LEARNING NONLOCAL REGULARIZATION OPERATORS

GERNOT HOLLE
Institute of Mathematics and Scientific Computing
University of Graz
Heinrichstrasse 36, 8010 Graz, Austria

KARL KUNISCH
Institute of Mathematics and Scientific Computing
University of Graz
Heinrichstrasse 36, 8010 Graz, Austria
Radon Institute
Austrian Academy of Sciences, RICAM Linz
Altenbergerstrasse 69, 4040 Linz, Austria

(Communicated by Harbir Antil)

Abstract. A learning approach for determining which operator from a class of nonlocal operators is optimal for the regularization of an inverse problem is investigated. The considered class of nonlocal operators is motivated by the use of squared fractional order Sobolev seminorms as regularization operators. First fundamental results from the theory of regularization with local operators are extended to the nonlocal case. Then a framework based on a bilevel optimization strategy is developed which allows to choose nonlocal regularization operators from a given class which i) are optimal with respect to a suitable performance measure on a training set, and ii) enjoy particularly favorable properties. Results from numerical experiments are also provided.

1. Introduction. In this work we discuss the use of a family of nonlocal energy seminorms for the regularization of inverse problems governed by partial differential equations. The archetypes for the considered family are Sobolev seminorms $|u|_{H^s(\Omega)}$ of fractional order $s \in (0, 1)$. The corresponding regularized inverse problems are

$$
\min_{u \in \text{dom}(S) \cap H^s(\Omega)} \| S(u) - y_\delta \|_{L^2(\Omega)}^2 + \nu |u|_{H^s(\Omega)}^2.
$$

Here $S$: dom$(S) \subseteq L^2(\Omega) \to L^2(\Omega)$ is the given forward operator, $\nu > 0$ is a regularization parameter, and $y_\delta \in L^2(\Omega)$ is the given measurement. The considered family of nonlocal energy seminorms $(|\cdot|_{\gamma,s})_{\gamma \in W_{ad}}$ will differ from Sobolev seminorms only by additional weighting terms $\gamma \in W_{ad}$. A precise definition of the nonlocal energy seminorms and the set of admissible weights $W_{ad}$ will be given in Section 2 below. The corresponding regularized inverse problem for a particular weight $\gamma \in W_{ad}$ is

$$
\min_{u \in \text{dom}(S) \cap H^s(\Omega)} \| S(u) - y_\delta \|_{L^2(\Omega)}^2 + |u|_{\gamma,s}^2.
$$

2020 Mathematics Subject Classification. Primary: 49J20, 49N45, 35R11; Secondary 35J05.

Key words and phrases. Nonlocal operators, optimal control, inverse problems, bilevel optimization, fractional order regularization, learning.

* Corresponding author: Gernot Holler.
The use of nonlocal weights was triggered from the experience with regularizers involving arrays of filters as described for instance in [28] in the context of image reconstruction. To cast this setting into a functional analytic setting the use of nonlocal operators and nonconstant weights arises as a natural choice. Moreover the choice of $s > 0$ and sufficiently large can also be important for the well-posedness of the forward mapping $S$.

1.1. The learning problem. To determine which element from this family of nonlocal energy seminorms is particularly suitable for a given problem we use a learning approach: We assume to be given ground truth data and noisy measurements, i.e. a set $(y_i^\dagger, u_i^\dagger, y_{\delta i})_{1 \leq i \leq N_{\text{Train}}}$ such that

$$S(u_i^\dagger) = y_i^\dagger,$$

and $y_{\delta i}$ are noisy measurements of $y_i^\dagger$ for $1 \leq i \leq N_{\text{Train}}$. We then determine a weight $\gamma^*$ such that solutions to the corresponding inverse problems represent the ground truth data particularly well. This is done by choosing $\gamma^* \in W_{\text{ad}}$ as a solution to

$$\min_{\gamma \in W_{\text{ad}}, u \in H^{s}(\Omega)} \frac{1}{2N_{\text{Train}}} \sum_{i=1}^{N_{\text{Train}}} \|u_i - u_i^\dagger\|_{L^2(\Omega)}^2 + R(\gamma)$$

subject to $u_i \in \arg \min_{u \in \text{dom}(S) \cap H^{s}(\Omega)} \|S(u) - y_{\delta i}\|_{L^2(\Omega)}^2 + |u|_{\gamma,s}^2.$

Here, $R: W_{\text{ad}} \to \mathbb{R}$ is an added regularization operator. We will favor the choice $R$ as the $L^1$ norm. This has the effect that nonlocality is only utilized if its effect is sufficiently strong, otherwise it is set to zero. As a side effect of this procedure, we obtain that in the regularized inverse problem the system matrices, which tend to be densely populated in the context of fractional order regularization, in fact become more sparse. Except for the numerical experiments, we only consider the case $N_{\text{Train}} = 1$. However, generalization of the analytical results to the case of multiple data vectors is straightforward using product spaces.

