Weierstraß-Institut
für Angewandte Analysis und Stochastik
Leibniz-Institut im Forschungsverbund Berlin e. V.

Preprint ISSN 2198-5855

Optimal selection of the regularization function in a generalized total variation model. Part II: Algorithm, its analysis and numerical tests

Michael Hintermüller[1,2], Carlos N. Rautenberg[2],Tao Wu[2], Andreas Langer[3]

submitted: March 3, 2016

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

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

[3] Department of Mathematics
University of Stuttgart
Pfaffenwaldring 57
70569 Stuttgart, Germany
E-Mail: langer@mathematik.uni-stuttgart.de

No. 2236
Berlin 2016

2010 *Mathematics Subject Classification.* 94A08, 68U10, 49K20, 49K30, 49K40, 49M37, 65K15.

Key words and phrases. Image restoration, generalized total variation regularization, spatially distributed regularization weight, Fenchel predual, bilevel optimization, variance corridor, projected gradient method, convergence analysis.

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.L. 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 [12] suitable for the present context, performed numerical tests concerning the choice of the upper level objective and the initial choice of α = 2.5$\cdot 10^{-3}$ when solving the bilevel problem. He also provided the original source images used in Figures 4.6 and 4.8.

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

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

Abstract

Based on the generalized total variation model and its analysis pursued in [22], 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.

1. Introduction. The following novel duality based bilevel optimization framework is proposed in [22] 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 $f \in L^2(\Omega)$:

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

where problem $\mathrm{D}(\alpha)$ is given by

$$(\mathrm{D}(\alpha)) \qquad \begin{array}{l} \text{minimize} \quad J_D(\mathbf{p}) := \frac{1}{2}|\operatorname{div}\mathbf{p} + K^* f|_B^2 \quad \text{over } \mathbf{p} \in H_0(\text{div}) \\ \text{s.t. } \mathbf{p} \in \mathbf{K}(\alpha) := \{\mathbf{q} \in H_0(\text{div}) \ : \ |\mathbf{q}(x)|_\infty \leq \alpha(x) \text{ f.a.a. } x \in \Omega\}, \end{array}$$

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 $\Omega \subset \mathbb{R}^\ell$ 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 $\mathbb{R}^\ell$. Here an below $(\cdot,\cdot)_{L^2(\Omega)}$ denotes 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

$$(\mathbf{v},\mathbf{w})_{H_0(\text{div})} := (\mathbf{v},\mathbf{w})_{L^2} + (\text{div}\,\mathbf{v}, \text{div}\,\mathbf{w}),$$

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

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

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. 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 [22] it is argued that $(\text{D}(\alpha))$ is the Fenchel pre-dual problem of the following generalized total variation problem:

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

where $BV(\Omega) := \{u \in L^1(\Omega) : \mathcal{D}u \in \mathbf{M}(\Omega,\mathbb{R}^\ell)\}$, 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 $(\mathbb{P})$ falls into the realm of mathematical programs with equilibrium constraints (MPECs) (in function space); see, e.g., [31,34] for an account of MPECs in $\mathbb{R}^n$, [5,20,24] for infinite dimensional settings, and [25,30,36] 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., [41]) 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 $(\mathbb{P})$ is studied in [22]:

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

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

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

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 section 4.2.1 below for details. Moreover, R

$$R(v)(x) := \int_\Omega w(x,y)\,(KB^{-1}v - (KB^{-1}K^* - I)f)^2(y)\mathrm{d}y \tag{1.2}$$

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 $\operatorname{div}\mathbf{p} = Bu - K^*f$, where u is the solution to (P) (see [22, Theorem 3.4]). This implies that

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

where the right hand side represents a convolved version of the image residual $Ku - f$. The quantity $\mathcal{P}_\delta$ penalizes violations of $\mathbf{p} \in \mathbf{K}(\alpha)$, i.e., $\mathcal{P}_\delta(\cdot, \alpha) : V \to \mathbb{R}_0^+$ is defined as

$$\mathcal{P}_\delta(\mathbf{p}, \alpha) := \int_\Omega \sum_{i=1}^{\ell} \big(G_\delta(-(p_i + \alpha)) + G_\delta(p_i - \alpha)\big)\mathrm{d}x, \tag{1.3}$$

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

$$G_\delta(r) = \begin{cases} \frac{1}{2}r^2 - \frac{\delta}{2}r + \frac{\delta^2}{6}, & r \geq \delta\ ; \\ r^3/6\delta, & r \in (0,\delta)\ ; \\ 0, & r \leq 0\ , \end{cases} \tag{1.4}$$

for $\delta > 0$. For $\delta = 0$, we use $r \mapsto G_0(r) := r^2/2$ for $r \geq 0$ and $G_0(r) := 0$ otherwise.

Utilizing [41], an optimal solution $(\mathbf{p}^*, \alpha^*) \in H_0^1(\Omega)^\ell \times \mathcal{A}_{\mathrm{ad}}$ of $(\tilde{\mathbb{P}})$ can be characterized by an adjoint state (a Lagrange multiplier) $\mathbf{q}^* \in H_0^1(\Omega)^\ell$ such that

$$\begin{aligned} &(J_0'(\operatorname{div}\mathbf{p}^*), \operatorname{div}\mathbf{p}) + \langle -\beta\mathbf{\Delta}\mathbf{q}^* + \gamma\mathbf{q}^* + A\mathbf{q}^* \\ &\qquad + \frac{1}{\epsilon}D_1 P_\delta(\mathbf{p}^*, \alpha^*)\mathbf{q}^*, \mathbf{p}\rangle_{H^{-1}, H_0^1} = 0, \end{aligned} \tag{1.5a}$$

$$\langle \lambda(-\Delta + I)\alpha^* + \frac{1}{\epsilon}\left(D_2 P_\delta(\mathbf{p}^*, \alpha^*)\right)^\top \mathbf{q}^*, \alpha - \alpha^*\rangle_{H^1(\Omega)^*, H^1(\Omega)} \geq 0, \tag{1.5b}$$

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

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

see [22, Thm. 6.3].

Besides characterizing stationarity, another benefit of (1.5) 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

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

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

$$\hat{J}'(\alpha) = \lambda(-\Delta + I)\alpha + \frac{1}{\epsilon}\left(D_2 P_\delta(\mathbf{p}(\alpha), \alpha)\right)^\top \mathbf{q}(\alpha), \tag{1.6}$$

where $\alpha \mapsto \mathbf{q}(\alpha)$ solves (1.5a) for $\mathbf{p}^* = \mathbf{p}(\alpha)$ and $\alpha^* = \alpha$.

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.

In order to study regularity properties of the solutions of H^1-projections onto $\mathcal{A}_{\text{ad}}$, in the following section 2 we higher order regularity results for solutions of elliptic variational inequality problems are proven. The projected gradient method is defined in section 3, and global convergence results are established. Section 4.1 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 [28] and [3]. Moreover, spatially adapted regularization or data weighting has been studied in [2, 6, 14, 15, 17, 23, 27]. For a brief discussion of these references we refer to part I of this work; see [22].

