OPTIMAL SELECTION OF THE REGULARIZATION FUNCTION IN A WEIGHTED TOTAL VARIATION MODEL. PART I: MODELLING AND THEORY*

MICHAEL HINTERMÜLLER† AND CARLOS N. RAUTENBERG‡

Abstract. A weighted total variation model with a spatially varying regularization weight is considered. Existence of a solution is shown, and the associated Fenchel-predual problem is derived. For automatically selecting the regularization function, a bilevel optimization framework is proposed. In this context, the lower-level problem, which is parameterized by the regularization weight, is the Fenchel predual of the weighted total variation model and the upper-level objective penalizes violations of a variance corridor. The latter object relies on a localization of the image residual as well as on lower and upper bounds inspired by the statistics of the extremes.

Key words. Image restoration, weighted total variation regularization, spatially distributed regularization weight, Fenchel predual, bilevel optimization, variance corridor.

AMS subject classifications. 94A08, 68U10, 49K20, 49K30, 49K40, 49M37, 65K15

1. Introduction.

Image restoration is concerned with estimating a (true) image from degraded data. The degradation process considered in this work results from some (linear) transformation and a subsequent addition of noise (due to, e.g., data transmission). The latter is assumed to be of Gaussian type, i.e., it can be described as a highly oscillatory function defined over the image domain, with zero mean and a known quadratic deviation from the mean. In cases where the aforementioned image transformation is just the identity, the restoration process is termed image denoising.

In this paper, we consider variational approaches to image restoration which aim at minimizing a suitably chosen energy, which is composed by adding a weighted filter term to a data fidelity term. A properly selected weight (regularization parameter) then regulates the compromise between data misfit and removal of noise; see, e.g., [8] and references therein. It is known that a too large weight leads to an oversmoothing of the data and cartoon-like reconstruction results. On the other hand, a too small weight implies overfitting in the sense that the reconstruction contains unwanted noise effects (while retaining image details). In general, the choice of a scalar weight or regularization parameter represents a global compromise between a sufficient noise removal versus a good data fit.

As images typically contain regions with large homogeneous features, which admit strong filtering without compromising the reconstruction results, and regions with image details, which require some carefully chosen low filtering in order to preserve details, rather than choosing a scalar weight one is interested in selecting a distributed weight or *regularization function*. The latter choice, however, is notoriously difficult due to the typical ill-conditioning or ill-posedness of the underlying image reconstruction problem.

---

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

†Weierstrass Institute, Mohrenstrasse 39, 10117 Berlin, Germany and Department of Mathematics, Humboldt-University of Berlin, Unter den Linden 6, 10099 Berlin, Germany (michael.hintermueller@wias-berlin.de).

‡Department of Mathematics, Humboldt-University of Berlin, Unter den Linden 6, 10099 Berlin, Germany. (rautenberg@math.hu-berlin.de).

In this paper we propose an automated choice rule for the regularization function in an extension of the renowned total variation (TV) regularization model for image restoration; see the seminal work [53] by Rudin, Osher and Fatemi for the latter model with a scalar regularization parameter.

In fact, our generalization of the TV-model consists in considering a non-negative distributed filtering weight (or regularization function) $\alpha : \Omega \to \mathbb{R}$, with $\operatorname{ess\,inf} \alpha \geq \underline{\alpha} > 0$. The associated variational problem then reads

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

where

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

with an image domain $\Omega \subset \mathbb{R}^\ell$, given $f \in L^2(\Omega)$ and linear and continuous operator $K : L^2(\Omega) \to L^2(\Omega)$, i.e., $K \in \mathscr{L}(L^2(\Omega))$, and $\int_\Omega \alpha(x)|\mathcal{D}u|$ denoting the α-weighted total variation of u. Note that problem (P) requires the minimization of a non-differentiable objective, and it is posed in the non-reflexive Banach space $BV(\Omega)$. For the definition of the involved function spaces and further basic tools we refer to Section 2.

For constant values of α, it has been found that the Fenchel predual version of (P) (i.e., the optimization problem whose dual is (P)) is amenable to the development of efficient solvers; [35]. More specifically, the predual problem consists in minimizing a smooth quadratic objective subject to pointwise norm constraints. Regarding our localized regularization framework, in this paper we show that the Fenchel predual of (P) is given by

$$\text{(D)} \qquad \begin{aligned} &\text{minimize} \quad J_D(\mathbf{p}) := \frac{1}{2} |\operatorname{div} \mathbf{p} + K^* f|_B^2 \quad \text{over } \mathbf{p} \in H_0(\operatorname{div}) \\ &\text{subject to (s.t.)} \quad \mathbf{p} \in \mathbf{K}(\alpha), \end{aligned}$$

for

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

where $|w|_B^2 := (w, B^{-1}w)_{L^2(\Omega)}$ with $B = K^*K$, which–for simplicity–is assumed invertible, and $|\cdot|_\infty$ denotes the maximum norm on $\mathbb{R}^\ell$. Here, K^* denotes the adjoint of the operator K. Further, vector-valued quantities are written in bold font, "s.t." and "f.a.a." stand for "subject to" and "for almost all", respectively, and further definitions are deferred until the next section. Sometimes, when refering to (D), we even write (D(α)) to emphasize the dependence of the Fenchel (pre)dual on the filtering weight. Note that (D), in contrast to (P), is posed in a Hilbert space and its objective functional is differentiable. The structure of the constraints render the problem suitable for its rapid numerical solution by efficient semismooth Newton methods. This solver class has been successfully applied before in the case of constant α; see [35].

For the choice of α we propose a duality-based *bilevel optimization* framework, where, for a suitably chosen upper-level objective J, (D(α)) acts as a constraint which is parametrized by the function α, i.e., we seek to solve the problem

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

Here, $\mathcal{A}_{\mathrm{ad}}$ denotes the set of admissible filtering weights. We note that solving ($\mathbb{P}$) yields a fully automated selection mechanism for α. For its practical realization, in part II of this paper [38] an iterative solution algorithm is introduced, analyzed and numerically tested for a range of problems including Fourier and wavelet inpainting.

Regarding the existing literature, we point out that while the classical TV-model, i.e. (P) with $\alpha \equiv \mathit{const.} > 0$, has received a considerable amount of attention in image processing (see, e.g., [1, 16–18, 22, 40, 53] as well as the monograph [58] and the many references therein, to mention only a few), the literature on the generalized version with a distributed regularization weight is very scarce. The few available contributions include the PhD thesis [44, Section 2.4.4] and the associated report [45], as well as [6] where a weighted TV-regularization for vortex density models is considered, and the recent work in [15].

When it comes to automated choice rules for the regularization function $\alpha : \Omega \to \mathbb{R}$, the literature is essentially void. We note, however, that distributed data fidelity weights have been considered in [23, 24] for restoring gray-scale and color images subject to blur and Gaussian noise, respectively, and in [39] for problems involving random-valued impulse or salt-and-pepper noise. In these contributions, the adjustment of the fidelity weight is based on local statistical estimators and the statistics of the extremes. In finite dimensions, a technique based on a statistical multiresolution criterion can be found in [29, 43], and in [3] a statistical approach with variance estimators different from the ones in [23] is pursued. Finally, we mention [11] where a deterministic choice rule utilizing a pre-segmentation of the image and a piecewice constant fidelity weight is considered.

In the framework of (supervised) parameter learning in imagining models, a bilevel approach has been considered in [47] and [55], where well-posedness of the problem and optimality systems are derived. In the aforementioned contributions, a training set of images is used to obtain an objective, and the lower-level problem is not considered in its dual formulation leading to a different set of challenges. Concerning spatially dependent parameter learning a recent work can be found on [15] where the objective is also determined by a training set of images. Further contributions in this direction can be found in [14, 20, 28, 46, 50, 54], the PhD thesis [19], and references therein.

Besides studying (P) from a Fenchel-duality point of view, the motivation for this work is twofold: (i) Considering a choice rule for the regularization weight $\alpha : \Omega \to \mathbb{R}$ rather than for a data fidelity weight as in [3, 23, 24, 29, 43] allows for a large variety of data spaces including Fourier or wavelet domains; and (ii) our desire to utilize a bilevel optimization framework for the choice of α (compare ($\mathbb{P}$)) stems from the fact that such a formulation provides a unifying variational approach to both the restoration of u as well as the *optimal* choice of the regularization parameter. These aspects contrast the available literature.

Obviously, the quality of the obtained α (and consequently of u) depends significantly on the upper-level objective J in our bilevel approach. In this paper we select J motivated by the statistics of the extremes (in a discrete setting) and an associated acceptable local variance corridor.

From an optimization theoretic point of view, bilevel optimization falls into the realm of *mathematical programming with equilibrium constraints* (or MPEC, for short). This problem class typically suffers from notoriously degenerate constraints and requires sophisticated (non-smooth) analysis tools–other than classical Karush-Kuhn-Tucker (KKT) theory–for the derivation of stationarity conditions. For a rather general account of this problem class in finite dimensions we refer to the monographs [48,51], and to the selected works and monographs [9,34,41] for infinite dimensional settings. In mathematical image processing, bilevel problems have been used recently in [21,47,55] for parameter learning, and in [42] for calibrating point spread functions in blind deconvolution.

The rest of the paper is organized as follows. In the next section we fix notation and preliminaries. Properties of (P) and its Fenchel predual (D) are studied in the following section 3. Section 4 contains the formulation of the bilevel problem for the optimal choice of $\alpha:\Omega\to\mathbb{R}$ and an existence proof. A regularized version of the lower-level problem (D) along with differential stability of its solution as a function of the regularization weight α are the subjects of section 5. The associated bilevel problem is studied in section 6, where we also provide its first-order optimality conditions are provided.

2. Notations and preliminaries. Throughout this paper, $\Omega\subset\mathbb{R}^\ell$, with $\ell=1,2$, is a bounded connected open set with Lipschitz boundary $\partial\Omega$. As noted earlier, we assume $K\in\mathscr{L}(L^2(\Omega))$ with K^*K invertible and K^* the adjoint operator of K. When K^*K is singular, then one may require $K1\neq 0$ or add an additional regularization term $\frac{a}{2}\int_\Omega |u|^2\mathrm{d}x$ to J_P with $a>0$ in order to have (P) well-posed. The data $f\in L^2(\Omega)$ is supposed to be obtained as $f=Ku_{\text{true}}+\eta$, with u_{true} the desired recovery target and $\eta\in L^2(\Omega)$ an oscillatory function, with $\int_\Omega\eta\mathrm{d}x=0$, and $\int_\Omega|\eta|^2\mathrm{d}x=\sigma^2|\Omega|$. In a discrete setting, η can be regarded as white Gaussian noise with zero mean and standard deviation $\sigma>0$.

By $\mathbf{M}(\Omega,\mathbb{R}^N)$ we denote the space of N-valued Borel measures, which is the dual of $C_c(\Omega;\mathbb{R}^N)$, the space of continuous $\mathbb{R}^N$-valued functions with compact support in Ω. For $u:\Omega\to\mathbb{R}$ we denote by $\mathcal{D}u$ its distributional gradient, and the space of functions of bounded variation is defined by

$$BV(\Omega)=\{u\in L^1(\Omega):\mathcal{D}u\in\mathbf{M}(\Omega,\mathbb{R}^\ell)\}.$$

Furthermore, for $\boldsymbol{\mu}\in\mathbf{M}(\Omega,\mathbb{R}^N)$, $|\boldsymbol{\mu}|$ is the smallest nonnegative scalar Borel measure ν such that $|\boldsymbol{\mu}(B)|_1\le\nu(B)$ for all Borel sets B. Here, $|\cdot|_1$ denotes the ℓ^1-norm on $\mathbb{R}^N$. The norm on $\mathbf{M}(\Omega,\mathbb{R}^N)$ is defined as $|\boldsymbol{\mu}|_{\mathbf{M}(\Omega,\mathbb{R}^N)}=|\boldsymbol{\mu}|(\Omega)$, and from measure theory it is know that $|\boldsymbol{\mu}|(\Omega)=\int_\Omega\mathrm{d}|\boldsymbol{\mu}|$. Equivalently, this norm may be defined via duality by

$$|\boldsymbol{\mu}|(\Omega)=\sup_{\boldsymbol{\varphi}\in C_c(\Omega;\mathbb{R}^N)}\{\langle\boldsymbol{\mu},\boldsymbol{\varphi}\rangle_{C_c',C_c}:\ -\mathbf{1}\le\boldsymbol{\varphi}(x)\le\mathbf{1}\text{ a.e. on }\Omega\},$$

where $\mathbf{1}=(1,\dots,1)^\top\in\mathbb{R}^N$, "a.e." stands for "almost everywhere" in the sense of the Lebesgue measure and the supremum can be equivalently taken over $\boldsymbol{\varphi}\in C_0(\Omega;\mathbb{R}^N)$. Note that this definition implies for $\boldsymbol{\mu}=\{\mu_i\}_{i=1}^N$ with $\mu_i\in\mathbf{M}(\Omega,\mathbb{R})$, that $|\boldsymbol{\mu}|_{\mathbf{M}(\Omega,\mathbb{R}^N)}=|\mu_1|_{\mathbf{M}(\Omega,\mathbb{R})}+\cdots+|\mu_N|_{\mathbf{M}(\Omega,\mathbb{R})}$. Then, the BV-norm is defined as $|u|_{BV(\Omega)}:=|u|_{L^1(\Omega)}+|\mathcal{D}u|(\Omega)$. Below, the space of non-negative Borel measures is denoted by $\mathbf{M}^+(\Omega)$.

Before we commence our analysis, a few words on (P) are in order. From now on, the expression $\int_\Omega\alpha(x)|\mathcal{D}u|$ is denoted by $\int_\Omega\alpha|\mathcal{D}u|$, and it stands for the integral of α on Ω with respect to the measure $|\mathcal{D}u|$, i.e., $\int_\Omega\alpha\,\mathrm{d}|\mathcal{D}u|$. Hence, α *needs* to be a $|\mathcal{D}u|$-measurable function in order for $\int_\Omega\alpha|\mathcal{D}u|$ to be correctly defined, i.e., α is measurable with respect to the σ-algebra determined by the $\mathcal{D}u$-completion of the Borel σ-algebra. A sufficient condition for this is given by $\alpha\in C(\Omega)$, the space of continuous functions on Ω. In fact, for $\alpha \in C(\overline{\Omega})$ with $\alpha > 0$ on $\overline{\Omega}$ and $u \in BV(\Omega)$ the duality based representation implies

$$\int_\Omega \alpha|\mathcal{D}u| := \sup_{\mathbf{p}\in H_0(\mathrm{div})} \left\{ \int_\Omega u(x)\operatorname{div}\mathbf{p}(x)\mathrm{d}x : \ -\mathbf{1}\alpha(x) \le \mathbf{p}(x) \le \mathbf{1}\alpha(x) \text{ a.e. on } \Omega \right\},$$

as shown in Lemma 3.3 below. If $u \in W^{1,1}(\Omega)$, then $\int_\Omega \alpha|\mathcal{D}u| = \int_\Omega \alpha(x)|\nabla u(x)|_1 \mathrm{d}x$, where ∇u is the weak gradient of u.

Additionally, if $\alpha > 0$ is a constant, then $\int_\Omega \alpha|\mathcal{D}u|$ reduces to $\alpha|\mathcal{D}u|(\Omega)$ and problem (P) becomes the well-known TV-model [17, 53]:

$$\text{(TV)} \qquad \text{minimize} \quad \frac{1}{2}\int_\Omega |Ku - f|^2 \mathrm{d}x + \alpha \int_\Omega |\mathcal{D}u| \quad \text{over } u \in BV(\Omega).$$

The total variation $\int_\Omega |\mathcal{D}u|$ is known to preserve edges while filtering noise when solving (TV). The choice of $\alpha > 0$ represents a trade-off between filtering and data fitting. Indeed, (too) large values of $\alpha > 0$ strongly remove noise, but also filter details. Also the intensity value in regions containing homogeneous image features gets changed. On the other hand, (too) small α tends to recover details, but noise possibly remains in the reconstruction. While scalar α affects the quality of the reconstruction globally in Ω, in this paper we seek to automatically determine a spatially distributed function $\alpha : \Omega \to \mathbb{R}$ such that α is relatively large in homogeneous image regions for proper denoising and relatively small values in regions where details need to be recovered.

We still need some more notation. In fact, for a Banach space X its norm is written as $|\cdot|_X$. Given a sequence $\{x_n\}$ in X, strong convergence of the sequence to an element $x \in X$ is denoted by "$x_n \to x$" and weak convergence by "$x_n \rightharpoonup x$". The topological dual of X is written as X', and for $f \in X', x \in X$, the duality pairing between X' and X is $\langle f, x\rangle_{X',X}$. For two Banach spaces X_1 and X_2, we write $X_1 \hookrightarrow X_2$ when X_1 is continuously embedded into X_2.