This paper is organized as follows. In Section 2 the necessary background is provided, and a stability property for solutions to Poisson-type nonlocal equations, which will be frequently needed throughout this work, is derived. Moreover, the class of weights considered in this work is introduced. Section 3 is concerned with the case of a linear forward problem. After deriving some basic properties of the regularized inverse problem, existence of solutions to the learning problem is proven and an optimality system is derived. In Section 4 we discuss the nonlinear case. After providing some results, which can be applied to general nonlinear functions, we discuss in detail the problem of estimating the convection term in an elliptic PDE. Finally, in Section 5 results from numerical experiments are presented which demonstrate the feasibility of our approach.

1.2. Related work. Note that (BP) is a bilevel optimization problem, i.e. an optimization problem, where the constraint involves another optimization problem (referred to as the lower level problem). A standard reference on bilevel optimization is [16]. Nonlocal operators have recently received a significant amount of attention in the literature, see e.g. [22, 18, 13, 2]. A learning problem for determining optimal filter parameters for nonlocal regularization operators in the context of image denoising problems was recently investigated in [12]. As a particular instance of
nonlocal regularization operators, fractional-type regularization operators are considered in [3, 4]. In terms of learning theory, the problem of learning regularization operators can be viewed as a supervised learning problem. The problem of choosing regularization operators from a parametrized class of functions based on training data, is studied in [23]. Optimal spectral filters for finite dimensional inverse problems are learned in [10]. Learning strategies for choosing regularization parameters in the context of multi-penalty Tikhonov regularization are investigated e.g. in [28, 14, 11, 25]. The problem of learning the discrepancy function is considered in [15]. In many of the mentioned references, the lower level problem is not differentiable, which in turn complicates the derivation of optimality conditions. This issue is then often overcome by smoothing the lower level problem. A different approach is presented in [29], where instead of smoothing the lower level problem, it is suggested to replace the lower level problem constraint by a differentiable update rule, which is given as the n-th step in an iterative procedure to determine approximate solutions to the lower level problem. In [24] a bilevel optimization approach is analysed where the lower level problem involves the bounded variation seminorm with a spatially dependent weight function, which is the optimization variable for the upper level problem, which is chosen on the basis of statistical reasoning.

2. Nonlocal energy spaces.

2.1. Preliminaries. Throughout this work, unless otherwise stated, we let $s \in (0, 1)$ and let $\Omega$ denote a nonempty, open, connected, and bounded Lipschitz domain in $\mathbb{R}^N$, where $N \in \mathbb{N}$. Furthermore, $|\cdot|$ denotes the Euclidian norm of a vector in $\mathbb{R}^N$. Following [18, Section 4], we introduce the notion of a nonlocal energy seminorm.

**Definition 2.1 (nonlocal energy).** Let $\gamma \in L^\infty(\Omega \times \Omega)$ be nonnegative and symmetric a.e. on $\Omega \times \Omega$. For $u \in L^2(\Omega)$ define a nonlocal energy seminorm by

$$|u|_{\gamma,s} := \left( \int_{\Omega \times \Omega} \left| \frac{u(y) - u(x)}{|x - y|^{N+2s}} \gamma(x, y) \right|^2 dx \right)^{1/2}.$$ 

The corresponding nonlocal energy space is defined by

$$V^{\gamma,s}(\Omega) := \{ v \in L^2(\Omega) : |v|_{\gamma,s} < \infty \}$$

and endowed with the norm

$$\|u\|_{V^{\gamma,s}(\Omega)} := \left( \|u\|^2_{L^2(\Omega)} + |u|^2_{\gamma,s} \right)^{1/2}. \quad (3)$$

**Remark 1.** If $\gamma$ is equal to 1 almost everywhere on $\Omega \times \Omega$, we write

$$H^s(\Omega) := V^{1,s}(\Omega), \quad |u|_{H^s(\Omega)} := |u|_{\gamma,s}, \quad \text{and} \quad \|u\|_{H^s(\Omega)} := \|u\|_{V^{1,s}(\Omega)}.$$ 

With this notation, $H^s(\Omega)$ coincides with the usual Sobolev space of fractional order $s$ (also known as Sobolev-Slobodeckij space), see e.g. [17] and [21, Definition 1.3.2.1].

