sobolev spaces. J. Math. Anal. Appl. **426**(1), 585–593 (2015)
33. Hintermüller, M., Rincon-Camacho, M.: Expected absolute value estimators for a spatially adapted regularization parameter choice rule in L1-TV-based image restoration. Inverse Probl. **26**(8), 085005 (2010)
34. Hintermüller, M., Surowiec, T.M., Mordukhovich, B.S.: Several approaches for the derivation of stationarity conditions for elliptic MPECs with upper-level control constraints. Math. Program. **146**(1–2), 555–582 (2014)
35. Hintermüller, M., Wu, T.: Bilevel optimization for calibrating point spread functions in blind deconvolution. Inverse Probl. Imaging **9**(4), 1139–1169 (2015)
36. Hinze, M., Pinnau, R., Ulbrich, M., Ulbrich, S.: Optimization with PDE Constraints. Mathematical Modelling: Theory and Applications, volume 23. Springer, New York (2009)
37. Hotz, T., Marnitz, P., Stichtenroth, R., Davies, L., Kabluchko, Z., Munk, A.: Locally adaptive image denoising by a statistical mul-

tiresolution criterion. Comput. Stat. Data Anal. **56**(3), 543–558 (2012)
38. Jalalzai, K.: Regularization of inverse problems in image processing. Ph.D. thesis, Ecole Polytechnique (2012)
39. Kinderlehrer, D., Stampacchia, G.: An introduction to Variational Inequalities and Their Applications. SIAM, Philadelphia (2000)
40. Kunisch, K., Pock, T.: A bilevel optimization approach for parameter learning in variational models. SIAM J. Imaging Sci. **6**, 938–983 (2012)
41. Luo, T., Pang, J.-S., Ralph, D.: Mathematical Programs with Equilibrum Constraints. Cambridge University Press, Cambridge (1996)
42. Nittka, R.: Elliptic and parabolic problems with Robin boundary conditions on Lipschitz domains. Ph.D. thesis, Universität Ulm (2010)
43. Nittka, R.: Quasilinear elliptic and parabolic Robin problems on Lipschitz domains. NoDEA Nonlinear Differ. Equ. Appl. **20**(3), 1125–1155 (2013)
44. Outrata, J., Kocvara, M., Zowe, J.: Nonsmooth Approach to Optimization Problems with Equilibrium Constraints. Nonconvex Optimization and its Applications, vol. 28. Kluwer Academic, Dordrecht (1998)
45. Pesquet, J.-C., Benazza-Benyahia, A., Chaux, C.: A SURE approach for digital signal/image deconvolution problems. IEEE Trans. Signal Process. **57**(12), 4616–4632 (2009)
46. Rodrigues, J.F.: Obstacle Problems in Mathematical Physics. North-Holland, Amsterdam (1987)
47. Schönlieb, C., De Los Reyes, J.C.: Image denoising: learning noise distribution via PDE-constrained optimisation. Inverse Probl. Imaging **7**(4), 1183–1214 (2013)
48. Serrin, J.: Local behavior of solutions of quasi-linear equations. Acta Math. **111**, 247–302 (1964)
49. Showalter, R.E.: Hilbert Space Methods for Partial Differential Equations. (Monographs and Studies in Mathematics.). Pitman, London (1977)
50. Showalter, R.E.: Monotone Operators in Banach Space and Nonlinear Partial Differential Equations. American Mathematical Society, Providence (1997)
51. Tröltzsch, F.: Optimal Control of Partial Differential Equations: Theory, Methods and Applications. Volume 112 of Graduate Studies in Mathematics. American Mathematical Society, Providence (2010). Translated from the 2005 German original by Jürgen Sprekels
52. Zowe, J., Kurcyusz, S.: Regularity and stability for the mathematical programming problem in Banach spaces. Appl. Math. Optim. **5**(1), 49–62 (1979)

Michael Hintermüller is currently the Director of the Weierstrass-Institute for Applied Analysis and Stochastics Berlin and Full Professor at Humboldt-Universität zu Berlin. He received his PhD from the Johannes Kepler University of Linz in Austria and was Assistant and Associate Professor at the Karl-Franzens-University of Graz from 1998 to 2007. He held a Chair in Applied Mathematics at the University of Sussex and in 2008 he accepted a position as a MATHEON-Research Professor at Humboldt-Universität zu Berlin. He is the Speaker of the Einstein Center for Mathematics Berlin, a member of the board of the research center MATHEON, a member of the Young Academy of the Austrian Academy of Sciences, and a member of the Class of SIAM Fellows of 2016. His research interests include modeling, analysis and solver design for problems in mathematical image processing, optimal control of partial differential equations and quasi-variational inequalities, as well as shape and topology optimization.

Carlos N. Rautenberg obtained his Ph.D. in Mathematics at Virginia Tech in 2010 on the topic of optimal filtering and estimation of dynamical systems. Subsequently, he spent four years as postdoc at the Karl-Franzens-Universität Graz in Austria where he worked on Quasi-Variational Inequalities (QVIs) and regularization schemes for ill-posed problems. Since 2014, he is an ECMath Junior Research Group Leader in Numerics and Optimization of Robust Equilibria at Humboldt-Universität zu Berlin. His research interests include: quasi-equilibrium problems, nonlinear PDEs, sensor placement, Riccati equations, and mathematical imaging.

Tao Wu received his Bachelor and Master degrees from the Chinese University of Hong Kong, and his PhD from Karl-Franzens-University of Graz. After his PhD, he has worked at the Humboldt University of Berlin and Weierstrass Institute for Applied Analysis and Stochastics. Since 2016 he is a post-doc at the Technical University of Munich.

Andreas Langer Andreas Langer studied Mathematics at the Johannes Kepler University Linz (Austria) and received his Ph.D. in Linz in 2011. From 2011 to 2014, he worked as a postdoc at the Karl-Franzens-University Graz (Austria). Currently, he is a postdoc at the Chair of Numerical Mathematics for High Performance Computing at the University of Stuttgart (Germany). His research interests include approximation techniques for partial differential equations, variational methods, and subspace correction methods, with applications in image processing.