SAND2020-4564R

BILEVEL PARAMETER OPTIMIZATION FOR LEARNING NONLOCAL IMAGE DENOISING MODELS

M. D'ELIA[‡], J.C. DE LOS REYES*[§], AND A. MINIGUANO TRUJILLO[§]

Abstract. We propose a bilevel optimization approach for the estimation of parameters in nonlocal image denoising models. The parameters we consider are both the space-dependent fidelity weight and weights within the kernel of the nonlocal operator. In both cases we investigate the differentiability of the solution operator in function spaces and derive a first order optimality system that characterizes local minima. For the numerical solution of the problems, we propose a second-order trust-region algorithm in combination with a finite element discretization of the nonlocal denoising models and we introduce a computational strategy for the solution of the resulting dense linear systems. Several experiments illustrate the applicability and effectiveness of our approach.

1. Introduction. Nonlocal image denoising has emerged in the last years as an important alternative in image processing, due to the fact that it enables the reconstruction of important image features by considering similar intensity patterns between pixels or patches in a given spatial neighbourhood or all over the whole image domain. Although, originally, research was focused on the design of direct nonlocal noise filters [41, 42, 44], more complex approaches based on energy functionals were proposed afterwards for the treatment of the denoising task [21, 22, 29]. This variational framework enabled the employment of additional modeling and analysis tools that have been used for image reconstruction tasks in a partial differential equation (PDE) setting. A similar variational framework is also employed in more recent works (see, e.g., [2]), where the authors use energies induced by (nonlocal) fractional differential operators.

Nonlocal denoising operators are characterized by kernels; the use of different kernels leads to different outcomes, and tuning their parameters is a difficult task. In recent years bilevel optimization has been successfully utilized for the identification of optimal parameters in image processing [12,13,26]; this attempt includes analytical as well as numerical studies, using both finite-dimensional [25,26] and PDE-constrained optimization approaches [12, 13, 24].

In this paper we aim at extending the bilevel optimization methodology to nonlocal operators with integrable kernels. Similar to previous contributions, we consider a supervised learning framework and assume existence of a training set of clean and noisy images we can learn from. Using a variational setting similar to the one developed in [16] and [19], we analyze the differentiability properties of the solution mapping and derive necessary optimality conditions of Karush-Kuhn-Tucker type.

To our knowledge, this is the first paper on bilevel optimization for nonlocal operators. In particular, the second part of the paper addresses the problem of nonlocal kernel identification, now subject of great interest in the nonlocal community, and provides an alternative to neural-networks-based algorithms [33] [**?**]. As such, the impact of this work goes beyond image processing, providing a useful tool in the context of nonlocal optimization and control for a wide range of applications including fracture mechanics [23, 27, 40], anomalous subsurface transport [5, 37, 38], phase transitions [4,14,20], multiscale and multiphysics systems [1,3], magnetohydrodynam-

*Corresponding author, email:juan.delosreyes@epn.edu.ec

‡Sandia National Laboratories, New Mexico, USA.

§Research Center on Mathematical Modelling (MODEMAT), Escuela Politécnica Nacional, Quito, Ecuador

ics [36], and stochastic processes [8, 15, 31, 32].

The paper is organized as follows. In Section 2 we briefly summarize results of the nonlocal vector calculus that will be useful throughout the paper and introduce nonlocal operators for image denoising. In Section 3 we consider a bilevel optimization approach to optimize the spatially dependent fidelity weight for a general denoising problem. In Section 4, we introduce and analyze the problem of finding optimal weights of a *nonlocal means* kernel. Finally, in Section 5, we introduce a second-order optimization algorithm for the solution of the bilevel problems and give insights of implementation aspects and numerical performance. Several numerical tests illustrate our theoretical findings.

2. Preliminaries in nonlocal imaging. Let Ω be a bounded domain in $\mathbb{R}^d$. We use the standard notation $(\cdot,\cdot)_{0,\Omega}$ and $\|\cdot\|_{0,\Omega}$ for the inner product and the norm in $L^2(\Omega)$, the space of square integrable functions on Ω.

2.1. Nonlocal vector calculus. The nonlocal models considered in this paper are analyzed using the nonlocal vector calculus [18]. We recall the basic concepts of such calculus that will be used in this paper. Given the functions $u(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}$ and $\boldsymbol{\nu}(\mathbf{x},\mathbf{y}) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ we let $\boldsymbol{\alpha}(\mathbf{x},\mathbf{y}) = -\boldsymbol{\alpha}(\mathbf{y},\mathbf{x}) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ be an anti-symmetric vector function and $\gamma(\mathbf{x},\mathbf{y}) = \gamma(\mathbf{y},\mathbf{x}) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a symmetric positive kernel, integrable over Ω. We define the nonlocal divergence of $\boldsymbol{\nu}$ as a mapping $\mathcal{D}\boldsymbol{\nu} : \mathbb{R}^d \to \mathbb{R}$ such that

$$\mathcal{D}\boldsymbol{\nu}(\mathbf{x}) := \int_{\mathbb{R}^d} \big(\boldsymbol{\nu}(\mathbf{x},\mathbf{y}) + \boldsymbol{\nu}(\mathbf{y},\mathbf{x})\big) \cdot \boldsymbol{\alpha}(\mathbf{x},\mathbf{y})\, d\mathbf{y}, \qquad \mathbf{x} \in \mathbb{R}^d, \tag{2.1}$$

and the nonlocal gradient of u as a mapping $\mathcal{G}u : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ such that

$$\mathcal{G}u(\mathbf{x},\mathbf{y}) := \big(u(\mathbf{y}) - u(\mathbf{x})\big)\boldsymbol{\alpha}(\mathbf{x},\mathbf{y}), \quad \forall\, \mathbf{x},\mathbf{y} \in \mathbb{R}^d. \tag{2.2}$$

The paper [18, §3.2] shows that the adjoint $\mathcal{D}^* = -\mathcal{G}$, as in the local case. The composition of nonlocal divergence and gradient gives

$$\mathcal{D}\big(\mathcal{G}u\big)(\mathbf{x}) = 2\int_{\mathbb{R}^d} \big(u(\mathbf{y}) - u(\mathbf{x})\big)\big(\boldsymbol{\alpha}(\mathbf{x},\mathbf{y}) \cdot \boldsymbol{\alpha}(\mathbf{x},\mathbf{y})\big)\, d\mathbf{y}.$$

With the identification $\gamma(\mathbf{x},\mathbf{y}) := \boldsymbol{\alpha}(\mathbf{x},\mathbf{y}) \cdot \boldsymbol{\alpha}(\mathbf{x},\mathbf{y})$ we define the nonlocal diffusion of u as the operator $\mathcal{L}u : \mathbb{R}^d \to \mathbb{R}$ such that

$$\mathcal{L}u(\mathbf{x}) := \mathcal{D}\big(\mathcal{G}u\big)(\mathbf{x}) = 2\int_{\mathbb{R}^d} \big(u(\mathbf{y}) - u(\mathbf{x})\big)\, \gamma(\mathbf{x},\mathbf{y})\, d\mathbf{y}, \quad \forall\, \mathbf{x} \in \mathbb{R}^d.$$

Then, we define the interaction domain Ω_I of a bounded region Ω as the set of points outside of the domain that interact with points inside of the domain, i.e.

$$\Omega_I = \{\mathbf{y} \in \mathbb{R}^d \backslash \Omega : \ \gamma(\mathbf{x},\mathbf{y}) \neq 0, \ \text{ for some}\, \mathbf{x} \in \Omega\}.$$

This set is the nonlocal counterpart of the boundary $\partial\Omega$ of a domain in a local setting. In this work we consider localized kernels, i.e. γ is such that for $\mathbf{x} \in \Omega$

$$\begin{cases} \gamma(\mathbf{x},\mathbf{y}) \geqslant 0 & \forall\, \mathbf{y} \in B_\varepsilon(\mathbf{x}), \\ \gamma(\mathbf{x},\mathbf{y}) = 0 & \forall\, \mathbf{y} \in \mathbb{R}^d \backslash B_\varepsilon(\mathbf{x}), \end{cases} \tag{2.3}$$

where $B_\varepsilon(\mathbf{x}) = \{\mathbf{y} \in \mathbb{R}^d : \ |\mathbf{x}-\mathbf{y}|_\infty \leqslant \varepsilon\}$, for all $\mathbf{x} \in \Omega$ and $\varepsilon > 0$ is referred to as interaction radius[1]. For such kernels, we can rewrite the interaction domain as

$$\Omega_I = \{\mathbf{y} \in \mathbb{R}^d \backslash \Omega : \ |\mathbf{y}-\mathbf{x}|_\infty \leqslant \varepsilon, \ \text{ for some } \mathbf{x} \in \Omega\}.$$

We define the nonlocal energy semi-norm, the nonlocal energy space and the constrained nonlocal energy space as follows

$$\begin{aligned} &\|v\|_V^2 := \int_{\Omega\cup\Omega_I}\int_{\Omega\cup\Omega_I} \big(v(\mathbf{x})-v(\mathbf{y})\big)^2 \gamma(\mathbf{x},\mathbf{y})\, d\mathbf{y} d\mathbf{x}, \quad \forall\, v \in L^2(\Omega), \\ &V(\Omega\cup\Omega_I) := \{v \in L^2(\Omega\cup\Omega_I) : \|v\|_V < \infty\}, \\ &V_c(\Omega\cup\Omega_I) := \{v \in V(\Omega\cup\Omega_I) : v|_{\widetilde{\Omega}} = 0\}, \ \text{ for some } \widetilde{\Omega} \subset \Omega_I. \end{aligned} \tag{2.4}$$

The paper [18, §4.3.2] proves that for integrable localized kernels as in (2.3) the constrained energy space $V_c(\Omega\cup\Omega_I)$ is equivalent to

$$L_c^2(\Omega\cup\Omega_I) := \{v \in L^2(\Omega\cup\Omega_I) : v|_{\widetilde{\Omega}} = 0\}$$

and that $\|\cdot\|_V \sim \|\cdot\|_{L^2(\Omega\cup\Omega_I)}$. Unless necessary, we drop the dependence of V and V_c on $\Omega\cup\Omega_I$.

Nonlocal volume constrained problems We consider the solution of nonlocal elliptic problems, i.e. the nonlocal counterpart of elliptic PDEs. Due to nonlocality, when solving a nonlocal problem, boundary conditions (i.e. conditions on the solution for $\mathbf{x} \in \partial\Omega$) do not guarantee the uniqueness of the solution, which can only be achieved by providing conditions on the interaction domain Ω_I [18]. As an illustrative example, we consider the following *nonlocal diffusion-reaction equation* for the scalar function u:

$$-\mathcal{L}u + \lambda u = f \qquad \mathbf{x} \in \Omega, \tag{2.5}$$

for some $f \in L^2(\Omega)$ and $\lambda \in L^\infty(\Omega)$ such that $\lambda : \Omega \to \mathbb{R}^+$. Uniqueness of u is guaranteed provided the following condition is satisfied [18]:

$$u = g \qquad \text{for } \mathbf{x} \in \Omega_I, \tag{2.6}$$

where g is some known function in the *trace space*

$$\widetilde{V}(\Omega_I) = \{z : \exists\, v \in V \ \text{ s.t. } \ v|_{\Omega_I} = z\}.$$

Without loss of generality, in our analysis we consider $g = 0$ so that $u \in V_c(\Omega\cup\Omega_I)$ with $\widetilde{\Omega} = \Omega_I$. The corresponding weak form is obtained in the same way as in the local setting by multiplying (2.5) by a test function and integrating over Ω, i.e.

$$\int_\Omega (-\mathcal{L}u + \lambda u - f)v\, d\mathbf{x} = \int_{\Omega\cup\Omega_I}\int_{\Omega\cup\Omega_I} \mathcal{G}u\, \mathcal{G}v\, d\mathbf{x}\, d\mathbf{y} + \int_\Omega (\lambda u - f)v\, d\mathbf{x} = 0, \tag{2.7}$$

where the equality follows from the nonlocal Green's identity [18]. Note that, by definition of $\mathcal{G}$, (2.7) is equivalent to

$$\int_{\Omega\cup\Omega_I}\int_{\Omega\cup\Omega_I} (u(x)-u(y))(v(x)-v(y))\gamma(\mathbf{x},\mathbf{y})\, d\mathbf{x}\, d\mathbf{y} + \int_\Omega (\lambda u - f)v\, d\mathbf{x} = 0. \tag{2.8}$$

[1]Note that, in general, nonlocal neighborhoods are Euclidean balls. However, the nonlocal calculus still holds for more general balls such as those induced by the ℓ-infinity norm (see an application in [10]).

2.2. Nonlocal denoising formulation. In order to use the nonlocal vector calculus for image denoising models, we consider the variational viewpoint proposed in [21] and study the following kernels: for $\delta > 0$,

- Yaroslavsky kernel:

$$\gamma_1(\mathbf{x},\mathbf{y}) = \exp\left\{-\frac{\big(f(\mathbf{x})-f(\mathbf{y})\big)^2}{\delta^2}\right\}\,\mathcal{X}(\mathbf{y}\in B_\varepsilon(\mathbf{x})),$$

- Nonlocal Means kernel:

$$\gamma_2(\mathbf{x},\mathbf{y}) = \exp\left\{-\frac{1}{\delta^2}\int_{\mathbb{R}^2} w(\mathbf{t})\big(f(\mathbf{x}+\mathbf{t})-f(\mathbf{y}+\mathbf{t})\big)^2\,d\mathbf{t}\right\}\,\mathcal{X}\big(\mathbf{y}\in B_\varepsilon(\mathbf{x})\big),$$

- Combination of the previous two kernels:

$$\begin{aligned}\gamma_C(\mathbf{x},\mathbf{y}) = \exp\Big\{&-\frac{\tau_1}{\delta^2}\big(f(\mathbf{x})-f(\mathbf{y})\big)^2\\ &-\frac{\tau_2}{\delta^2}\int_{\mathbb{R}^2} w(\mathbf{t})\big(f(\mathbf{x}+\mathbf{t})-f(\mathbf{y}+\mathbf{t})\big)^2\,d\mathbf{t}\Big\}\,\mathcal{X}\big(\mathbf{y}\in B_\varepsilon(\mathbf{x})\big),\end{aligned}\quad (2.9)$$

where f is a given noisy image and $\mathcal{X}(\cdot\in B)$ is the indicator function over the set B.

In [7] it is shown that nonlocal means presents advantages in presence of textures or periodic structures, whereas neighborhood filters, e.g. the Yaroslavsky filter, may perform better for the preservation of particular edges. As a consequence, a kernel that considers a combination of both contributions, as in (2.9), may provide an increased denoising capability. We refer to [6] for more details on these and other nonlocal kernels.

For a given kernel function, we formulate the nonlocal denoising problem as the following energy minimization problem.

$$\min_{u\in V_c} J(u,\lambda) = \frac{\mu}{2}\|u\|_V^2 + \frac{1}{2}\int_\Omega \lambda\,(u-f)^2\,d\mathbf{x}, \quad (2.10)$$

where $f\in L^\infty(\Omega\cup\Omega_I)$ stands for the noisy image and λ is a weight that balances the fidelity term against the nonlocal regularizer. The weight λ can be either a (positive) real number or a spatially dependent quantity.

As an example, for a given noisy image, in Figure 2.1 we report for the nonlocal means kernel the contour lines of a loss function associated with a scalar λ and a scalar weight w. This two-dimensional plot shows the difficulties related to the optimization, in fact, these complex banana-shaped contour lines are a challenge for several minimization algorithms, especially first-order ones.