The linear space of infinitely differentiable functions with compact support in Ω is denoted by $\mathscr{D}(\Omega)$. We make use of the usual real Lebesgue and Sobolev spaces $L^p(\Omega)$ and $W^{1,p}(\Omega)$, with $1 \le p \le \infty$, respectively, with norms $|v|_{L^p(\Omega)} = (\int_\Omega |v(x)|^p \mathrm{d}x)^{1/p}$ and $|w|_{W^{1,p}(\Omega)} = |w|_{L^p(\Omega)} + |\nabla w|_{L^p(\Omega)^N}$. The (standard) inner product in $L^2(\Omega)$ is $(\cdot,\cdot)$. Additionally, $W_0^{1,p}(\Omega)$ denotes the subspace of $W^{1,p}(\Omega)$ of functions whose trace is zero on $\partial\Omega$; together with $|v|_{W_0^{1,p}(\Omega)} = |\nabla v|_{L^p(\Omega)^N}$, it is a Banach space.

The Hilbert space $H_0(\mathrm{div})$ is defined as

$$\text{(2.1)} \qquad H_0(\mathrm{div}) := \{\mathbf{v} \in L^2(\Omega)^N : \operatorname{div}\mathbf{v} \in L^2(\Omega) \ \text{ and } \ \mathbf{v}\cdot\boldsymbol{n}|_{\partial\Omega} = 0\},$$

where $\boldsymbol{n}$ denotes the outer unit normal vector and the boundary condition is taken in the $H^{-1/2}(\partial\Omega)$-sense as the trace operator can be proven to be continuous from $H(\mathrm{div}) := \{\mathbf{v} \in L^2(\Omega)^N : \operatorname{div}\mathbf{v} \in L^2(\Omega)\}$ to $H^{-1/2}(\partial\Omega)$; see [30, Theorem 2.5].

We also need to handle convergence of closed, convex and non-empty subsets of a reflexive Banach space. For this matter we use Mosco convergence [49, 52].

Definition 2.1 (Mosco convergence). *Let $\mathbf{K}$ and $\mathbf{K}_n$, for each $n \in \mathbb{N}$, be non-empty, closed and convex subsets of a reflexive Banach space X. We say that the sequence $\{\mathbf{K}_n\}$ converges to $\mathbf{K}$ in the sense of Mosco as $n \to \infty$, written as*

$$\mathbf{K}_n \xrightarrow{\mathrm{M}} \mathbf{K},$$

if the following two conditions hold:

(i) *For all $v \in \mathbf{K}$, there exists $\{v_n\}$ such that $v_n \in \mathbf{K}_n$ and $v_n \to v$ in X.*
(ii) *If $v_n \in \mathbf{K}_n$ and $v_n \rightharpoonup v$ in X along a subsequence, then $v \in \mathbf{K}$.*

The main application of Mosco convergence lies in the following result on the stability of minimizers for convex constrained optimization problems. For its statement, let X, X_0, X_1 be reflexive Banach spaces such that $|x|_X := |Dx|_{X_0} + |x|_{X_1}$ where $D : X \to X_0$ is linear. Note that continuity of D is inferred since $|Dx|_{X_0} \leq |x|_X$.

PROPOSITION 2.2. *Let $J : X \to \mathbb{R}$ be convex, bounded from below, Fréchet differentiable with $\langle J'(x) - J'(y), x - y\rangle_{X',X} \geq c|D(x-y)|^2_{X_0}$ for some $c > 0$, and such that $J(x) \to +\infty$ if $|Dx|_{X_0} \to +\infty$. Suppose that $\mathbf{K}$ and $\mathbf{K}_n$, for each $n \in \mathbb{N}$, are closed and convex subsets of X containing the vector 0 and such that there exists $r > 0$ for which $|x|_{X_1} \leq r$ if $x \in \mathbf{K}$ or $x \in \mathbf{K}_n$.*

Then, if $x_n \in \arg\min_{x\in\mathbf{K}_n} J(x)$ and $\mathbf{K}_n \xrightarrow{\mathrm{M}} \mathbf{K}$, we have:

$$x_n \rightharpoonup x^* \quad \text{in } X \qquad \text{and} \qquad Dx_n \to Dx^* \quad \text{in } X_0,$$

along a subsequence not denoted specifically, where $x^* \in \arg\min_{x\in\mathbf{K}} J(x)$.

Proof. Since J is convex and continuous (for being Fréchet differentiable), it is weakly lower semicontinuous. Let $\{x_k\}$ be an infimizing sequence for $\inf_{x\in\mathbf{K}} J(x)$. Since $\{J(x_k)\}$ is bounded, it follows that $\{|Dx_k|_{X_0}\}$ is also bounded. Additionally, since $x_k \in \mathbf{K}$, we have $|x_k|_{X_1} \leq r$ which implies that $\{x_k\}$ is bounded in X, since the norm in X is given by $|\cdot|_X = |D(\cdot)|_{X_0} + |\cdot|_{X_1}$. Hence, there is a subsequence that converges weakly to a minimizer of J on $\mathbf{K}$ (note that $\mathbf{K}$ is weakly closed as it is norm closed).

Similarly we prove that there exists a sequence $\{x_n\}$ with $x_n \in \arg\min_{x\in\mathbf{K}_n} J(x)$. Since $0 \in \mathbf{K}_n$, $J(x_n) \leq J(0)$ and $x_n \in \mathbf{K}_n$, it follows that $\{x_n\}$ is bounded in X. Hence, it weakly converges along a subsequence to some $x^* \in X$. The convergence $\mathbf{K}_n \xrightarrow{\mathrm{M}} \mathbf{K}$ implies that $x^* \in \mathbf{K}$, and that for an arbitrary $y \in \mathbf{K}$, there is a sequence $\{y_n\}$ with $y_n \in \mathbf{K}_n$ such that $y_n \to y$ in X. Therefore,

$$J(x^*) \leq \varliminf_{n\to\infty} J(x_n) \leq \varliminf_{n\to\infty} J(y_n) = J(y).$$

As $y \in \mathbf{K}$ was arbitrary, we find $x^* \in \arg\min_{x\in\mathbf{K}} J(x)$.

Since $x_n \in \arg\min_{x\in\mathbf{K}_n} J(x)$ and J is convex and differentiable, we equivalenty get $\langle J'(x_n), x - x_n\rangle_{X',X} \geq 0$ for all $x \in \mathbf{K}_n$. We also have $\langle J'(x) - J'(y), x - y\rangle_{X',X} \geq c|D(x-y)|^2_{X_0}$. Again, using that $\mathbf{K}_n \xrightarrow{\mathrm{M}} \mathbf{K}$, there is $\{z_n\}$ with $z_n \to x^*$ in X and $z_n \in \mathbf{K}_n$. Hence, we obtain

$$\langle J'(z_n), z_n - x_n\rangle_{X',X} \geq \langle J'(x_n) - J'(z_n), x_n - z_n\rangle_{X',X} \geq c|D(x_n - z_n)|^2_{X_0}.$$

However, $X \ni x \mapsto J'(x) \in X'$ is continuous (since J is Fréchet differentiable) and $z_n - x_n \rightharpoonup 0$ in X, which implies that $|D(x_n - z_n)|_{X_0} \to 0$. Since $z_n \to x^*$ in X, we obtain $Dx_n \to Dx^*$ in X_0. ☐

3. Problems (P) **and** (D)**.** As mentioned already in the introduction, for a constant α the Fenchel predual problem of (P) is given by (D); see [35]. Now we establish the same result for a spatially varying α, provided certain regularity assumptions hold true. In part II of this work [38], the predual of (P) is then taken as the starting point for the design of efficient solution algorithms. In particular, the predual (D) consists in minimizing a smooth objective over a Hilbert space subject to pointwise norm constraints, which can be tackled via Moreau-Yosida regularization and a semismooth Newton solver.

Analytically, it should be noted that the duality result hinges on dense embeddings of closed convex sets in function space. In particular, we need to known whether smooth vector fields satisfying the aforementioned pointwise constraints are dense in the set of functions bounded by the same pointwise constraints and of (lower) regularity dictated by the predual problem; see the proof of Lemma 3.3 and Proposition 4.1 below. A word of caution is in order here! In fact, such embedding results are delicate as they *can not* be inferred from the dense embedding of the underlying spaces (see [36], and [37] for a counterexample).

For our analysis it is convenient to introduce the following set valued map $\mathbf{K}$.

DEFINITION 3.1. *Let X be a set of $\mathbb{R}^M$-valued functions on $\Omega \subset \mathbb{R}^\ell$. Then we define*

$$\mathbf{K}(\alpha, X) := \{\mathbf{p} \in X : -\mathbf{1}\alpha(x) \le \mathbf{p}(x) \le \mathbf{1}\alpha(x) \text{ a.e. on } \Omega\},$$

where $\mathbf{1} = \{1\}_{i=1}^M$ and $\alpha : \Omega \to \mathbb{R}$. Whenever X is omitted, then we refer to $X = H_0(\mathrm{div})$, i.e., $\mathbf{K}(\alpha) := \mathbf{K}(\alpha, H_0(\mathrm{div}))$.

For $|v|_B^2 := (v, B^{-1}v)$ and $B = K^*K$, we recall (D), i.e.,

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

Due to convexity, (D) is equivalent to the variational inequality: Find $\mathbf{p} \in \mathbf{K}(\alpha)$ such that

$$\langle A\mathbf{p} + \mathbf{f}, \mathbf{q} - \mathbf{p}\rangle_{H_0(\mathrm{div})^*, H_0(\mathrm{div})} \ge 0, \quad \forall \mathbf{q} \in \mathbf{K}(\alpha), \tag{3.1}$$

where $A : H_0(\mathrm{div}) \to H_0(\mathrm{div})^*$ is defined as $A\mathbf{p} := -\nabla B^{-1}\operatorname{div}\mathbf{p}$, with $\mathbf{p} \in H_0(\mathrm{div})$ and $\mathbf{f} = -\nabla B^{-1}K^*f \in H_0(\mathrm{div})^*$.

The next result clarifies the existence and uniqueness properties of a solution of (P) as well as (D), respectively.

PROPOSITION 3.2. *The following existence results hold true:*

(a) *If $\alpha \in C(\overline{\Omega})$ and $\alpha(x) > 0$ for all $x \in \overline{\Omega}$, then* (P) *admits a unique solution.*

(b) *If $\alpha : \Omega \to \mathbb{R}$ is measurable, non-negative and bounded, then* (D) *has a solution. For any two solutions $\mathbf{p}_1$ and $\mathbf{p}_2$ to* (D)*, one has $\operatorname{div}\mathbf{p}_1 = \operatorname{div}\mathbf{p}_2$.*

Proof. Let $\{u_n\}$ in $BV(\Omega)$ be an infimizing sequence for $J_P(\cdot, \alpha)$. Since $\ell \le 2$, the embedding $BV(\Omega) \hookrightarrow L^2(\Omega)$ is continuous (see for example [7, Theorem 10.1.3.]) and

$$c|\mathcal{D}u|(\Omega) \le \int_\Omega \alpha|\mathcal{D}u| \le C|\mathcal{D}u|(\Omega), \tag{3.2}$$

with $c := \min_{x\in\overline{\Omega}} \alpha(x) > 0$ and $C := \max_{x\in\overline{\Omega}} \alpha(x) > 0$, we observe that $c|\mathcal{D}u_n|(\Omega) \le J_P(u_n, \alpha) < \infty$. Therefore, $\{|\mathcal{D}u_n|(\Omega)\}$ is bounded. Since $u_n \in BV(\Omega) \subset L^1(\Omega)$, it can be written as $u_n = \bar{u}_n + w_n$ with $\bar{u}_n = \frac{1}{|\Omega|}\int u_n \mathrm{d}x$ and $\int_\Omega w_n \mathrm{d}x = 0$. By the Sobolev inequality (see [31]), we have $|u_n - \bar{u}_n|_{L^2} \le c_1|\mathcal{D}u_n|(\Omega)$ with some $c_1 > 0$. Hence, $\{|w_n|_{L^2}\}$ and $\{|Kw_n|_{L^2}\}$ are bounded. Since $\{u_n\}$ is an infimizing sequence, $\{|Ku_n|_{L^2}\}$ is bounded, and then $Ku_n = \bar{u}_n K1 + Kw_n$ yields that $\{|\bar{u}_n|\}$ is bounded (note that $K1 \ne 0$ since K is one-to-one). Hence, $\{u_n\}$ is bounded in $L^1(\Omega)$ and, thus, $\{u_n\}$ is bounded in $BV(\Omega)$, as well.

The above arguments yield the existence of $u^* \in BV(\Omega)$ such that $u_n \rightharpoonup u^*$ in $L^2(\Omega)$, $u_n \to u^*$ in $L^1(\Omega)$ and $|\mathcal{D}u^*|(\Omega) \le \underline{\lim}\,|\mathcal{D}u_n|(\Omega)$ along a subsequence. This follows from the compact embedding $BV(\Omega) \hookrightarrow L^1(\Omega)$; see Theorem 10.1.4 and

Proposition 10.1.1 in [7]. We also have $Ku_n \rightharpoonup Ku^*$ in $L^2(\Omega)$. Further, by Lemma 3.3 and since $\mathbf{K}(\alpha, \mathscr{D}(\Omega)^\ell)$ is $H_0(\operatorname{div})$-dense in $\mathbf{K}(\alpha)$ (see [36]), we observe

$$\int_\Omega \alpha|\mathcal{D}u^*| = \sup_{\mathbf{p}\in\mathbf{K}(\alpha,\mathscr{D}(\Omega)^\ell)} (u^*, -\operatorname{div}\mathbf{p}) = \sup_{\mathbf{p}\in\mathbf{K}(\alpha,\mathscr{D}(\Omega)^\ell)} \Big(\varliminf_{n\to\infty} (u_n, -\operatorname{div}\mathbf{p})\Big)$$
$$\leq \varliminf_{n\to\infty} \Big(\sup_{\mathbf{p}\in\mathbf{K}(\alpha,\mathscr{D}(\Omega)^\ell)} (u_n, -\operatorname{div}\mathbf{p})\Big) = \varliminf_{n\to\infty} \int_\Omega \alpha|\mathcal{D}u_n|.$$

Therefore, for some subsequence of $\{u_n\}$, which we also denote by $\{u_n\}$, we have

$$\frac{1}{2}\int_\Omega |Ku^* - f|^2 \mathrm{d}x + \int_\Omega \alpha|\mathcal{D}u^*| \leq \varliminf_{n\to\infty} \left(\frac{1}{2}\int_\Omega |Ku_n - f|^2 \mathrm{d}x + \int_\Omega \alpha|\mathcal{D}u_n| \right),$$

i.e., u^* is a minimizer.

Let $u, v \in BV(\Omega)$, then $\mathcal{D}w = \sigma_1\mathcal{D}u + \sigma_2\mathcal{D}v$ for $w = \sigma_1 u + \sigma_2 v$ and for some $\sigma_1, \sigma_2 \geq 0$. Hence, $\nu = \sigma_1|\mathcal{D}u| + \sigma_2|\mathcal{D}v| \in \mathbf{M}^+(\Omega)$, and by definition $|\mathcal{D}w(B)| \leq \nu(B)$ for all Borel sets B. However, $|\mathcal{D}w| \in \mathbf{M}^+(\Omega)$ is the smallest measure μ such that $|\mathcal{D}w(B)| \leq \mu(B)$ for all Borel sets B. We therefore have $|\mathcal{D}w|(B) \leq \nu(B)$ for all Borel sets. As a consequence we get

$$\int_\Omega \alpha|\mathcal{D}(\sigma_1 u + \sigma_2 v)| \leq \sigma_1 \int_\Omega \alpha|\mathcal{D}u| + \sigma_2 \int_\Omega \alpha|\mathcal{D}v|.$$

In particular, this implies that $u \mapsto \int_\Omega \alpha|\mathcal{D}u|$ is convex. Additionally, since $B = K^*K$ is invertible, then K is injective and we have that $u \mapsto \frac{1}{2}\int_\Omega |Ku - f|^2\mathrm{d}x$ is strictly convex. Hence, $u \mapsto J_P(u, \alpha)$ is strictly convex and the uniqueness of the minimizers follows from classical arguments (see for example [1]).

Note that B and B^{-1} are bounded, self-adjoint and invertible. Hence, $(v, w)_B$ and $(v, w)_{L^2}$ are equivalent inner products. In fact, $c_1|v|_{L^2} \leq |v|_B \leq c_2|v|_{L^2}$ for some $0 < c_1 \leq c_2$, and for all $v \in L^2(\Omega)$. The set $\mathbf{K}(\alpha)$ is a closed, convex and non-empty subset of $H_0(\operatorname{div})$. Let $\mathbf{p} \mapsto \mathrm{I}_{\mathbf{K}(\alpha)}(\mathbf{p})$ be its indicator function (i.e., $\mathrm{I}_{\mathbf{K}(\alpha)}(\mathbf{p}) = 0$ if $\mathbf{p} \in \mathbf{K}(\alpha)$ and $\mathrm{I}_{\mathbf{K}(\alpha)}(\mathbf{p}) = \infty$ otherwise). Then, $\mathbf{p} \mapsto J_D(\mathbf{p}) + \mathrm{I}_{\mathbf{K}(\alpha)}(\mathbf{p})$ is lower semicontinuous, convex and coercive (note that α is bounded, and then $\mathbf{K}(\alpha)$ is bounded in $L^2(\Omega)$). Therefore, the existence of solutions to (D) follow immediately.