We now provide a set of assumptions on the weight $\gamma$, under which the nonlocal energy norm $\|\cdot\|_{V^{\gamma,s}(\Omega)}$ defined by (3) is equivalent to the fractional order Sobolev norm $\|\cdot\|_{H^s(\Omega)}$, which in turn implies that the corresponding nonlocal energy space $V^{\gamma,s}(\Omega)$ coincides with the fractional order Sobolev space $H^s(\Omega)$.

**Assumption 2.1.** The weight $\gamma \in L^\infty(\Omega \times \Omega)$ is nonnegative and symmetric a.e. on $\Omega \times \Omega$. Furthermore, there exist constants $\gamma_1, \gamma_2, \delta > 0$ such that for almost all $(x, y) \in \Omega \times \Omega$ the following statements hold:
Hence, the norms established in Corollary 1 now implies that $\|\cdot\|_{\gamma,s}(\Omega)$ that induces the norm $\gamma(x,y)$.

Remark 2. If, given any function $\sigma: [0, \infty) \to [0, \infty)$ satisfying $\gamma_1 \leq \sigma(z) \leq \gamma_2$ for all $z \in [0, \infty)$, we define $\gamma$ by

$$
\gamma(x,y) := \begin{cases} 
\sigma(|x-y|) & \text{if } |x-y| \leq \delta, \\
0 & \text{else,}
\end{cases}
$$

for almost all $(x,y) \in \Omega \times \Omega$,

then $\gamma$ satisfies Assumption 2.1. Note that - except for boundary effects - the regularizer utilizing $\sigma$ is invariant under translation and rotation, which is not the case for general functions $\gamma$.

The following result is a combination of Lemmas 4.1 and 4.2 from [18].

**Lemma 2.2.** Let $\gamma \in L^\infty(\Omega \times \Omega)$ satisfy Assumption 2.1 and let $u \in L^2(\Omega)$. Then

$$
|u|_{\gamma,s}^2 \leq \gamma_1 |u|_{H^{\gamma,s}(\Omega)}^2 \quad \text{and} \quad |u|_{H^{\gamma,s}(\Omega)}^2 \leq \gamma_1^{-1} |u|_{\gamma,s}^2 + 4|\Omega|\delta^{-N-s}||u||_{L^2(\Omega)}^2. \quad (4)
$$

Here, $|\Omega|$ denotes the Lebesgue-measure of $\Omega$.

As a direct corollary of Theorem 2.2 we obtain that if $\gamma \in L^\infty(\Omega \times \Omega)$ satisfies Assumption 2.1, then the corresponding nonlocal energy space is topologically equivalent to the fractional order Sobolev space $H^s(\Omega)$.

**Corollary 1** (Equivalence of norms). There exist constants $m, M > 0$ such that for all $\gamma \in L^\infty(\Omega \times \Omega)$ satisfying Assumption 2.1 we have

$$
m||u||_{H^{\gamma,s}(\Omega)} \leq ||u||_{V^{\gamma,s}(\Omega)} \leq M||u||_{H^{\gamma,s}(\Omega)}, \quad \text{for all } u \in L^2(\Omega).
$$

In particular, the norms on $H^s(\Omega)$ and $V^{\gamma,s}(\Omega)$ are equivalent.

It is straightforward to verify that

$$
\langle u, v \rangle_{\gamma,s} := + \int_{\Omega \times \Omega} \frac{(u(x) - u(y))(v(x) - v(y))}{|x-y|^{N+2s}} \gamma(x,y) \, dx \, dy
$$

defines a symmetric and positive semidefinite bilinear form on $V^{\gamma,s}(\Omega)$. Moreover, if we let $\langle \cdot, \cdot \rangle_{V^{\gamma,s}(\Omega)} := \langle \cdot, \cdot \rangle_{L^2(\Omega)} + \langle \cdot, \cdot \rangle_{\gamma,s}$, then $\langle \cdot, \cdot \rangle_{V^{\gamma,s}(\Omega)}$ is an inner product on $V^{\gamma,s}(\Omega)$ that induces the norm $\| \cdot \|_{V^{\gamma,s}(\Omega)}$. Since $H^s(\Omega)$ is complete, the equivalence of norms established in Corollary 1 now implies that $V^{\gamma,s}(\Omega)$ is also complete. Hence, $V^{\gamma,s}(\Omega)$ is a Hilbert space. We let

$$
\Pi^0(\Omega) := \{ f \in L^2(\Omega) \mid \exists c \in \mathbb{R} : f(x) = c \quad \text{a.e. on } \Omega \}
$$