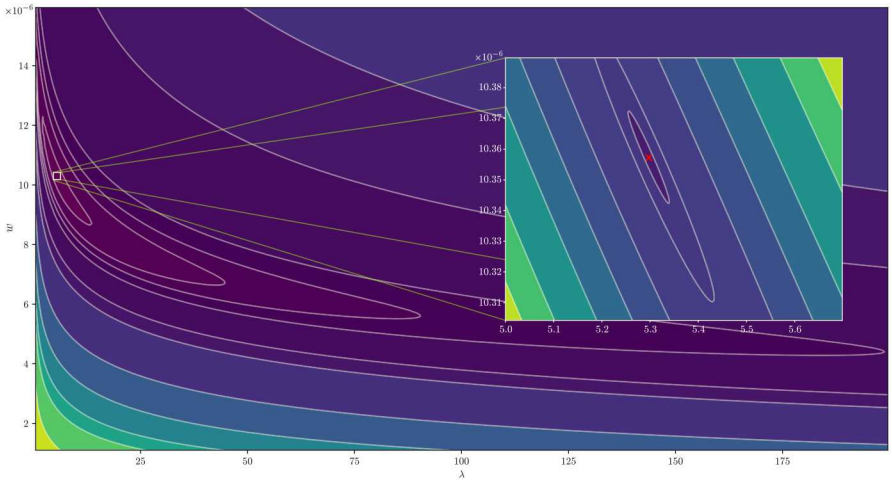


Fig. 2.1: Contour plot of a loss function for different parameters λ and w for the nonlocal means kernel.

3. Optimization with respect to λ. We study the problem of identifying the optimal spatially dependent λ in the "lower-level" denoising model (2.10). First, we analyze the existence of a solution to the lower-level problem with fixed parameters. Then, we state the bi-level problem for the identification of the optimal λ and study the differentiability of the solution operator and the reduced cost functional. We also derive a first order optimality system for the estimation of the optimal parameter. The case $\lambda \in \mathbb{R}^+$ is studied at the end of the section as a particular instance.

3.1. Lower-level problem. We recall the energy-based formulation of the nonlocal denoising problem:

$$\min_{u \in V_c} J(u, \lambda) = \frac{\mu}{2}\|u\|_V^2 + \frac{1}{2}\int_\Omega \lambda(\mathbf{x})(u-f)^2\, d\mathbf{x}, \tag{3.1}$$

where the energy norm $\|\cdot\|_V$ is defined as in (2.4) and, in particular, is induced by the scalar product

$$\begin{aligned}(u,v)_V &= \int_{\Omega\cup\Omega_I}\int_{\Omega\cup\Omega_I} \big(u(\mathbf{x})-u(\mathbf{y})\big)\big(v(\mathbf{x})-v(\mathbf{y})\big)\gamma(\mathbf{x},\mathbf{y})\, d\mathbf{y}d\mathbf{x} \\ &= 2\int_\Omega\int_{\Omega\cup\Omega_I} \big(u(\mathbf{x})-u(\mathbf{y})\big)v(\mathbf{x})\gamma(\mathbf{x},\mathbf{y})\, d\mathbf{y}d\mathbf{x}, \quad \forall\, u, v \in V_c.\end{aligned} \tag{3.2}$$

In what follows we refer to (3.1) as the *lower-level problem* and we study its well-posedness as well as a necessary and sufficient conditions for the characterization of its minima.

THEOREM 3.1. *For every $\lambda \in L^\infty(\Omega)$, such that $\lambda(\mathbf{x}) \geqslant 0$ a.e., there exists a unique solution $u \in V_c$ for the lower-level problem* (3.1).

Proof. Since the functional J is bounded from below, there exists a minimizing sequence $\{u_n\} \subset V_c$. Thanks to the coercivity in V_c of the energy term, the sequence is bounded in V_c; thus, there exists a subsequence, still denoted by $\{u_n\}$, that weakly converges in V_c, i.e. $u_n \rightharpoonup u^*$. Since J is convex and continuous with respect to the

energy norm, it is weak lower semi-continuous. Therefore,

$$J(u^*) \leqslant \lim \inf_{n\to\infty} J(u_n).$$

The uniqueness of the solution follows from the strict convexity of the functional. □

We now consider the parameter space $\mathcal{U} := H^1(\Omega) \cap L^\infty(\Omega)$ and the admissible set $\mathcal{U}_{ad} := \{\eta \in \mathcal{U} : \eta(\mathbf{x}) \geqslant 0 \text{ a.e.}\}$. For $\lambda \in \mathcal{U}_{ad}$, a necessary and sufficient optimality condition for the lower-level problem is given by the nonlocal variational equality

$$\mu(u,\psi)_V + \big(\lambda(u-f),\psi\big)_{0,\Omega} = 0, \quad \forall \psi \in V_c. \tag{3.3}$$

By choosing $\psi = u$ in (3.3) we have

$$\mu\|u\|_V^2 \leqslant \mu\|u\|_V^2 + \int_\Omega \lambda u^2 = \int_\Omega \lambda f u \leqslant \|\lambda\|_{H^1}\|f\|_\infty \|u\|_{0,\Omega}.$$

Using the equivalence of the energy and L^2 norms, we then obtain the following *a-priori* estimate

$$\|u\|_{0,\Omega} \leqslant K\|\lambda\|_{H^1(\Omega)}. \tag{3.4}$$

We will use this estimate in the analysis of the differentiability properties of the solution operator.

3.2. Bilevel problem. We consider the following bilevel optimization problem

$$\begin{aligned} &\min_{\lambda\in\mathcal{C}} \quad \mathcal{J}(u,\lambda) = \ell(u) + \frac{\beta}{2}\|\lambda\|^2_{H^1(\Omega)} \\ &\text{s.t.} \quad u = \arg\min_{u\in V_c} J(u,\lambda) = \frac{\mu}{2}\|u\|_V^2 + \int_\Omega \lambda(u-f)^2\, d\mathbf{x}, \end{aligned} \tag{3.5}$$

where the feasible set is $\mathcal{C} = \big\{\lambda \in H^1(\Omega) : b \geqslant \lambda(\mathbf{x}) \geqslant 0\big\}$ is a subset of the control space $\mathcal{U}$.

The loss function $\ell(u)$ is assumed to be strictly convex and continuous with respect to u. The simplest case corresponds to the Peak Signal-to-Noise Ratio-related loss function $\ell(u) := \frac{1}{2}\|u - u^T\|^2_{0,\Omega}$, which arises from a supervised learning framework, where u^T corresponds to the ground truth image and f to the corrupted one. In such framework, the training set is typically large (i.e. we assume several pairs (u^T, f) are available) and the number of lower-level problems increases accordingly, but analytical difficulties are the same. For this reason we restrict our attention to a single (u^T, f) pair which corresponds to a single lower-level problem. Alternative loss functions based on the image statistics have also been recently proposed [24] and may also been considered in our framework.

THEOREM 3.2. *The bilevel optimization problem* (3.5) *admits a solution* $\lambda \in \mathcal{C}$.

Proof. Since the functional $\mathcal{J}$ is bounded from below, there exists a minimizing sequence $\{\lambda_n\} \subset \mathcal{C}$ such that $\mathcal{J}\big(u(\lambda_n),\lambda_n\big) \to \mathcal{J}\big(u(\lambda^*),\lambda^*\big)$. Also, the Tikhonov term guarantees that this sequence is bounded in $H^1(\Omega)$. Thus, there exists a subsequence, still denoted by $\{\lambda_n\}$, that converges strongly in L^2.

Let $u_n \in V_c$ be the unique (see Theorem 3.1) optimal solution to the lower-level problem (3.1) corresponding to λ_n. From the stability estimate (3.4) we have that

$$\|u_n\|_{0,\Omega} \leqslant K\|\lambda_n\|_{H^1(\Omega)} \leqslant \overline{K},$$

i.e. $\{u_n\}$ is uniformly bounded in V_c. Thus, there exists a subsequence, that we still denote by $\{u_n\}$, that weakly converges in V_c (and L^2_c, because of the equivalence of spaces) to u^*. Next, we show that $u^* = u(\lambda^*)$, i.e. the limit of $\{u_n\}$ is the optimal solution of the lower-level problem corresponding to λ^*. Formally,

$$\mathcal{J}(u^*, \lambda^*) \leqslant \liminf_{n\to\infty} \mathcal{J}(u_n, \lambda_n). \tag{3.6}$$

We treat the first two terms in J as we did in Theorem 3.1 for the lower-level problem. For the third term, we have

$$\int_\Omega \lambda^*(u^* - f)^2 \, d\mathbf{x} \leqslant \liminf_{n\to\infty} \int_\Omega \lambda^*(u_n - f)^2 \, d\mathbf{x},$$

which, because of the strong convergence of λ_n in $L^2(\Omega)$, implies

$$\int_\Omega \lambda^*(u^* - f)^2 \, d\mathbf{x} \leqslant \liminf_{n\to\infty} \int_\Omega \lambda_n(u_n - f)^2 \, d\mathbf{x}.$$

We conclude that

$$\mathcal{J}(u^*, \lambda^*) \leqslant \liminf_{n\to\infty} \mathcal{J}(u_n, \lambda_n).$$

□ The strict convexity of the denoising functional implies that the constraint (3.1) has a unique minimizer and it can be replaced by its necessary and sufficient optimality condition. We obtain the following nonlocal-equation-constrained optimization problem:

$$\min_{\lambda\in\mathcal{C}} \ \ell(u) + \frac{\beta}{2}\|\lambda\|^2_{H^1(\Omega)} \tag{3.7a}$$

$$\text{s.t.} \quad \mu(u, \psi)_V + \big(\lambda(u - f), \psi\big)_{0,\Omega} = 0, \quad \forall \psi \in V_c, \tag{3.7b}$$

for $u \in V_c$. Note that (3.7b) is obtained by taking variations of the lower-level problem; its well-posedness follows from Theorem 3.1 as well as from the coercivity of the bilinear form $(\cdot, \cdot)_V$.

3.3. Differentiability of the Solution Operator. In this section we analyze the differentiability properties of the solution mapping. The presence of the weak norm in the denoising model, does not allow us to obtain Fréchet differentiability results. Fortunately, first-order optimality conditionsonly require Gâteaux differentiability, which is proved in the following theorem.

THEOREM 3.3. *Let $\mathcal{V}$ be an ϵ-neighbourhood containing $\mathcal{C}$ and $S_\lambda \colon \mathcal{V} \longrightarrow V_C$ be the solution operator, which assigns to each λ the corresponding solution to* (3.7b)*. Then, the operator S_λ is Gâteaux differentiable.*

Proof. Let $h \in \mathcal{U}$, and u_t and u be the unique solutions to (3.7b) corresponding to $\lambda + th$ and λ, respectively. For ϵ and t small enough, equation (3.7b) is well-posed. Throughout the proof, we let $C > 0$ denote a generic positive constant.

By taking the difference of the equations corresponding to $\lambda + th$ and λ, we have

$$\mu(u_t - u, \psi)_V + \big((\lambda + th)(u_t - f) - \lambda(u - f), \psi\big)_{0,\Omega} = 0$$

or, equivalently,

$$\mu \int_{\Omega\cup\Omega_I}\int_{\Omega\cup\Omega_I} \big((u_t-u)(\mathbf{x})-(u_t-u)(\mathbf{y})\big)\big(\psi(\mathbf{x})-\psi(\mathbf{y})\big)\gamma(\mathbf{x},\mathbf{y})\,d\mathbf{x}\,d\mathbf{y} + \int_\Omega \lambda(\mathbf{x})(u_t-u)(\mathbf{x})\psi(\mathbf{x})\,d\mathbf{x} + t\int_\Omega h(\mathbf{x})u_t(\mathbf{x})\psi(\mathbf{x})\,d\mathbf{x} = t\int_\Omega h(\mathbf{x})f(\mathbf{x})\psi(\mathbf{x})\,d\mathbf{x}. \tag{3.8}$$

By choosing $\psi = u_t - u$ and since $\lambda \in \mathcal{V}$, we have

$$\begin{aligned}\|u_t-u\|_V^2 &\leqslant C\left|t\int_\Omega h(x)(f(x)-u_t(x))(u_t-u)(x)\,d\mathbf{x}\right| \\ &\leqslant Ct\|h\|_\infty\|f-u_t\|_{0,\Omega}\|u_t-u\|_{0,\Omega},\end{aligned}$$

which implies that

$$\begin{aligned}\|u_t-u\|_{0,\Omega} &\leqslant Ct\|h\|_\infty\big\{\|f\|_{0,\Omega}+\|u_t\|_{0,\Omega}\big\} \\ &\leqslant Ct\|h\|_\infty\big\{\|f\|_{0,\Omega}+K(\|\lambda\|_\infty+t\|h\|_\infty)\big\}.\end{aligned}$$

Therefore, the sequence $\{z_t\}_{t>0}$, with $z_t := (u_t-u)/t$, is bounded and there exists a subsequence (still denoted by $\{z_t\}$) such that $z_t \rightharpoonup z$ weakly in V. From (3.8) we have

$$\int_{\Omega\cup\Omega_I}\int_{\Omega\cup\Omega_I} \mu\,\mathcal{G}\left(\frac{u_t-u}{t}\right)\mathcal{G}\psi + \int_\Omega \lambda\left(\frac{u_t-u}{t}\right)\psi + \int_\Omega h(u_t-u)\psi = -\int_\Omega h(u-f)\psi,$$

which implies that

$$\int_{\Omega\cup\Omega_I}\int_{\Omega\cup\Omega_I} \mu\,\mathcal{G}z_t\,\mathcal{G}\psi + \int_\Omega \lambda z_t\psi + \int_\Omega h(u_t-u)\psi = -\int_\Omega h(u-f)\psi.$$

Taking the limit as $t\to 0$, we obtain

$$\int_{\Omega\cup\Omega_I}\int_{\Omega\cup\Omega_I} \mu\,\mathcal{G}z\,\mathcal{G}\psi + \int_\Omega \lambda\, z\,\psi = -\int_\Omega h(u-f)\,\psi,$$

which implies that

$$\mu\mathcal{L}z + \lambda z = h(f-u).$$

By subtracting the equations for the state and the linearized state we have

$$\mu\left(\frac{u_t-u}{t}-z,\psi\right)_V + \int_\Omega \lambda\left(\frac{u_t-u}{t}-z\right)\psi = -\int_\Omega h(u_t-u)\psi.$$

Finally, by choosing $\psi = \frac{u_t-u}{t} - z$ we obtain

$$\left\|\frac{u_t-u}{t}-z\right\|_V \leqslant C\|h\|_{L^\infty}\|u_t-u\|_{0,\Omega}.$$

The continuity of the solution operator implies that $\left\|\frac{u_t-u}{t}-z\right\|_V \to 0$ as $t\to 0$, which concludes the proof. □

3.4. Optimality system. Thanks to the Gâteaux differentiability of the solution operator we are able to derive an optimality system for the characterization of optimal solutions of (3.7).