2. An obstacle problem and projection results.

The following 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 (3.1). The $H^2(\Omega)$-regularity for a single obstacle and with a C^∞-boundary was established by Brézis in [9]. Similar and related partial results can also be found in the classical texts by Rodrigues [35] and Kinderlehrer and Stampacchia [29]. 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]), and for dimensions $\ell > 3$, the $C^{0,r}(\overline{\Omega})$-regularity is obtained from estimates due to Serrin; see [37].

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 reguarization weight generated by some projection-based descent method.

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

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

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

$$\text{Find } u \in \mathcal{A} : \quad \int_\Omega \nabla u \cdot \nabla(v - u) + (u - f)(v - u)\mathrm{d}x \ge 0, \quad \forall v \in \mathcal{A}. \tag{2.1}$$

In addition u^* solves uniquely:

$$\text{Find } u \in \mathcal{A} \text{ and } \frac{\partial u}{\partial \boldsymbol{\nu}} = 0 \text{ on } \partial\Omega : \quad (Lu - f, v - u)_{H^1(\Omega)^*, H^1(\Omega)} \ge 0, \quad \forall v \in \mathcal{A}, \tag{2.2}$$

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

$$\max(|u^*|_{C^{0,r}(\overline{\Omega})}, |u^*|_{H^2(\Omega)}) \leq C(|f|_{L^2(\Omega)} + |L\underline{\alpha}|_{L^2(\Omega)} + |L\overline{\alpha}|_{L^2(\Omega)}). \tag{2.3}$$

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

$$a(u, w) + (F_\rho(u) - f, w) = 0, \quad \forall w \in H^1(\Omega), \tag{2.4}$$

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

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

Note that (2.4) is the first-order optimality condition for the problem:

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

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 (2.4) is the variational form of a semilinear Neumann problem, i.e., the solution u^*_ρ to (2.4) satisfies

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

see [38, 39] 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 [18] 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 [37], [33] or Theorem 3.1.5 in [32]) 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 $\tilde{C}_2$ independ on f_ρ. Therefore, we have

$$|u^*_\rho|_{H^2(\Omega)} \leq \tilde{C}_1 \left(|f|_{L^2(\Omega)} + \left|\frac{1}{\rho}(u^*_\rho - \overline{\alpha})^+\right|_{L^2(\Omega)} + \left|\frac{1}{\rho}(u^*_\rho - \underline{\alpha})^-\right|_{L^2(\Omega)} \right), \tag{2.5}$$

and

$$\begin{aligned} |u^*_\rho|_{C^{0,r}(\overline{\Omega})} &\leq \tilde{C}_2 \left(|u^*_\rho|_{L^2(\Omega)} + |f|_{L^2(\Omega)} + \left|\frac{1}{\rho}(u^*_\rho - \overline{\alpha})^+\right|_{L^2(\Omega)} + \left|\frac{1}{\rho}(u^*_\rho - \underline{\alpha})^-\right|_{L^2(\Omega)} \right) \\ &\leq 2\max(\tilde{C}_2, \tilde{C}_1) \left(|f|_{L^2(\Omega)} + \left|\frac{1}{\rho}(u^*_\rho - \overline{\alpha})^+\right|_{L^2(\Omega)} + \left|\frac{1}{\rho}(u^*_\rho - \underline{\alpha})^-\right|_{L^2(\Omega)} \right). \end{aligned} \tag{2.6}$$

Note that by Green's theorem, $a(v, w) = (Lv, w)_{H^1(\Omega)^*, H^1(\Omega)} + \int_{\partial\Omega}(\frac{\partial v}{\partial\boldsymbol{\nu}})(w)\mathrm{d}S$. Then, by taking $w = \frac{1}{\rho}(u^*_\rho - \overline{\alpha})^+ \in H^1(\Omega)$ in (2.4), we observe that

$$\frac{1}{\rho}a(u^*_\rho - \overline{\alpha}, (u^*_\rho - \overline{\alpha})^+) + \left|\frac{1}{\rho}(u^*_\rho - \overline{\alpha})^+\right|^2_{L^2(\Omega)} = (f - L\overline{\alpha}, \frac{1}{\rho}(u^*_\rho - \overline{\alpha})^+), \tag{2.7}$$

where we have used that $L\overline{\alpha} \in L^2(\Omega)$, $\partial\overline{\alpha}/\partial\boldsymbol{\nu} = 0$ and $\underline{\alpha} \leq \overline{\alpha}$. Furthermore,

$$a(u_\rho^* - \overline{\alpha}, (u_\rho^* - \overline{\alpha})^+)_{H^1(\Omega)^*,H^1(\Omega)} = \left|(u_\rho^* - \overline{\alpha})^+\right|^2_{L^2(\Omega)} + \left|\nabla(u_\rho^* - \overline{\alpha})^+\right|^2_{L^2(\Omega)^\ell}.$$

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 (2.4), 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 (2.3) 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 (2.4) (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 [13,35] implies that $\tilde{u}$ solves (2.1) 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 semicontinuous, $|\partial\tilde{u}/\partial\boldsymbol{\nu}|_{H^{1/2}(\partial\Omega)} \leq \liminf_{\rho\to 0} \left|\partial u_\rho^*/\partial\boldsymbol{\nu}\right|_{H^{1/2}(\partial\Omega)} = 0$. 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 (2.2), as well. □

Remark 2.1. *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 2.2. *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 $\omega \in H^1(\Omega)$,*

$$P_\mathcal{A}(\omega) := \operatorname*{arg\,min}_{\alpha\in\mathcal{A}} \frac{1}{2}|\alpha - \omega|^2_{H^1(\Omega)}. \tag{2.8}$$

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

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

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 (2.8) 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 varational 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 tht 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., [7,8] for further details.

Recall that our duality result in [22, Thm. 3.4] requires $C(\overline{\Omega})$-regularity of the regularization weight. Below, α_{n+1} represents a suitable approximation. Since it results from an $H^1(\Omega)$-projection, and $H^1(\Omega) \not\hookrightarrow C(\overline{\Omega})$, 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 $\alpha_{n+1} \in C^{0,r}(\overline{\Omega})$ for some $r \in (0,1)$, and thus the required regularity property.