Let $\mathbf{p}_1$ and $\mathbf{p}_2$ be two different solutions to (D). Since J_D is convex, $\sigma\mathbf{p}_1 + (1-\sigma)\mathbf{p}_2$ is also a solution to (D) for $0 \leq \sigma \leq 1$. Additionally, since $B = K^*K$ is invertible, the norm $v \mapsto |v|_B$ is strictly convex which implies that $v \mapsto \frac{1}{2}|v + K^*f|_B^2$ is also strictly convex. Consequently, $\operatorname{div}\mathbf{p}_1 = \operatorname{div}\mathbf{p}_2$ a.e. on Ω. ☐

It turns out that (D) and (P) are dual to each other as specified below in Theorem 3.4, an extension of [35, Theorem 2.2]. Its proof crucially depends on the fact that $\mathbf{K}(\alpha)$ contains dense (in the sense of $H_0(\operatorname{div})$) subsets of more regular functions. It should be noted that such a density result cannot hold true in general. Indeed, we refer to [36] for a counterexample. The following lemma, which relies on our density result, is needed in the proof of Theorem 3.4.

Lemma 3.3. *Let $v \in L^2(\Omega)$, $\alpha \in C(\overline{\Omega})$ and $\alpha(x) > 0$ for all $x \in \overline{\Omega}$. Then,*

$$\sup_{\mathbf{p}\in\mathbf{K}(\alpha)} (v, -\operatorname{div}\mathbf{p}) = \begin{cases} +\infty, & \text{if } v \notin BV(\Omega), \\ \int_\Omega \alpha|\mathcal{D}v|, & \text{if } v \in BV(\Omega). \end{cases} \tag{3.3}$$

Proof. The set $\mathbf{K}(\alpha, C_0^1(\Omega)^\ell)$ is dense, in the sense of $H(\mathrm{div})$, in $\mathbf{K}(\alpha)$, cf. [36]. Hence, the supremum in (3.3) can be taken over $\mathbf{K}(\alpha, C_0^1(\Omega)^\ell)$ without changing its value. Next, let $\hat\alpha = \inf_{x\in\overline{\Omega}} \alpha(x) > 0$ and $v \notin BV(\Omega)$. Then

$$\sup_{\mathbf{p}\in\mathbf{K}(\alpha, C_0^1(\Omega)^\ell)} (v, -\operatorname{div}\mathbf{p}) \geq \sup_{\mathbf{p}\in\mathbf{K}(\hat\alpha, C_0^1(\Omega)^\ell)} (v, -\operatorname{div}\mathbf{p}) = \hat\alpha \sup_{\mathbf{p}\in\mathbf{K}(1, C_0^1(\Omega)^\ell)} (v, -\operatorname{div}\mathbf{p}) = +\infty,$$

where we have used that $\sup_{\mathbf{p}\in\mathbf{K}(1,C_0^1(\Omega)^\ell)}(v, -\operatorname{div}\mathbf{p}) = \infty$ if $v \notin BV(\Omega)$; see [4, Prop. 3.6, page 120] or [31].

Also, the set $\mathbf{K}(\alpha, \mathscr{D}(\Omega)^\ell)$ is dense, in the sense of $H(\mathrm{div})$, in $\mathbf{K}(\alpha)$ [36]. Thus, $\mathbf{K}(\alpha)$ can be replaced by $\mathbf{K}(\alpha, \mathscr{D}(\Omega)^\ell)$ in (3.3). Furthermore , $\mathbf{K}(1, \mathscr{D}(\Omega)^\ell)$ is dense, in the sense of C_0, in $\mathbf{K}(1, C_0(\Omega)^\ell)$ [36]. Then

$$\begin{aligned}
\sup_{\mathbf{p}\in\mathbf{K}(\alpha,\mathscr{D}(\Omega)^\ell)} (v, -\operatorname{div}\mathbf{p}) &= \sup_{\mathbf{p}\in\mathbf{K}(\alpha,\mathscr{D}(\Omega)^\ell)} \langle \mathcal{D}v, \mathbf{p}\rangle_{C_0', C_0} = \sup_{\mathbf{p}\in\mathbf{K}(1,\mathscr{D}(\Omega)^\ell)} \langle \mathcal{D}v, \alpha\mathbf{p}\rangle_{C_0', C_0} \\
&= \sup_{\mathbf{p}\in\mathbf{K}(1,\mathscr{D}(\Omega)^\ell)} \langle \alpha\mathcal{D}v, \mathbf{p}\rangle_{C_0', C_0} = \sup_{\mathbf{p}\in\mathbf{K}(1,C_0(\Omega)^\ell)} \langle \alpha\mathcal{D}v, \mathbf{p}\rangle_{C_0', C_0} \\
&= |\alpha\mathcal{D}v|(\Omega).
\end{aligned}$$

Note, however, that $|\alpha\mathcal{D}v|(\Omega) = \int_\Omega \alpha|\mathcal{D}v|$, (see [26, Theorem 20, III.2]). We also observe that the aforementioned reference applies to real-valued measures. Its extension to $\mathbb{R}^\ell$-valued ones follows from an analogous proof. □

For the ease of reference, we briefly recall the Fenchel duality result [10, 27] adapted to our setting. Let V and Y be Banach spaces, with V^* and Y^* their respective topological duals. Suppose that $\Lambda \in \mathscr{L}(V, Y)$, and let $\mathcal{F} : V \to \mathbb{R}\cup\{+\infty\}$ and $\mathcal{G} : Y \to \mathbb{R}\cup\{+\infty\}$ be convex lower semicontinuous functionals not identically $+\infty$. Assume there exists $v_0 \in V$ such that $\mathcal{F}(v_0) < +\infty$, $\mathcal{G}(\Lambda v_0) < +\infty$, and $\mathcal{G}$ is continuous at Λv_0. Then, one has

$$\inf_{v\in V} \mathcal{F}(v) + \mathcal{G}(\Lambda v) = \sup_{w\in Y^*} -\mathcal{F}^*(\Lambda^* w) - \mathcal{G}^*(w), \tag{3.4}$$

where $\mathcal{F}^* : V^* \to \mathbb{R}\cup\{+\infty\}$ and $\mathcal{G}^* : Y^* \to \mathbb{R}\cup\{+\infty\}$ denote the conjugates of $\mathcal{F}$ and $\mathcal{G}$, respectively, i.e.,

$$\mathcal{F}^*(v^*) := \sup_{v\in V} \langle v, v^*\rangle - \mathcal{F}(v),$$

and analogously for $\mathcal{G}^*$. Further, $(\hat v, \hat w)$ are solutions to the two optimization problems in (3.4) if and only if

$$-\hat w \in \partial\mathcal{G}(\Lambda\hat v) \qquad \text{and} \qquad \Lambda^*\hat w \in \partial\mathcal{F}(\hat v), \tag{3.5}$$

where $\partial\mathcal{F}$ and $\partial\mathcal{G}$ denote the subdifferentials of $\mathcal{F}$ and $\mathcal{G}$, respectively.

We now state our duality theorem.

THEOREM 3.4. *Let $\alpha \in C(\overline{\Omega})$ with $\alpha(x) > 0$ for all $x \in \overline{\Omega}$. Then the Fenchel dual of* (D) *is given by* (P). *If $\hat{\mathbf{p}}$ is a solution to* (D) *and $\hat u$ is the solution to* (P), *then they satisfy*

$$B\hat u = \operatorname{div}\hat{\mathbf{p}} + K^* f, \ \text{ and } \langle(-\operatorname{div})^*\hat u, \mathbf{p} - \hat{\mathbf{p}}\rangle_{H_0(\mathrm{div})^*, H_0(\mathrm{div})} \leq 0, \ \forall \mathbf{p}\in\mathbf{K}(\alpha). \tag{3.6}$$

Proof. The Fenchel duality result is applied with $V = H_0(\mathrm{div})$, $Y = Y^* = L^2(\Omega)$, $\Lambda = -\operatorname{div}$, $\mathcal{G} : Y \to \mathbb{R}$ defined as $\mathcal{G}(v) = \frac{1}{2}|v - K^* f|_B^2$ and with its convex conjugate

$\mathcal{G}^* : Y \to \mathbb{R}$ given by $\mathcal{G}^*(v) = \frac{1}{2}|Kv+f|^2_{L^2} - \frac{1}{2}|f|^2_{L^2}$; compare [35]. For $\mathcal{F} : V \to \mathbb{R}$ given by $\mathcal{F}(\mathbf{p}) = \mathrm{I}_{\mathbf{K}(\alpha)}(\mathbf{p})$ its convex conjugate $\mathcal{F}^* : V^* \to \mathbb{R}$ satisfies

$$\mathcal{F}^*((-\operatorname{div})^* v) = \sup_{\mathbf{p}\in\mathbf{K}(\alpha)} (v, -\operatorname{div}\mathbf{p}).$$

As a consequence $\mathcal{F}^*(\Lambda^* v) = \int_\Omega \alpha|\mathcal{D}v|$ if $v \in BV(\Omega)$ and $\mathcal{F}^*(\Lambda^* v) = +\infty$ if $v \notin BV(\Omega)$, by Lemma 3.3. This proves that the Fenchel dual of (D) is given by (P).

The relations (3.6) are obtained directly from an application of (3.5). ∎

4. The bilevel optimization problem for choosing α. The main goal of this paper is to provide an optimization-theoretic framework for automatically choosing the distributed regularization weight $\alpha : \Omega \to \mathbb{R}^+$. In contrast to [25], where instead of α a distributed data fidelity weight λ is chosen by an external update mechanism, we suggest a *bilevel optimization* problem for selecting α.

Typically, the classic TV-models can be formulated in terms of a global (image) residual; see [17]. In what follows, we are interested in a localized residuals, as these will induce a regularization function α. Our starting point is similar to [11], where, given a normalized weight $w \in L^\infty(\Omega\times\Omega)$ with $\int_\Omega\int_\Omega w(x,y)\,dxdy = 1$, the image residual (also associated with the noise variance) at u is localized by $S : L^2(\Omega) \to L^\infty(\Omega)$ with

$$S(u)(x) := \int_\Omega w(x,y)\,(Ku-f)^2(y)\mathrm{d}y.$$

Since $B = K^*K \in \mathscr{L}(L^2(\Omega))$ is assumed to be invertible, at a primal-dual solution $(u,\mathbf{p})$ fulfilling (3.6) the localized residual may be expressed as

$$R(v)(x) := \int_\Omega w(x,y)\,(KB^{-1}v + (KB^{-1}K^* - I)f)^2(y)\mathrm{d}y \tag{4.1}$$

with $v = \operatorname{div}\mathbf{p}$. Note that, like S, R maps from $L^2(\Omega)$ into $L^\infty(\Omega)$.

In what follows, we focus on the Fenchel predual problem (D) for setting up a bilevel optimization problem for fixing α. In contrast to proceeding with a primal formulation, this has two immediate advantages: (i) As (D) becomes a constraint of the overall problem, we may handle a problem of obstacle type giving rise to a variational inequality (VI) of the first kind, rather than a VI of the second kind representing the first-order optimality condition of (P). (ii) Compared to the approach in [25], the requirements on w may be weakened as coercivity of $z \mapsto \int_\Omega\int_\Omega w(x,y)z^2(y)\mathrm{d}y\mathrm{d}x$ in $L^2(\Omega)$ is no longer required. Besides these analytical benefits, the dual approach also has advantages in the design of efficient numerical solution algorithms.

We continue by specifying a suitable upper-level objective. In fact, motivated by the statistics of the extremes in a discrete setting [25, 32, 33], we consider a current selection of α acceptable whenever the image residual resides within a suitably choosen variance corridor; otherwise we penalize violations of this feasibility corridor. For this purpose, we define $F : L^2(\Omega) \to \mathbb{R}^+_0$ with

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

where $0 < \underline{\sigma} \le \overline{\sigma} < \infty$ are typically chosen using statistical properties involving the noise contained in the measurement f. In fact, practical choices of $\overline{\sigma}, \underline{\sigma}$ based on the statistics of the extreme are discussed in detail in part II of this work [38]. The previous choice of objective contrasts with existing literature choices that, in the context of parameter learning problems, consider objectives involving (among others) least squares terms, and regularized total variation costs (see, for example, [21, 22]).

Now we state our bilevel problem.

Problem ($\mathbb{P}$). Let $\lambda > 0$ and $p > \max(2, \ell)$. Consider

$$(\mathbb{P}) \quad \begin{aligned} &\text{minimize} \quad J(\mathbf{p}, \alpha) := F \circ R(\operatorname{div} \mathbf{p}) + \frac{\lambda}{p}|\alpha|^p_{W^{1,p}(\Omega)} \text{ over } (\mathbf{p}, \alpha) \in H_0(\operatorname{div}) \times \mathcal{A}_{\mathrm{ad}}, \\ &\text{s.t. } \mathbf{p} \in \arg\min\{J_D(\mathbf{q}) : \mathbf{q} \in \mathbf{K}(\alpha)\}, \end{aligned}$$

where the non-empty set of admissible weight functions is given by

$$\mathcal{A}_{\mathrm{ad}} := \{\alpha \in H^1(\Omega) : \underline{\alpha} \leq \alpha \leq \overline{\alpha}, \text{ a.e. on } \Omega\}, \tag{4.2}$$

for $\underline{\alpha}, \overline{\alpha} \in L^2(\Omega)$ with $0 < \epsilon_0 \leq \underline{\alpha}(x) < \overline{\alpha}(x) - \epsilon_1$ a.e. on Ω for some $\epsilon_1 > 0$.

A few words on Problem ($\mathbb{P}$) are in order. Suppose that $\mathbf{p} \in H_0(\operatorname{div})$ is a solution to (D) for some positive $\alpha \in C(\overline{\Omega})$. Then the solution $u \in BV(\Omega)$ of (P) for that α satisfies $Bu = \operatorname{div} \mathbf{p} + K^* f$ (see (3.6)). This implies

$$KB^{-1} \operatorname{div} \mathbf{p} + (KB^{-1}K^* - I)f = Ku - f.$$

Hence, we have $F \circ R(\operatorname{div} \mathbf{p}) = F \circ S(u)$. Since F penalizes violations above $\overline{\sigma}^2$ and below $\underline{\sigma}^2$, we are interested in residuals $S(u)$ which satisfy $\underline{\sigma}^2 \leq S(u) \leq \overline{\sigma}^2$. Note that $S(u)(x) = \int_\Omega w(x, y)(Ku - f)^2(y)\mathrm{d}y$ for $x \in \Omega$ may be interpreted as a local variance [25] and recall that $f = Ku_{\mathrm{true}} + \eta$ where $\int_\Omega |\eta|^2 \mathrm{d}x = \sigma^2|\Omega|$. Consequently, if for some α^* we have $u(\alpha^*) = u_{\mathrm{true}}$, then it is expected that $S(u) \simeq \sigma^2$. Thus, choosing $\underline{\sigma} < \sigma < \overline{\sigma}$ we expect $F \circ (S(u)) = F \circ R(\operatorname{div} \mathbf{p}) \simeq 0$. Moreover, the existence of solutions to ($\mathbb{P}$) is closely related to continuity properties of the map $\alpha \mapsto \mathbf{K}(\alpha)$, which, in turn, requires Mosco convergence of $\mathbf{K}(\alpha_n)$ for a sequence $\{\alpha_n\}$ in $\mathcal{A}_{\mathrm{ad}}$. This latter issue is connected to the question whether a set of functions of higher regularity is densely contained in $\mathbf{K}(\alpha)$, which we address next.

PROPOSITION 4.1. *Let $\alpha \in W^{1,p}(\Omega) \cap \mathcal{A}_{ad}$. If $p = \ell = 1$ or $p > \ell$, then we have*

$$\overline{\mathbf{K}(\alpha, H^1_0(\Omega)^\ell)}^{H_0(\operatorname{div})} = \mathbf{K}(\alpha).$$

Proof. First recall that $\Omega \subset \mathbb{R}^\ell$. It is known that for $p = \ell = 1$ or $p > \ell$, $W^{1,p}(\Omega)$ embeds continuously into $C(\overline{\Omega})$. Since $\alpha \in \mathcal{A}_{\mathrm{ad}}$ and $\alpha > 0$ on $\overline{\Omega}$, the closure of $\mathbf{K}(\alpha, \mathscr{D}(\Omega)^\ell)$ in the $H_0(\operatorname{div})$-norm is $\mathbf{K}(\alpha)$ (see [36]). Since $\mathbf{K}(\alpha, \mathscr{D}(\Omega)^\ell) \subset \mathbf{K}(\alpha, H^1_0(\Omega)^\ell)$, the result follows. □

As the reader will have noticed, ($\mathbb{P}$) contains $\mathbf{p}$ and α as optimization variables with α also parameterizing the contraint set $\mathbf{K}(\alpha)$. For the proof of existence of a solution one, thus, needs to study the convergence of $\mathbf{K}(\alpha_n)$ for a sequence α_n converging in some sense towards α. This naturally leads to the notion of Mosco convergence. For this purpose, we next adapt a result of Boccardo and Murat (see [12] or [13]) for double obstacle constraints to $H^1_0(\Omega)^\ell$ and $H_0(\operatorname{div})$.