THEOREM 3.4. *Let $(u, \lambda) \in V_c \times \mathcal{C}$ be an optimal solution to problem (3.7). There exists an adjoint state $p \in L^2_c(\Omega \cup \Omega_I)$ and Lagrange multipliers $\mu^+_\Omega, \mu^-_\Omega \in L^2(\Omega)$ and $\mu^+_\Gamma, \mu^-_\Gamma \in H^{1/2}(\Gamma)$ such that the following optimality system is satisfied.*

$$\mu(u, \psi)_V + \big(\lambda(u - f), \psi\big)_{0,\Omega} = 0, \forall \psi \in V, \tag{3.9a}$$

$$\mu(p, \phi)_V + (\lambda\, p, \phi)_{0,\Omega} = -(\nabla \ell(u), \phi)_{0,\Omega}, \forall \phi \in V, \tag{3.9b}$$

$$\begin{aligned} -\beta \Delta \lambda + \beta \lambda &= \mu^+_\Omega - \mu^-_\Omega \quad \textit{in } \Omega, \\ \beta \frac{\partial \lambda}{\partial \vec{n}} &= \mu^+_\Gamma - \mu^-_\Gamma \quad \textit{on } \Gamma, \end{aligned} \tag{3.9c}$$

$$\begin{aligned} &0 \leqslant \mu^+_\Omega(\mathbf{x}) \perp \lambda(\mathbf{x}) \geqslant 0, \quad 0 \leqslant \mu^-_\Omega(\mathbf{x}) \perp (b - \lambda(\mathbf{x})) \geqslant 0, \quad \forall \mathbf{x} \in \Omega, \\ &0 \leqslant \mu^+_\Gamma(\mathbf{x}) \perp \lambda(\mathbf{x}) \geqslant 0, \quad 0 \leqslant \mu^-_\Gamma(\mathbf{x}) \perp (b - \lambda(\mathbf{x})) \geqslant 0, \quad \forall \mathbf{x} \in \Gamma. \end{aligned} \tag{3.9d}$$

Proof. Let us consider the reduced cost functional

$$j(\lambda) := \ell(u(\lambda)) + \frac{\beta}{2} \|\lambda\|^2_{H^1}, \tag{3.10}$$

where $u(\lambda)$ is the unique solution to the state equation (3.7b) corresponding to λ. Taking the derivative of the reduced cost with respect to λ, we have

$$j'(\lambda)h = \big(\nabla \ell(u(\lambda)), u'(\lambda)h\big)_{0,\Omega} + \beta(\lambda, h)_{H^1}, \quad \forall h \in \mathcal{U}, \tag{3.11}$$

where $u'(\lambda)h$ is the unique solution of the linearized equation

$$\mu\big(u'(\lambda)h, \psi\big)_V + \big(\lambda u'(\lambda)h, \psi\big)_{0,\Omega} = -\big(h(u - f), \psi\big)_{0,\Omega}, \quad \forall \psi \in V_C. \tag{3.12}$$

Using the adjoint equation

$$\mu(p, \phi)_V + (\lambda p, \phi)_{0,\Omega} = -(\nabla \ell(u), \phi)_{0,\Omega}, \quad \forall \phi \in V_c, \tag{3.13}$$

which is uniquely solvable by the same arguments as in Theorem 3.1, we obtain that

$$j'(\lambda)h = -\mu\big(u'(\lambda)h, p\big)_V - \big(\lambda u'(\lambda)h, p\big)_{0,\Omega} + \beta(\lambda, h)_{H^1}. \tag{3.14}$$

By using the linearized equation we then obtain

$$j'(\lambda)h = \int_\Omega (u - f)ph \; dx + \beta(\lambda, h)_{H^1}. \tag{3.15}$$

The box constraints on the parameter λ imply that the a first order necessary optimality condition is given by the following variational inequality:

$$j'(\lambda)(h - \lambda) = \int_\Omega (u - f)p(h - \lambda) \; dx + \beta(\lambda, h - \lambda)_{H^1} \geqslant 0, \quad \forall h \in \mathcal{C}. \tag{3.16}$$

The latter corresponds to an obstacle problem with bilateral bounds. Integration by parts then yields

$$\begin{aligned}(\lambda, v)_{H^1} &= (\lambda, v)_{0,\Omega} + (\nabla\lambda, \nabla v)_{0,\Omega} \\ &= (\lambda, v)_{0,\Omega} + \int_\Gamma \frac{\partial\lambda}{\partial\vec{n}} v \; d\Gamma - (\Delta\lambda, v)_{0,\Omega} \;\; \forall\, v \in H^1(\Omega),\end{aligned}$$

where the extra regularity, i.e. $\lambda \in H^2(\Omega)$, follows from [43, Thm. 5.2]. Consequently, the variational inequality (3.16) can be written in strong form as

$$\begin{aligned}-\beta\Delta\lambda + \beta\lambda &= \mu_\Omega \quad \text{in } \Omega, \\ \beta\frac{\partial\lambda}{\partial\vec{n}} &= \mu_\Gamma \quad \text{on } \Gamma,\end{aligned}$$

where the multipliers $\mu_\Omega \in L^2(\Omega)$ and $\mu_\Gamma \in H^{1/2}(\Gamma)$ satisfy

$$(\mu_\Omega, v - \lambda) \geqslant 0, \;\forall v \in \mathcal{C} \quad \text{and} \quad (\mu_\Gamma, v - \lambda) \geqslant 0, \;\forall v \in \mathcal{C},$$

or, equivalently,

$$\begin{aligned}\big(\mu_\Omega(\mathbf{x}), v - \lambda(\mathbf{x})\big) \geqslant 0, \quad \forall v \in [0, b] \subset \mathbb{R}, \text{ a.e. in } \Omega \\ \big(\mu_\Gamma(x), v - \lambda(x)\big) \geqslant 0, \quad \forall v \in [0, b] \subset \mathbb{R}, \text{ a.e. in } \Gamma.\end{aligned}$$

By decomposing μ_Ω and μ_Γ in its positive and negative parts, we have

$$\begin{aligned}&\mu_\Omega = \mu_\Omega^+ - \mu_\Omega^-, \\ &\mu_\Omega^+ \geqslant 0, \qquad\qquad \lambda(\mathbf{x}) \geqslant 0, \qquad\qquad \mu_\Omega^+(x)\lambda(x) = 0, \\ &\mu_\Omega^- \geqslant 0, \qquad\qquad \lambda(x) \leqslant b, \qquad\qquad \mu_\Omega^-(x)\big(b - \lambda(x)\big) = 0,\end{aligned}$$

and similarly for μ_Γ. □

3.5. The scalar parameter case. When $\lambda \in \mathbb{R}^+$, the Tikhonov regularization is no longer required and the bilevel problem is given by

$$\min_{0 \leqslant \lambda \leqslant b} \; \ell(u) \tag{3.17a}$$

$$\text{s.t.} \mu(u, \psi)_V + \lambda(u - f, \psi)_{0,\Omega} = 0, \quad \forall\, \psi \in V_c. \tag{3.17b}$$

Let the Lagrangian and its derivative with respect to u be defined as

$$\mathbb{L}(u, \lambda, p) := \ell(u) + \mu(u, p)_V + \lambda(u - f, p)_{0,\Omega}$$

and

$$\mathbb{L}_u(v) = \big(\nabla\ell(u), v\big)_{0,\Omega} + \mu(p, v)_V + \lambda(p, v)_{0,\Omega} = 0.$$

It follows that

$$\mu(p, v)_V + \lambda(p, v)_{0,\Omega} = -\big(\nabla\ell(u), v\big)_{0,\Omega}, \quad \forall\, v \in V_c.$$

On the other hand, the derivative of $\mathbb{L}$ with respect to λ is given by

$$\mathbb{L}_\lambda(h - \lambda) = (u - f, p)(h - \lambda) = (h - \lambda)\int_\Omega (u - f)p \; dx \geqslant 0, \quad \forall\, h \in [0, b].$$

Thus, the optimality system reads as follows.

$$\mu(u,\psi)_V + \lambda(u-f,\psi)_{0,\Omega} = 0, \qquad \forall \psi \in V_c, \tag{3.18a}$$

$$\mu(p,v)_V + \lambda(p,v)_{0,\Omega} = -\big(\nabla \ell(u), v\big)_{0,\Omega}, \qquad \forall v \in V_c, \tag{3.18b}$$

$$P_{[0,b]}\left(\lambda - c\int_\Omega (u-f)p\; dx\right) = \lambda, \qquad \forall c > 0, \tag{3.18c}$$

where $P_{[0,b]}$ is the standard projection operator onto the interval $[0,b]$.

4. Optimization with respect to the weights. In this section we introduce and analyze the bilevel problem for the identification of the optimal weight in the nonlocal means kernel. We consider a modified nonlocal means kernel where we restrict the integral to a bounded region, i.e.

$$\gamma_w(\mathbf{x},\mathbf{y}) = \exp\Big\{-\int_{B_\rho(\mathbf{0})} w(\boldsymbol{\tau})\big(f(\mathbf{x}+\boldsymbol{\tau}) - f(\mathbf{y}+\boldsymbol{\tau})\big)^2 d\boldsymbol{\tau}\Big\}, \tag{4.1}$$

where the ℓ^∞ ball B_ρ is the patch of the image explored within the integration and $w \in \mathcal{U}_{ad} := \big\{v \in L^2(B_\rho(\mathbf{0})) : 0 \leqslant w(\mathbf{t}) \leqslant W, \forall\, \mathbf{t} \in B_\rho(\mathbf{0})\big\}$ is the spatial weight (see equation (2.9)). This type of ℓ^∞ (or square) patches have been also used in [35]. Note that, to simplify the notation in the analysis, we embedded the parameter δ in the weight w; in Section 5, for each numerical test, we provide more details on the choice of kernel parameters, including δ.

Although our analysis is focused on a specific kernel, it can be extended to any exponential-type kernel and, more in general, to kernels whose energy space i equivalent to L^2.

4.1. Lower-level problem. For a given $\lambda \in \mathbb{R}^+$ and $w \in \mathcal{U}_{ad}$ we consider the following denoising problem

$$\min_{u \in V_c^w} J(u,\lambda) = \frac{\mu}{2}\|u\|_{V^w}^2 + \frac{\lambda}{2}\int_\Omega (u-f)^2\, d\mathbf{x}, \tag{4.2}$$

where the weight-depended energy space is defined as $V_c^w = \big\{v \in L^2(\Omega \cup \Omega_I) : \|v\|_{V^w} < \infty\big\}$ with

$$\|v\|_{V^w}^2 := \int_{\Omega\cup\Omega_I}\int_{\Omega\cup\Omega_I} \big(v(\mathbf{x}) - v(\mathbf{y})\big)^2 \gamma_w(\mathbf{x},\mathbf{y})\, d\mathbf{y} d\mathbf{x}.$$

Note that the spaces V_c^w are also equivalent to L_c^2 for all w. By proceeding in a similar manner as in Theorem 3.1, it can be readily verified that, for every $w \in \mathcal{U}_{ad}$, there exists a unique solution $u \in V^w$ for the lower-level problem (3.1). Moreover, the strict convexity and differentiability of the fidelity term yields the following necessary and sufficient optimality conditions

$$\mu(u,\psi)_{V^w} + \big(\lambda(u-f),\psi\big)_{0,\Omega} = 0, \quad \forall \psi \in V_c^w, \tag{4.3}$$

where

$$(u,\psi)_{V^w} = \int_{\Omega\cup\Omega_I}\int_{\Omega\cup\Omega_I} \big(v(\mathbf{x}) - v(\mathbf{y})\big)\big(\psi(\mathbf{x}) - \psi(\mathbf{y})\big)\gamma_w(\mathbf{x},\mathbf{y})\, d\mathbf{y} d\mathbf{x}, \tag{4.4}$$

and the following a-priori bound

$$\|u\|_{V^w} \leqslant C_\lambda \|f\|_{0,\Omega}. \tag{4.5}$$

4.2. Bilevel problem. We consider following bilevel optimization problem for the estimation of the optimal weight in (4.1).

$$\min_{(w,u)\in\mathcal{T}_{ad}} \ell(u), \tag{4.6}$$

where the feasible set is given by $\mathcal{T}_{ad} := \big\{(u,w) : w \in \mathcal{U}_{ad} \text{ and (4.3) holds}\big\}$.

THEOREM 4.1. *The bilevel problem* (4.6) *admits a solution* $(u^*, w^*) \in \mathcal{T}_{ad}$.

Proof. Since the functional is bounded from below, the box constraints and the a-priori estimate (4.5) imply that there exists a minimizing sequence $\{(w_n, u_n)\} \in \mathcal{T}_{ad}$ that is uniformly bounded. Moreover, the box constraints and the equivalence of spaces imply that the sequence $\{u_n\}$ is also bounded in $L^2(\Omega \cup \Omega_I)$. Thus, there exists a weakly convergent subsequence, that we still denote by $\{(w_n, u_n)\}$, and a limit point $(w^*, u^*) \in L^2(B_\rho(\mathbf{0})) \times L^2(\Omega \cup \Omega_I)$ such that $w_n \rightharpoonup w^*$ weakly in $L^2(B_\rho(\mathbf{0}))$ and $u_n \rightharpoonup u^*$ weakly in $L^2(\Omega \cup \Omega_I)$.

We next show that $(u^*, w^*) \in \mathcal{T}_{ad}$. Since $\mathcal{U}_{ad}$ is weakly closed, w^* satisfies the box constraints. Moreover, since $f \in L^\infty(\Omega \cup \Omega_I)$, it follows that

$$\int\limits_{B_\rho(\mathbf{0})} -w_n(\boldsymbol{\tau})\big(f(\mathbf{x}+\boldsymbol{\tau}) - f(\mathbf{y}+\boldsymbol{\tau})\big)^2 d\boldsymbol{\tau} \to \int\limits_{B_\rho(\mathbf{0})} -w^*(\boldsymbol{\tau})\big(f(\mathbf{x}+\boldsymbol{\tau}) - f(\mathbf{y}+\boldsymbol{\tau})\big)^2 d\boldsymbol{\tau}.$$

Hence, $\gamma_{w_n}(\mathbf{x},\mathbf{y}) \to \gamma_{w^*}(\mathbf{x},\mathbf{y})$, $\forall \mathbf{x},\mathbf{y}$. In addition, note that each pair (u_n, w_n) solves

$$\mu \int\limits_{\Omega\cup\Omega_I}\int\limits_{\Omega\cup\Omega_I} (u_n - u_n')(v - v')\gamma_{w_n}(\mathbf{x},\mathbf{y})d\mathbf{y}\, d\mathbf{x} + \lambda \int\limits_{\Omega} u_n\, v\, d\mathbf{x} = \lambda \int\limits_{\Omega} f\, v\, d\mathbf{x}, \tag{4.7}$$

where, to simplify the notation, we used $v := v(\mathbf{x})$ and $v' := v(\mathbf{y})$. Thus,

$$\mu \int\limits_{\Omega\cup\Omega_I}\int\limits_{\Omega\cup\Omega_I} (u_n - u^* - u_n' + u^{*\prime})(v - v')\gamma_{w_n}(\mathbf{x},\mathbf{y})d\mathbf{y}\, d\mathbf{x} + \lambda \int\limits_{\Omega} u_n\, v\, d\mathbf{x}$$

$$- \lambda \int\limits_{\Omega} f\, v\, d\mathbf{x} + \mu \int\limits_{\Omega\cup\Omega_I}\int\limits_{\Omega\cup\Omega_I} (u^* - u^{*\prime})(v - v')\gamma_{w_n}(\mathbf{x},\mathbf{y})d\mathbf{y}\, d\mathbf{x} = 0.$$

Note that the first term goes to 0 as $n \to \infty$. In fact, the uniform bound $|\gamma_w(\mathbf{x},\mathbf{y})| \leqslant 1$, $\forall w \in \mathcal{U}_{ad}$, and the weak L^2 convergence of $\{u_n\}$ imply that

$$\left| \int\limits_{\Omega\cup\Omega_I}\int\limits_{\Omega\cup\Omega_I} (u_n - u^* - u_n' + u^{*\prime})(v - v')\gamma_{w_n}(\mathbf{x},\mathbf{y})d\mathbf{y}\, d\mathbf{x} \right|$$

$$\leqslant \left| \int\limits_{\Omega\cup\Omega_I}\int\limits_{\Omega\cup\Omega_I} (u_n - u^* - u_n' + u^{*\prime})(v - v')d\mathbf{y}\, d\mathbf{x} \right| \to 0 \ \text{ as } \ n \to \infty.$$

The Lebesgue dominated convergence theorem, yields

$$\lim_{n\to\infty} \int\limits_{\Omega\cup\Omega_I}\int\limits_{\Omega\cup\Omega_I} (u^* - u^{*\prime})(v - v')\gamma_{w_n}(\mathbf{x},\mathbf{y})d\mathbf{y}\, d\mathbf{x}$$

$$= \int\limits_{\Omega\cup\Omega_I}\int\limits_{\Omega\cup\Omega_I} (u^* - u^{*\prime})(v - v')\gamma_{w^*}(\mathbf{x},\mathbf{y})d\mathbf{y}\, d\mathbf{x}.$$

Consequently, as $n \to \infty$, $(u^*, w^*) \in \mathcal{T}_{ad}$, i.e.,

$$\mu \int_{\Omega\cup\Omega_I} \int_{\Omega\cup\Omega_I} (u^* - u^{*\prime})(v - v')\gamma_{w^*}(\mathbf{x}, \mathbf{y}) d\mathbf{y}\, d\mathbf{x} + \lambda \int_{\Omega} (u^* - f) v\, d\mathbf{x} = 0. \tag{4.8}$$

Finally, since the loss function is convex and continuous, it is weakly lower semi-continuous, and, thus, (u^*, w^*) is a solution of (4.6). ∎

4.2.1. Differentiability of the solution operator. We first prove a lemma that will be useful in the proof of differentiability.