THEOREM 3.1. *Let $\{\alpha_n\}$ be generated by Algorithm 1. Then, $\alpha_n \in H^2(\Omega) \cap C^{0,r}(\overline{\Omega})$ for all $n \in \mathbb{N}$, every limit point α^* of $\{\alpha_n\}$ is stationary for $(\tilde{\mathbb{P}}_{\text{red}})$, i.e., $\alpha^* = P_{\mathcal{A}_{ad}}(\alpha^* - \nabla\hat{J}(\alpha^*))$, and belongs to $H^2(\Omega) \cap C^{0,r}(\overline{\Omega})$. 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 $(\mathbf{p}^*, \alpha^*) \in H_0^1(\Omega)^\ell \times \mathcal{A}_{\text{ad}}$ be a solution to problem $(\tilde{\mathbb{P}})$. Setting $K(\mathbf{p}^*, \alpha^*) :=$

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

$\frac{1}{\epsilon} D_2 P_\delta(\mathbf{p}^*, \alpha^*)$, by [22, Prop. 6.3] (compare (1.5)) there exists an adjoint state $\mathbf{q}^* \in H_0^1(\Omega)^\ell$ satisfying

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

Let $\boldsymbol{G}'_\delta$ be the Nemytskii operator induced (component wise) by $r \mapsto G'_\delta(r) = (r)_\delta^+$. Since $G'_\delta(r) \in C^2(\mathbb{R})$ with $|G''_\delta|_{L^\infty(\mathbb{R})}, |G'''_\delta|_{L^\infty(\mathbb{R})} \leq \max(1, \delta)$ it follows that $K(\mathbf{p}^*, \alpha^*)^\mathrm{T} \mathbf{q}^* \in W^{1,1}(\Omega) \cap L^2(\Omega)$ since $(\mathbf{p}^*, \alpha^*) \in H_0^1(\Omega)^\ell \times H^1(\Omega)$. The application of Theorem 2.1 yields $\alpha^* \in H^2(\Omega) \cap C^{0,r}(\overline{\Omega})$. Given that $L^2(\Omega) \ni \alpha \mapsto \mathbf{p}(\alpha) \in H_0^1(\Omega)^\ell$ is Lipschitz continuous, note also that the map $H^1(\Omega) \ni \alpha \mapsto K(\mathbf{p}(\alpha), \alpha) \in H^1(\Omega) \hookrightarrow L^4(\Omega)^\ell$ for $\ell \leq 4$ is Lipschitz continuous, too (see [22, Prop. 6.2]), 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 [40] 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

$$\langle \hat{J}'(\alpha), \omega \rangle_{H^1(\Omega)', H^1(\Omega)} = \int_\Omega (\lambda(-\Delta\alpha + \alpha) - K(\mathbf{p}(\alpha), \alpha)^\top \mathbf{q}(\alpha)) \omega \mathrm{d}x, \tag{3.3}$$

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 2.2 in appendix 2 yields $P_{\mathcal{A}_{\mathrm{ad}}}(\alpha - \tau \nabla \hat{J}(\alpha)) \in H^2(\Omega) \cap C^{0,r}(\overline{\Omega})$ and that it satisfies homogeneous Neumann boundary conditions. By induction one shows $\alpha_n \in H^2(\Omega) \cap C^{0,r}(\overline{\Omega})$ 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 [8]) 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 [26]). 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 (1.5)) denote the states and adjoint states associated with α_1 and α_2 in $\mathcal{A}_{\mathrm{ad}}$, 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 proven 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 (see the proof of [22, Thm. 5.2]), and analogously as done in the proof of [22, Thm. 5.2], 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.

THEOREM 3.2. *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 $\mu_n = \tau_{n-1}$ and bounded from below, 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 $H^1(\Omega) \ni \alpha \mapsto K(\mathbf{p}(\alpha), \alpha)^{\mathrm{T}} \mathbf{q}(\alpha) \in L^2(\Omega)$, 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

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

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, $\mathbf{p}(\alpha, f) \to 0$ in $H_0^1(\Omega)^\ell$ as $f \downarrow 0$ in $L^2(\Omega)$ by [22, Thm. 5.1] and the remark at the end of the proof. An analogous proof to the one of [22, Thm. 5.1] 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 $L^2(\Omega)$.

If the stepsize τ_n is non-decreasing and bounded below, 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 $\Psi = P_{\mathcal{A}_{\mathrm{ad}}}(\psi - \tilde{\tau} \nabla \hat{J}(\psi))$ and $\Theta = P_{\mathcal{A}_{\mathrm{ad}}}(\theta - \tilde{\tau} \nabla \hat{J}(\theta))$ for some $\psi, \theta \in \mathcal{A}_{\mathrm{ad}}$. Then, using that the projection map $P_{\mathcal{A}_{\mathrm{ad}}}$ is non-expansive, $\nabla \hat{J}(\alpha) = (-\Delta + I)^{-1} \hat{J}'(\alpha) = \mathcal{R}^{-1} \hat{J}'(\alpha)$ 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

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

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

$$\begin{cases} \text{minimize} \quad J(\mathbf{p},\alpha) \quad \text{over } \mathbf{p} \in (\mathbb{R}^{|\Omega_h|})^2, \ \alpha \in \mathcal{A}_{\mathrm{ad}}, \\ \text{s.t. } \ g(\mathbf{p},\alpha) = 0, \end{cases} \tag{4.1}$$

where we set $\Omega_h := \{1,2,...,n_1\} \times \{1,2,...,n_2\}$ and define the mesh size $h := \sqrt{1/(n_1n_2)}$. Assuming constant bounds in $\mathcal{A}_{\mathrm{ad}}$, the discrete admissible set, again denoted by $\mathcal{A}_{\mathrm{ad}}$, is given by

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

The discrete objective reads

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

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-$n_{(\mathrm{w})}$, 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

$$(\operatorname{div}\mathbf{p})_{(j_1,j_2)} = \frac{1}{h}\left(\mathbf{p}^1_{(j_1,j_2)} - \mathbf{p}^1_{(j_1-1,j_2)} + \mathbf{p}^2_{(j_1,j_2)} - \mathbf{p}^2_{(j_1,j_2-1)}\right), \quad \forall (j_1,j_2) \in \Omega_h,$$

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 $\mathbf{\Delta}$ is defined by $\mathbf{\Delta}\mathbf{p} = (\Delta_{(\mathrm{D})}\mathbf{p}^1, \Delta_{(\mathrm{D})}\mathbf{p}^2)$ 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

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

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 $H_0(\operatorname{div})$-norm as

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

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 [21]. 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.

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 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.