LEMMA 4.2. *Let $\{\alpha_n\}$ be a sequence in $\mathcal{A}_{\mathrm{ad}}$ such that $\alpha_n \rightharpoonup \alpha^*$ in $W^{1,p}(\Omega)$. Then, if $p > 2$ we have*

$$\mathbf{K}(\alpha_n, H^1_0(\Omega)^\ell) \xrightarrow{\mathrm{M}} \mathbf{K}(\alpha^*, H^1_0(\Omega)^\ell).$$

If $p > \max(2, \ell)$, then

$$\mathbf{K}(\alpha_n) \xrightarrow{\mathrm{M}} \mathbf{K}(\alpha^*).$$

Proof. The result of Boccardo and Murat (see [13, p.87] or [12]) implies that if $\alpha_n \rightharpoonup \alpha^*$ in $W^{1,p}(\Omega)$ with $p > 2$ and $\alpha_n \in \mathcal{A}_{\mathrm{ad}}$ for all $n \in \mathbb{N}$, then $K^-(\alpha_n) \to K^-(\alpha^*)$ in the sense of Mosco for $K^-(\alpha) := \{q \in H_0^1(\Omega) : -\alpha(x) \le q(x) \text{ a.e. on } \Omega\}$. The same is true for $K^+(\alpha) := \{q \in H_0^1(\Omega) : q(x) \le \alpha(x) \text{ a.e. on } \Omega\}$. Furthermore, for $K(\alpha) := \{q \in H_0^1(\Omega) : -\alpha(x) \le q(x) \le \alpha(x) \text{ a.e. on } \Omega\}$, and since $W^{1,p}(\Omega)$ and $H_0^1(\Omega)$ both embed compactly into $L^2(\Omega)$, it follows that if $v_n \in K(\alpha_n)$ for $n \in \mathbb{N}$ and $v_n \rightharpoonup v^*$ in $H_0^1(\Omega)$ as $n \to \infty$ then $v^* \in K(\alpha^*)$. Let $p^* \in K(\alpha^*)$. Then $p^* \in K^+(\alpha^*)$ and $p^* \in K^-(\alpha^*)$, and there exist sequences $\{p_n^+\}$ and $\{p_n^-\}$ such that $p_n^\pm \in K^\pm(\alpha_n)$ for all $n \in \mathbb{N}$ and $p_n^\pm \to p^*$ in $H_0^1(\Omega)$. Since the maps $\min(\cdot,\cdot), \max(\cdot,\cdot) : H_0^1(\Omega) \times H_0^1(\Omega) \to H_0^1(\Omega)$ are continuous in the strong and weak topologies, we have that $p_n := \max(p_n^+, 0) + \min(p_n^-, 0)$ satisfies $p_n \in K(\alpha_n)$ and $p_n \to p^*$. Therefore, $K(\alpha_n) \xrightarrow{\mathrm{M}} K(\alpha^*)$ and the extension to the multidimensional case $\mathbf{K}(\alpha_n, H_0^1(\Omega)^\ell) \xrightarrow{\mathrm{M}} \mathbf{K}(\alpha^*, H_0^1(\Omega)^\ell)$ is direct.

Let $\mathbf{p}_n \in \mathbf{K}(\alpha_n)$ for $n \in \mathbb{N}$ and $\mathbf{p}_n \rightharpoonup \mathbf{p}^*$ in $H_0(\mathrm{div})$. Then we prove first that $\mathbf{p}^* \in \mathbf{K}(\alpha^*)$: By the compact embedding of $W^{1,p}(\Omega)$ into $L^2(\Omega)$, we have that, along a subsequence, $\alpha_n \to \alpha^*$ in $L^2(\Omega)$ and also pointwise almost everywhere. Since also $\mathbf{p}_n \rightharpoonup \mathbf{p}^*$ in $L^2(\Omega)^\ell$, by Mazur's Lemma we have that $\tilde{\mathbf{p}}_n = \sum_{k=n}^{N(n)} \lambda(n)_k \mathbf{p}_k$, with $\sum_{k=n}^{N(n)} \lambda(n)_k = 1$ and $\lambda(n)_k \ge 0$, converges strongly in $L^2(\Omega)^\ell$ to $\mathbf{p}^*$ (and hence pointwise a.e. along a further subsequence) for some $\{\lambda(n)_k\}_{k=n}^{N(n)}$ and some $N(n) \in \mathbb{N}$. But, $\mathbf{p}_n \in \mathbf{K}(\alpha_n)$ yields

$$-\mathbf{1} \sum_{k=n}^{N(n)} \lambda(n)_k \alpha_k(x) \le \tilde{\mathbf{p}}_n(x) \le \mathbf{1} \sum_{k=n}^{N(n)} \lambda(n)_k \alpha_k(x), \qquad \text{a.e. on } \Omega$$

and $\sum_{k=n}^{N(n)} \lambda(n)_k \alpha_k \to \alpha^*$ pointwise a.e. so that $\mathbf{p}^* \in \mathbf{K}(\alpha^*)$.

Let $\mathbf{p}^* \in \mathbf{K}(\alpha^*)$ be arbitrary and suppose that $p > \max(2, \ell)$. Then, by Proposition 4.1 there exists $\{\mathbf{q}^j\}$ in $\mathbf{K}(\alpha^*, H_0^1(\Omega)^\ell)$ such that $|\mathbf{p}^* - \mathbf{q}^j|_{H_0(\mathrm{div})} \le 1/j$. Since $\mathbf{q}^j \in \mathbf{K}(\alpha^*, H_0^1(\Omega)^\ell)$ for all $j \in \mathbb{N}$, and $\mathbf{K}(\alpha_k, H_0^1(\Omega)^\ell) \xrightarrow{\mathrm{M}} \mathbf{K}(\alpha^*, H_0^1(\Omega)^\ell)$, there exists a sequence $\{\mathbf{q}_k^j\}$ such that $\mathbf{q}_k^j \in \mathbf{K}(\alpha_k, H_0^1(\Omega)^\ell) \subset \mathbf{K}(\alpha_k)$ and $|\mathbf{q}^j - \mathbf{q}_k^j|_{H_0^1(\Omega)^\ell} \to 0$ as $k \to \infty$. Let $\mathbf{p}_j := \mathbf{q}_{k_j}^j$ with k_j so that $|\mathbf{q}^j - \mathbf{p}_j|_{H_0^1(\Omega)^\ell} \le 1/j$. Then,

$$\overline{\lim_{k\to\infty}} |\mathbf{p}^* - \mathbf{p}_j|_{H_0(\mathrm{div})} \le \overline{\lim_{k\to\infty}} |\mathbf{p}^* - \mathbf{q}^j|_{H_0(\mathrm{div})} + \overline{\lim_{k\to\infty}} |\mathbf{q}^j - \mathbf{p}_j|_{H_0(\mathrm{div})} = 0.$$

This concludes the proof. $\square$

We are now ready to prove the main result of this section concerning the existence of solutions to problem ($\mathbb{P}$).

Theorem 4.3. *Problem* ($\mathbb{P}$) *has at least one solution.*

Proof. Let $\mathcal{S} : \mathcal{A}_{\mathrm{ad}} \to 2^{H_0(\mathrm{div})}$ be the set-valued solution map of the lower-level problem, i.e., $\mathcal{S}(\alpha) := \arg\min_{\mathbf{q} \in \mathbf{K}(\alpha)} J_D(\mathbf{q})$, and note that $\operatorname{div} \mathcal{S}(\alpha)$ is a singleton; see Proposition 3.2. Further, let $\{\mathbf{p}_n, \alpha_n\}$ be an infimizing sequence for ($\mathbb{P}$). Then $\mathbf{p}_n \in \mathcal{S}(\alpha_n)$ and $\operatorname{div} \mathbf{p}_n = \operatorname{div} \mathcal{S}(\alpha_n)$. Since $0 \le J(\mathbf{p}_n, \alpha_n) \le M$ for some $M > 0$ independent of n, $\{\alpha_n\}$ is uniformly bounded in $W^{1,p}(\Omega)$. Hence, $\alpha_n \rightharpoonup \alpha^*$ in $W^{1,p}(\Omega)$ along a subsequence (also denoted by $\{\alpha_n\}$) for some α^* in $W^{1,p}(\Omega) \cap \mathcal{A}_{\mathrm{ad}}$.

Since $\alpha_n \rightharpoonup \alpha^*$ in $W^{1,p}(\Omega)$, then $\mathbf{K}(\alpha_n) \xrightarrow{\mathrm{M}} \mathbf{K}(\alpha^*)$ by Lemma 4.2. Therefore, $\mathbf{p}_n \rightharpoonup \mathbf{p}^* \in \mathcal{S}(\alpha^*)$ in $H_0(\mathrm{div})$ and further $\mathrm{div}\,\mathbf{p}_n \to \mathrm{div}\,\mathcal{S}(\alpha^*)$ in $L^2(\Omega)$ (see Proposition 2.2). Hence, $(\alpha^*, \mathbf{p}^*)$ is feasible for $(\mathbb{P})$. Finally, note that

$$J(\mathbf{p}^*, \alpha^*) \leq \varliminf_{n\downarrow 0} J(\mathbf{p}_n, \alpha_n) \leq J(\mathbf{p}(\alpha), \alpha), \quad \forall \alpha \in \mathcal{A}_{\mathrm{ad}},$$

i.e., $(\mathbf{p}^*, \alpha^*)$ is a minimizer for Problem $(\mathbb{P})$. ◻

5. Solution stability and differentiability of a smoothed lower-level problem. Problem $(\mathbb{P})$ falls into the realm of elliptic mathematical programs with equilibrium constraints; see, e.g., [34]. This problem class typically suffers from degenerate feasible sets preventing KKT-systems for characterizing stationarity. An indicator of this degeneracy is the non-differentiability of $\alpha \mapsto \mathbf{p}(\alpha)$ with $\mathbf{p}(\alpha) \in \arg\min\{J_D(\mathbf{q}) : \mathbf{q} \in \mathbf{K}(\alpha)\}$. As a consequence, the derivation of stationarity conditions for a primal-dual characterization of a solution does not follow from KKT-theory in Banach space [59], and designing solution algorithms is delicate. As a remedy, in this section we study a family of problems $(\tilde{\mathrm{D}})$ approximating (D) with respect to their (differential) stability with respect to α. The family $(\tilde{\mathrm{D}})$, which is associated to a minimization problem via (5.3) below and convexity, is defined as follows.

Problem $(\tilde{\mathrm{D}})$**.** Let $0 \leq \alpha \in L^2(\Omega)$, $\beta, \gamma, \delta \in \mathbb{R}_0^+$, $\epsilon \in \mathbb{R}^+$, and either $V := H_0^1(\Omega)^\ell$ if $\beta > 0$, or $V := H_0(\mathrm{div})$ if $\beta = 0$. Then problem $(\tilde{\mathrm{D}})$ reads: Find $\mathbf{p} \in V$ such that

$$(\tilde{\mathrm{D}}) \qquad \langle -\beta \boldsymbol{\Delta} \mathbf{p} + \gamma \mathbf{p} + A\mathbf{p} + \mathbf{f}, \mathbf{v}\rangle_{V^*,V} + \frac{1}{\epsilon}(P_\delta(\mathbf{p}, \alpha), \mathbf{v}) = 0, \quad \forall \mathbf{v} \in V.$$

The operator $\boldsymbol{\Delta} : H_0^1(\Omega)^\ell \to H^{-1}(\Omega)^\ell$ is the vectorial Laplacian, i.e., for $\mathbf{w}, \mathbf{v} \in H_0^1(\Omega)^\ell$, $\langle -\boldsymbol{\Delta}\mathbf{w}, \mathbf{v}\rangle_{V^*,V} := (\boldsymbol{\nabla}\mathbf{w}, \boldsymbol{\nabla}\mathbf{v}) = \sum_{k=1}^{\ell} \int_\Omega \nabla w_k \cdot \nabla v_k \mathrm{d}x$. The map $A : H_0(\mathrm{div}) \to H_0(\mathrm{div})^*$ and $\mathbf{f} \in H_0(\mathrm{div})^*$ are as in (3.1), and $P_\delta : H_0^1(\Omega) \times L^2(\Omega) \to L^2(\Omega)^\ell$ is given by

$$P_\delta(\mathbf{p}, \alpha) := (\mathbf{p} - \alpha \mathbf{1})_\delta^+ - (\mathbf{p} + \alpha \mathbf{1})_\delta^-, \tag{5.1}$$

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

$$(r)_\delta^+ = \begin{cases} r - \delta/2, & r \geq \delta\ ; \\ r^2/2\delta, & r \in (0, \delta)\ ; \\ 0, & r \leq 0\ . \end{cases} \tag{5.2}$$

The map $r \mapsto (r)_\delta^+$ is a differentiable approximation of the positive part $r \mapsto (r)^+ := \max(r, 0)$ and similarly, for $(r)_\delta^- := (-r)_\delta^+$ and the negative part $(r)^- := (-r)^+$. Further, for $\delta = 0$, $(r)_\delta^+ := (r)^+$ and $(r)_\delta^- := (r)^-$. Note that for $0 \leq \delta_1 \leq \delta_2$, we have that $0 \leq (r)_{\delta_2}^+ \leq (r)_{\delta_1}^+$ for all $r \in \mathbb{R}$. For $\mathbf{r} \in \mathbb{R}^\ell$, $(\mathbf{r})_\delta^+$ is defined component-wise, i.e., $(\mathbf{r})_\delta^+ = ((r_1)_\delta^+, (r_1)_\delta^+, , \ldots, (r_l)_\delta^+)$ and $(\mathbf{r})_\delta^-$ analogously.

The existence of a solution to $(\tilde{\mathrm{D}})$ follows from monotone operator theory [56]. We note that for $\beta > 0$ or $\gamma > 0$ the solution is unique. Since $\langle A\mathbf{p}, \mathbf{p}\rangle_{V^*,V} = |\,\mathrm{div}\,\mathbf{p}|_B^2$, we have that for $\beta = \gamma = 0$ all solutions to $(\tilde{\mathrm{D}})$ have the same divergence: If $\mathbf{p}_1$ and $\mathbf{p}_2$ are both solutions, then by considering $\mathbf{v} = \mathbf{p}_2 - \mathbf{p}_1$ in $(\tilde{\mathrm{D}})$ and subtracting the equation associated to $\mathbf{p}_1$ to the one of $\mathbf{p}_2$ we have $|\,\mathrm{div}(\mathbf{p}_2 - \mathbf{p}_1)|_B^2 = \langle A(\mathbf{p}_2 - \mathbf{p}_1), (\mathbf{p}_2 - \mathbf{p}_1)\rangle_{V^*,V} \leq 0$ given that $\mathbf{p} \mapsto P_\delta(\mathbf{p}, \alpha)$ is monotone.

One readily argues that $\mathbf{p}$ of $(\tilde{\mathrm{D}})$ solves the following convex optimization problem:

$$\text{minimize} \quad \mathcal{J}(\mathbf{p}, \alpha), \quad \text{over } \mathbf{p} \in V, \tag{5.3}$$

where

$$\mathcal{J}(\mathbf{p},\alpha) := \frac{\beta}{2}|\mathbf{p}|_V^2 + \frac{\gamma}{2}|\mathbf{p}|^2_{L^2(\Omega)} + J_D(\mathbf{p}) + \frac{1}{\epsilon}\mathcal{P}_\delta(\mathbf{p},\alpha),$$

with $(\tilde{\mathrm{D}})$ representing the associated Euler-Lagrange system. Above, the functional $\mathcal{P}_\delta(\cdot,\alpha): V \to \mathbb{R}_0^+$ is defined as

$$\mathcal{P}_\delta(\mathbf{p},\alpha) := \int_\Omega \sum_{i=1}^{\ell} \big(G_\delta(-(p_i+\alpha)) + G_\delta(p_i-\alpha)\big)\mathrm{d}x, \tag{5.4}$$

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

$$G_\delta(r) = \begin{cases} \frac{1}{2}r^2 - \frac{\delta}{2}r + \frac{\delta^2}{6}, & r \geq \delta\ ; \\ r^3/6\delta, & r \in (0,\delta)\ ; \\ 0, & r \leq 0\ , \end{cases} \tag{5.5}$$

for $\delta > 0$. For $\delta = 0$, we use $r \mapsto G_0(r) := r^2/2$ for $r \geq 0$ and $G_0(r) := 0$ otherwise. Note that $\frac{\mathrm{d}}{\mathrm{d}r}G_\delta(r) = (r)_\delta^+$ and $\frac{\mathrm{d}}{\mathrm{d}r}G_\delta(-r) = -(-r)_\delta^+ = -(r)_\delta^-$, and further $\frac{\mathrm{d}^2}{\mathrm{d}r^2}G_\delta(r), \frac{\mathrm{d}^2}{\mathrm{d}r^2}G_\delta(-r) \geq 0$ for all $r \in \mathbb{R}$. Additionally, for $0 \leq \delta_0 \leq \delta_1$ we have $0 \leq (r)^+_{\delta_1} \leq (r)^+_{\delta_0}$ and also

$$0 \leq G_{\delta_1}(r) \leq G_{\delta_0}(r) \leq G_0(r), \qquad \forall r \in \mathbb{R}. \tag{5.6}$$

We now study asymptotic properties of $(\tilde{\mathrm{D}})$. More specifically, we establish stability with respect to perturbations of the parameters $(\alpha,\beta,\gamma,\delta)$. Additionally, we study vanishing ϵ and show that the associated limit of a sequence of solutions to $(\tilde{\mathrm{D}})$ solves (D).