LEMMA 4.2. *Let $w \in \mathcal{U}_{ad}$ and $h \in L^2(B_\rho(\mathbf{0}))$ be a feasible direction, i.e., there exists some $t \in \mathbb{R}^+$ such that $w + th \in \mathcal{U}_{ad}$. Then, the weak solution of problem*

$$\mathcal{L}_t u_t + \lambda(u_t - f) = 0 \tag{4.9}$$

with

$$\mathcal{L}_t v(\mathbf{x}) = 2\mu \int_{\Omega\cup\Omega_I \cap B_\varepsilon(\mathbf{x})} (u' - u)\gamma_{w+th}(\mathbf{x}, \mathbf{y}) d\mathbf{y}, \tag{4.10}$$

satisfies the estimate

$$\mu \|u_t\|^2_{V^w} + \lambda \|u_t\|^2_{L^2(\Omega\cup\Omega_I)} \leqslant \lambda \|f\|_{L^2(\Omega\cup\Omega_I)} \|u_t\|_{L^2(\Omega\cup\Omega_I)}. \tag{4.11}$$

Proof. The weak formulation of (4.9) reads

$$\mu \int_{\Omega\cup\Omega_I} \int_{\Omega\cup\Omega_I} (u_t - u_t')(v - v')\gamma_{w+th}(\mathbf{x}, \mathbf{y}) d\mathbf{y}\, d\mathbf{x} + \lambda \int_{\Omega} (u_t - f) v\, d\mathbf{x} = 0, \quad \forall\, v \in V_t, \tag{4.12}$$

where V_t is the energy space induced by using the weight $(w + th)$. For $v = u_t$, the result follows from

$$\mu \int_{\Omega\cup\Omega_I} \int_{\Omega\cup\Omega_I} (u_t - u_t')^2 \gamma_{w+th}(\mathbf{x}, \mathbf{y}) d\mathbf{y}\, d\mathbf{x} + \lambda \|u_t\|^2_{0,\Omega\cup\Omega_I}$$

$$= \mu \|u_t\|^2_{V_t} + \lambda \|u_t\|^2_{0,\Omega\cup\Omega_I} \leqslant \lambda \|f\|_{L^2(\Omega\cup\Omega_I)} \|u_t\|_{0,\Omega\cup\Omega_I}.$$

∎

REMARK 1. *The last result implies that*

$$\mu \|u_t\|^2_{V^w} \leqslant \lambda \|f\|_{L^2(\Omega\cup\Omega_I)} \|u_t\|_{0,\Omega\cup\Omega_I} \quad \textit{and} \quad \|u_t\|_{L^2(\Omega\cup\Omega_I)} \leqslant \|f\|_{L^2(\Omega\cup\Omega_I)},$$

which implies that $\|u_t\|_{V^w} \leqslant C(\lambda) \|f\|_{L^2(\Omega\cup\Omega_I)}$.

Next, we prove that the sequence $\{z_t\} = \left\{\frac{u_t - u}{t}\right\}$ has a bounded L^2 norm and, thus, contains a weakly convergent subsequence.

LEMMA 4.3. *Let $w \in \mathcal{U}_{ad}$ and $h \in L^2(B_\rho(\mathbf{0}))$ be a feasible direction. The sequence $\{z_t\} = \big\{(u_t - u)/t\big\}$, where u and u_t are the solutions of (4.3) and (4.9), is bounded in L^2.*

Proof. By subtracting the weak forms (4.3) and (4.12), and using the equivalence of norms, we obtain

$$\mu \int\limits_{\Omega\cup\Omega_I}\int\limits_{\Omega\cup\Omega_I} (u_t-u_t')(v-v')\gamma_{w+th}(\mathbf{x},\mathbf{y})d\mathbf{y}\,d\mathbf{x}$$
$$-\mu \int\limits_{\Omega\cup\Omega_I}\int\limits_{\Omega\cup\Omega_I} (u-u')(v-v')\gamma_{w}(\mathbf{x},\mathbf{y})d\mathbf{y}\,d\mathbf{x}+\ \lambda\int\limits_{\Omega}(u_t-u)v\,d\mathbf{x}=0,\quad \forall\, v\in V. \tag{4.13}$$

Thus,

$$\mu \int\limits_{\Omega\cup\Omega_I}\int\limits_{\Omega\cup\Omega_I} (u_t-u-u_t'+u')(v-v')\gamma_{w}(\mathbf{x},\mathbf{y})d\mathbf{y}\,d\mathbf{x}+$$
$$\mu \int\limits_{\Omega\cup\Omega_I}\int\limits_{\Omega\cup\Omega_I} (u_t-u_t')(v-v')\Big[\gamma_{w+th}(\mathbf{x},\mathbf{y})-\gamma_{w}(\mathbf{x},\mathbf{y})\Big]d\mathbf{y}\,d\mathbf{x}$$
$$+\lambda\int\limits_{\Omega}(u_t-u)v\,d\mathbf{x}=0. \tag{4.14}$$

By choosing $v=u_t-u$ and dividing all expressions by t, we have

$$\frac{\mu}{t}\|u_t-u\|^2_{V^w}+\frac{\lambda}{t}\|u_t-u\|^2_{0,\Omega\cup\Omega_I}+\mu\int\limits_{\Omega\cup\Omega_I}\int\limits_{\Omega\cup\Omega_I}(u_t-u_t')(u_t-u-u_t'+u')$$
$$\frac{1}{t}\Big[\gamma_{w+th}(\mathbf{x},\mathbf{y})-\gamma_{w}(\mathbf{x},\mathbf{y})\Big]d\mathbf{y}\,d\mathbf{x}=0.$$

By using the differentiability of the exponential function as a superposition operator and the equivalence of norms, we obtain

$$\frac{C}{t}\|u_t-u\|^2_{0,\Omega\cup\Omega_I}+\frac{\lambda}{t}\|u_t-u\|^2_{0,\Omega\cup\Omega_I}$$
$$\leqslant \tilde{C}\left(\|h\|_{L^2(B_\rho(\mathbf{0}))}\|f\|^2_{\infty,\Omega}+\frac{o(t)}{t}\right)\|u_t\|_{0,\Omega\cup\Omega_I}\|u_t-u\|_{0,\Omega\cup\Omega_I},$$

which, combined with (4.11), implies that

$$\left\|\frac{u_t-u}{t}\right\|_{0,\Omega\cup\Omega_I}\leqslant C(\lambda,h)\|f\|^3_{\infty}.$$

□

The lemma above guarantees existence of a weakly convergent subsequence and of a limit point z^* such that $z_t\rightharpoonup z^*$ in $L^2(\Omega\cup\Omega_I)$. In the following lemma we derive the equation for z^*.

LEMMA 4.4. *Let z^* be such that $z_t\rightharpoonup z^*$ in $L^2(\Omega\cup\Omega_I)$. Then, z^* corresponds to the unique solution of the linearized equation*

$$\mu(z^*,v)_V+\mu(u,v)_{\tilde{V}}+\lambda(z^*,v)_{0,\Omega}=0,\quad \forall v\in V, \tag{4.15}$$

with $\widetilde{V} := \{u \in L^2(\Omega \cup \Omega_I) : \|u\|_{\widetilde{V}} < \infty\}$, *where* $\|\cdot\|_{\widetilde{V}}$ *is the energy norm induced by the linearized kernel*

$$\widetilde{\gamma}_h(\mathbf{x},\mathbf{y}) = \gamma_w(\mathbf{x},\mathbf{y}) \int_{B_\rho(\mathbf{0})} -h(\boldsymbol{\tau})\big(f(\mathbf{x}+\boldsymbol{\tau}) - f(\mathbf{y}+\boldsymbol{\tau})\big)^2 \, d\boldsymbol{\tau}. \tag{4.16}$$

Proof. By (4.13) we have

$$\mu \int_{\Omega\cup\Omega_I}\int_{\Omega\cup\Omega_I} (u_t - u_t')(v-v')\gamma_{w+th}(\mathbf{x},\mathbf{y})d\mathbf{y}\,d\mathbf{x}$$

$$- \mu \int_{\Omega\cup\Omega_I}\int_{\Omega\cup\Omega_I} (u-u')(v-v')\gamma_w(\mathbf{x},\mathbf{y})d\mathbf{y}\,d\mathbf{x} + \lambda\int_\Omega (u_t-u)v\,d\mathbf{x} = 0.$$

Adding and subtracting $(u,v)_{V_t}$ and dividing both sides by t, we have

$$\mu \int_{\Omega\cup\Omega_I}\int_{\Omega\cup\Omega_I} \frac{1}{t}\big((u_t-u)-(u_t'-u')\big)(v-v')\gamma_{w+th}(\mathbf{x},\mathbf{y})d\mathbf{y}\,d\mathbf{x}$$

$$+\mu \int_{\Omega\cup\Omega_I}\int_{\Omega\cup\Omega_I} (u-u')(v-v')\frac{1}{t}\big[\gamma_{w+th}(\mathbf{x},\mathbf{y}) - \gamma_w(\mathbf{x},\mathbf{y})\big]d\mathbf{y}\,d\mathbf{x}$$

$$+\frac{\lambda}{t}(u_t-u,v)_{0,\Omega} = 0,$$

The weak convergence $\frac{u_t-u}{t} \rightharpoonup z^*$ in $L^2(\Omega\cup\Omega_I)$, the strong convergence $w+th \to w$, and the continuity and differentiability of the exponential function as superposition operator, imply that the limit for $t\to 0$ of the previous equation is given by

$$\mu(z^*,v)_V + \lambda(z^*,v)_{0,\Omega} \quad + \mu \int_{\Omega\cup\Omega_I}\int_{\Omega\cup\Omega_I} (u-u')(v-v')\widetilde{\gamma}_h(\mathbf{x},\mathbf{y})d\mathbf{y}\,d\mathbf{x} = 0.$$

Uniqueness follows as for the state equation. ∎

The following theorem finalizes the differentiability result.

THEOREM 4.5. *Let* $S_w\colon \mathcal{U}_{ad} \longrightarrow V$ *be the solution operator which maps* w *into the corresponding solution to equation* (4.3). *Then the operator* S_w *is Gâteaux differentiable.*

Proof. In addition to Lemmas 4.2–4.4, it only remains to prove that

$$\left\| \frac{u_t-u}{t} - z^* \right\|_{L^2(\Omega\cup\Omega_I)} \to 0 \quad \text{as } t\to 0.$$

From equations (4.14) and (4.15) we obtain that the difference $\zeta := \frac{u_t-u}{t} - z^*$ is solution of the equation

$$\mu(\zeta,v)_V + \frac{\mu}{t}(u_t,v)_{V_t} - \frac{\mu}{t}(u_t,v)_V - \mu(u_t,v)_{\widetilde{V}} + \mu(u_t,v)_{\widetilde{V}} - \mu(u,v)_{\widetilde{V}} + \lambda\int_\Omega \zeta v = 0,$$

or, equivalently,

$$\mu(\zeta,v)_V + \mu \int_{\Omega\cup\Omega_I}\int_{\Omega\cup\Omega_I} (u_t-u_t')(v-v')\left[\frac{1}{t}\gamma_{w+th}(\mathbf{x},\mathbf{y}) - \frac{1}{t}\gamma_w(\mathbf{x},\mathbf{y}) - \widetilde{\gamma}_h(\mathbf{x},\mathbf{y})\right]d\mathbf{y}\,d\mathbf{x}$$

$$+\mu(u_t-u,v)_{\widetilde{V}} + \lambda\int_\Omega \zeta v = 0.$$

By choosing $v = \zeta$ we have

$$\mu\|\zeta\|_V^2 + \lambda\int_\Omega \zeta^2 = -\mu(u_t - u, \zeta)_{\tilde{V}} \\ - \mu\int\limits_{\Omega\cup\Omega_I}\int\limits_{\Omega\cup\Omega_I} (u_t - u_t')(\zeta - \zeta')\left[\frac{1}{t}\gamma_{w+th}(\mathbf{x},\mathbf{y}) - \frac{1}{t}\gamma_w(\mathbf{x},\mathbf{y}) - \widetilde{\gamma}_h(\mathbf{x},\mathbf{y})\right] d\mathbf{y}\, d\mathbf{x}.$$

The nonlocal Poincaré inequality, the convergence $u_t \to u$ in $L^2(\Omega\cup\Omega_I)$ and the differentiability of the exponential function, allow us to take the limit as $t \to \infty$, which yields the result. □

4.3. Optimality system. The differentiability of the solution operator allows us to derive an optimality system that characterizes the optimal solution of (4.6).