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 [16]. 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,

Algorithm 2 Discretized projected gradient method.

```
Require: $\underline{\alpha}, \overline{\alpha}, \overline{\sigma}, \underline{\sigma}, \lambda, \beta, \gamma, \mu, \epsilon, \delta, \tau^0, \mathrm{tol}_{(\mathrm{b})} > 0$, $0 < c < 1$, $0 < \theta_- < 1 \leq \theta_+$, $n_{(\mathrm{w})} \in \mathbb{N}$.
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 $\mathbf{q}^k$:
        $$- \nabla(\mu I + K^*K)^{-1} \operatorname{div} \mathbf{q}^k - \beta \boldsymbol{\Delta} \mathbf{q}^k + \gamma \mathbf{q}^k + \frac{1}{\epsilon} \operatorname{diag}\Big(G_\delta''(\mathbf{p}^k - \alpha^k \mathbf{1}) + G_\delta''(-\mathbf{p}^k - \alpha^k \mathbf{1})\Big)\mathbf{q}^k$$
        $$= \nabla(\mu I + K^*K)^{-1} K^* \operatorname{diag}(Ku^k - f)\Big(w * \big((R(\operatorname{div} \mathbf{p}^k) - \overline{\sigma}^2)^+ - (\underline{\sigma}^2 - R(\operatorname{div} \mathbf{p}^k))^+\big)\Big).$$
    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 $\kappa^k := \Big|P_{\mathcal{A}_{\mathrm{ad}}}(\alpha^k - \nabla \hat{J}(\alpha^k)) - \alpha^k\Big|_{H^1(\Omega)}$.
    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.
```

$\alpha^0 = 2.5 \times 10^{-3}$ seems universally good for all examples. In particular, our choice of α^0 will be illustrated for the denoising example in Figure 4.3.

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 4.1. 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

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

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 $\chi^2(n_{(\mathrm{w})}^2)$ denote the chi-squared distribution with $n_{(\mathrm{w})}^2$ degrees of freedom. Ideally,

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 $|g(\mathbf{p}^l, \alpha; \epsilon^l)|_{H_0(\mathrm{div})^*} \geq \mathrm{tol}_{(\mathrm{l})} \left|g(\mathbf{p}^{\tilde{l}}, \alpha; \epsilon^l)\right|_{H_0(\mathrm{div})^*}$ do
    Compute the Newton step $\delta\mathbf{p}^l$ by solving
        $-\nabla(\mu I + K^*K)^{-1} \operatorname{div} \delta\mathbf{p}^l - \beta\boldsymbol{\Delta}\delta\mathbf{p}^l + \gamma\delta\mathbf{p}^l + \frac{1}{\epsilon^l}\operatorname{diag}\left(G''_\delta(\mathbf{p}^l - \alpha\mathbf{1}) + G''_\delta(-\mathbf{p}^l - \alpha\mathbf{1})\right)\delta\mathbf{p}^l$
        $= -g(\mathbf{p}^l, \alpha; \epsilon^l)$.
    Update $\mathbf{p}^{l+1} := \mathbf{p}^l + \delta\mathbf{p}^l$.
    if $|g(\mathbf{p}^{l+1}, \alpha; \epsilon^l)|_{H_0(\mathrm{div})^*} < \mathrm{tol}_{(\mathrm{l})} \left|g(\mathbf{p}^{\tilde{l}}, \alpha; \epsilon^l)\right|_{H_0(\mathrm{div})^*}$ 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 $\mathbf{p}^l$.
```

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 component-wise (see [16]), 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 [19]). The upper bound $\overline{\sigma}$ was previously established in [16]. Under conditions analogous to the ones in [16], 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 [19], eq. 31' on p. 133 and eq. '31 on p. 135), the limiting distributions of the maximum and minimum value $\mathfrak{f}_{\max}$ and $\mathfrak{f}_{\min}$ are given by

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

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 $r^0 := (I - \Delta_{(\mathrm{N})})(\alpha^0 - \tilde{\alpha}) + \frac{1}{\epsilon_\alpha}\big((\alpha^0 - \overline{\alpha})^+ - (\alpha^0 - \underline{\alpha})^+\big)$.
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
        $\xi^l_j = \begin{cases} 1 & \text{if } \alpha^l_j > \overline{\alpha} \text{ or } \alpha^l_j < \underline{\alpha}, \\ 0 & \text{otherwise.} \end{cases}$
    Update
        $\alpha^{l+1} := \alpha^l + \delta\alpha^l,$
        $r^{l+1} := (I - \Delta_{(\mathrm{N})})(\alpha^{l+1} - \tilde{\alpha}) + \frac{1}{\epsilon_\alpha}\left((\alpha^{l+1} - \overline{\alpha})^+ - (\alpha^{l+1} - \underline{\alpha})^+\right).$
    Set $l := l + 1$.
until $|r^l|_{H^1(\Omega)^*} < \mathrm{tol}_{(\mathrm{p})}\,|r^0|_{H^1(\Omega)^*}$.
Return $\alpha^l$.
```

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

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

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

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

see eq. 32' on p. 133 and eq. '32 on p. 135 in [19]. The corresponding expectations ($\mathfrak{E}$) and standard deviations ($\mathfrak{d}$) for $y_{\max}(T_{\max})$ and $y_{\min}(T_{\min})$ are given by

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

where $\kappa \simeq 0.577215$ is the Euler-Mascheroni constant (see [19], p. 141). It follows from the standardizations of $T_{\max}$ and $T_{\min}$ that

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

It can be straightforwardly proven (see [16]) that

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

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

$$\begin{aligned}\mathrm{P}\left(T_{\min} \geq \mathfrak{E}(T_{\min}) - \mathfrak{d}(T_{\min})\right) &= 1 - \mathrm{P}\left(T_{\min} \leq \mathfrak{E}(T_{\min}) - \mathfrak{d}(T_{\min})\right)\\ &= 1 - (1 - e^{-e^{-k-\frac{\pi}{\sqrt{6}}}}) \simeq 0.86.\end{aligned}$$

Furthermore, although it is not possible to obtain closed-form expressions for $\mathrm{P}\left(T_{\max} \leq \mathfrak{E}(T_{\min}) - \mathfrak{d}(T_{\min})\right)$ and $\mathrm{P}\left(T_{\min} \geq \mathfrak{E}(T_{\max}) + \mathfrak{d}(T_{\max})\right)$, 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

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

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

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

Second choice of $\overline{\sigma}$ and $\underline{\sigma}$. Our second choice of the local variance bounds are 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

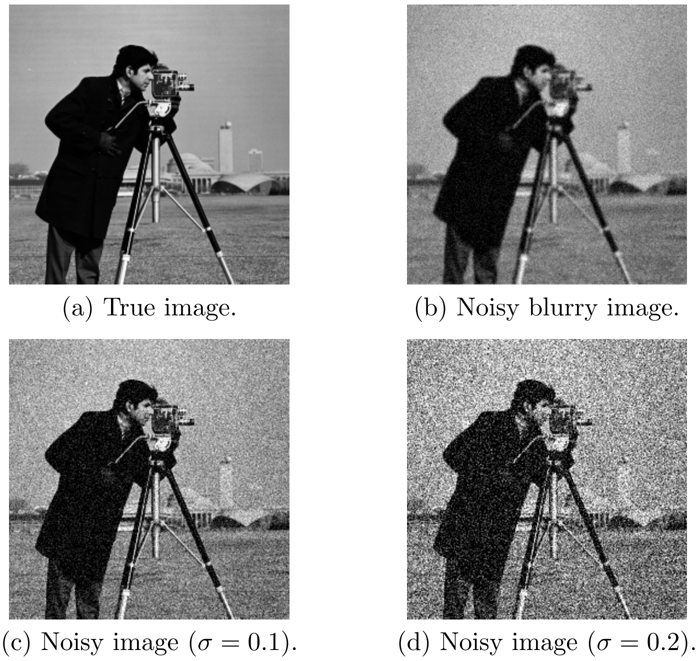
$$\mathfrak{E}(\chi^2(n^2_{(\mathrm{w})})) = n^2_{(\mathrm{w})}, \qquad \mathfrak{d}(\chi^2(n^2_{(\mathrm{w})})) = \sqrt{2}n_{(\mathrm{w})}.$$

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

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

4.3. Experiments on denoising. We first test our method on a denoising problem. The observed image is generated by adding Gaussian white noise of standard deviation 0.1 to the test image “Cameraman”; see subplots (a) and (c) in Figure 4.1. 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 Figure 4.2, 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) [16].