Theorem 5.1. *Let $\{\beta_n\}$, $\{\gamma_n\}$, $\{\delta_n\}$ and $\{\epsilon_n\}$ satisfy $\beta_n, \gamma_n, \delta_n \geq 0$, $\epsilon_n > 0$ for all $n \in \mathbb{N}$, and $(\beta_n, \gamma_n, \epsilon_n) \to (\beta^*, \gamma^*, \epsilon^*)$. For $n \in \mathbb{N}$, suppose that α_n and α^* are non-negative functions on Ω, and let $\{\mathbf{p}_n\}$ be a sequence of solutions to* $(\tilde{\mathrm{D}})$ *for $\alpha = \alpha_n$ and $(\beta,\gamma,\delta,\epsilon) = (\beta_n,\gamma_n,\delta_n,\epsilon_n)$. Then the following statements hold true.*

(i) Consistency. *Suppose that $\alpha_n \rightharpoonup \alpha^*$ in $W^{1,p}(\Omega)$, $p > \max(2,\ell)$, $\alpha_n, \alpha^* \in \mathcal{A}_{ad}$, $\sup \delta_n < \infty$, and $(\beta^*, \gamma^*, \epsilon^*) = (0,0,0)$. Then,*

$$\operatorname{div}\mathbf{p}_n \to \operatorname{div}\mathbf{p}^* \text{ in } L^2(\Omega),$$

where $\mathbf{p}^*$ is any solution to (D) *for $\alpha = \alpha^*$.*

(ii) Stability. *Suppose that $\alpha_n \to \alpha^*$ in $L^2(\Omega)$, $\delta_n \to \delta^*$ and $\beta^*, \epsilon^* > 0$. Then,*

$$\mathbf{p}_n \to \mathbf{p}^* \text{ in } H_0^1(\Omega)^\ell,$$

where $\mathbf{p}^*$ is the solution to $(\tilde{\mathrm{D}})$ *for $\alpha = \alpha^*$ and $(\beta,\gamma,\delta,\epsilon) = (\beta^*,\gamma^*,\delta^*,\epsilon^*)$.*

Remark 5.1. *In the case* **(ii)** *the above results can be extended as follows. Let $f_n \to f^*$ in $L^2(\Omega)$ and assume that $\{\mathbf{p}_n\}$ is a sequence of solutions to* $(\tilde{\mathrm{D}})$ *with $f = f_n$, $n \in \mathbb{N}$, and $\mathbf{p}^*$ solves* $(\tilde{\mathrm{D}})$ *with $f = f^*$. Then the assertion of* **(ii)** *in Theorem 5.1 remains true since $\mathbf{f}_n = -\nabla B^{-1}K^* f_n$ satisfies $\mathbf{f}_n \to \mathbf{f}^* := -\nabla B^{-1}K^* f^*$ in $H^{-1}(\Omega)$.*

Proof. (of Theorem 5.1) We split the proof into several steps. *Step 1: Boundedness of $\tilde{\mathbf{p}} = \tilde{\mathbf{p}}(\alpha,\beta,\gamma,\epsilon,\delta)$, a solution of* $(\tilde{\mathrm{D}})$*, in $H_0(\mathrm{div})$ if $\epsilon > 0$, $\delta \geq 0$ and $0 \leq \alpha \in L^2(\Omega)$ are in bounded sets, and in $H_0^1(\Omega)^\ell$ if β^{-1} is in a bounded set.*

Setting $\mathbf{v} := \tilde{\mathbf{p}}$ in ($\tilde{\text{D}}$), we obtain from the monotonicity of $\mathbf{p} \mapsto P_\delta(\mathbf{p}, \alpha)$ and $-\beta\boldsymbol{\Delta} + \gamma I + A : V \to V^*$, where $V = H_0(\mathrm{div})$ if $\beta = 0$ and $V = H_0^1(\Omega)^\ell$ if $\beta > 0$, that

$$\begin{aligned} |\operatorname{div}\tilde{\mathbf{p}}|_B^2 + \beta|\tilde{\mathbf{p}}|_V^2 &\le \langle -\beta\boldsymbol{\Delta}\mathbf{p} + \gamma\tilde{\mathbf{p}} + A\tilde{\mathbf{p}}, \tilde{\mathbf{p}}\rangle_{V^*,V} + \frac{1}{\epsilon}(P_\delta(\tilde{\mathbf{p}}, \alpha), \tilde{\mathbf{p}}) = -\langle \mathbf{f}, \tilde{\mathbf{p}}\rangle_{V^*,V} \\ &\le |K^* f|_B |\operatorname{div}\tilde{\mathbf{p}}|_B, \end{aligned}$$

where we have used that $(v, w) \mapsto (v, B^{-1}w)$ is an inner product in $L^2(\Omega)$. Hence, $\{\operatorname{div}\tilde{\mathbf{p}}(\alpha, \beta, \gamma, \epsilon, \delta)\}_{\alpha,\beta,\gamma,\epsilon,\delta}$ is bounded in $L^2(\Omega)$, and, if β^{-1} is in a bounded set, $\{\tilde{\mathbf{p}}(\alpha, \beta, \gamma, \epsilon, \delta)\}_{\alpha,\beta,\gamma,\epsilon,\delta}$ is bounded in $H_0^1(\Omega)^\ell$.

Since $\tilde{\mathbf{p}}$ solves (5.3), we obtain from $\mathcal{J}(\tilde{\mathbf{p}}, \alpha) \le \mathcal{J}(\mathbf{0}, \alpha)$ that

$$\mathcal{P}_\delta(\tilde{\mathbf{p}}, \alpha) \le \frac{\epsilon}{2}|K^* f|_B^2. \tag{5.7}$$

When $\delta, \epsilon > 0$ are in bounded sets, then this implies that $(\tilde{\mathbf{p}} - \alpha\mathbf{1})_\delta^+$ and $(\tilde{\mathbf{p}} + \alpha\mathbf{1})_\delta^-$ are bounded in $L^2(\Omega)^\ell$. Hence, since $0 \le \alpha \in L^2(\Omega)$ is in a bounded set, we have $\{\tilde{\mathbf{p}}\}_{\alpha,\gamma,\epsilon,\delta}$ is bounded in $L^2(\Omega)^\ell$. Thus, if $\delta \ge 0$, $\epsilon > 0$ and $0 \le \alpha \in L^2(\Omega)$ are in bounded sets, then $\{\tilde{\mathbf{p}}(\alpha, \beta, \gamma, \epsilon, \delta)\}_{\alpha,\beta,\gamma,\epsilon,\delta}$ is bounded in $H_0(\mathrm{div})$.

Step 2: Weak convergence in $H_0(\mathrm{div})$ *in* **(i)** *or in* $H_0^1(\Omega)^\ell$ *in* **(ii)** *of the approximating sequence* $\{\tilde{\mathbf{p}}(\alpha_n, \gamma_n, \epsilon_n, \delta_n)\}$ *to* $\mathbf{p}^*$. Consider first **(i)**. Let $\{\alpha_n\}$, $\{\beta_n\}$, $\{\gamma_n\}$, $\{\delta_n\}$ and $\{\epsilon_n\}$ be as in our hypothesis, and $\{\tilde{\mathbf{p}}_n\}$ denote a subsequence of $\{\tilde{\mathbf{p}}(\alpha_n, \gamma_n, \epsilon_n, \delta_n)\}$ which converges weakly in $H_0(\mathrm{div})$ to some $\mathbf{p}^* \in H_0(\mathrm{div})$. Then, $p_i^n \rightharpoonup p_i^*$ in $L^2(\Omega)$ where $\tilde{\mathbf{p}}_n = (p_1^n, p_2^n, \dots, p_\ell^n)$ and $\mathbf{p}^* = (p_1^*, p_2^*, \dots, p_\ell^*)$, so that $(p_i^n \pm \alpha_n) \rightharpoonup (p_i^* \pm \alpha^*)$ in $L^2(\Omega)$. Let $\hat{\delta} = \sup \delta_n < \infty$. Then, by (5.6) we observe

$$\begin{aligned} \mathcal{P}_{\hat{\delta}}(\mathbf{p}^*, \alpha^*) &\le \sum_{i=1}^{\ell} \Big(\varliminf_{n\to\infty} \int_\Omega G_{\delta_n}(-(p_i^n + \alpha_n))\mathrm{d}x + \varliminf_{n\to\infty} \int_\Omega G_{\delta_n}(p_i^n - \alpha_n)\mathrm{d}x \Big) \\ &\le \varliminf_{n\to\infty} \int_\Omega \sum_{i=1}^{\ell} \big(G_{\delta_n}(-(p_i^n + \alpha_n)) + G_{\delta_n}(p_i^n - \alpha_n) \big)\mathrm{d}x \\ &= \varliminf_{n\to\infty} \mathcal{P}_{\delta_n}(\tilde{\mathbf{p}}_n, \alpha_n) \le \varliminf_{n\to\infty} \frac{\epsilon_n}{2}|K^* f|_B^2 = 0, \end{aligned}$$

where we have used (5.7) for the last inequality. This shows $\mathbf{p}^* \in \mathbf{K}(\alpha^*)$. Next we prove that $\mathbf{p}^*$ is the solution to (D). For this purpose let $\mathbf{r} \in \mathbf{K}(\alpha^*, H_0^1(\Omega)^\ell)$ be arbitrary. Since $\alpha_n \rightharpoonup \alpha^*$ in $W^{1,p}(\Omega)$, Lemma 4.2 yields that $\mathbf{K}(\alpha_n, H_0^1(\Omega)^\ell)$ converges in the sense of Mosco to $\mathbf{K}(\alpha^*, H_0^1(\Omega)^\ell)$. Hence, there is a sequence $\{\mathbf{r}_n\}$ with $\mathbf{r}_n \in \mathbf{K}(\alpha_n, H_0^1(\Omega)^\ell)$ and $\mathbf{r}_n \to \mathbf{r}$ in $H_0^1(\Omega)^\ell$. Since $\mathcal{J}(\mathbf{p}_n, \alpha_n) \le \mathcal{J}(\mathbf{r}_n, \alpha_n)$ and $\mathbf{r}_n \in \mathbf{K}(\alpha_n, H_0^1(\Omega)^\ell)$, we observe that

$$\begin{aligned} &\frac{1}{2}|\operatorname{div}\mathbf{p}_n + K^* f|_B^2 + \frac{\beta_n}{2}|\mathbf{p}_n|_V^2 + \frac{\gamma_n}{2}|\mathbf{p}_n|_{L^2(\Omega)}^2 + \frac{1}{\epsilon_n}\mathcal{P}_{\delta_n}(\mathbf{p}_n, \alpha_n) \\ &\qquad\qquad \le \frac{\beta_n}{2}|\mathbf{r}_n|_V^2 + \frac{\gamma_n}{2}|\mathbf{r}_n|_{L^2(\Omega)}^2 + \frac{1}{2}|\operatorname{div}\mathbf{r}_n + K^* f|_B^2. \end{aligned}$$

By (5.7), $\frac{1}{\epsilon_n}\mathcal{P}_{\delta_n}(\mathbf{p}_n, \alpha_n)$ is bounded, and since $\mathcal{J}(\mathbf{p}_n, \alpha_n) \le \mathcal{J}(\mathbf{0}, \alpha_n)$ we have that $\frac{\beta_n}{2}|\mathbf{p}_n|_V^2$ and $\frac{\gamma_n}{2}|\mathbf{p}_n|_{L^2(\Omega)}^2$ are also bounded for all $n \in \mathbb{N}$. Further, since $(\beta_n, \gamma_n) \to$

$(0,0)$ and $\mathbf{r}_n \to \mathbf{r}$ in $H_0^1(\Omega)^\ell$, we have

$$\begin{aligned}\frac{1}{2}&|\operatorname{div}\mathbf{p}^* + K^*f|_B^2 + c\\ &\le \varliminf_{n\to\infty}\left(\frac{1}{2}|\operatorname{div}\mathbf{p}_n + K^*f|_B^2 + \frac{\beta_n}{2}|\mathbf{p}_n|_V^2 + \frac{\gamma_n}{2}|\mathbf{p}_n|^2_{L^2(\Omega)} + \frac{1}{\epsilon_n}\mathcal{P}_{\delta_n}(\mathbf{p}_n,\alpha_n)\right)\\ &\le \frac{1}{2}|\operatorname{div}\mathbf{r} + K^*f|_B^2,\end{aligned}$$

where $c = \varliminf_{n\to\infty}(\frac{\beta_n}{2}|\mathbf{p}_n|_V^2 + \frac{\gamma_n}{2}|\mathbf{p}_n|^2_{L^2(\Omega)} + \frac{1}{\epsilon_n}\mathcal{P}_{\delta_n}(\mathbf{p}_n,\alpha_n))$. Here, we have used that $\operatorname{div}\mathbf{p}_n \rightharpoonup \operatorname{div}\mathbf{p}^*$ in $L^2(\Omega)$. Finally, since $\mathbf{r}\in\mathbf{K}(\alpha^*, H_0^1(\Omega)^\ell)$ is arbitrary and $\mathbf{K}(\alpha^*, H_0^1(\Omega)^\ell)$ is dense (in the $H_0(\operatorname{div})$-topology) in $\mathbf{K}(\alpha^*)$, we can choose $\mathbf{r}=\mathbf{r}_k$ with $\mathbf{r}_k\to\mathbf{p}^*$ in $H_0(\operatorname{div})$. This shows that $c=0$. Furthermore, let $\mathbf{p}\in\mathbf{K}(\alpha^*)$ be arbitrary. Then there exists $\mathbf{r}=\mathbf{r}_k\in\mathbf{K}(\alpha^*, H_0^1(\Omega)^\ell)$ such that $\mathbf{r}_k\to\mathbf{p}$ in $H_0(\operatorname{div})$. Using this fact in the above inequality together with $c=0$, we have

$$\frac{1}{2}|\operatorname{div}\mathbf{p}^* + K^*f|_B^2 \le \frac{1}{2}|\operatorname{div}\mathbf{p} + K^*f|_B^2,$$

i.e., $\mathbf{p}^*$ solves (D).

Consider next the case **(ii)**, and let $\{\alpha_n\}$, $\{\beta_n\}$, $\{\gamma_n\}$, $\{\delta_n\}$ and $\{\epsilon_n\}$ be as in our hypothesis. Let $\{\mathbf{p}_n\}$ be a subsequence of $\{\mathbf{p}(\alpha_n,\gamma_n,\epsilon_n,\delta_n)\}$ which converges weakly in $H_0^1(\Omega)^\ell$ (and strongly in $L^2(\Omega)^\ell$) to some $\mathbf{p}^*\in H_0^1(\Omega)^\ell$. Additionally, we have $\alpha_n\to\alpha^*$ and $\operatorname{div}\mathbf{p}_n\rightharpoonup\operatorname{div}\mathbf{p}^*$ in $L^2(\Omega)$. Since $(\beta_n,\gamma_n,\delta_n,\epsilon_n)\to(\beta^*,\gamma^*,\delta^*,\epsilon^*)$ with $\beta^*,\epsilon^*>0$, taking the lim inf on both sides of the inequality $\mathcal{J}(\mathbf{p}_n,\alpha_n)\le\mathcal{J}(\mathbf{p},\alpha_n)$ for arbitrary $\mathbf{p}$, it readily follows that $\mathbf{p}^*$ solves ($\tilde{\mathrm{D}}$) for $\alpha=\alpha^*$ and $(\beta,\gamma,\delta,\epsilon)=(\beta^*,\gamma^*,\delta^*,\epsilon^*)$.