THEOREM 4.6. *Let (u, w) be an optimal solution to problem (4.6). Then, there exists a Lagrange multiplier $p \in L_c^2(\Omega\cup\Omega_I)$ such that the following optimality system is satisfied.*

$$\begin{aligned} &\mu(u,\psi)_{V^w} + \big(\lambda(u-f),\psi\big)_{0,\Omega} = 0, && \forall\psi\in V^w,\\ &\mu(p,\phi)_{V^w} + (\lambda\, p,\phi)_{0,\Omega} + \big(\nabla\ell(u),\phi\big)_{0,\Omega} = 0, && \forall\phi\in V^w,\\ &\mu\int\limits_{\Omega\cup\Omega_I}\int\limits_{\Omega\cup\Omega_I}\Big[\big(u(w)-u(w)'\big)(p-p')\widetilde{\gamma}_{(h-w)}(\mathbf{x},\mathbf{y})\Big]d\mathbf{y}\,d\mathbf{x}\geqslant 0, && \forall h\in\mathcal{U}_{ad}. \end{aligned} \tag{4.17}$$

Proof. Let the reduced cost functional be defined as

$$j(w) := \ell\big(u(w)\big), \tag{4.18}$$

where $u(w)$ is the unique solution to the state equation (4.3) corresponding to w. By taking the derivative of the reduced cost with respect to w, in direction h, we have

$$j'(w)h = \big(\nabla\ell(u(w)), u'(w)h\big)_{0,\Omega},$$

where $u'(w)h$ is the unique solution to the linearized equation

$$\mu\big(u'(w)h,\psi\big)_{V^w} + \mu\big(u(w),\psi\big)_{\tilde{V}} + \lambda\big(u'(w)h,\psi\big)_{0,\Omega} = 0, \quad \forall\psi\in V^w.$$

As in (3.18b), the adjoint equation is given by

$$\mu(p,v)_{V^w} + \lambda(p,v)_{0,\Omega} = -\big(\nabla\ell(u),v\big)_{0,\Omega}, \quad \forall v\in V^w, \tag{4.19}$$

which is uniquely solvable by the same arguments as in Theorem 3.1. Thus we obtain

$$j'(w)h = -\mu\big(u'(w)h,p\big)_{V^w} - \lambda\big(u'(w)h,p\big)_{0,\Omega}.$$

Using the linearized equation and considering the box constraints on w, we then get the first order necessary optimality condition

$$j'(w)(h-w) = \mu\int\limits_{\Omega\cup\Omega_I}\int\limits_{\Omega\cup\Omega_I}\Big[\big(u(w)-u(w)'\big)(p-p')\widetilde{\gamma}_{(h-w)}(\mathbf{x},\mathbf{y})\Big]d\mathbf{y}\,d\mathbf{x}\geqslant 0,$$

for all $h \in \mathcal{U}_{ad}$. □

REMARK 2. *When w is a scalar parameter, the last expression in the optimality system may be replaced with*

$$P_{[0,W]}\Big(w - c\big(u,p\big)_{\hat{V}}\Big) = w, \quad \forall c > 0,$$

where $P_{[0,W]}$ is the standard projection operator onto the interval $[0,W]$ and

$$(u,p)_{\hat{V}} :=$$

$$\mu \int\limits_{\Omega\cup\Omega_I} \int\limits_{\Omega\cup\Omega_I} \big(u-u'\big)(p-p')\gamma_w(\mathbf{x},\mathbf{y}) \left[\int\limits_{B_\rho(\mathbf{0})} -\big(f(\mathbf{x}+\boldsymbol{\tau}) - f(\mathbf{y}+\boldsymbol{\tau})\big)^2 \, d\boldsymbol{\tau} \right] d\mathbf{y}\, d\mathbf{x}.$$

This enables the use of projection algorithms for solving the bilevel problem.

5. Numerical tests. In this section we propose a numerical algorithm for the solution of the bilevel problem and illustrate our theory with several numerical tests. First, we describe the discretization of the system of optimality conditions and the technique used to evaluate the kernel. Then, we present the optimization algorithm which relies on a trust-region scheme with active set prediction and limited memory BFGS matrices. The section concludes with results of numerical computations illustrating the main features of the proposed approach.

5.1. Discretization. As we are interested in developing a second-order algorithm to solve the optimality system (3.18), we consider an H^1–Riesz representation of the derivative, $j'(\lambda)h$ of the reduced cost, where

$$j'(\lambda)h = \int_\Omega (u-f)ph \, dx + \beta(\lambda,h)_{H^1}. \tag{5.1}$$

Note that the functional above solves the following variational equation

$$(y,h)_{0,\Omega} + (\nabla y, \nabla h)_{0,\Omega} = \big((u-f)p, h\big)_{0,\Omega} + \beta(\lambda,h)_{H^1}, \tag{5.2}$$

or, equivalently, the following PDE

$$\begin{aligned} -\Delta y + y &= -\beta\Delta\lambda + \beta\lambda + (u-f)p && \text{in } \Omega, \\ \frac{\partial y}{\partial \vec{n}} &= \beta\frac{\partial \lambda}{\partial \vec{n}} && \text{on } \partial\Omega. \end{aligned} \tag{5.3}$$

Since we are interested in possibly discontinuous nonlocal solutions we use two finite element bases to approximate u and p in $L^2(\Omega)$ and y, λ in $H^1(\Omega)$. Specifically, we consider piecewise constant and piecewise linear elements, respectively, defined over a partition of $\Omega \cup \Omega_I$, which we denote as $\mathcal{T}^h$. Throughout this section we denote discretized quantities by using the superscript h, and we fix $\mu = \frac{1}{2}$.

First, we consider the nonlocal systems involved in the discretization with respect to λ; in this case we let $w(\mathbf{t}) = \delta^{-2}$, for all $\mathbf{t}$ in (4.1). The discrete analogue of the nonlocal variational equations in (3.9) is given by

$$\widetilde{\lambda}_i^h u_i^h + \eta_i u_i^h - \sum_{T_j\in\mathcal{T}} u_j^h \gamma_{i,j}^h = \widetilde{\lambda}_i^h f_i^h, \qquad \forall T_i \in \mathcal{T}^h, \tag{5.4a}$$

$$\widetilde{\lambda}_i^h p_i^h + \eta_i p_i^h - \sum_{T_j\in\mathcal{T}} p_j^h \gamma_{i,j}^h = u_i^T - u_i^h, \qquad \forall T_i \in \mathcal{T}^h; \tag{5.4b}$$

where u_i^h, p_i^h, $\widetilde{\lambda}_i^h$, and f_i^h are the values of the approximate nonlocal state, nonlocal adjoint, fidelity weight, and forcing term at triangle T_i, and $\gamma_{i,j}^h$ is the value of the approximate kernel for $\mathbf{x} \in T_i$ and $\mathbf{y} \in T_j$, and $\eta_i := \sum_{T_j} \gamma_{i,j}^h$. The evaluation of f_i^h, $\widetilde{\lambda}_i^h$ and $\gamma_{i,j}^h$ depends on the location of the pixels and is described in detail in the next section.

We rewrite system (5.4) in a more compact form as follows

$$\big(\operatorname{diag}(\boldsymbol{\lambda}) + \operatorname{diag}(\boldsymbol{\eta}) - \Gamma\big)\mathbf{u} = \boldsymbol{\lambda} \circ \mathbf{f}, \tag{5.5a}$$

$$\big(\operatorname{diag}(\boldsymbol{\lambda}) + \operatorname{diag}(\boldsymbol{\eta}) - \Gamma\big)\mathbf{p} = \mathbf{u}^T - \mathbf{u}, \tag{5.5b}$$

where, for a vector $\mathbf{v}$, $\operatorname{diag}(\mathbf{v})$ is the $n \times n$ diagonal matrix whose diagonal entries are the components of $\mathbf{v}$, and where $\circ$ is the Hadamard product, i.e., $\mathbf{v} \circ \mathbf{w} := \big(v_1 w_1, \ldots, v_n w_n\big)^\top$ for any two vectors $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$. The bold notation refers to the vectors whose components are the values of the variables at the DOFs. The matrix Γ is such that $\Gamma_{i,j} := \gamma_{i,j}^h$.

Equation (5.3) for the gradient of the reduced cost functional for problem (3.7) is discretized as

$$(A + B)\mathbf{y} = F(\mathbf{u}, \boldsymbol{\lambda}), \qquad \text{where}$$

$$A_{i,j} = \int_\Omega \nabla\phi_i \nabla\phi_j d\mathbf{x}, \qquad B_{i,j} = \int_\Omega \phi_i \phi_j d\mathbf{x}, \qquad \text{and}$$

$$F(\mathbf{u}, \boldsymbol{\lambda})_i = \int_\Omega \big((u^h - f^h)p^h + \beta\lambda^h\big)\phi_i d\mathbf{x} + \sum_{j=1}^{n} \lambda_i^h \int_\Omega \nabla\phi_i \cdot \nabla\phi_j d\mathbf{x}$$

being ϕ_i and ϕ_j elements of the finite element basis associated with triangles T_i and T_j, respectively, and λ_i^h the values of the finite element solution λ^h at the degrees of freedom.

Second, we consider the optimization with respect to w; we let $\lambda \in \mathbb{R}^+$ and $w \in \mathbb{R}^+$. Thus, (5.5) becomes

$$\big(\lambda I_n + \operatorname{diag}(\boldsymbol{\eta}_w) - \Gamma_w\big)\mathbf{u}_w = \lambda\mathbf{f} \tag{5.6a}$$

$$\big(\lambda I_n + \operatorname{diag}(\boldsymbol{\eta}_w) - \Gamma_w\big)\mathbf{p}_w = \mathbf{u}^T - \mathbf{u}_w, \tag{5.6b}$$

where we use the sub-index w to indicate the dependency of the kernel on w. The gradient of the reduced cost functional for problem (4.6) is discretized as

$$j'(w)^h = \mathbf{p}_w \cdot \big(\operatorname{diag}(\widehat{\boldsymbol{\eta}}) - \widehat{\Gamma}_w\big)\mathbf{u}_w. \tag{5.7}$$

for which $\widehat{\boldsymbol{\eta}}_i := \sum_{T_j} \widehat{\Gamma}_{w,i,j}$ and $\widehat{\Gamma}_{w,i,j} = \widehat{\gamma}_{w,i,j}^h$ is a discretization of $\widetilde{\gamma}_{(1)}$, defined in (4.16).

Finally, we mention that in our numerical experiments we precondition the nonlocal systems by the following diagonal preconditioner P: let $a_{i,j}$ be the entries of the matrix of the system for u^h, we have:

$$P_{i,i} := \left(\sum_j a_{i,j}^2\right)^{-1/2}.$$

Using this preconditioner, we solved the nonlocal systems with the Loose Generalized Minimal Residual Method (LGMRES).

5.1.1. Evaluating the modified nonlocal means kernel. For a given image $f : \Omega \mapsto [0, 255]$, where $\Omega = [0, N] \times [0, M]$, the evaluation of the modified nonlocal means kernel requires the identification of several patches within the image. In this section, we describe how to define those regions and efficiently evaluate the kernel. Since the kernel decays exponentially away from the origin, we consider a relatively small radius ρ. Also, recall that for both the optimization with respect to λ and w we only consider constant weights, i.e. $w(\mathbf{t}) = w \in \mathbb{R}^+$.

By definition, $\Omega_I = [-\varepsilon, N + \varepsilon] \times [-\varepsilon, M + \varepsilon] \setminus \Omega$; we extend f to zero outside Ω. We assume that the pixels are uniformly distributed and ordered over the domain and label them by the integer $\boldsymbol{i}$. As illustrated in Figure 5.1, pixel $\boldsymbol{i}$ is located at the upper-left corner of the element $T_i \in \mathcal{T}^h$ so that every element is associated with one pixel.

In this setting, the approximation $\widetilde{\lambda}_i^h$ introduced above consists in the value of λ^h at the pixel corresponding to element T_i, i.e. pixel $\boldsymbol{i}$.

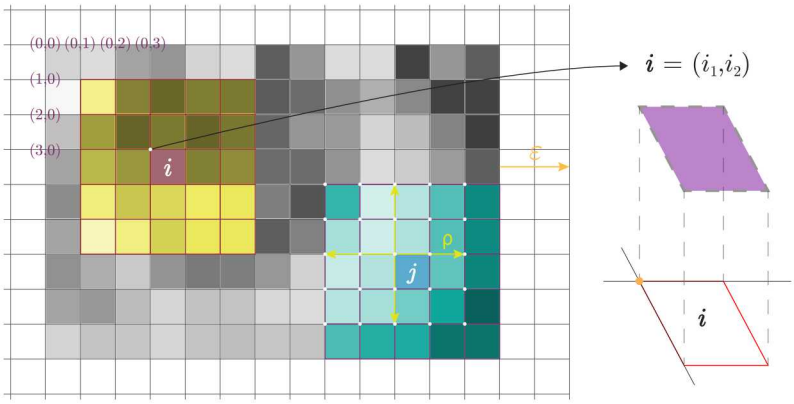


Fig. 5.1: Description of patches and pixel discretization in relation to the finite element grid.

Let f^h be an approximation of f in the computational domain such that $f_i^h = f(\boldsymbol{i})$, i.e., the value of the approximate image over the element T_i corresponds to the value of the image at pixel $\boldsymbol{i}$. A *patch* $\mathbf{P}_{\boldsymbol{i}}(f)$ is a sub-image of f^h around pixel $\boldsymbol{i}$ given by

$$\mathbf{P}_{\boldsymbol{i}}(f)(\boldsymbol{t}) = f^h(\boldsymbol{i} + \boldsymbol{t}), \qquad \forall \boldsymbol{t} \in [-\rho : \rho]^2,$$

where the interval $[a : b]$ denotes the closed interval of integers from a to b.

We refer to the sum of the image values within a patch as the *patch measure* and denote it by $\mathbf{p}_{\boldsymbol{i}}(f)$. Notice that a patch will have at most $(2\rho + 1)^2 =: |\mathcal{P}|$ pixels. We approximate the value of the kernel in (4.1) at points corresponding to pixels $(\boldsymbol{i}, \boldsymbol{j})$ as follows:

$$\gamma_{\boldsymbol{i},\boldsymbol{j}}^h \approx \exp\left\{ - w\left(\mathbf{p}_{\boldsymbol{i}}(f^2) + \mathbf{p}_{\boldsymbol{j}}(f^2) - 2 \sum_{\boldsymbol{t} \in [-\rho:\rho]^2} \mathbf{P}_{\boldsymbol{i}}(f)(\boldsymbol{t}) \circ \mathbf{P}_{\boldsymbol{j}}(f)(\boldsymbol{t}) \right) \right\} \chi\big(\boldsymbol{j} \in B_\varepsilon(\boldsymbol{i}) \cap \gamma_{\boldsymbol{i},\boldsymbol{j}}^h > \iota\big). \tag{5.8}$$

This serves as an approximation of $\gamma_{i,j}^h$ in (5.4), where elements are associated with the corresponding pixels.

As noted in [39], high dissimilarity values between each pair of patches do not provide meaningful information to the resulting image restoration process, therefore in (5.8) we introduce the threshold parameter $\iota > 0$ that acts as an acceptance tolerance between patches. Furthermore, we consider large interaction radii, i.e. $\varepsilon \gg 1$. This constraint induces a multi-banded matrix approximation of the nonlocal operator with $\varepsilon - 1$ bands yielding at most $(2\varepsilon+1)^2 =: \mathcal{E}$ neighbors per pixel. These two constraints ensure that only close and similar regions of the image are compared and, at the same time, it reduces memory allocation and computational cost. As a consequence, the nonlocal kernel can be evaluated in $\mathcal{O}\Big(|\mathcal{P}|\big[NM(1+\mathcal{E}) - \mathcal{E}\big]\Big)$ operations.

5.2. Optimization algorithm. The reduced objective functionals (3.10) and (4.18) are not necessarily convex since equations (5.5) and (5.6) are not linear in terms of λ and w, respectively. Thus, we resort to trust-region methods with a quadratic cost involving limited memory BFGS matrices. First, we summarize the limited memory approach and then briefly introduce the projected trust-region algorithm developed in [45] for the solution of general nonlinear box-constrained optimization problems.