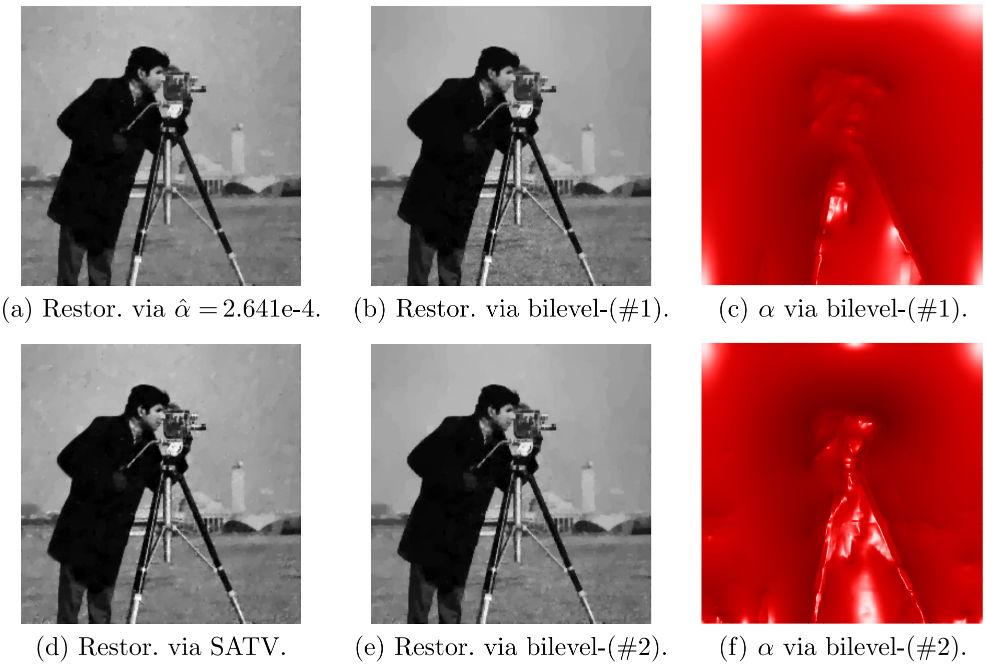
Subplot (a) in Figure 4.2 indicates that the scalar $\hat{\alpha}$ can not simultaneously recover, to visual satisfaction, the detail regions (e.g. where the camera and the tripod are placed) and the homogenous regions (e.g. the background sky). The SATV restoration yields significant improvement in this respect. Our bilevel restorations in subplots



(a) True image. (b) Noisy blurry image.

(c) Noisy image ($\sigma = 0.1$). (d) Noisy image ($\sigma = 0.2$).

Fig. 4.1: "Cameraman" image.



(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).

Fig. 4.2: Denoising: $\sigma = 0.1$.

(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 4.1, 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 section 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 $\hat{\alpha}$.

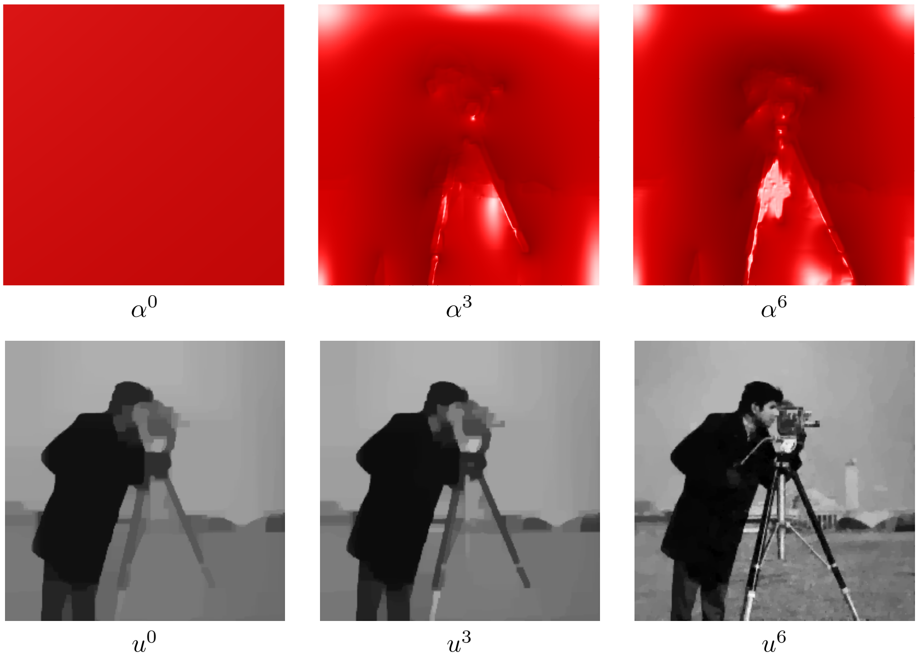
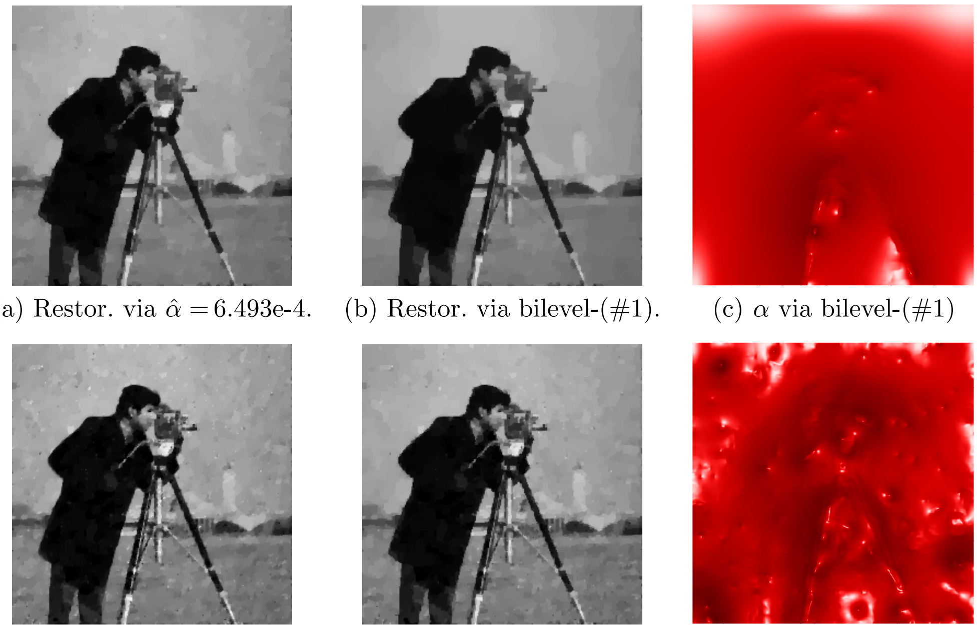