denote the space of functions in $L^2(\Omega)$ which are constant a.e. on $\Omega$. We denote by $Q^0 : L^2(\Omega) \to \Pi^0(\Omega)$ the $L^2$-orthogonal projection on $\Pi^0(\Omega)$. For $u \in L^2(\Omega)$ we have

$$
Q^0 u(x) = c, \quad \text{where } c := (1/|\Omega|) \int_{\Omega} u(x) dx.
$$

**Lemma 2.3.** Let $\gamma \in L^\infty(\Omega \times \Omega)$ satisfy Assumption 2.1. Then for $u \in L^2(\Omega)$ we have $|u|_{\gamma,s} = 0$ if and only if $u \in \Pi^0(\Omega)$.

**Proof.** Using Fubini’s theorem, it is straightforward to verify that for $u \in \Pi^0(\Omega)$ we have $|u|_{\gamma,s} = 0$. Conversely, if for $u \in H^s(\Omega)$ it holds that $|u|_{\gamma,s} = 0$, then

$$
\int_{\Omega} \frac{|u(x) - u(y)|^2}{|x-y|^{N+2s}} \gamma(x,y) \, dy = 0 \quad \text{for almost all } x \in \Omega.
$$

This implies that $u$ is constant a.e. on $B_\delta(x) \cap \Omega$ for almost all $x \in \Omega$. Since $\Omega$ is connected, the claim follows by standard arguments.

\[\square\]
We show that for all $\sigma$,

**Proof.**

Then $u$ is an accumulation point of $(\gamma, s)$ with respect to the strong topology on $L^2(\Omega)$ such that

$$Q^0 u^n = 0, \quad \|u^n\|_{L^2(\Omega)} = 1, \quad \text{and} \quad |u^n|_{\gamma_{\min}, s} \leq 1/n \quad \text{for all } n \in \mathbb{N}.$$ Using the equivalence of norms established in Corollary 1, it follows that $(u^n)$ is bounded in $H^s(\Omega)$. Since $H^s(\Omega)$ is reflexive, $(u^n)$ has an accumulation point $u \in H^s(\Omega)$ with respect to the weak topology on $H^s(\Omega)$. Since $H^s(\Omega)$ is compactly embedded in $L^2(\Omega)$ (see [21, Theorem 1.4.3.2]), it follows that $u$ is also an accumulation point of $(u^n)$ with respect to the strong topology on $L^2(\Omega)$. Using the continuity of $Q^0$ on $L^2(\Omega)$ and the weak lower semi continuity of the nonlocal energy seminorm we deduce that $Q^0 u = 0$ and $|u|_{\gamma_{\min}, s} = 0$. From Theorem 2.3 it is clear that this implies $u = 0$. However, since $u$ is an accumulation point of $(u^n)$ with respect to the strong topology on $L^2(\Omega)$, we must also have $\|u\|_{L^2(\Omega)} = \lim_{n \to \infty} \|u^n\|_{L^2(\Omega)} = 1$, which is a contradiction. Hence, the proof is finished. 

**Remark 3.** The requirement that $\Omega$ is connected is essential to ensure that for every $\sigma$ satisfying Assumption 2.1 and every $u \in L^2(\Omega)$ the seminorm $|u|_{\sigma, s}$ is zero if and only if $u$ is constant almost everywhere. The fractional order Sobolev seminorm $|u|_{\gamma(\Omega)}$, however, has this property for all open sets $\Omega$, connected or not. The reason for this is that while the weight is equal to 1 almost everywhere for the fractional order Sobolev norm, in general $\sigma(|x-y|)$ might be zero for $|x-y| > \delta$.

**Lemma 2.4** (Poincare-Wirtinger inequality for nonlocal energy spaces). There exists a constant $C > 0$ such that for every $\gamma \in L^\infty(\Omega \times \Omega)$ satisfying Assumption 2.1 and every $u \in H^s(\Omega)$ with $Q^0 u = 0$ it holds that

$$\|u\|_{L^2(\Omega)} \leq C|u|_{\gamma, s}.$$ (5)