5.2.1. Limited memory BFGS. For large-scale optimization problems, limited memory methods are known to be effective techniques, as they require minimal storage and provide easy-to-compute second order information, often resulting in a fast local superlinear convergence rate [28]. The limited memory BFGS method approximates the inverse of the Hessian of a functional j at iteration $k+1$, say H_{k+1}, without storing the dense matrices H_k at each iteration. Instead, it stores m correction pairs $\{q_i, d_i\}_{i\in[k-1:k-m]} \subset \mathbb{R}^{n,2}$, where

$$q_i := \mathbf{x}_{i+1} - \mathbf{x}_i \qquad \text{and} \qquad d_i := \nabla j(\mathbf{x}_{i+1}) - \nabla j(\mathbf{x}_i),$$

that contain information related to the curvature of j. Paper [9] introduces a compact form to define the limited memory matrix $B_k = H_k^{-1}$ in terms of the $n \times m$ correction matrices

$$S_k := \begin{pmatrix} q_{k-m} & \cdots & q_{k-1} \end{pmatrix} \qquad \text{and} \qquad Y_k := \begin{pmatrix} d_{k-m} & \cdots & d_{k-1} \end{pmatrix}.$$

The main idea of the algorithm is that the matrix $S_k^\top Y_k$ can be written as the sum of the following three matrices (used within the algorithm):

$$S_k^\top Y_k = L_k + D_k + R_k,$$

where L_k is strictly lower triangular, D_k is diagonal, and R_k is strictly upper triangular. For $\theta > 0$, if the correction pairs satisfy $q_i^\top d_i > 0$, then the matrix obtained by updating θI_n with the BFGS formula and the correction pairs after k-times can be written as

$$B_k := \theta I_n - W_k M_k W_k^\top, \tag{5.9a}$$

where W_k and M_k are the block matrices given by

$$W_k := \begin{pmatrix} Y_k & \theta S_k \end{pmatrix}, \tag{5.9b}$$

$$M_k := \begin{pmatrix} -D_k & L_k^\top \\ L_k & \theta S_k^\top S_k \end{pmatrix}^{-1}. \tag{5.9c}$$

Note that, as M_k is a $2m \times 2m$ matrix, the cost of computing the inverse in (5.9c) is negligible. Hence, using the compact representation (5.9a), various computations involving B_k become inexpensive, as is the case of the product of B_k times a vector.

One aspect of the BFGS method is that each update is positive definite. As the limited memory formula (5.9a) can also be stated as

$$B_k = V_k^\top H_k V_k + p_k q_k q_k^\top, \tag{5.10}$$

with $p_k := \left(q_k^\top d_k\right)^{-1}$ and $V_k := I_n - p_k d_k q_k^\top$. Thus, we can guarantee positive definiteness using Powell's method [34] in which d_k is redefined as

$$d_k := \begin{cases} d_k & \text{if } q_k^\top d_k \geqslant 0.2 q_k^\top B_k d_k, \\ \alpha_k d_k + (1-\alpha_k) B_k q_k & \text{otherwise,} \end{cases} \tag{5.11}$$

where $\alpha_k := \frac{0.8 q_k^\top B_k q_k}{q_k^\top B_k q_k - q_k^\top d_k}$. If the updated $q_k^\top d_k$ is too close to zero, to maintain numerical stability, the limited memory matrix is not updated.

5.2.2. Active set estimation and search direction. Because of the box constraints, second order information is only relevant far from the bounds. Following [45], and assuming 0 and b as lower and upper bounds, respectively, we introduce the quantity $\xi_k := \min\left\{\beta_k, c\|\nabla j(\mathbf{x}_k)\|\right\}$, where β_k and c are positive constants such that $0 < \beta_k < \frac{b}{2}$, and define the strongly-active and inactive index sets by

$$A_k := \left\{i \in \{1,\dots,n\} : \mathbf{x}_{k,i} \leqslant \xi_k \vee \mathbf{x}_{k,i} \geqslant b - \xi_k\right\}, \tag{5.12a}$$

$$I_k := \{1,\dots,n\} \backslash A_k = \left\{i \in \{1,\dots,n\} : \xi_k < \mathbf{x}_{k,i} < b - \xi_k\right\}, \tag{5.12b}$$

respectively, where $\mathbf{x}_{k,i}$ is the i-th element of $\mathbf{x}_k$. Now, suppose the current trust region radius is $\widehat{\Delta} > 0$, with its maximum value $\Delta_{\max} > 0$, and let $\kappa > 0$. We can obtain a search direction at step x_k as follows:

- **Projected gradient direction:** Compute

$$d_{*k}^G(\widehat{\Delta}) := \max\left\{0, \min\left\{b,\, \mathbf{x}_k - \frac{\widehat{\Delta}}{\Delta_{\max}} \kappa_k \nabla\eta(\mathbf{x}_k)\right\}\right\} - \mathbf{x}_k. \tag{5.13}$$

- **Projected trust-region direction:** We look for a direction $d_{*k}^{\mathrm{tr}}(\widehat{\Delta})$ defined for each index of the sets A_k and I_k, respectively. We begin with A_k, for which we let $v_k^{A_k}$ be the subvector

$$v_k^{A_k} := \begin{cases} \mathbf{x}_{k,i} & \text{if } \mathbf{x}_{k,i} \leqslant \xi_k, \\ b - \mathbf{x}_{k,i} & \text{if } x_{k,i} \geqslant b - \xi_k. \end{cases}$$

  Then we define the subvector

$$d_{*k}^{A_k}(\widehat{\Delta}) := \min\left\{1, \frac{\widehat{\Delta}}{\|v_k^{A_k}\|}\right\} v_k^{A_k}. \tag{5.14}$$

  For the inactive set I_k we solve a reduced trust-region subproblem. Here, let B_k be partitioned into two submatrices $B_k^{A_k}$ and $B_k^{I_k}$, obtained by taking columns of B_k indexed by A_k and I_k, respectively. Let $d_{*k}^{I_k}(\widehat{\Delta})$ be a solution of the following TR–subproblem

$$\begin{aligned} &\min\ d^\top\left[\left(B_k^{I_k}\right)^\top\left(\nabla j(\mathbf{x}_k) + B_k^{A_k} d_{*k}^{A_k}\right)\right] + \frac{1}{2} d^\top \left(B_k^{I_k}\right)^\top B_k^{I_k} d \\ &\text{s.t. } \|d\| \leqslant \widehat{\Delta}. \end{aligned} \tag{5.15}$$

The projected trust-region direction is then defined as

$$d^{\mathrm{tr}}_{*k}(\widehat{\Delta}) := \max\left\{0, \min\left\{b,\, x_k + \begin{pmatrix} d^{A_k}_{*k}(\widehat{\Delta}) \\ d^{I_k}_{*k}(\widehat{\Delta}) \end{pmatrix}\right\}\right\} - \mathbf{x}_k. \tag{5.16}$$

Since this direction may not be a descent direction for j for far iterates, we use a convex combination with the gradient direction as follows.

- **Search direction:** Let

$$d_{*k}(\widehat{\Delta}) := t_{*k} d^{G}_{*k}(\widehat{\Delta}) + (1 - t_{*k}) d^{\mathrm{tr}}_{*k}(\widehat{\Delta}), \tag{5.17}$$

where t_{*k} is a solution of the following one-dimensional problem

$$\min_{t\in[0,1]} j\big(x_k + t d^{G}_{*k}(\widehat{\Delta}) + (1-t) d^{\mathrm{tr}}_{*k}(\widehat{\Delta})\big). \tag{5.18}$$

5.2.3. Algorithm. We state next the projected trust-region algorithm with L-BFGS update as described in [45].

5.3. Experimental results. We present the results of the bilevel optimization with respect to λ and w using the modified nonlocal means kernel. The results are organized as follows: for each test we report a figure and a table. The figure displays the clean image u^T from a database, four noisy images, (a)–(d), and the corresponding (optimal) denoised images, (e)–(h). Values of the Structural Similarity Index (SSIM), which measure the similarity of the recovered image against u^T, are also included (rounded up to two digits). In the table we report optimal SSIM values and output parameters of the optimization.

For our computations, we use images from the USC-SIPI Image Database and the FVC2000 Database, which are padded with a border of zeroes of width ε, in order to deal with information in Ω_I. For each image, a sample of four noisy images is obtained by adding different levels of Gaussian noise with standard deviation σ; that is $f = u^T + \eta$ with $\eta \sim \mathcal{N}(0, \sigma^2)$. The values of σ are taken according to Table 5.1. We use the constant patch radius $\rho = 5$, i.e., each patch contains 121 pixels and we set the interaction radius ε so that there are at most $5\min\{N, M\}$ neighbors per pixel. The problem dimension N and M will be specified below for each experiment.

Table 5.1: Parameters associated to noisy data

Sample	(a)	(b)	(c)	(d)
σ^2	$10^{1.5}$	$10^{2.0}$	$10^{2.5}$	$10^{3.0}$
Filtering δ	5×10^2	10^2	3.16×10^2	5.13×10^2

5.3.1. Optimizing the fidelity parameter λ. We consider both the case of constant $\lambda \in [0, b]$ and space-dependent $\lambda \in \mathcal{C} = \big\{\lambda \in H^1(\Omega) : b \geqslant \lambda(\mathbf{x}) \geqslant 0\big\}$. The upper bound b is set to 10^5 and the acceptance tolerance is set to $\iota = 10^{-9}$. Recall that in this case we set $w(\mathbf{t}) = \delta^{-2}$; values for each image are reported in Table 5.1.

Algorithm 1 Projected Trust-Region Algorithm with L-BFGS Update

1: Choose x_0 and a symmetric positive definite matrix H_0. Let constants satisfy $0 < \beta_0 < \frac{b}{2}$, $c > 0$, $0 < \nu_1 < 1 < \nu_2$, $0 < \tau_1 < \tau_2 < 1$, $\upsilon \in (0,1)$, $\omega \in \mathbb{R}$, $\Delta_0 > 0$, and $\Delta_{\max} > \Delta_{\min} > 0$. Now set $m \in \mathbb{N}$, $k = 0$, and $B_0 = H_0^{-1}$.
2: **repeat**
3: Let $\Delta_k := \min\left\{\Delta_{\max}, \max\{\Delta_{\min}, \Delta_k\}\right\}$ and $\widetilde{\Delta} = \Delta_k$.
4: Determine index sets A_k and I_k by (5.12a) and (5.12b).
5: Find $d_{*k}^{\mathrm{tr}}(\widetilde{\Delta})$ by determining $d_{*k}^{A_k}(\widetilde{\Delta})$ and $d_{*k}^{I_k}(\widetilde{\Delta})$ as in (5.14) and (5.15).
6: Set

$$\kappa_k := \min\left\{1, \frac{\Delta_{\max}}{\|\nabla j(x_k)\|}, \frac{\omega}{\|\nabla j(x_k)\|}\right\}.$$

7: Compute $d_{*k}^{G}(\widetilde{\Delta})$ and t_{*k} as in (5.13) and (5.18), respectively. Let

$$d_{*k}(\widehat{\Delta}) := t_{*k} d_{*k}^{G}(\widehat{\Delta}) + (1 - t_{*k}) d_{*k}^{\mathrm{tr}}(\widehat{\Delta}).$$

8: Compute

$$r_{*k} := \frac{j(x_k + d_{*k}) - j(x_k)}{\nabla j(x_k)^{\top} d_{*k}(\widehat{\Delta}) + \frac{1}{2}\, d_{*k}(\widehat{\Delta})^{\top} B_k d_{*k}(\widehat{\Delta})}$$

9: **if** $j(x_k) - j\big(x_k + d_{*k}(\widehat{\Delta})\big) \geqslant -\upsilon \nabla j(x_k)^{\top} d_{*k}^{G}(\widehat{\Delta})$ and $r_{*k} \geqslant \tau_1$ hold **then**
10: Let $q_k := d_{*k}$, $x_{k+1} := x_k + d_{*k}$, $\beta_k = \widehat{\Delta}$, and

$$\Delta_{k+1} := \begin{cases} \widehat{\Delta} & \text{if } \tau_1 < r_{*k} < \tau_2, \\ \nu_2 \widehat{\Delta} & \text{if } r_{*k} \geqslant \tau_2. \end{cases}$$

11: Let $\widehat{m} := \min\{k+1, m\}$.
12: Update B_k with the $n \times \widehat{m}$ matrices S_k and Y_k to get B_{k+1}.
13: Let $k = k + 1$ and return to step 1.3.
14: **else**
15: Let $\widehat{\Delta} = \tau_1 \widehat{\Delta}$.
16: Return to step 1.5.
17: **end if**
18: **until** x_k satisfies a stopping criteria.

Constant parameter. We initialize the TR algorithm with $\lambda_0 = 100$ and we note that, for $\lambda \in [0, b]$, the gradient of the reduced cost functional (3.10) reduces to $\nabla j(\lambda) = (\mathbf{u} - \mathbf{f}) \cdot \mathbf{p}$.

The results are displayed in Figure 5.2 and Table 5.2. In the latter we report, for each clean image and its corresponding noisy sample, the optimal λ, its SSIM value, the number of iterations of the TR algorithm, and the dimensions of the image. From Figure 5.2 we note that, after the optimization, there is a significant increase in the SSIM values. Moreover, as expected, the nonlocal means kernel allows regularization of each sample while preserve the textures (see, e.g., [21]). Hence, discontinuities are preserved and restored without blurring. Furthermore, in Table 5.2 it is noticeable that the the best solution found for each noisy image is located in the interior of the

convex set.

We also note that, at each iteration, the objective function decreases monotonically and the radius of the trust-region decreases around the solution.

Fig. 5.2: Resulting images of scalar parameter optimization

Table 5.2: Results of scalar optimization

	Sample	Best λ	SSIM	Iteration Count	(N, M)
Lena	(a)	81.39288608477064	0.9067	6	$(256, 256)$
	(b)	3.711999777944892	0.8731	16	
	(c)	199.3507646217158	0.7452	14	
	(d)	250.1708380247259	0.5508	18	
cameraman	(a)	221.1940317431237	0.9246	14	$(256, 256)$
	(b)	21.55437493757969	0.8765	14	
	(c)	170.63141217270092	0.7811	13	
	(d)	171.0574733520938	0.6469	13	
monarch	(a)	81.27508954688543	0.9529	6	$(256, 171)$
	(b)	3.448376834317583	0.9153	21	
	(c)	111.9243249470581	0.8194	9	
	(d)	165.9950203556561	0.6363	14	

Spatially dependent parameter. The optimization with respect to a space-dependent λ is a large scale nonconvex problem; thus, to prevent stagnation in regions far from local minima, we restart the optimization in the following two cases [11, 30]:

1. The trust region radius Δ_k becomes sufficiently small. Whenever $\Delta_k < \Delta_{\min}$, we set $\Delta_k = \Delta_{\text{reset}}$ with $\Delta_{\text{reset}} \in (0, \Delta_{\max}]$ and continue iterating if there is a decrease in the objective function. This is done in order to prevent algorithm to halt at a non-stationary point, whenever the trust region radius decreases too quickly.
2. The value $q_k^\top d_k$ is too close to zero. If $q_k^\top d_k < \varsigma \ll 1$, then all the stored pairs $\{q_i, d_i\}$ are removed and both S_k and Y_k are rebuilt from scratch. This prevents the occurrence of ill-conditioned updates.

Moreover, after each successful update of the limited memory pairs, we modify the L-BFGS initialization parameter θ in (5.9), by setting $\theta_k = {}^{\|d_k\|}/_{\|q_k\|}$ [17]. We set the maximum number of iterations to 10^3 initializing with the constant candidate $\lambda_0 = 200$.

The results are displayed in Figure 5.3 and Table 5.3. In addition to the noisy sample and its corresponding set of solutions to each image u^T of the database, we also include a third row of images displaying the optimal $\lambda(\mathbf{x})$. In the table we report the optimal SSIM value, the number of iterations of the TR algorithm, the order of magnitude of the 2–norm, and the dimensions of the image.