Fig. 4.3: Evolution of α^k and u^k in bilevel-(#2).

We also illustrate the evolution of α^k and u^k along the iterations of the projected gradient algorithm in Figure 4.3. As instructed by the guideline at the end of section 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 Figure 4.2.

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 Figures 4.4. From these results, a general observation is that detection of spatial patterns in α becomes more challenging as the



(a) Restor. via $\hat{\alpha} = $6.493e-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).

Fig. 4.4: Denoising: $\sigma = 0.2$.

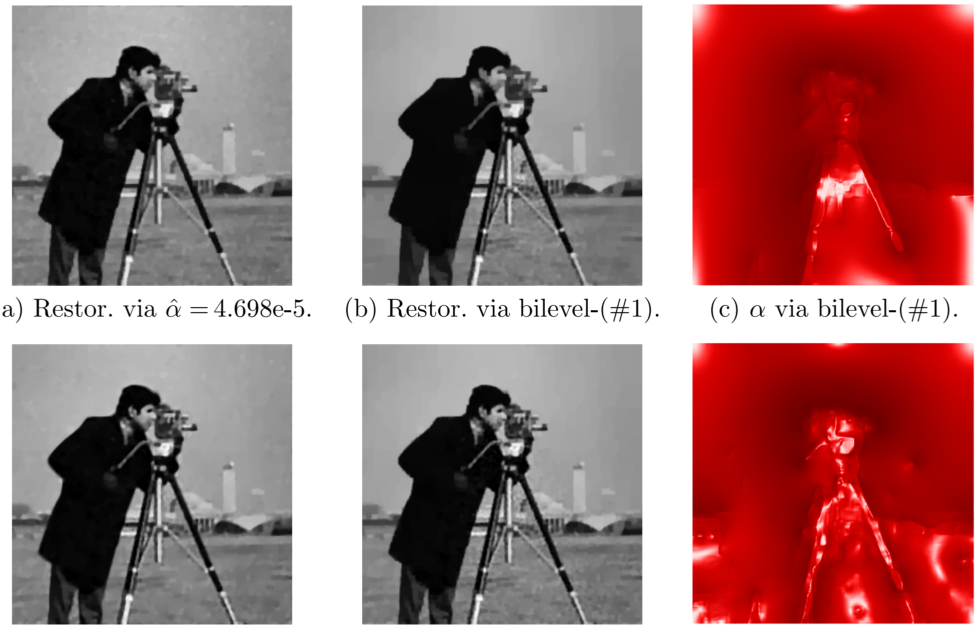
noise level increases. For relatively loose bounds such as $\underline{\sigma}^2_{(l)}$ and $\overline{\sigma}^2_{(l)}$, the pattern in the spatially variant α becomes less significant. On the other hand, artifacts due to strong noise tend to appear in α via relatively tight bounds such as $\underline{\sigma}^2_{(t)}$ and $\overline{\sigma}^2_{(t)}$. Nevertheless, the restorations via the bilevel approaches seem never worse off than the restorations via scalar $\hat{\alpha}$ or SATV, both visually and quantitatively.

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

Table 4.1: Comparison with respect to PSNR and SSIM.

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 Figure 4.1(b). 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 Figure 4.5, 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 best scalar $\hat{\alpha}$ in PSNR and SSIM; see Table 4.1. 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).



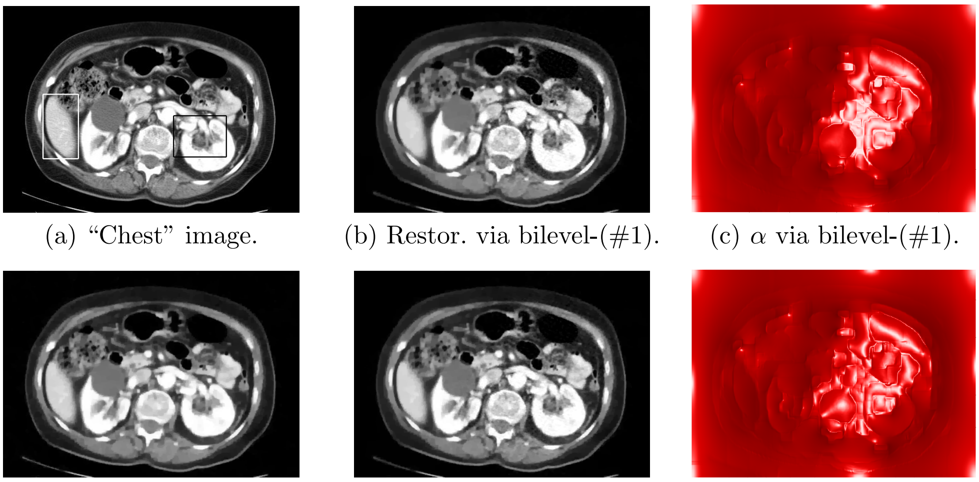
(a) Restor. via $\hat{\alpha} = 4.698\text{e-}5$. (b) Restor. via bilevel-(#1). (c) α via bilevel-(#1).

(d) Restor. via SATV. (e) Restor. via bilevel-(#2). (f) α via bilevel-(#2).

Fig. 4.5: Deblurring.

4.5. Experiments on Fourier inpainting. Now we consider Fourier inpainting, which is typically encountered in parallel magnetic resonance imaging. For the test image "Chest" in Figure 4.6(a), the corresponding data f is 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 is coded in the frequency domain rather than the image domain. This renders the SATV method [16] inapplicable to Fourier inpainting.

The results via bilevel-(#1) and bilevel-(#2) are displayed in Figure 4.6, where the corresponding local variance bounds are given by $\underline{\sigma}^2_{(\mathrm{l})} = 0.00077$, $\overline{\sigma}^2_{(\mathrm{l})} = 0.00570$, $\underline{\sigma}^2_{(\mathrm{t})} = 0.00199$, $\overline{\sigma}^2_{(\mathrm{t})} = 0.00301$. 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 Figure 4.7 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 $\hat{\alpha}$.