**Proof.** Let $\gamma_{\min} := \gamma \chi_A$, where $\chi_A$ denotes the characteristic function of the set $A := \{(x, y) \in \Omega \times \Omega \mid |x-y| \geq \delta\}$. It is easy to show that $|u|_{\gamma_{\min}, s} \leq |u|_{\gamma, s}$ for every $u \in H^s(\Omega)$ and every $\gamma$ satisfying Assumption 2.1. Consequently, it suffices to prove the claim for $\gamma = \gamma_{\min}$. We argue by contradiction. If the claim is wrong, then there is a sequence $(u^n)$ in $H^s(\Omega)$ such that

$$Q^0 u^n = 0, \quad \|u^n\|_{L^2(\Omega)} = 1, \quad \text{and} \quad |u^n|_{\gamma_{\min}, s} \leq 1/n \quad \text{for all } n \in \mathbb{N}.$$ Using the equivalence of norms established in Corollary 1, it follows that $(u^n)$ is bounded in $H^s(\Omega)$. Since $H^s(\Omega)$ is reflexive, $(u^n)$ has an accumulation point $u \in H^s(\Omega)$ with respect to the weak topology on $H^s(\Omega)$. Since $H^s(\Omega)$ is compactly embedded in $L^2(\Omega)$ (see [21, Theorem 1.4.3.2]), it follows that $u$ is also an accumulation point of $(u^n)$ with respect to the strong topology on $L^2(\Omega)$. Using the continuity of $Q^0$ on $L^2(\Omega)$ and the weak lower semi continuity of the nonlocal energy seminorm we deduce that $Q^0 u = 0$ and $|u|_{\gamma_{\min}, s} = 0$. From Theorem 2.3 it is clear that this implies $u = 0$. However, since $u$ is an accumulation point of $(u^n)$ with respect to the strong topology on $L^2(\Omega)$, we must also have $\|u\|_{L^2(\Omega)} = \lim_{n \to \infty} \|u^n\|_{L^2(\Omega)} = 1$, which is a contradiction. Hence, the proof is finished. 

**2.2. A stability property.**

**Lemma 2.5** (Stability). Let $1 \leq p, q \leq \infty$ be such that $1/p + 1/q = 1$ and $H^s(\Omega)$ is compactly embedded in $L^q(\Omega)$. Let $u^n \rightharpoonup u$ in $H^s(\Omega)$, $\gamma^n \rightharpoonup \gamma$ in $L^\infty(\Omega \times \Omega)$, and $p^n \to p \in L^p(\Omega)$, where $(u^n)$ in $H^s(\Omega)$, $(\gamma^n)$ in $L^\infty(\Omega \times \Omega)$, and $(p^n)$ in $L^p(\Omega)$ are sequences related by

$$\langle u^n, v \rangle_{\gamma^n, s} = \langle p^n, v \rangle_{L^p(\Omega), L^q(\Omega)} \quad \text{for all } v \in H^s(\Omega) \quad \text{and all } n \in \mathbb{N}. \quad (6)$$

Then

$$\langle u, v \rangle_{\gamma, s} = \langle p, v \rangle_{L^p(\Omega), L^q(\Omega)} \quad \text{for all } v \in H^s(\Omega) \quad \text{and} \quad \lim_{n \to \infty} |u^n|_{\gamma^n, s}^2 = |u|_{\gamma, s}^2. \quad (7)$$

**Proof.** The proof is divided into three steps.

**Step 1.** We show that for all $v \in C^\infty_0(\Omega)$

$$\frac{(u^n(x) - u^n(y))(v(x) - v(y))}{|x-y|^{N+2s}} \to \frac{(u(x) - u(y))(v(x) - v(y))}{|x-y|^{N+2s}} \quad \text{in } L^1(\Omega \times \Omega).$$

Since $s < 1$ we can find $\varepsilon > 0$ such that $s + \varepsilon < 1$ and $\varepsilon' := s - \varepsilon > 0$. Since $\varepsilon' < s$, we have $H^{\varepsilon'}(\Omega)$ is compactly embedded in $H^s(\Omega)$ (see [21, Theorem 1.4.3.2]) and
thus $u^n \to u$ in $H^s(\Omega)$. Using Hölder’s inequality for the first, and the mean value theorem for the second inequality below, we estimate
\[
\int_{\Omega \times \Omega} \frac{|(u^n(x) - u^n(y)) - (u(x) - u(y))|}{|x - y|^{N+2s}} \, dx \, dy \\
\leq |u^n - u|_{H^s(\Omega)} \left( \int_{\Omega \times \Omega} \frac{(v(x) - v(y))^2}{|x - y|^{N+2(2s-s')}} \, dx \, dy \right)^{1/2} \\
\leq |u^n - u|_{H^s(\Omega)} \|Dv\|_{L^\infty(\Omega)} \left( \int_{\Omega \times \Omega} \frac{1}{|x - y|^{N+2(2s-s')-1}} \, dx \, dy \right)^{1/2},
\]
where $C < \infty$ since $2s - s' - 1 = s + \varepsilon - 1 < 0$.