Step 3: $\operatorname{div}\mathbf{p}_n\to\operatorname{div}\mathbf{p}^*$ *in* $L^2(\Omega)$ *in* **(i)**, *and* $\mathbf{p}_n\to\mathbf{p}^*$ *in* $H_0^1(\Omega)^\ell$ *in* **(ii)**. Consider first **(i)**. Let $\mathbf{v}^h\in\mathbf{K}(\alpha^*, H_0^1(\Omega)^\ell)$ be such that $|\mathbf{v}^h-\mathbf{p}^*|_{H_0(\operatorname{div})}\le h$ for $h>0$ and let $\mathbf{v}_n^h\in\mathbf{K}(\alpha_n, H_0^1(\Omega)^\ell)$ with $\mathbf{v}_n^h\to\mathbf{v}^h$ in $H_0^1(\Omega)^\ell$. Note that the existence of $\mathbf{v}^h$ is guaranteed since $\mathbf{K}(\alpha^*, H_0^1(\Omega)^\ell)$ is dense in $\mathbf{K}(\alpha^*)$ in the $H_0(\operatorname{div})$-topology. The existence of $\{\mathbf{v}_n^h\}$ is guaranteed as $\mathbf{K}(\alpha_n, H_0^1(\Omega)^\ell)$ converges in the sense of Mosco to $\mathbf{K}(\alpha^*, H_0^1(\Omega)^\ell)$. From the strong monotonicity of the operator $\mathcal{A}_n(\cdot):=(-\beta_n\mathbf{\Delta}+\gamma_n I+A)(\cdot)$, and the fact that $\mathbf{p}_n$ solves ($\tilde{\mathrm{D}}$) we have that

$$\begin{aligned}\frac{1}{2}|\operatorname{div}(\mathbf{p}_n-\mathbf{v}_n^h)|_B^2 + \frac{\beta_n}{2}|\mathbf{p}_n-\mathbf{v}_n^h|_V^2 + \frac{\gamma_n}{2}|\mathbf{p}_n-\mathbf{v}_n^h|_{L^2(\Omega)} &= \langle\mathcal{A}_n(\mathbf{p}_n-\mathbf{v}_n^h),\mathbf{p}_n-\mathbf{v}_n^h\rangle_{V^*,V}\\ &\le\left\langle\frac{1}{\epsilon_n}P_{\delta_n}(\mathbf{p}_n,\alpha_n)+\mathbf{f}-\beta_n\mathbf{\Delta}\mathbf{v}_n^h+\gamma_n\mathbf{v}_n^h+A\mathbf{v}_n^h,\mathbf{v}_n^h-\mathbf{p}_n\right\rangle_{V^*,V}.\end{aligned}$$

Since $\mathbf{v}_n^h\in\mathbf{K}(\alpha_n, H_0^1(\Omega)^\ell)$, then $P_{\delta_n}(\mathbf{v}_n^h,\alpha_n)=0$ and given that $\mathbf{p}\mapsto P_{\delta_n}(\mathbf{p},\alpha_n)$ is monotone, we observe $\langle P_{\delta_n}(\mathbf{p}_n,\alpha_n),\mathbf{v}_n^h-\mathbf{p}_n\rangle\le 0$. Hence, the above inequality implies

$$\begin{aligned}\frac{1}{2}|\operatorname{div}(\mathbf{p}_n-\mathbf{v}_n^h)|_B^2 + \frac{\beta_n}{2}|\mathbf{p}_n-\mathbf{v}_n^h|_V^2 + \frac{\gamma_n}{2}|\mathbf{p}_n-\mathbf{v}_n^h|_{L^2(\Omega)}&\\ \le\left\langle\mathbf{f}-\beta_n\mathbf{\Delta}\mathbf{v}_n^h+\gamma_n\mathbf{v}_n^h+A\mathbf{v}_n^h,\mathbf{v}_n^h-\mathbf{p}_n\right\rangle_{V^*,V}&.\end{aligned}\tag{5.8}$$

From $\mathcal{J}(\mathbf{p}_n,\alpha_n)\le\mathcal{J}(\mathbf{0},\alpha_n)$ we obtained that $\sqrt{\beta_n}|\mathbf{p}_n|_V$ and $\sqrt{\gamma_n}|\mathbf{p}_n|_{L^2(\Omega)}$ are bounded. Since $(\beta_n,\gamma_n)\to(0,0)$, and $\mathbf{v}_n^h\to\mathbf{v}^h$ in $H_0^1(\Omega)^\ell$, we have that $\varlimsup_{n\to\infty}\langle-\beta_n\mathbf{\Delta}\mathbf{v}_n^h+\gamma_n\mathbf{v}_n^h,\mathbf{v}_n^h-\mathbf{p}_n\rangle_{V^*,V}\le 0$ and since $\operatorname{div}\mathbf{p}_n\rightharpoonup\operatorname{div}\mathbf{p}^*$ in $L^2(\Omega)$ we observe that

$$\varlimsup_{n\to\infty}\langle\mathbf{f}-\beta_n\mathbf{\Delta}\mathbf{v}_n^h+\gamma_n\mathbf{v}_n^h+A\mathbf{v}_n^h,\mathbf{v}_n^h-\mathbf{p}_n\rangle_{V^*,V}\le(K^*f+\operatorname{div}\mathbf{v}^h,\operatorname{div}(\mathbf{v}^h-\mathbf{p}^*))_B.$$

As $|\mathbf{v}^h - \mathbf{p}^*|_{H_0(\mathrm{div})} \le h$ and from (5.8), we infer

$$\overline{\lim_{n\to\infty}} \frac{1}{2}|\operatorname{div}(\mathbf{p}_n - \mathbf{v}_n^h)|_B^2 = O(h) \text{ as } h \to 0.$$

This completes the proof, since $\mathbf{v}_n^h \to \mathbf{v}^h$ in $H_0^1(\Omega)^\ell$ with $|\mathbf{v}^h - \mathbf{p}^*|_{H_0(\mathrm{div})} \le h$ and $h > 0$ is arbitrary.

Next consider **(ii)**. From the inequality

$$\mathcal{J}(\mathbf{p}^*, \alpha^*) \le \underline{\lim}\, \mathcal{J}(\mathbf{p}_n, \alpha_n) \le \overline{\lim}\, \mathcal{J}(\mathbf{p}_n, \alpha_n) \le \overline{\lim}\, \mathcal{J}(\mathbf{p}^*, \alpha_n) = \mathcal{J}(\mathbf{p}^*, \alpha^*),$$

and since $\mathbf{p}_n \to \mathbf{p}^*$ in $L^2(\Omega)^\ell$, $\alpha_n \to \alpha^*$ in $L^2(\Omega)$ and $\operatorname{div}\mathbf{p}_n \rightharpoonup \operatorname{div}\mathbf{p}^*$ in $L^2(\Omega)$, we have that

$$\lim_{n\to\infty} \left(\frac{1}{2}|\operatorname{div}\mathbf{p}_n|_B^2 + \frac{\beta}{2}|\mathbf{p}_n|^2_{H_0^1(\Omega)^\ell} \right) = \frac{1}{2}|\operatorname{div}\mathbf{p}^*|_B^2 + \frac{\beta}{2}|\mathbf{p}^*|^2_{H_0^1(\Omega)^\ell}.$$

Further, $\mathbf{p} \mapsto \frac{1}{2}|\operatorname{div}\mathbf{p}|_B^2 + \frac{\beta}{2}|\mathbf{p}|^2_{H_0^1(\Omega)^\ell}$ yields a norm which is equivalent to the usual one in $H_0^1(\Omega)^\ell$. This implies that $|\mathbf{p}_n|_{H_0^1(\Omega)^\ell} \to |\mathbf{p}^*|_{H_0^1(\Omega)^\ell}$, given $\mathbf{p}_n \rightharpoonup \mathbf{p}^*$ in $H_0^1(\Omega)^\ell$. Hence, $\mathbf{p}_n \to \mathbf{p}^*$ in $H_0^1(\Omega)^\ell$. ∎

We are left with establishing differentiability properties of the mapping $\alpha \mapsto \mathbf{p}(\alpha)$. For this purpose, we require $\beta > 0$ in ($\tilde{\mathrm{D}}$). We remark already here that a corresponding differentiability results for $\beta = 0$ seems elusive. The main reason for this limitation stems from the fact that $H_0(\mathrm{div})$ does not embed continuously into any $L^{2+\xi}(\Omega)^\ell$ for any $\xi > 0$, but $H_0^1(\Omega)$ does so for $\xi \in (0, \bar{\xi}(\ell)]$ with $\bar{\xi}(\ell) = +\infty$ for $\ell = 1$, $\bar{\xi}(\ell) \in [0, +\infty)$ for $\ell = 2$, and $\bar{\xi}(\ell) = 2\ell/(\ell-2)$ for $\ell > 2$.

THEOREM 5.2. *Let $\xi \in (0, \bar{\xi}(\ell)]$ and $\mathbf{p}(\cdot) : L^{2+\xi}(\Omega) \to H_0^1(\Omega)^\ell$ be the solution mapping of ($\tilde{\mathrm{D}}$) with $\beta > 0$. Then $\alpha \mapsto \mathbf{p}(\alpha)$ is Gâteaux differentiable, and the Gâteaux differential of $\mathbf{p}$ at α^* in direction ω is the unique solution to: Find $\mathbf{s} \in H_0^1(\Omega)^\ell$ such that*

$$(5.9)\qquad \langle -\beta\boldsymbol{\Delta}\mathbf{s} + \gamma\mathbf{s} + A\mathbf{s}, \mathbf{v}\rangle_{H^{-1},H_0^1} + \frac{1}{\epsilon}(P_\delta'(\mathbf{p}^*, \alpha^*)(\mathbf{s}, \omega), \mathbf{v}) = 0, \quad \forall \mathbf{v} \in H_0^1(\Omega)^\ell,$$

with $\mathbf{p}^* := \mathbf{p}(\alpha^*)$ and $P_\delta' : H_0^1(\Omega)^\ell \times L^{2+\xi}(\Omega) \to \mathscr{L}(H_0^1(\Omega)^\ell \times L^{2+\xi}(\Omega), H^{-1}(\Omega)^\ell)$ denotes the Fréchet derivative of P_δ.

Proof. We divide the proof into several steps. *Step 1.* Let $\mathbf{p}^* := \mathbf{p}(\alpha^*)$ and $\mathbf{p}_t := \mathbf{p}(\alpha_t)$ with $\alpha_t := \alpha^* + t\omega$, i.e., $\mathbf{p}^*$ and $\mathbf{p}_t$ are the solutions of ($\tilde{\mathrm{D}}$) for $\alpha = \alpha^*$ and $\alpha = \alpha^* + t\omega$, respectively. Then, by taking the difference of the corresponding equations ($\tilde{\mathrm{D}}$) for $\mathbf{p}^*$ and $\mathbf{p}_t$, it follows that

(5.10)

$$\langle -\beta\boldsymbol{\Delta}(\mathbf{p}_t - \mathbf{p}^*) + \gamma(\mathbf{p}_t - \mathbf{p}^*) + A(\mathbf{p}_t - \mathbf{p}^*), \mathbf{v}\rangle_{H^{-1},H_0^1} + \frac{1}{\epsilon}(P_\delta(\mathbf{p}_t, \alpha_t) - P_\delta(\mathbf{p}^*, \alpha^*), \mathbf{v}) = 0,$$

for $\mathbf{v} \in H_0^1(\Omega)^\ell$.

For $\mathbf{v} = \mathbf{p}_t - \mathbf{p}^*$ in (5.10) and with $Q_t(\mathbf{p}) := P_\delta(\mathbf{p}, \alpha_t)$ and $Q_0(\mathbf{p}) := P_\delta(\mathbf{p}, \alpha^*)$ we obtain

$$\beta|\mathbf{p}_t - \mathbf{p}^*|^2_{H_0^1(\Omega)^\ell} + \langle A(\mathbf{p}_t - \mathbf{p}^*), \mathbf{p}_t - \mathbf{p}^*\rangle_{H^{-1},H_0^1} + \gamma|\mathbf{p}_t - \mathbf{p}^*|^2_{L^2(\Omega)^\ell} +$$
$$\frac{1}{\epsilon}(Q_t(\mathbf{p}_t) - Q_t(\mathbf{p}^*), \mathbf{p}_t - \mathbf{p}^*) = -\frac{1}{\epsilon}(Q_t(\mathbf{p}^*) - Q_0(\mathbf{p}^*), \mathbf{p}_t - \mathbf{p}^*).$$

The functions $\mathbb{R} \ni r \mapsto (r)^+_\delta, (r)^-_\delta$ are Lipschitz continuous with Lipschitz constant bounded by a constant independent of δ. Hence, it follows that $|Q_t(\mathbf{p}) - Q_0(\mathbf{p})|_{L^2(\Omega)^\ell} \le L_Q|\alpha_t - \alpha|_{L^2(\Omega)} = L_Q|t||\omega|_{L^2(\Omega)}$ with some constant $L_Q > 0$. Therefore,

$$(Q_t(\mathbf{p}^*) - Q_0(\mathbf{p}^*), \mathbf{p}_t - \mathbf{p}^*) \le$$
$$L_Q C|t||\omega|_{L^2(\Omega)} \left(\beta|\mathbf{p}_t - \mathbf{p}^*|^2_{H^1_0(\Omega)^\ell} + |\operatorname{div}(\mathbf{p}_t - \mathbf{p}^*)|^2_B + \gamma|\mathbf{p}_t - \mathbf{p}^*|^2_{L^2(\Omega)^\ell}\right)^{1/2},$$

where $C = 1/\sqrt{\gamma}$. Furthermore, $(Q_t(\mathbf{p}) - Q_t(\mathbf{q}), \mathbf{p} - \mathbf{q}) \ge 0$ for all $\mathbf{p}, \mathbf{q} \in L^2(\Omega)^\ell$, and since $\langle A\mathbf{p}, \mathbf{p}\rangle_{H^{-1},H^1_0} = |\operatorname{div}\mathbf{p}|^2_B$ we infer

(5.11)
$$\left(\beta|\mathbf{p}_t - \mathbf{p}^*|^2_{H^1_0(\Omega)^\ell} + |\operatorname{div}(\mathbf{p}_t - \mathbf{p}^*)|^2_B + \gamma|\mathbf{p}_t - \mathbf{p}^*|^2_{L^2(\Omega)^\ell}\right)^{1/2} \le |t|\left(\frac{CL_Q}{\epsilon}\right)|\omega|_{L^2(\Omega)}.$$

Let $\{t_n\}$ be a sequence of reals such that $t_n \to 0$. Then, by (5.11) there exists a subsequence (denoted the same) such that

$$\mathbf{s}^*_n := \frac{\mathbf{p}_{t_n} - \mathbf{p}^*}{t_n} \rightharpoonup \mathbf{s}^* \quad \text{in } H^1_0(\Omega)^\ell, \tag{5.12}$$

for some $\mathbf{s}^* \in H^1_0(\Omega)^\ell$.

Step 2: $\mathbf{s}^*$ *solves* (5.9) *uniquely.* We have that $H^1_0(\Omega)^\ell \times L^{2+\xi}(\Omega) \ni (\mathbf{p}, \alpha) \mapsto P_\delta(\mathbf{p}, \alpha) \in H^{-1}(\Omega)^\ell$ is continuously Fréchet differentiable (see Proposition 6.2 below and the remarks at the end of its statement). Then, for any $\mathbf{v} \in H^1_0(\Omega)^\ell$ and by the mean value theorem we get

(5.13)
$$(P_\delta(\mathbf{p}_t, \alpha_t) - P_\delta(\mathbf{p}^*, \alpha^*), \mathbf{v}) = ({P_\delta}'(\mathbf{p}^* + \zeta_t(\mathbf{p}_t - \mathbf{p}^*), \alpha^* + \zeta_t(\alpha_t - \alpha^*))\begin{pmatrix}\mathbf{p}_t - \mathbf{p}^* \\ \alpha_t - \alpha^*\end{pmatrix}, \mathbf{v}),$$

for some $t \mapsto \zeta_t \in (0, 1)$. From (5.11), we have that $\mathbf{p}_t \to \mathbf{p}^*$ in $H^1_0(\Omega)^\ell$ as $t \to 0$ and by definition $\alpha_t - \alpha^* = t\omega$. It follows that $(\mathbf{p}_t, \alpha_t) \to (\mathbf{p}^*, \alpha^*)$ in $V \times L^{2+\xi}(\Omega)$. Additionally, $P'_\delta : H^1_0(\Omega)^\ell \times L^{2+\xi}(\Omega) \to \mathscr{L}(H^1_0(\Omega)^\ell \times L^{2+\xi}(\Omega), H^{-1}(\Omega)^\ell)$ is continuous (since P_δ is continuously Fréchet differentiable), and together with (5.12) this implies

$$\lim_{n\to\infty}\left(\frac{P_\delta(\mathbf{p}_{t_n}, \alpha_{t_n}) - P_\delta(\mathbf{p}^*, \alpha^*)}{t_n}, \mathbf{v}\right) = (P'_\delta(\mathbf{p}^*, \alpha^*)(\mathbf{s}^*, \omega), \mathbf{v}).$$

One readily shows that $\langle -\beta\boldsymbol{\Delta}\mathbf{s}^*_n + A_\gamma\mathbf{s}^*_n, \mathbf{v}\rangle_{H^{-1},H^1_0} \to \langle -\beta\boldsymbol{\Delta}\mathbf{s}^* + A_\gamma\mathbf{s}^*, \mathbf{v}\rangle_{H^{-1},H^1_0}$, where $A_\gamma = A + \gamma I$. Therefore, dividing (5.10) by t_n and taking $n \to \infty$, we obtain that $\mathbf{s}^*$ solves (5.9) .