(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).

Fig. 4.6: Fourier inpainting: "Chest".

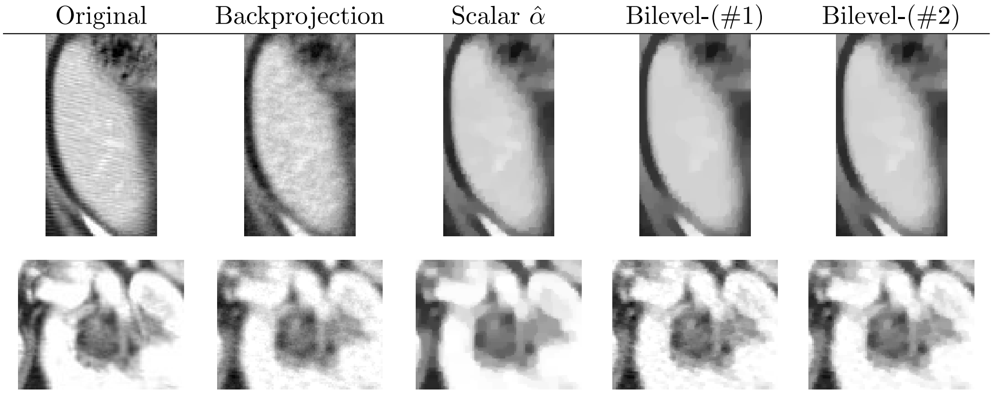
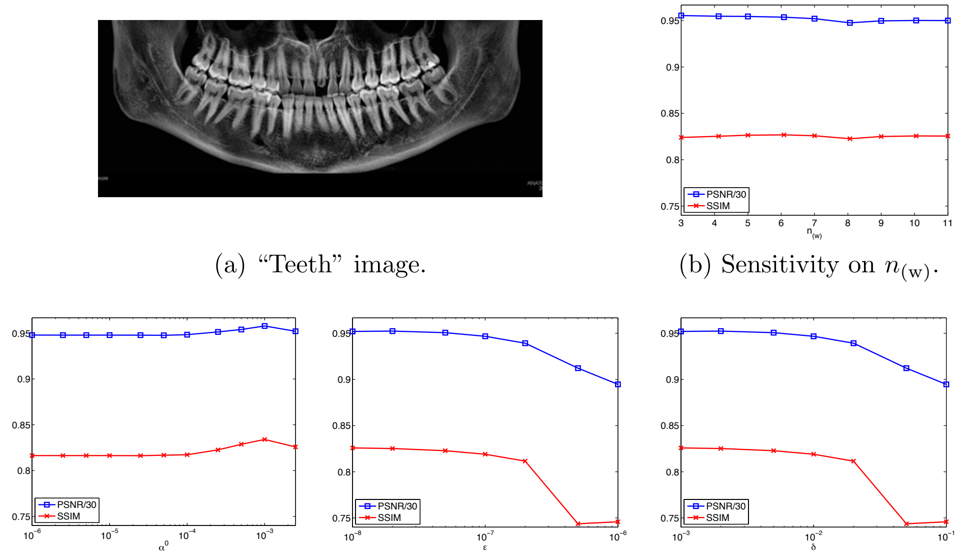


Fig. 4.7: "Chest": zoomed views.

We also test on another medical image "Teeth", see Figure 4.8(a), under the 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 $n_{(\mathrm{w})}$, α^0, ϵ, δ, and λ. Here the parameter $n_{(\mathrm{w})}$ 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 4.8 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 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 section 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.



(a) "Teeth" image.

(b) Sensitivity on $n_{(\mathrm{w})}$.

(c) Sensitivity on α^0.

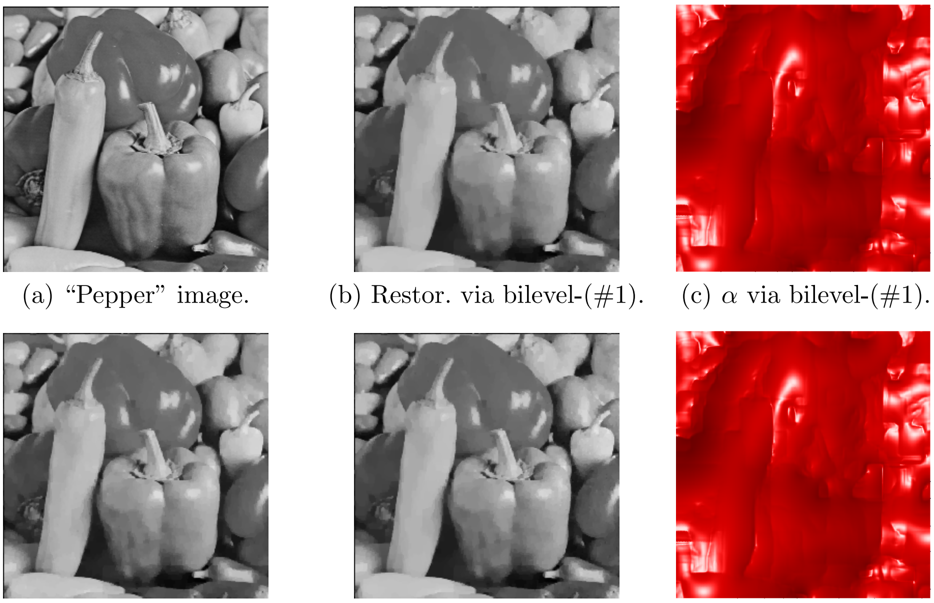
(d) Sensitivity on ϵ.

(e) Sensitivity on δ.

Fig. 4.8: Sensitivity tests on "Teeth".

4.6. Experiments on Wavelet inpainting. We conclude this section by a wavelet inpainting problem on the "Pepper" image; see Figure 4.9. Our task is to "inpaint" the missing Haar wavelet coefficients due to lossy image transmission or communication; see [10, 11] for more background information. The given data is 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 is 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 (#2) 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 Figure 4.9. 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 4.1.



(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).

Fig. 4.9: Wavelet inpainting: "Pepper".

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 to 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, etc.) 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] R. A. Adams and J. J. F. Fournier. *Sobolev Spaces.* Academic Press, second edition, 2003.

[2] A. Almansa, C. Ballester, V. Caselles, and G. Haro. A tv based restoration model with local constraints. *J. Sci. Comput.*, 34(3):209–236, 2008.

[3] P. Athavale, R. Jerrard, M. Novaga, and G. Orlandi. Weighted tv minimization and applications to vortex density models. Technical report, University of Pisa, Department of Mathematics, 2015.

[4] H. Attouch, G. Buttazzo, and G. Michaille. *Variational Analysis in Sobolev and BV Spaces.* MPS-SIAM, 2006.