**Step 2.** We compute
\[
\langle p, v \rangle_{L^p(\Omega), L^q(\Omega)} = \lim_{{n \to \infty}} \langle p^n, v \rangle_{L^p(\Omega), L^q(\Omega)} \\
= \lim_{{n \to \infty}} \int_{\Omega \times \Omega} \frac{(u^n(x) - u^n(y))(v(x) - v(y))\gamma_n(x,y)}{|x - y|^{N+2s}} \, dx \, dy \\
= \int_{\Omega \times \Omega} \frac{(u(x) - u(y))(v(x) - v(y))\gamma(x,y)}{|x - y|^{N+2s}} \, dx \, dy \text{ for all } v \in C_\infty(\Omega). \tag{8}
\]
where the result from the first step justifies the third equality. Recalling that $C_\infty(\bar{\Omega})$ is dense in $H^s(\Omega)$ (see [21, Theorem 1.4.2.1]) and observing that (8) is continuous with respect to $v$ on $H^s(\Omega)$, it follows that (8) holds for all $v \in H^s(\Omega)$. This proves the first equality in (7).

**Step 3.** It remains to prove the second equality in (7). Subtracting the first equality in (7) with $v = u$ from (6) with $v = u^n$, we obtain
\[
|u^n|^2_{\gamma_n,s} - |u|^2_{\gamma,s} = \langle p^n, u^n \rangle_{L^p(\Omega), L^q(\Omega)} - \langle p, u \rangle_{L^p(\Omega), L^q(\Omega)} \tag{9}
\]
Since $u^n \to u$ in $L^q(\Omega)$ and $p^n \to p$ in $L^p(\Omega)$ it follows that the right hand side of (9) tends to zero as $n \to \infty$. This finishes the proof. $\square$

2.3. Distance dependent weights. Let $d$ denote the diameter of $\Omega$, i.e. $d := \sup_{x,y \in \Omega} |x - y|$. From now on we restrict ourselves to $\gamma \in L^\infty(\Omega \times \Omega)$ of the form
\[
\gamma(x,y) = \sigma(|x - y|) \quad \text{a.e. on } \Omega \times \Omega, \tag{10}
\]
where $\sigma \in L^\infty((0,d))$ satisfies the following conditions:

(A1) $0 \leq \sigma(t) \leq \gamma_2$ a.e. on $(0,d)$,

(A2) $0 < \gamma_1 \leq \sigma(t)$ a.e. on $(0,\delta)$.

To simplify notation, if $\sigma$ and $\gamma$ are related by (10), then we write $|u|_{\sigma,s} := |u|_{\gamma,s}$. The set of feasible weights is defined by
\[
W_{ad} := \left\{ \sigma \in L^\infty((0,d)) \mid \sigma \text{ satisfies } (A1) \text{ and } (A2) \right\}.
\]
Some care must be taken, since it is not immediately clear, although intuitively reasonable, that for every $\sigma \in L^\infty((0,d))$ there is $\gamma \in L^\infty(\Omega \times \Omega)$ satisfying (10). The difficulty stems from the fact that $L^\infty((0,d))$ consists only of equivalence classes of functions coinciding in the almost everywhere sense on $(0,d)$. We emphasize that (10) must be understood in the sense that it holds for all representatives of the equivalence classes $\sigma$ and $\gamma$. In the following proposition we confirm that the assumption that $\gamma \in L^\infty(\Omega \times \Omega)$ is well-defined by (10) for any $\sigma \in L^\infty((0,d))$ is
indeed justified. To avoid confusion between equivalence classes of functions and their representatives, in the following proposition, we use the special notation \([\cdot]\) to denote equivalence classes of functions.

**Proposition 1.** For every \([\sigma] \in L^\infty((0,d))\) there exists a unique \([\gamma] \in L^\infty(\Omega \times \Omega)\) such that for all \(\sigma \in [\sigma]\) and \(\gamma \in [\gamma]\)

\[
\gamma(x, y) = \sigma(|x - y|) \quad \text{a.e. on } \Omega \times \Omega. \tag{11}
\]

Moreover, it holds that

\[
||[\gamma]||_{L^\infty(\Omega \times \Omega)} \leq ||[\sigma]||_{L^\infty((0,d))}. \tag{12}
\]

**Proof.** First, we take a particular representative \(\sigma: (0,d) \rightarrow \mathbb{R}\) of the equivalence class of measurable functions \([\sigma] \in L^\infty((0,d))\) to define the equivalence class of functions \([\gamma]\) as the set of all measurable functions \(\gamma: \Omega \times \Omega \rightarrow \mathbb{R}\) satisfying

\[
\gamma(x, y) = \sigma(|x - y|) \quad \text{for almost all } (x, y) \in \Omega \times \Omega.
\]