In Figure 5.3, we note that the optimal SSIM is much higher than the one associated with the noisy image and that there is a significant improvement compared to results obtained with a constant λ. We also observe that the optimal parameter is able to catch discontinuities and noise, see in particular (c) and (d).

Table 5.3: Results of optimization with respect to $\lambda(\mathbf{x})$.

	Sample	SSIM	Iteration Count	(N, M)
Lena	(a)	0.9311	1000	$(256, 256)$
	(b)	0.8922	2	
	(c)	0.8172	518	
	(d)	0.6947	295	
cameraman	(a)	0.9430	626	$(256, 256)$
	(b)	0.8690	2	
	(c)	0.8626	296	
	(d)	0.8092	679	
monarch	(a)	0.9683	81	$(256, 171)$
	(b)	0.9248	13	
	(c)	0.8854	14	
	(d)	0.7711	1000	

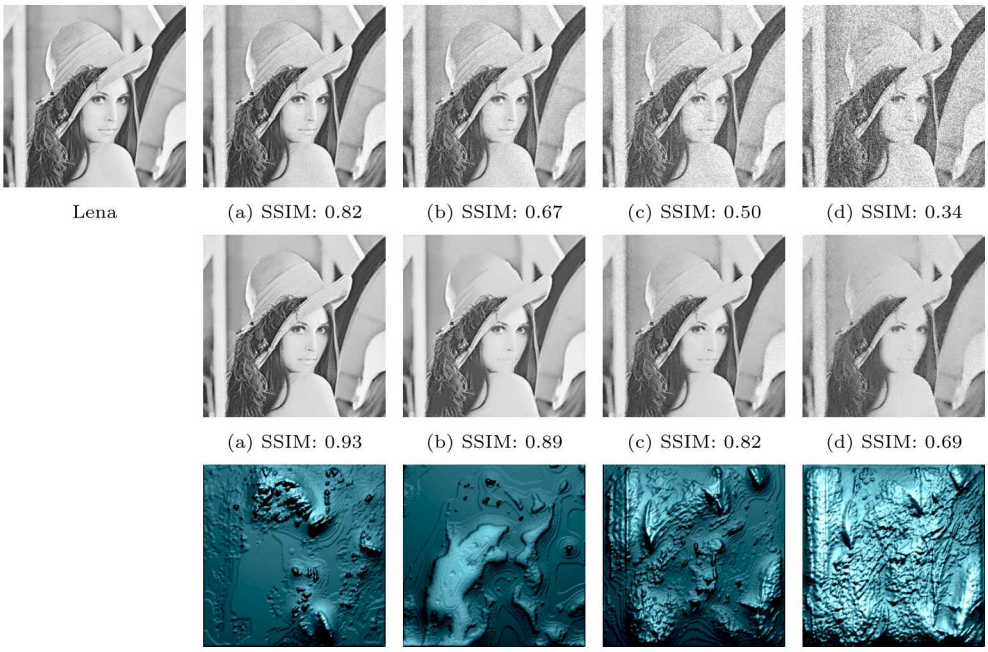


Fig. 5.3: Resulting images of spatial parameter optimization

Fig. 5.3: Resulting images of spatial parameter optimization

Training set. We consider a batch learning approach: we are interested in learning the fidelity constant λ from several images with the same noise level by solving the following coupled optimization problem:

$$\min_{0\leqslant\lambda\leqslant b} \; L(U) \tag{5.19a}$$

$$\text{s.t.} \quad \mu(u_i,\psi)_{V_i} + \lambda(u_i - f_i,\psi)_{0,\Omega} = 0, \quad \forall \psi \in V_i, i \in I, \tag{5.19b}$$

where $U = \{u_i\}_{i\in I}$ is a set of reconstructed images from a noisy sample $F = \{f_i\}_{i\in I}$ and L is a generalization of the loss function ℓ, defined as

$$L(U) = \frac{1}{|I|}\sum_{i\in I}\ell(u_i) = \frac{1}{2|I|}\sum_{i\in I}\|u_i - u_i^T\|_{0,\Omega}^2, \tag{5.19c}$$

being $U^T = \{u_i^T\}_{i\in I}$ a set of clean images.Finally, V_i is the function space associated with the kernel generated by the noisy image f_i.

Problem (5.19a) has the drawback that for every nonlocal system in (5.19c), we need to compute and evaluate a different kernel, which results in a costly and long computation. Therefore, we replaced spaces $\{V_i\}_{i\in I}$ from (5.19c) with the single space $V_{\bar{F}}$, whose associated kernel $\gamma_{\bar{F}}$ corresponds to the noisy "mean image" $\bar{F} = \frac{1}{|I|}\sum_{i\in I} f_i$.

Following the analysis carried out for problem (3.17), we can further get a system of adjoint equations and a reduced derivative, resulting in the following optimality system

$$\mu(u_i,\psi)_{V_{\bar{F}}} + \lambda(u_i - f_i,\psi)_{0,\Omega} = 0, \qquad \forall \psi \in V_{\bar{F}}, i \in I, \tag{5.20a}$$

$$\mu(p_i,\phi)_{V_{\bar{F}}} + \lambda(p_i,\phi)_{0,\Omega} = \frac{1}{|I|}(u_i^T - u_i,\phi)_{0,\Omega}, \qquad \forall \phi \in V_{\bar{F}}, \tag{5.20b}$$

$$P_{[0,b]}\left(\lambda - c\sum_{i\in I}(u_i - f_i, p_i)\right) = \lambda, \qquad \forall c > 0. \tag{5.20c}$$

The training set of clean and noisy images is constructed as follows: we select ten images $\{u_i^T\}_{i\in I}$ from the FVC2000 Database and resize them down to 203×190 pixels. Then, pixelwise, we add Gaussian noise of variance $\sigma^2 = 10^3$ to obtain the noisy data $\{f_i\}_{i\in I}$.

We initialize the TR algorithm with $\lambda_0 = 100$, $b = 10^5$, and set the weight of the kernel to $w = 0.1$. The results are displayed in Figure 5.4 and Table 5.4, respectively. In the latter, we report the SSIM value of the reconstruction for each image. After 15 iterations of the algorithm, the optimal value of λ was 4.44416033704.

In Figure 5.4 we note an increase in the SSIM values of the denoised images, compared to the clean ones. However, it is noticeable that the shape of $\bar{F}$ is affecting the denoising process as most results present a noisy boundary around centered portions of the fingerprint (e.g. items B,F, and G) or higher regularization in non-centered portions of the fingerprint (e.g. items B, E, and F). This behavior is expected as intensity values are highest around each fingerprint, yet inside they are lowest. This propagates for the averaged fingerprint in $\bar{F}$, and thus the kernel does not regularize the overlapping areas of images with non-centered fingerprints.

Finally, in Figure 5.5 we show the outcome of the denoising algorithm (lower-level problem) for one noisy image that does not belong to the training set. The lower-level problem is solved in correspondence of the optimal (trained) λ. The behaviour described in the paragraph above replicates for the denoised validation image.

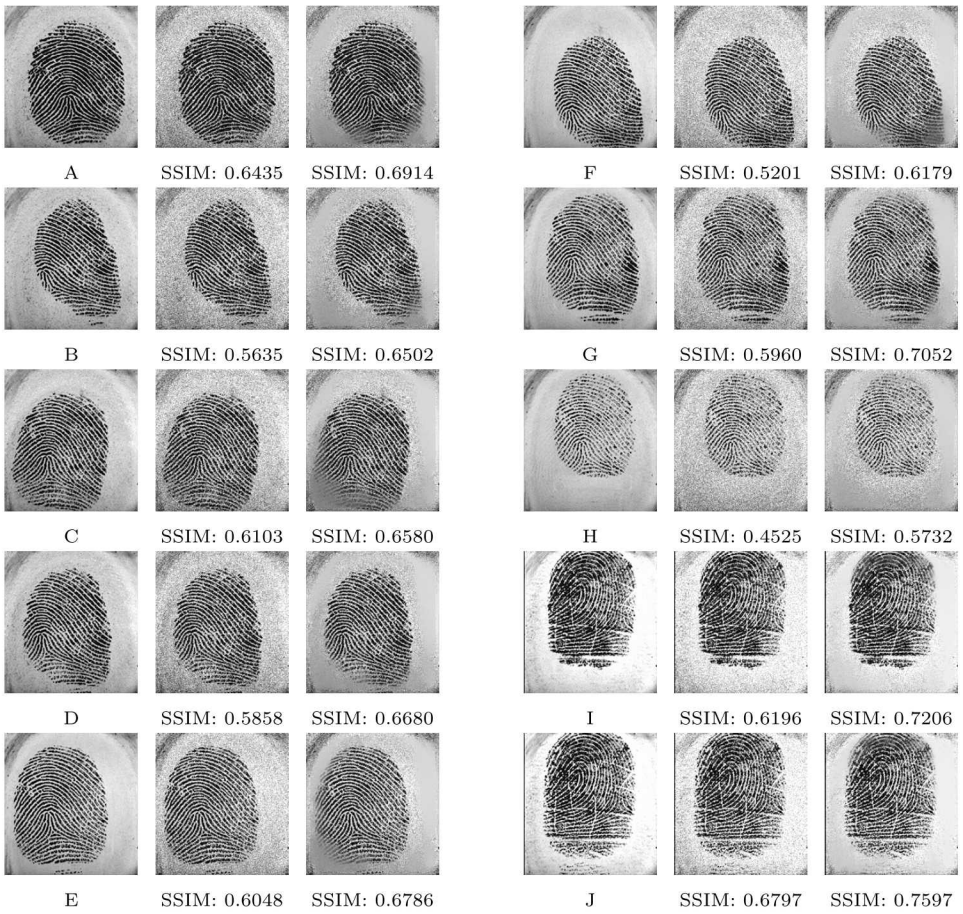


Fig. 5.4: Resulting images of scalar parameter training

Fig. 5.5: Validation image with trained parameter

Table 5.4: Results of batch training

Sample	SSIM	Sample	SSIM
A	0.6914	F	0.6179
B	0.6502	G	0.7052
C	0.6580	H	0.5732
D	0.6680	I	0.7206
E	0.6786	J	0.7597

5.4. Optimizing with respect to the weight w. We solve problem (4.6) with $w \in \mathcal{C} = [0, W]$. Note that every evaluation of the reduced objective functional (4.18) requires the numerical solution of (4.3), and hence requires updating γ_w. However, by definition, we have that $\gamma_{w_1} = \gamma_{w_0}^{w_1/w_0}$, which provides a fast way to get a new kernel for any $w_1, w_0 \in \mathcal{C}$.

Additionally, the gradient of the reduced objective functional (4.18) requires the computation of the linearized kernel $\widehat{\gamma}_w^h$, see (5.7). According to (5.8), we have

$$\begin{aligned} \widehat{\gamma}^h_{w,\boldsymbol{i},\boldsymbol{j}} &= \widetilde{\gamma}^h_{(1),\boldsymbol{i},\boldsymbol{j}} \\ &= \gamma^h_{w,\boldsymbol{i},\boldsymbol{j}} \cdot \left(-\mathbf{p}_{\boldsymbol{i}}(f^2) - \mathbf{p}_{\boldsymbol{j}}(f^2) + 2 \sum_{\boldsymbol{t} \in [-\rho:\rho]^2} \mathbf{P}_{\boldsymbol{i}}(f)(\boldsymbol{t}) \circ \mathbf{P}_{\boldsymbol{j}}(f)(\boldsymbol{t}) \right) \end{aligned} \tag{5.21}$$

for all pixels $\boldsymbol{i}, \boldsymbol{j} \in \mathcal{T}^h$. Letting $\breve{\gamma}^h_{i,j} = -\mathbf{p}_{\boldsymbol{i}}(f^2) - \mathbf{p}_{\boldsymbol{j}}(f^2) + 2\sum_{\boldsymbol{t}\in[-\rho:\rho]^2} \mathbf{P}_{\boldsymbol{i}}(f)(\boldsymbol{t}) \circ \mathbf{P}_{\boldsymbol{j}}(f)(\boldsymbol{t})$, $\widehat{\gamma}^h_w$ can be easily computed by the Hadamard product $\gamma^h_w \circ \breve{\gamma}^h$. Furthermore, $\breve{\gamma}^h$ depends on the noisy image f only. Thus, it is computed once and for all and we have $\gamma^h_w = \exp\{-w \cdot \breve{\gamma}^h\}$.

As numerically, the exponential function has a limited exponent range which prevents the effects of underflowing and overflowing, care has to be taken whenever choosing W and ι. Considering that the entries of γ^h_w are in $[0, 1]$, here we focus on avoiding underflow. This numerical condition occurs for images with high levels of noise, i.e., patches are highly dissimilar, resulting in a matrix with entries close to 0. Hence, if W is high, then the formula $\gamma_{w_1} = \gamma_{w_0}^{w_1/w_0}$ can return a constant matrix with no further possible updates. In contrast, if we make W small, then optimizing in images with low levels of noise, i.e., close-to-one patch distance, will result in an underestimation of the optimal value for w. This comes as W can be taken as low such that it is accepted as optimal, whereas the best image reconstruction could require $w \geqslant W$. In order to avoid this behavior, we set $W = K \max\{{}^{300}\!/_{\max \breve{\gamma}^h}, {}^{5}\!/_{\kappa} \times 10^{-5}\}$ with K given as in Table 5.5 and κ is a scaling parameter introduced below. This value is chosen so that whenever the entries of $\breve{\gamma}^h$ are small due to low levels of noise, cases (a) and (b), then w can be taken as big as some multiple of 300 that avoids underflow; and if the entries of $\breve{\gamma}^h$ are big due to high levels of noise, cases (c) and (d), then the values of w will be again limited to avoid a constant matrix. Now, for the acceptance tolerance we set $\iota = 10^{-10}$ which will be applied once for an initial kernel of parameter $w_{-1} = 10^{-6}$. This allows us to keep entries that could be deleted whenever $w > w_{-1}$, yet still remove entries with high dissimilarity values. Additionally, as in practice the numerical range $\mathcal{C}$ is small, we scale the argument of the objective function in order

to further avoid cancellation errors whenever reaching a local minimum. For this, we set the scaling parameter $\kappa = 10^{-6}$. Finally, we set $\lambda = 100$.

Table 5.5: Parameters for weight optimization

Sample	(a)	(b)	(c)	(d)
K	2	1	1	1
w_0	2×10^{-5}	1×10^{-5}	5×10^{-6}	2×10^{-6}

We initialize w with w_0 as reported in Table 5.5. The corresponding results are presented in Figure 5.6 and Table 5.6. For each clean image and its corresponding noisy sample, we report the optimal w, the corresponding SSIM, the number of iterations of the TR algorithm, and the dimensions of the image. We again observe a significant increase in the SSIM values and that the optimal parameters are within the interior of the convex set $\mathcal{C}$.