[5] V. Barbu. *Optimal control of variational inequalities*, volume 100 of *Res. Notes Math.* Pitman, London, United Kingdom, 1984.

[6] M. Bertalmio, V. Caselles, B. Rougé, and A. Solé. Tv based image restoration with local constraints. *J. Sci. Comput.*, 19:95–122, 2003.

[7] D. P. Bertsekas. On the Goldstein-Levitin-Polyak gradient projection method. *IEEE Trans. Automatic Control*, AC-21(2):174–184, 1976.

[8] D. P. Bertsekas and E. M. Gafni. Convergence of a gradient projection method. Report P-121, Laboratory for Information and Decision Systems Report, Laboratory for Information and Decision Syste Massachusetts Institute of Technology, Cambridge, MA, 1982.

[9] H. Brézis. *Problèmes Unilatéraux.* PhD thesis, Sc. math. Paris VI. 1971., 1972.

[10] R. H. Chan, J. Yang, and X. Yuan. Alternating direction method for image inpainting in wavelet domain. *SIAM J. Imaging Sci.*, 4:807–826, 2011.

[11] T. F. Chan, J. Shen, and H.-M. Zhou. Total variation wavelet inpainting. *J. Math. Imaging Vis.*, 25:107–125, 2006.

[12] K. Chen, Y. Dong, and M. Hintermüller. A nonlinear multigrid solver with line Gauss–Seidel-semismooth-Newton smoother for the Fenchel pre-dual in total variation based image restoration. *Inverse Prob. Imaging*, 5:323–339, 2011.

[13] M. Chipot. *Variational Inequalities and Flow in Porous Media.* Springer-Verlag New York Inc., 1984.

[14] Y. Dong, M. Hintermüller, and M. Rincon-Camacho. Automated regularization parameter selection in multi-scale total variation models for image restoration. *Journal of Mathematical Imaging and Vision*, 40(1):82–104, 2011.

[15] Y. Dong, M. Hintermüller, and M. Rincon-Camacho. A multi-scale vectorial l^τ-tv framework for color image restoration. *International Journal of Computer Vision*, 92(3):296–307, 2011.

[16] Y. Dong, M. Hintermüller, and M. M. Rincon-Camacho. Automated regularization parameter selection in multi-scale total variation models for image restoration. *J. Math. Imaging Vision*, 40(1):82–104, 2011.

[17] K. Frick, P. Marnitz, and A. Munk. Statistical multiresolution dantzig estimation in imaging: Fundamental concepts and algorithmic framework. *Electronic Journal of Statistics*, 6:231–268, 2012.

[18] P. Grisvard. Elliptic problems in nonsmooth domains. volume 24 of monographs and studies in mathematics, 1985.

[19] E. Gumbel. Les valeurs extrêmes des distributions statistiques. *Ann. Inst. H. Poincaré*, 5(2):115–158, 1935.

[20] M. Hintermüller and I. Kopacka. 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.

[21] M. Hintermüller and K. Kunisch. Path-following methods for a class of constrained minimization problems in function space. *SIAM J. Optim.*, 17(1):159–187 (electronic), 2006.

[22] M. Hintermüller and C. N. Rautenberg. Optimal selection of the regularization function in a generalized total variation model. Part I: Modelling and theory. Technical report, Humboldt-Univeristät zu Berlin, 2016.

[23] M. Hintermüller and M. Rincon-Camacho. Expected absolute value estimators for a spatially adapted regularization parameter choice rule in l1-tv-based image restoration. *Inverse Problems*, 26(8), 2010.

[24] M. Hintermüller, T. M. Surowiec, and B. S. Mordukhovich. Several approaches for the derivation of stationarity conditions for elliptic mpecs with upper-level control constraints. *Mathematical Programming*, 146(1-2):555–582, 2014.

[25] M. Hintermüller and T. Wu. Bilevel optimization for calibrating point spread functions in blind deconvolution. *Inverse Problems and Imaging*, 9(4):1139–1169, 2015.

[26] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. *Optimization with PDE constraints*, volume 23 of *Mathematical Modelling: Theory and Applications*. Springer, New York, 2009.

[27] T. Hotz, P. Marnitz, R. Stichtenroth, L. Davies, Z. Kabluchko, and A. Munk. Locally adaptive image denoising by a statistical multiresolution criterion. *Computational Statistics and Data Analysis*, 56(3):543–558, 2012.

[28] K. Jalalzai. *Regularization of inverse problems in image processing.* PhD thesis, Ecole Polytechnique, 2012.

[29] D. Kinderlehrer and G. Stampacchia. *An Introduction to Variational Inequalities and Their Applications.* SIAM, 2000.

[30] K. Kunisch and T. Pock. A bilevel optimization approach for parameter learning in variational models. *SIAM Journal on Imaging Sciences*, 6:938–983, 2012.

[31] T. Luo, J.-S. Pang, and D. Ralph. *Mathematical Programs with Equilibrum Constraints.* Cambridge University Press, Cambridge, United Kingdom, 1996.

[32] R. Nittka. *Elliptic and Parabolic Problems with Robin Boundary Conditions on Lipschitz Domains.* PhD thesis, Universität Ulm, 2010.

[33] R. Nittka. Quasilinear elliptic and parabolic Robin problems on Lipschitz domains. *NoDEA Nonlinear Differential Equations Appl.*, 20(3):1125–1155, 2013.

[34] J. Outrata, M. Kocvara, and J. Zowe. *Nonsmooth Approach to Optimization Problems with Equilibrium Constraints*, volume 28 of *Nonconvex Optimization and its Applications.* Kluwer Academic Publishers, Dordrecht, Netherlands, 1998.

[35] J. F. Rodrigues. *Obstacle Problems in Mathematical Physics.* North-Holland, 1987.

[36] C. Schönlieb and J. C. De Los Reyes. Image denoising: Learning noise distribution via pde-constrained optimisation. *Inverse Problems and Imaging*, 7(4):1183–1214, 2013.

[37] J. Serrin. Local behavior of solutions of quasi-linear equations. *Acta Math.*, 111:247–302, 1964.

[38] R. E. Showalter. *Hilbert space methods for partial differential equations.* Pitman, London-San Francisco, Calif.-Melbourne, 1977. Monographs and Studies in Mathematics, Vol. 1.

[39] R. E. Showalter. *Monotone Operators in Banach Space and Nonlinear Partial Differential Equations.* American Mathematical Society, 1997.

[40] F. Tröltzsch. *Optimal control of partial differential equations*, volume 112 of *Graduate Studies in Mathematics.* American Mathematical Society, Providence, RI, 2010. Theory, methods and applications, Translated from the 2005 German original by Jürgen Sprekels.

[41] J. Zowe and S. Kurcyusz. Regularity and stability for the mathematical programming problem in Banach spaces. *Appl. Math. Optim.*, 5(1):49–62, 1979.