We now prove that \([\gamma \psi] \in L^1(\Omega \times \Omega)\) for every \([\psi] \in L^1(\Omega \times \Omega)\). Using that Fubini's theorem and polar coordinates can be employed for all nonnegative and measurable functions (see [20, Theorems 2.39 and 2.49]), this follows from the estimate

\[
||[\gamma \psi]||_{L^1(\Omega \times \Omega)} = \int_{\Omega} \int_0^d \int_{\partial B(x,r) \cap \Omega} |\sigma(|x - y|)| \psi(x, y)| dS(y) dr dx
\]

\[
= \int_0^d \int_{\Omega} |\sigma(r)| \int_{\partial B(x,r) \cap \Omega} |\psi(x, y)| dS(y) dr dx \leq ||[\sigma]||_{L^\infty((0,d))} ||[\psi]||_{L^1(\Omega \times \Omega)}.
\]

Here and in the sequel \(S\) denotes the \(N-1\) dimensional surface area on a sphere with radius \(r\). Having established that \([\gamma \psi] \in L^1(\Omega \times \Omega)\), we use Fubini's theorem and polar coordinates for real valued integrable functions to obtain

\[
\int_{\Omega \times \Omega} \gamma(x, y) \psi(x, y) dxdy = \int_{\Omega} \int_0^d \sigma(r) \int_{\partial B(x,r) \cap \Omega} \psi(x, y) dS(y) dr dx.
\]

Note that the right-hand side of this equation is independent of the particular representative of \([\sigma]\) used to define \([\gamma]\). Thus, since \([\psi] \in L^1(\Omega \times \Omega)\) was arbitrary, by the fundamental lemma of calculus of variations, the definition of \([\gamma]\) is independent of the particular representative of \([\sigma]\) chosen to define \([\gamma]\). This implies that \([\gamma]\) is well-defined by (11) as an equivalence class of functions on \(\Omega \times \Omega\). It remains to show (12). To do this, we argue as follows: If (12) does not hold, then there exists \(\varepsilon > 0\) and \(A \subset \Omega \times \Omega\) with \(|A| > 0\) such that for all \(\gamma \in [\gamma]\)

\[
|\gamma(x, y)| \geq ||[\sigma]||_{L^\infty((0,d))} + \varepsilon \quad \text{a.e. on } A,
\]

where \(|A|\) denotes the \(2N\)-dimensional Lebesgue measure of \(A\). It follows that

\[
(||[\sigma]||_{L^\infty((0,d))} + \varepsilon)|A|
\]

\[
\leq \int_{\Omega \times \Omega} |\gamma(x, y)| \chi_A(x, y) dxdy = \int_{\Omega} \int_0^d |\sigma(r)| \int_{\partial B(x,r) \cap \Omega} \chi_A(x, y) dS(y) dr dx
\]

\[
\leq ||[\sigma]||_{L^\infty((0,d))} \int_{\Omega} \int_0^d \int_{\partial B(x,r) \cap \Omega} \chi_A(x, y) dS(y) dr dx \leq ||[\sigma]||_{L^\infty((0,d))}|A|.
\]

However, this can only be true if \(|A| = 0\), which contradicts our assumption. □
Remark 4. Note that in the proof of Proposition 1 we use a representation in polar coordinates. Representation in polar coordinates is often seen as a special case of the coarea formula. Unfortunately, the coarea formula as given e.g. in [19] cannot be directly applied to \( \gamma \psi \), since it has a requirement that \( \gamma \psi \in L^1(\Omega \times \Omega) \), but this is exactly what we are proving in the first part of the proof.

It follows from similar arguments as in Proposition 1, that \( \gamma \in L^\infty(\Omega \times \Omega) \) defined as in (11) satisfies Assumption 2.1 for all \( \sigma \in W_{ad} \). Thus, using Corollary 1, there exist constants \( m, M > 0 \) such that for all \( \sigma \in W_{ad} \) it holds that

\[
m\|u\|_{H^r(\Omega)} \leq \|u\|_{V^r(\Omega)} \leq M\|u\|_{H^r(\Omega)}.
\]

Lemma 2.6. The mapping

\[
\Phi: L^\infty((0, d)) \to L^\infty(\Omega \times \Omega)
\]

\[
\sigma \mapsto \gamma, \text{ such that } \gamma(x, y) = \sigma(|y - x|) \text{ a.e. on } \Omega \times \Omega,
\]

is well-defined, linear, continuous, and sequentially weak*-to-weak* continuous.