Lena (a) SSIM: 0.82 (b) SSIM: 0.67 (c) SSIM: 0.50 (d) SSIM: 0.34

(e) SSIM: 0.90 (f) SSIM: 0.84 (g) SSIM: 0.75 (h) SSIM: 0.60

Fig. 5.6: Results for the kernel weight optimization

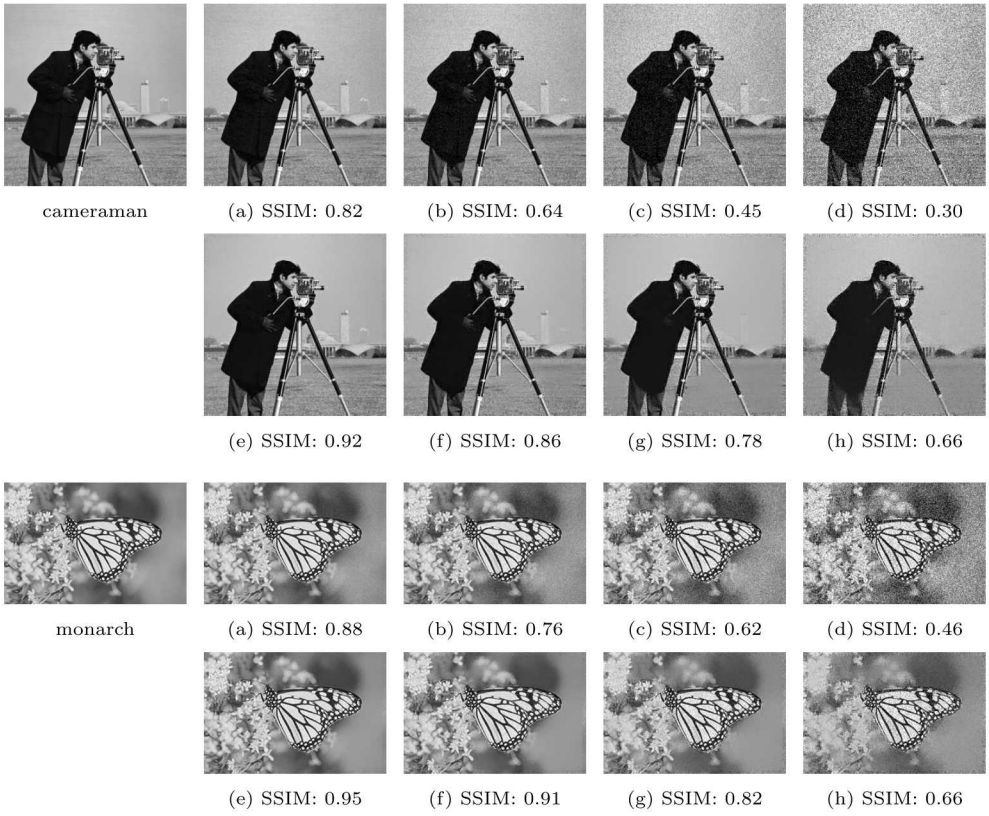


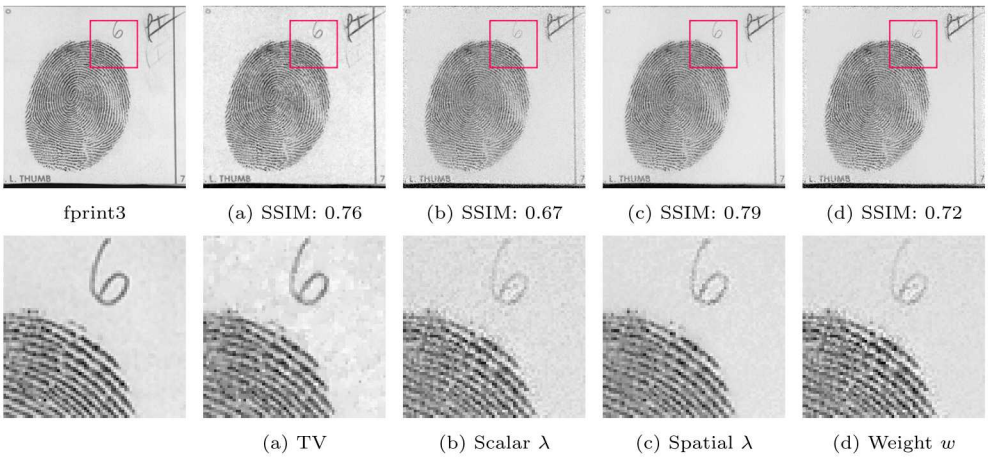
Fig. 5.6: Results for the kernel weight optimization

Table 5.6: Results of weight optimization

	Sample	**Best** w	**SSIM**	**Iteration Count**	$\frac{(N,M)}{\lvert g'(w)\rvert}$
Lena	(a)	0.00010058037591041884	0.9039	23	(256, 256)
	(b)	$3.479\,534\,200\,755 \times 10^{-5}$	0.8446	19	
	(c)	$1.365\,942\,222\,785 \times 10^{-5}$	0.7477	13	
	(d)	$6.010\,164\,391\,428 \times 10^{-6}$	0.5978	16	
cameraman	(a)	0.00016391632182323091	0.9249	29	(256, 256)
	(b)	$4.748\,817\,967\,295 \times 10^{-5}$	0.8638	19	
	(c)	$1.039\,828\,596\,144 \times 10^{-5}$	0.7764	11	
	(d)	$4.667\,619\,915\,317 \times 10^{-6}$	0.6591	11	
monarch	(a)	$9.386\,449\,960\,967 \times 10^{-5}$	0.9519	26	(256, 171)
	(b)	$1.062\,158\,555\,166 \times 10^{-5}$	0.9073	16	
	(c)	$2.500\,581\,037\,726 \times 10^{-5}$	0.8216	9	
	(d)	$4.946\,448\,008\,344 \times 10^{-6}$	0.6577	13	

5.5. Comparison between methods. Finally, we briefly compare the results obtained after optimizing problems (3.17), (3.7), and (4.6), and compare them with total variation denoising. For this purpose, we select a fingerprint image, named `fprint3`, and add Gaussian noise with standard deviation $\sigma = 10^3$. Each result is displayed in Figure 5.7, where the SSIM value of the optimal image is also provided. Moreover, a close-up of each image is plotted, in order to compare the graphical differences of each method.

Visually, it is clear that total variation denoising does not perform as well as most of the nonlocal approaches. The well-know staircasing effect of total variation is present in the fingerprint structure. In (b) and (d) the border of the fingerprint retains some noise, which comes from underfitting, and the intensity level of the number at the top is smoothed. In contrast, image (c) recovers the border of the fingerprint and the number is sharper.



fprint3 (a) SSIM: 0.76 (b) SSIM: 0.67 (c) SSIM: 0.79 (d) SSIM: 0.72

(a) TV (b) Scalar λ (c) Spatial λ (d) Weight w

Fig. 5.7: Comparison between local and nonlocal denoising methods

(a) Total Variation denoising, (b) Nonlocal denoising for scalar λ, (c) Nonlocal denoising for spatially dependent λ, (c) Nonlocal denoising for kernel scalar w.

6. Acknowledgments. MD was supported by Sandia National Laboratories (SNL), SNL is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energys National Nuclear Security Administration contract number DE-NA0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government. SAND report number xxx. This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research under Award Number DE-SC-0000230927.

REFERENCES

[1] B. Alali and R. Lipton. Multiscale dynamics of heterogeneous media in the peridynamic formulation. *Journal of Elasticity*, 106(1):71–103, 2012.

[2] H. Antil and S. Bartels. Spectral approximation of fractional PDEs in image processing and phase field modeling. *Computational Methods in Applied Mathematics*, 17(4):661–678, 2017.

[3] E. Askari. Peridynamics for multiscale materials modeling. *Journal of Physics: Conference Series, IOP Publishing*, 125(1):649–654, 2008.

[4] P.W. Bates and A. Chmaj. An integrodifferential model for phase transitions: Stationary solutions in higher space dimensions. *J. Statist. Phys.*, 95:1119–1139, 1999.

[5] D.A. Benson, S.W. Wheatcraft, and M.M. Meerschaert. Application of a fractional advection-dispersion equation. *Water Resources Research*, 36(6):1403–1412, 2000.

[6] Antoni Buades, Bartomeu Coll, and Jean-Michel Morel. A review of image denoising algorithms, with a new one. *Multiscale Modeling & Simulation*, 4(2):490–530, 2005.

[7] Antoni Buades, Bartomeu Coll, and Jean-Michel Morel. Self-similarity-based image denoising. *Communications of the ACM*, 54(5):109–117, 2011.

[8] N. Burch, M. D'Elia, and R. Lehoucq. The exit-time problem for a markov jump process. *The European Physical Journal Special Topics*, 223:3257–3271, 2014.

[9] Richard H. Byrd, Jorge Nocedal, and Robert B. Schnabel. Representations of quasi-Newton matrices and their use in limited memory methods. *Mathematical Programming*, 63(1-3):129–156, jan 1994.

[10] G. Capodaglio, M. D'Elia, P. Bochev, and M. Gunzburger. An energy-based coupling approach to nonlocal interface problems. *arXiv:2001.03696*, 2019.

[11] Coralia Cartis, Lindon Roberts, and Oliver Sheridan-Methven. Escaping local minima with derivative-free methods: a numerical investigation.

[12] Juan Carlos De los Reyes, C-B Schönlieb, and Tuomo Valkonen. Bilevel parameter learning for higher-order total variation regularisation models. *Journal of Mathematical Imaging and Vision*, 57(1):1–25, 2017.

[13] Juan Carlos De los Reyes and Carola-Bibiane Schönlieb. Image denoising: Learning the noise model via nonsmooth PDE-constrained optimization. *Inverse Problems & Imaging*, 7(4), 2013.

[14] A.H. Delgoshaie, D.W. Meyer, P. Jenny, and H. Tchelepi. Non-local formulation for multiscale flow in porous media. *Journal of Hydrology*, 531(1):649–654, 2015.

[15] M. D'Elia, Q. Du, M. Gunzburger, and R. Lehoucq. Nonlocal convection-diffusion problems on bounded domains and finite-range jump processes. *Computational Methods in Applied Mathematics*, 29:71–103, 2017.

[16] Marta D'Elia and Max Gunzburger. Optimal distributed control of nonlocal steady diffusion problems. *SIAM Journal on Control and Optimization*, 52(1):243–273, 2014.

[17] Alp Dener and Todd Munson. Accelerating Limited-Memory Quasi-Newton Convergence for Large-Scale Optimization. In João M. F. Rodrigues, Pedro J. S. Cardoso, Jânio Monteiro, Roberto Lam, Valeria V. Krzhizhanovskaya, Michael H. Lees, Jack J. Dongarra, and Peter M.A. Sloot, editors, *Computational Science – ICCS 2019*, pages 495–507, Cham, 2019. Springer International Publishing.

[18] Q. Du, M. D. Gunzburger, R. B. Lehoucq, and K. Zhou. Analysis and Approximation of Nonlocal Diffusion Problems with Volume Constraints. *SIAM Review*, 54(4):667–696, 2012.

[19] Marta D'Elia and Max Gunzburger. Identification of the diffusion parameter in nonlocal steady diffusion problems. *Applied Mathematics & Optimization*, 73(2):227–249, 2016.

[20] P. Fife. *Some nonclassical trends in parabolic and parabolic-like evolutions*, chapter Vehicular Ad Hoc Networks, pages 153–191. Springer-Verlag, New York, 2003.

[21] Guy Gilboa and Stanley Osher. Nonlocal linear image regularization and supervised segmentation. *Multiscale Modeling & Simulation*, 6(2):595–630, 2007.

[22] Guy Gilboa and Stanley Osher. Nonlocal operators with applications to image processing. *Multiscale Modeling & Simulation*, 7(3):1005–1028, 2008.

[23] Youn Doh Ha and Florin Bobaru. Characteristics of dynamic brittle fracture captured with peridynamics. *Engineering Fracture Mechanics*, 78(6):1156–1168, 2011.

[24] Michael Hintermüller, Carlos N Rautenberg, Tao Wu, and Andreas Langer. Optimal selection of the regularization function in a weighted total variation model. part ii: Algorithm, its analysis and numerical tests. *Journal of Mathematical Imaging and Vision*, 59(3):515–533, 2017.

[25] Michael Hintermüller and Tao Wu. Bilevel optimization for calibrating point spread functions in blind deconvolution. 2015.

[26] Karl Kunisch and Thomas Pock. A bilevel optimization approach for parameter learning in variational models. *SIAM Journal on Imaging Sciences*, 6(2):938–983, 2013.

[27] D. Littlewood. Simulation of dynamic fracture using peridynamics, finite element modeling, and contact. In *Proceedings of the ASME 2010 International Mechanical Engineering Congress and Exposition, Vancouver, British Columbia, Canada*, 2010.

[28] Dong C. Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(1-3):503–528, aug 1989.

[29] Yifei Lou, Xiaoqun Zhang, Stanley Osher, and Andrea Bertozzi. Image recovery via nonlocal operators. *Journal of Scientific Computing*, 42(2):185–197, 2010.

[30] Alvaro Maggiar, Andreas Wächter, Irina S. Dolinskaya, and Jeremy Staum. A Derivative-Free Trust-Region Algorithm for the Optimization of Functions Smoothed via Gaussian Convolution Using Adaptive Multiple Importance Sampling. *SIAM Journal on Optimization*, 28(2):1478–1507, jan 2018.

[31] M.M. Meerschaert and A. Sikorskii. *Stochastic models for fractional calculus*. Studies in mathematics, Gruyter, 2012.

[32] R. Metzler and J. Klafter. The random walk's guide to anomalous diffusion: a fractional dynamics approach. *Physics Reports*, 339(1):1–77, 2000.

[33] G. Pang, L. Lu, and G. E. Karniadakis. fpinns: Fractional physics-informed neural networks. *SIAM Journal on Scientific Computing*, 41(4):A2603–A2626, 2019.

[34] M. J. D. Powell. A fast algorithm for nonlinearly constrained optimization calculations. In *Lecture Notes in Mathematics*, pages 144–157. Springer Berlin Heidelberg, 1978.

[35] Joseph Salmon. On Two Parameters for Denoising with Non-Local Means. *IEEE Signal Processing Letters*, 17(3):269–272, mar 2010.

[36] A.A. Schekochihin, S.C. Cowley, and T.A. Yousef. Mhd turbulence: Nonlocal, anisotropic, nonuniversal? In *In IUTAM Symposium on computational physics and new perspectives in turbulence*, pages 347–354. Springer, Dordrecht, 2008.

[37] R. Schumer, D.A. Benson, M.M. Meerschaert, and B. Baeumer. Multiscaling fractional advection-dispersion equations and their solutions. *Water Resources Research*, 39(1):1022–1032, 2003.

[38] R. Schumer, D.A. Benson, M.M. Meerschaert, and S.W. Wheatcraft. Eulerian derivation of the fractional advection-dispersion equation. *Journal of Contaminant Hydrology*, 48:69–88, 2001.

[39] Mina Sharifymoghaddam, Soosan Beheshti, Pegah Elahi, and Masoud Hashemi. Similarity Validation Based Nonlocal Means Image Denoising. *IEEE Signal Processing Letters*, 22(12):2185–2188, dec 2015.

[40] S.A. Silling. Reformulation of elasticity theory for discontinuities and long-range forces. *Journal of the Mechanics and Physics of Solids*, 48:175–209, 2000.

[41] Stephen M Smith and J Michael Brady. Susan—a new approach to low level image processing. *International journal of computer vision*, 23(1):45–78, 1997.

[42] Carlo Tomasi and Roberto Manduchi. Bilateral filtering for gray and color images. In *Iccv*, volume 98, page 2, 1998.

[43] Giovanni Maria Troianiello. *Elliptic differential equations and obstacle problems*. Springer Science & Business Media, 2013.

[44] Leonid P Yaroslavsky. Digital picture processing: an introduction. *Applied Optics*, 25:3127, 1986.

[45] Gonglin Yuan, Zengxin Wei, and Maojun Zhang. An active-set projected trust region algorithm for box constrained optimization problems. *Journal of Systems Science and Complexity*, 28(5):1128–1147, nov 2014.