We are only left to prove that the solution to (5.9) is unique. Fix α and recall that $H^1_0(\Omega)^\ell \ni \mathbf{p} \mapsto P_\delta(\mathbf{p}, \alpha) \in H^{-1}(\Omega)^\ell$ is monotone and differentiable. Then, from the inequality $(P_\delta(\mathbf{p}^* + s\mathbf{r}, \alpha^*) - P_\delta(\mathbf{p}^*, \alpha^*), s\mathbf{r}) \ge 0$, we obtain that $(D_1P_\delta(\mathbf{p}^*, \alpha^*)\mathbf{r}, \mathbf{r}) \ge 0$, i.e., $D_1P_\delta(\mathbf{p}^*, \alpha^*)$ is a monotone, linear and bounded operator where $D_1P_\delta(\mathbf{p}, \alpha)$ and $D_2P_\delta(\mathbf{p}, \alpha)$ denote the Fréchet derivatives of $\mathbf{p} \mapsto P_\delta(\mathbf{p}, \alpha)$ and $\alpha \mapsto P_\delta(\mathbf{p}, \alpha)$, respectively. Since $P'_\delta(\mathbf{p}^*, \alpha^*)(\mathbf{s}, \omega) = D_1P_\delta(\mathbf{p}^*, \alpha^*)\mathbf{s} + D_2P_\delta(\mathbf{p}^*, \alpha^*)\omega$ (see Proposition 6.2 below) and $-\beta\boldsymbol{\Delta} + A + \gamma I$ is a strongly monotone, linear and bounded operator by the Lax-Milgram Lemma, the equation (5.9) has a unique solution. Thus, for every positive sequence $\{t_n\}$ such $t_n \to 0$, every convergent subsequence of the quotient (5.12) converges for the same limit and hence, the entire sequence converges to the same limit.

Step 3: $\mathbf{s}_n^* \to \mathbf{s}$ *in* $H_0^1(\Omega)^\ell$ *and conclusion .* For any $t > 0$, considering $\mathbf{v} = (\mathbf{p}_t - \mathbf{p}^*)/t$ in (5.13), leads to

$$\left(\frac{P_\delta(\mathbf{p}_t, \alpha_t) - P_\delta(\mathbf{p}^*, \alpha^*)}{t}, \frac{\mathbf{p}_t - \mathbf{p}^*}{t}\right) =$$

$$(P'_\delta(\mathbf{p}^* + \zeta_t(\mathbf{p}_t - \mathbf{p}^*), \alpha^* + \zeta_t(\alpha_t - \alpha^*)) \begin{pmatrix} \frac{\mathbf{p}_t - \mathbf{p}^*}{t} \\ \frac{\alpha_t - \alpha^*}{t} \end{pmatrix}, \frac{\mathbf{p}_t - \mathbf{p}^*}{t}). \tag{5.14}$$

Since $\mathbf{p}_{t_n} \to \mathbf{p}^*$ in $H_0^1(\Omega)^\ell$ and $\alpha_{t_n} \to \alpha^*$ in $L^{2+\xi}(\Omega)$, it follows that $\mathbf{p}^* + \zeta_{t_n}(\mathbf{p}_{t_n} - \mathbf{p}^*) \to \mathbf{p}^*$ in $H_0^1(\Omega)^\ell$ (particularly in $L^2(\Omega)^\ell$) and $\alpha^* + \zeta_{t_n}(\alpha_{t_n} - \alpha^*) \to \alpha^*$ in $L^{2+\xi}(\Omega)$. Additionally, we know that $(\mathbf{p}_{t_n} - \mathbf{p}^*)/t_n \rightharpoonup \mathbf{s}^*$ in $H_0^1(\Omega)^\ell$ and, by the Rellich-Kondrachov Theorem, $(\mathbf{p}_{t_n} - \mathbf{p}^*)/t_n \to \mathbf{s}^*$ in $L^2(\Omega)^\ell$ (as well as pointwise almost everywhere convergence on Ω) respectively along a subsequence (that we also denote by $\{\mathbf{p}_{t_n}\}$ for the ease of exposition). Moreover, we have $(\alpha_{t_n} - \alpha^*)/t_n = \omega$ for all $n \in \mathbb{N}$. Therefore, since $P'_\delta : V \times L^{2+\xi}(\Omega) \to \mathscr{L}(H_0^1(\Omega)^\ell \times L^{2+\xi}(\Omega), H^{-1}(\Omega)^\ell)$ is continuous, by (5.9) and (5.14) we obtain

$$\begin{aligned} \lim_{n\to\infty} -\frac{1}{\epsilon}\left(\frac{P_\delta(\mathbf{p}_{t_n}, \alpha_{t_n}) - P_\delta(\mathbf{p}^*, \alpha^*)}{t_n}, \frac{\mathbf{p}_{t_n} - \mathbf{p}^*}{t_n}\right) &= -\frac{1}{\epsilon}(P'_\delta(\mathbf{p}^*, \alpha^*)(\mathbf{s}^*, \omega), \mathbf{s}^*) \\ &= \langle -\beta\boldsymbol{\Delta}\mathbf{s}^* + A\mathbf{s}^* + \gamma\mathbf{s}^*, \mathbf{s}^*\rangle_{H^{-1}, H_0^1}. \end{aligned} \tag{5.15}$$

Note that $(\mathbf{u}, \mathbf{v})_\beta := \langle -\beta\boldsymbol{\Delta}\mathbf{u} + A\mathbf{u} + \gamma\mathbf{u}, \mathbf{v}\rangle_{H^{-1}, H_0^1}$ with $\mathbf{u}, \mathbf{v} \in H_0^1(\Omega)^\ell$ is an inner product in $H_0^1(\Omega)^\ell$ which is equivalent to the usual one and $|\mathbf{u}|_\beta^2 := (\mathbf{u}, \mathbf{u})_\beta^2 = \beta|\mathbf{u}|_V^2 + \gamma|\mathbf{u}|_{L^2(\Omega)^\ell}^2 + |\operatorname{div}\mathbf{u}|_B^2$. Then, by the lower-semicontinuity of the norm, (5.10) and (5.15), we find that

$$\begin{aligned} (\mathbf{s}^*, \mathbf{s}^*)_\beta &\le \varliminf_{n\to\infty} (\mathbf{s}_n^*, \mathbf{s}_n^*)_\beta \le \varlimsup_{n\to\infty} (\mathbf{s}_n^*, \mathbf{s}_n^*)_\beta = \varlimsup_{n\to\infty} \langle -\beta\boldsymbol{\Delta}\mathbf{s}_n^* + A\mathbf{s}_n^* + \gamma\mathbf{s}_n^*, \mathbf{s}_n^*\rangle_{H^{-1}, H_0^1} \\ &= \varlimsup_{n\to\infty} -\frac{1}{\epsilon}\left(\frac{P_\delta(\mathbf{p}_{t_n}, \alpha_{t_n}) - P_\delta(\mathbf{p}^*, \alpha^*)}{t_n}, \frac{\mathbf{p}_{t_n} - \mathbf{p}^*}{t_n}\right) \\ &= \langle -\beta\boldsymbol{\Delta}\mathbf{s}^* + A\mathbf{s}^* + \gamma\mathbf{s}^*, \mathbf{s}^*\rangle_{H^{-1}, H_0^1} = (\mathbf{s}^*, \mathbf{s}^*)_\beta, \end{aligned}$$

where $\mathbf{s}_n^* := \frac{\mathbf{p}_{t_n} - \mathbf{p}^*}{t_n}$, i.e., $|\mathbf{s}_n^*|_\beta \to |\mathbf{s}^*|_\beta$. Hence, since we already have that $\mathbf{s}_n^* \rightharpoonup \mathbf{s}^*$ in $H_0^1(\Omega)^\ell$, we conclude $\mathbf{s}_n^* \to \mathbf{s}^*$ in $H_0^1(\Omega)^\ell$. Finally, this implies that for every α, ω, there exists $\boldsymbol{\mu}(\alpha, \omega) \in H_0^1(\Omega)^\ell$ such that

$$\lim_{t\to 0} \frac{\mathbf{p}(\alpha + t\omega) - \mathbf{p}(\alpha)}{t} = \lim_{t\to 0} \frac{\mathbf{p}_t - \mathbf{p}_0}{t} = \boldsymbol{\mu}(\alpha, \omega), \qquad \text{in } H_0^1(\Omega)^\ell. \tag{5.16}$$

Consequently, $\boldsymbol{\mu}(\alpha, \omega)$ is the Gâteaux differential of $\mathbf{p}$ at α in direction ω. It follows additionally from (5.9) that the map $\omega \mapsto \boldsymbol{\mu}(\alpha, \omega)$ is linear and continuous, i.e., $\boldsymbol{\mu}(\alpha, \omega) = \boldsymbol{\mu}(\alpha)\omega$ with $\boldsymbol{\mu}(\alpha) \in \mathscr{L}(L^{2+\xi}(\Omega), H_0^1(\Omega)^\ell)$. Furthermore, since $P'_\delta : H_0^1(\Omega)^\ell \times L^{2+\xi}(\Omega) \to \mathscr{L}(H_0^1(\Omega)^\ell \times L^{2+\xi}(\Omega), H^{-1}(\Omega)^\ell)$, the map $\alpha \mapsto \boldsymbol{\mu}(\alpha)$ is continuous. Hence, it is the Fréchet derivative of the map $\alpha \mapsto \mathbf{p}(\alpha)$. $\square$

6. The regularized bilevel problem for choosing α. Now we introduce $(\tilde{\mathbb{P}})$, the regularized version of Problem $(\mathbb{P})$ by replacing the lower-level problem (D) in $(\mathbb{P})$

by $(\tilde{\mathrm{D}})$, and the admissible set $\mathcal{A}_{\mathrm{ad}}$ is defined with $\underline{\alpha}, \overline{\alpha} \in H^2$. In contrast to $(\mathbb{P})$, we pose $(\tilde{\mathbb{P}})$ over $\alpha \in W^{1,p}(\Omega)$ for $p \geq 2$ and not with $p > \max(2,\ell)$ as in $(\mathbb{P})$. However, the duality result of Theorem 3.4 requires the $C(\overline{\Omega})$-regularity of the filtering weight α. Analytically, as seen earlier, it can be obtained by requiring $p > \max(2,\ell)$, or it arises in an *a posteriori* way when solving $(\tilde{\mathbb{P}})$ by a projection method; see Theorem 3.1 of part II of this work [38].

Problem $(\tilde{\mathbb{P}})$: Let $\lambda, \beta, \delta > 0$, and $p \geq 2$. Consider the problem:

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

The set $\mathcal{A}_{\mathrm{ad}}$, is given by

$$\mathcal{A}_{\mathrm{ad}} := \{\alpha \in H^1(\Omega) : \underline{\alpha} \leq \alpha \leq \overline{\alpha}, \quad \text{a.e. on } \Omega\}, \tag{6.1}$$

for given $\underline{\alpha}, \overline{\alpha} \in H^2(\Omega)$, such $0 \leq \underline{\alpha} \leq \overline{\alpha}$ a.e. in Ω and $\frac{\partial \underline{\alpha}}{\partial \boldsymbol{\nu}} = \frac{\partial \overline{\alpha}}{\partial \boldsymbol{\nu}} = 0$ in $\partial\Omega$.

For $(\tilde{\mathbb{P}})$ we now establish the analogue of Theorem 4.3, which hinges on Theorem 5.1.

THEOREM 6.1. *Problem* $(\tilde{\mathbb{P}})$ *admits a solution.*

Proof. Take an infimizing sequence $\{(\mathbf{p}_n, \alpha_n)\}$ for $(\tilde{\mathbb{P}})$. We have that $\{\alpha_n\}$ is bounded in $W^{1,p}(\Omega)$ due to $\lambda > 0$. The embedding $W^{1,p}(\Omega) \hookrightarrow L^2(\Omega)$ is compact [2] yielding a subsequence of $\{\alpha_n\}$ (also denoted by $\{\alpha_n\}$) such that $\alpha_n \rightharpoonup \alpha^*$ in $W^{1,p}(\Omega)$ and $\alpha_n \to \alpha^*$ in $L^2(\Omega)$, for some $\alpha^* \in W^{1,p}(\Omega) \cap \mathcal{A}_{\mathrm{ad}}$ due to the weak closedness of $\mathcal{A}_{\mathrm{ad}}$ in $W^{1,p}(\Omega)$. Additionally, $\mathbf{p}_n = \mathbf{p}(\alpha_n)$, where $\alpha \mapsto \mathbf{p}(\alpha)$ is the solution map of $(\tilde{\mathrm{D}})$. Then, by Theorem 5.1, we have $\mathbf{p}_n \to \mathbf{p}^* := \mathbf{p}(\alpha^*)$ in $H^1_0(\Omega)^\ell$. The fact that $(\mathbf{p}^*, \alpha^*)$ is a minimizer follows from weak lower semicontinuity arguments. □

We now focus on the derivation of first-order necessary optimality conditions. For this matter, from now on we consider only the case $p = 2$.

Utilizing the solution map $\alpha \mapsto \mathbf{p}(\alpha)$ for the regularized lower-level problem, the reduced version of $(\tilde{\mathbb{P}})$ is given by

$$\text{minimize} \quad \hat{J}(\alpha) := J(\mathbf{p}(\alpha), \alpha) \quad \text{over } \alpha \in \mathcal{A}_{\mathrm{ad}}. \tag{6.2}$$

Since the map $H^1(\Omega) \ni \alpha \mapsto \mathbf{p}(\alpha) \in H^1_0(\Omega)^\ell$ is differentiable (recall Theorem 5.2), it follows that $H^1(\Omega) \ni \alpha \mapsto \hat{J}(\alpha) \in \mathbb{R}$ is differentiable as well, with derivative $\hat{J}'(\alpha) \in H^1(\Omega)^*$ and gradient $\nabla\hat{J}(\alpha) := \mathcal{R}^{-1}\hat{J}'(\alpha) \in H^1(\Omega)$ with $\mathcal{R}$ the Riesz map for $H^1(\Omega)$.

6.1. First-order optimality system.

The following first-order necessary optimality conditions will be the starting point for the conception of solution algorithms for $(\tilde{\mathbb{P}})$ and the associated adjoint equation is fundamental for efficient computations of the gradient of the reduced objective $\hat{J}(\cdot)$. The latter object is central for defining a projected gradient algorithm in part II of this work [38].

Set $X := V \times H^1(\Omega)$ with $\mathbf{x} := (\mathbf{p},\alpha) \in X$, $V := H^1_0(\Omega)^\ell$, $\mathcal{C} := V \times \mathcal{A}_{\mathrm{ad}}$, and $Y := V^*$. Let $T : X \to \mathbb{R}$ be the differentiable map with $T(\mathbf{x}) := J(\mathbf{p},\alpha)$, and let $g : X \to Y$ with

$$g(\mathbf{x}) := -\beta\boldsymbol{\Delta}\mathbf{p} + \gamma\mathbf{p} + A\mathbf{p} + \mathbf{f} + \frac{1}{\epsilon}P_\delta(\mathbf{p},\alpha).$$

Then, problem $(\tilde{\mathbb{P}})$ is equivalent to

$$\begin{aligned} &\text{minimize} \quad T(\mathbf{x}) \quad \text{over } \mathbf{x} \in X, \\ &\text{s.t.} \quad \mathbf{x} \in \mathcal{C} \quad \text{and} \quad g(\mathbf{x}) = 0, \end{aligned}$$

where $T(\mathbf{x}) := F \circ R(\operatorname{div} \mathbf{p}) + \frac{\lambda}{2}|\alpha|^2_{H^1(\Omega)}$.

As a first step towards first-order optimality conditions for this problem, we show that $g : X \to Y$ is differentiable. Let $X = V \times H^1(\Omega)$ with $\mathbf{x} = (\mathbf{p}, \alpha) \in X$, $V = H_0^1(\Omega)^\ell$, $\mathcal{C} = V \times \mathcal{A}_{\rm ad}$, and $Y = V^*$.

Proposition 6.2. *The map $g : X \to Y$ is differentiable at any $\mathbf{x} = (\mathbf{p}, \alpha) \in \mathcal{C}$. For $\mathbf{r} = (\mathbf{r}_1, r_2) \in X$, the derivative is given by*

$$g'(\mathbf{x})\mathbf{r} = -\beta \boldsymbol{\Delta} \mathbf{r}_1 + \gamma \mathbf{r}_1 + A\mathbf{r}_1 + \frac{1}{\epsilon} P'_\delta(\mathbf{p}, \alpha)(\mathbf{r}_1, r_2),$$

with $P'_\delta(\mathbf{p}, \alpha)(\mathbf{r}_1, r_2) = D_1 P_\delta(\mathbf{p}, \alpha)\mathbf{r}_1 + D_2 P_\delta(\mathbf{p}, \alpha) r_2$, where $D_1 P_\delta(\mathbf{p}, \alpha)$ and $D_2 P_\delta(\mathbf{p}, \alpha)$ denote the Fréchet derivatives of $\mathbf{p} \mapsto P_\delta(\mathbf{p}, \alpha)$ and $\alpha \mapsto P_\delta(\mathbf{p}, \alpha)$, respectively, and are given by

$$D_1 P_\delta(\mathbf{p}, \alpha)\mathbf{r}_1 := \big(\boldsymbol{G}''_\delta(\mathbf{p} - \alpha \mathbf{1}) + \boldsymbol{G}''_\delta(-\mathbf{p} - \alpha \mathbf{1})\big)\mathbf{r}_1, \tag{6.3a}$$

$$D_2 P_\delta(\mathbf{p}, \alpha) r_2 := \big(\boldsymbol{G}''_\delta(-\mathbf{p} - \alpha \mathbf{1}) - \boldsymbol{G}''_\delta(\mathbf{p} - \alpha \mathbf{1})\big)\mathbf{1} r_2, \tag{6.3b}$$

where $\boldsymbol{G}'_\delta : L^{2+\xi}(\Omega)^\ell \to L^2(\Omega)^\ell$ (for some sufficiently small $\xi > 0$) is given by $\boldsymbol{G}'_\delta(\mathbf{p}) = (G'_\delta(p_1), \ldots, G'_\delta(p_l))$ where $G'_\delta : L^{2+\xi}(\Omega) \to L^2(\Omega)$ is the Nemytskii operator induced by the real valued function $r \mapsto (r)^+_\delta$. Further, $\boldsymbol{G}''_\delta$ denotes the Fréchet derivative of $\boldsymbol{G}'_\delta$

The proof of the above proposition follows from standard results for the differentiability of superposition operators [5, 57]. In fact, we first note that from the Sobolev Imbedding Theorem (see [2]) for any $l \geq 1$ (recall that $\Omega \subset \mathbb{R}^\ell$) there exists $\xi > 0$, such that $H^1(\Omega) \hookrightarrow L^{2+\xi}(\Omega)$: In fact, for $l = 1$, $H^1(\Omega) \hookrightarrow C^{0,1/2}(\overline{\Omega})$, for $l = 2$, $H^1(\Omega) \hookrightarrow L^q(\Omega)$ with $2 \leq q < \infty$ and for $l \geq 3$, $H^1(\Omega) \hookrightarrow L^{2\frac{l}{l-2}}(\Omega)$. Therefore, since $H_0^1(\Omega) \hookrightarrow H^1(\Omega)$ and $L^2(\Omega) \hookrightarrow H^{-1}(\Omega)$, it suffices to prove that $L^{2+\xi}(\Omega)^\ell \times L^{2+\xi}(\Omega) \ni (\mathbf{p}, \alpha) \mapsto g(\mathbf{p}, \alpha) \in L^2(\Omega)^\ell$ is differentiable.

Note that the map $P_\delta : (W_1)^\ell \times W_2 \to W_3$ is differentiable when W_1 and W_2 are any of the spaces: $L^{2+\xi}(\Omega), H^1(\Omega)$ and $H_0^1(\Omega)$ and W_3 is $L^2(\Omega)^\ell$ or $H^{-1}(\Omega)^\ell$.

Further observe that for $\mathbf{p} = \{p^i\}_{i=1}^\ell$, $\mathbf{r}_1 = \{r_1^i\}_{i=1}^\ell$ with $p^i, r_1^i \in H_0^1(\Omega)$ for $i = 1, 2, \ldots, l$ and $\alpha, r_2 \in H^1(\Omega)$, the expressions for (6.3a) and (6.3b) are given explicitly by $D_1 P_\delta(\mathbf{p}, \alpha)\mathbf{r}_1 = \{((p^i - \alpha)^{+'}_\delta + (-p^i - \alpha)^{+'}_\delta) r_1^i\}_{i=1}^\ell$ and $D_2 P_\delta(\mathbf{p}, \alpha) r_2 = \{((-p^i - \alpha)^{+'}_\delta - (p^i - \alpha)^{+'}_\delta) r_2\}_{i=1}^\ell$. This implies that that $D_1 P_\delta(\mathbf{p}, \alpha), D_2 P_\delta(\mathbf{p}, \alpha) \in L^\infty(\Omega)^\ell$.

A direct application of KKT-theory in Banach space [59] finally yields the following characterization.

Proposition 6.3. *For an optimal solution $(\mathbf{p}^*, \alpha^*) \in H_0^1(\Omega)^\ell \times \mathcal{A}_{ad}$ of $(\tilde{\mathbb{P}})$, there exists an adjoint state (a Lagrange multiplier) $\mathbf{q}^* \in H_0^1(\Omega)^\ell$ such that*

$$\begin{aligned} (J_0'(\operatorname{div} \mathbf{p}^*), \operatorname{div} \mathbf{p}) + \langle -\beta \boldsymbol{\Delta} \mathbf{q}^* + \gamma \mathbf{q}^* + A\mathbf{q}^* + \frac{1}{\epsilon} D_1 P_\delta(\mathbf{p}^*, \alpha^*)\mathbf{q}^*, \mathbf{p} \rangle_{H^{-1}, H_0^1} = 0, \\ \langle \lambda(-\Delta + I)\alpha^* + \frac{1}{\epsilon}\left(D_2 P_\delta(\mathbf{p}^*, \alpha^*)\right)^\top \mathbf{q}^*, \alpha - \alpha^* \rangle_{H^1(\Omega)^*, H^1(\Omega)} \geq 0, \end{aligned}$$

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

$$-\beta \mathbf{\Delta} \mathbf{p}^* + \gamma \mathbf{p}^* + A\mathbf{p}^* + \mathbf{f} + \frac{1}{\epsilon} P_\delta(\mathbf{p}^*, \alpha^*) = 0,$$

in the $H^{-1}(\Omega)^\ell$*-sense.*

Note the differentiability of $J_0 = F \circ R$ follows by standard arguments and that "$Z^\top$" denotes the transpose of an operator Z. Further, $\mathbf{q}$ allows to compute $\hat{J}'$ at some α in an amenable way. In fact, we have

$$\hat{J}'(\alpha) = \lambda(-\Delta + I)\alpha + \frac{1}{\epsilon} \left(D_2 P_\delta(\mathbf{p}(\alpha), \alpha)\right)^\top \mathbf{q}(\alpha), \tag{6.4}$$

where $\alpha \mapsto \mathbf{q}(\alpha)$ solves the first equation in Proposition 6.3 for $\mathbf{p}^* = \mathbf{p}(\alpha)$ and $\alpha^* = \alpha$.

The above structural representation of $\hat{J}'(\alpha)$ combined with a regularity result for the metric projection in $H^1(\Omega)$ onto $\mathcal{A}_{\mathrm{ad}}$ provides the starting point for the design of a numerical solver for $(\tilde{\mathbb{P}})$. This development is the subject of the second part of this work [38].

7. Conclusion. The generalization of the total variation regularization filter proposed in this work involves a scalar weight function α. Its proper choice allows to locally adjust the strength of the filter and, hence, to recover image details while still significantly removing noise from homogeneous image regions. While we are able to address important analytical questions such as existence of solutions for the associated variational model for image reconstruction and the relation between the primal problem (P) and its predual problem (D), important issues such as a structural analysis of the implications of spatially weighted total variation regularization (*Depending on the local choice of* α*, which features do occur in reconstruction results, e.g., starting in 1D?*) remain for future work.

In order to identify a monolithic variational approach to both, the reconstruction of the image and the proper choice of α, a bilevel optimization model is proposed. While the image reconstruction problem in the lower level now serves merely the purpose of a provider of a feasible point for the bilevel problem, given a choice of α, the upper level problem, i.e., its objective along with constraints, aims to identify a *good* set of filtering weights. It should be noted that, due to the non-convexity of the bilevel problem, one should not expect this good set to be a singleton, in general. In order to address the non-convexity *and* non-smoothness of the problem when deriving stationarity conditions, a smoothing approach is pursued. In particular, this avoids the complications due to non-differentiability. The latter is particularly complicated as it arises from the solution set of a variational inequality with the relevant parameter occurring in the upper bound of a pointwise norm constraint on $\mathbf{p}$, i.e. the first-order condition for the lower level optimization problem. In this respect, several analytical questions remain for future work. These include a (generalized) differentiability sensitivity analysis of the map $\alpha \mapsto \mathbf{p}(\alpha)$, the passage to the limit with the regularization/penalization parameters δ/ϵ in the first-order system of Proposition 6.3. This, if possible at all, may identify a *useful* limiting stationarity system for the original bilevel problem $(\mathbb{P})$, which may then be the departure point for the design of associated numerical solvers. A similar path has been considered in [15] where a limiting system is obtained via vanishing regularization parameters.

In a second part of this work [38], we propose a projection-based descent algorithm for solving $(\tilde{\mathbb{P}})$. The method yields a approximating sequence of filtering weights

$\{\alpha_k\}$ in $C(\bar{\Omega})$ and is employed to denoising, deblurring and Fourier as well as wavelet inpainting problems.

REFERENCES

[1] R. Acar and C. Vogel. Analysis of bounded variation penalty methods for ill-posed problems. *Inverse Problems*, 10:1217–1229, 1994.

[2] R. A. Adams and J. J. F. Fournier. *Sobolev Spaces.* Academic Press, second edition, 2003.

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

[4] L. Ambrosio, N. Fusco, and D. Pallara. *Functions of bounded variation and free discontinuity problems.* Oxford Mathematical Monographs. The Clarendon Press Oxford University Press, New York, 2000.

[5] J. Appell and P. P. Zabrejko. *Nonlinear Superposition Operators.* Cambridge University Press, 1990.

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

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

[8] G. Aubert and P. Kornprobst. *Mathematical problems in image processing: partial differential equations and the calculus of variations*, volume 147. Springer Science & Business Media, 2006.

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

[10] V. Barbu and T. Precupanu. *Convexity and optimization in Banach spaces.* Springer Monographs in Mathematics. Springer, Dordrecht, fourth edition, 2012.

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

[12] L. Boccardo and F. Murat. Nouveaux résultats de convergence dans des problèmes unilatéraux. In *Nonlinear partial differential equations and their applications. Collège de France Seminar, Vol. II (Paris, 1979/1980)*, volume 60 of *Res. Notes in Math.*, pages 64–85, 387–388. Pitman, Boston, Mass.-London, 1982.

[13] L. Boccardo and F. Murat. Homogenization of nonlinear unilateral problems. In *Composite media and homogenization theory (Trieste, 1990)*, volume 5 of *Progr. Nonlinear Differential Equations Appl.*, pages 81–105. Birkhäuser Boston, Boston, MA, 1991.

[14] L. Calatroni, C. Chung, J. C. D. L. Reyes, C.-B. Schönlieb, and T. Valkonen. Bilevel approaches for learning of variational imaging models. 05 2015.

[15] V. C. Cao, J. C. De los Reyes, and C. B. Schoenlieb. Learning optimal spatially-dependent regularization parameters in total variation image restoration. *ArXiv e-prints*, Mar. 2016.

[16] A. Chambolle. An algorithm for total variation minimization and application. *J. Math. Imaging Vis.*, 20:89–97, 2004.

[17] A. Chambolle and P.-L. Lions. Image recovery via total variation minimization and related problems. *Numer. Math.*, 76(2):167–188, 1997.

[18] T. Chan, G. Golub, and P. Mulet. A nonlinear primal-dual method for total variation-based image restoration. *SIAM Journal on Scientific Computing*, 20(6):1964–1977, 1999.

[19] Y. Chen. *Learning fast and effective image restoration models.* PhD thesis, Graz University of Technology, 2014.

[20] Y. Chen, W. Yu, and T. Pock. On learning optimized reaction diffusion processes for effective image restoration. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[21] J. C. De los Reyes, C.-B. Schönlieb, and T. Valkonen. Bilevel parameter learning for higher-order total variation regularisation models. *Journal of Mathematical Imaging and Vision*, pages 1–25, 2016.

[22] J. C. De Los Reyes, C.-B. Schönlieb, and T. Valkonen. The structure of optimal parameters for image restoration problems. *J. Math. Anal. Appl.*, 434(1):464–500, 2016.

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

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

2011.

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

[26] N. Dunford and J. T. Schwartz. *Linear Operators. I. General Theory*. With the assistance of W. G. Bade and R. G. Bartle. Pure and Applied Mathematics, Vol. 7. Interscience Publishers, Inc., New York, 1958.

[27] I. Ekeland and R. Temam. *Convex Analysis and Variational Problems*. North-Holland and American Elsevier, 1976.

[28] J. Fehrenbach, M. Nikolova, G. Steidl, and P. Weiss. Bilevel image denoising using Gaussianity tests. In *Scale space and variational methods in computer vision*, volume 9087 of *Lecture Notes in Comput. Sci.*, pages 117–128. Springer, Cham, 2015.

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

[30] V. Girault and P.-A. Raviart. *Finite element methods for Navier-Stokes equations: theory and algorithms*, volume 5. Springer-Verlag, Berlin, 1986.

[31] E. Giusti. *Minimal Surfaces and Functions of Bounded Variation*. Birkhäuser, 1984.

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

[33] E. J. Gumbel. *Statistics of extremes*. Dover Publications, Inc., Mineola, NY, 2004. Reprint of the 1958 original [Columbia University Press, New York; MR0096342].

[34] M. Hintermüller and I. Kopacka. Mathematical programs with complementarity constraints in function space: *C*- and strong stationarity and a path-following algorithm. *SIAM J. Optim.*, 20(2):868–902, 2009.

[35] M. Hintermüller and K. Kunisch. Total bounded variation regularization as a bilaterally constrained optimization problem. *SIAM Journal of Applied Mathematics*, 64:1311–1333, 2004.

[36] M. Hintermüller and C. N. Rautenberg. On the density of classes of closed convex sets with pointwise constraints in Sobolev spaces. *J. Math. Anal. Appl.*, 426(1):585–593, 2015.

[37] M. Hintermüller, C. N. Rautenberg, and S. Rösel. Density of convex intersections and applications. *WIAS Preprint No. 2333*, 2016.

[38] M. Hintermüller, C. N. Rautenberg, T. Wu, and A. Langer. Optimal selection of the regularization function in a generalized total variation model. Part II: Algorithm, its analysis and numerical tests. WIAS Preprint No. 2236, (2016).

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

[40] M. Hintermüller and G. Stadler. An infeasible primal-dual algorithm for total bounded variation-based inf-convolution-type image restoration. *SIAM Journal on Scientific Computing*, 28(1):1–23, 2006.

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

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

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

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

[45] K. Jalalzai. Discontinuities of the minimizers of the weighted or anisotropic total variation for image reconstruction. Technical Report arXiv:1402.0026, arXiv.org Cornell University Library, 2014.

[46] T. Klatzer and T. Pock. Continuous hyper-parameter learning for support vector machines. In *Computer Vision Winter Workshop (CVWW)*, 2015.

[47] K. Kunisch and T. Pock. A bilevel optimization approach for parameter learning in variational models. *SIAM J. Imaging Sci.*, 6(2):938–983, 2013.

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

[49] U. Mosco. Convergence of convex setis and solutions of variational inequalities. *Advances in Mathematics*, 3 (4):510–585, 1969.

[50] P. Ochs, R. Ranftl, T. Brox, and T. Pock. Bilevel optimization with nonsmooth lower level problems. In *Scale space and variational methods in computer vision*, volume 9087 of *Lecture Notes in Comput. Sci.*, pages 654–665. Springer, Cham, 2015.
[51] J. Outrata, M. Kocvara, and J. Zowe. *Nonsmooth Approach to Optimization Problems with Equilibrium Constraints*, volume 28 of *Nonconvex Optimization and its Applications*. Kluwer Academic Publishers, Dordrecht, Netherlands, 1998.
[52] J. F. Rodrigues. *Obstacle Problems in Mathematical Physics*. North-Holland, 1987.
[53] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1â“4):259–268, 11 1992.
[54] U. Schmidt and S. Roth. Shrinkage fields for effective image restoration. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.
[55] C. Schönlieb and J. C. De Los Reyes. Image denoising: Learning noise distribution via pde-constrained optimisation. *Inverse Problems and Imaging*, 7(4):1183–1214, 2013.
[56] R. E. Showalter. *Monotone Operators in Banach Space and Nonlinear Partial Differential Equations*. American Mathematical Society, 1997.
[57] F. Tröltzsch. *Optimal control of partial differential equations*, volume 112 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2010. Theory, methods and applications, Translated from the 2005 German original by Jürgen Sprekels.
[58] C. Vogel. *Computational Methods for Inverse Problems*, volume 23 of *Frontiers Appl. Math.* SIAM, Philadelphia, USA, 2002.
[59] J. Zowe and S. Kurcyusz. Regularity and stability for the mathematical programming problem in Banach spaces. *Appl. Math. Optim.*, 5(1):49–62, 1979.