Proof. We have already seen in Proposition 1 that \( \Phi \) is well-defined and continuous. Clearly, \( \Phi \) is also linear. It remains to show weak*-to-weak* sequential continuity. To do this, we let \( (\sigma^n) \) be a sequence in \( L^\infty((0, d)) \) and \( \sigma \in L^\infty((0, d)) \) be such that

\[
\sigma^n \rightharpoonup^* \sigma \text{ in } L^\infty((0, d)).
\]

Let \( \gamma \in L^\infty(\Omega \times \Omega) \) and the sequence \( (\gamma^n) \) in \( L^\infty(\Omega \times \Omega) \) be defined by \( \gamma(x, y) := \sigma(|x - y|) \) and \( \gamma^n(x, y) := \sigma^n(|x - y|) \), respectively, for almost all \( (x, y) \in \Omega \times \Omega \) and all \( n \in \mathbb{N} \). We must show that \( \gamma^n \rightharpoonup^* \gamma \) in \( L^\infty(\Omega \times \Omega) \). For this purpose, let \( \psi \in L^1(\Omega \times \Omega) \) be arbitrary. Using Proposition 1 and Fubini’s theorem, the integrals

\[
\int_{\Omega} |\gamma^n(x, y)||\psi(x, y)| \, dy \quad \text{and} \quad \int_{\Omega} |\psi(x, y)| \, dy
\]

exist and are finite for almost all \( x \in \Omega \) for all \( n \in \mathbb{N} \). Employing polar coordinates and Fubini’s theorem, we obtain that for almost all \( x \in \Omega \)

\[
\int_{\Omega} \gamma^n(x, y)\psi(x, y) \, dy = \int_{0}^{d}\int_{\partial B(x, s) \cap \Omega} \gamma^n(x, y)\psi(x, y)\, dS(y) \, ds
\]

\[
= \int_{0}^{d} \sigma^n(s) \int_{\partial B(x, s) \cap \Omega} \psi(x, y)\, dS(y) \, ds,
\]

and that the map \( s \mapsto \int_{\partial B(x, s) \cap \Omega} \psi(x, y)\, dS(y) \) is in \( L^1((0, d)) \). Using that \( \sigma^n \rightharpoonup^* \sigma \) in \( L^\infty((0, d)) \), we deduce that

\[
\lim_{n \to \infty} \int_{\Omega} \gamma^n(x, y)\psi(x, y) \, dy \, dx = \int_{\Omega} \gamma(x, y)\psi(x, y) \, dy \, dx
\]

(13)

for almost all \( x \in \Omega \). We now define

\[
f^n(x) := \int_{\Omega} \gamma^n(x, y)\psi(x, y) \, dy \quad \text{and} \quad f(x) := \int_{\Omega} \gamma(x, y)\psi(x, y) \, dy.
\]

Equation (13) shows that \( f^n(x) \to f(x) \) a.e. on \( \Omega \). Since moreover

\[
|f^n(x)| \leq \|\sigma^n\|_{L^\infty((0, d))} \int_{\Omega} |\psi(x, y)| \, dx \, dy \leq C \int_{\Omega} |\psi(x, y)| \, dx \, dy
\]
and \( x \mapsto \int_{\Omega} |\psi(x, y)| \, dy \) is integrable, Lebesgue’s dominated convergence theorem asserts that

\[
\lim_{n \to \infty} \int_{\Omega \times \Omega} \gamma^n(x, y) \psi(x, y) \, dx \, dy = \int_{\Omega \times \Omega} \gamma(x, y) \psi(x, y) \, dx \, dy.
\]

Since \( \psi \in L^1(\Omega \times \Omega) \) was chosen arbitrarily, this is precisely what we needed to show.

**Remark 5.** It can be of interest to take a note of the interpretation of nonlocal energy seminorms for \( \Omega = \mathbb{R} \) using Fourier analysis. Here, \( \mathcal{F} : L^2(\mathbb{R}) \to L^2(\mathbb{R}) \) denotes the usual Fourier transform, see e.g. [5, Section 4.12]. Similarly as in [17, Proposition 3.4], for \( u \in C_c(\mathbb{R}) \) we compute

\[
|u|_{\sigma,s}^2 = \int_{\mathbb{R}} \mathcal{F} \left( \frac{|u(z + \cdot) - u(\cdot)|}{|z|^{1/2 + s}} \sigma(|z|)^{1/2} \right) \| \mathcal{F}(u(\xi)) \|^2_{L^2(\mathbb{R})} \, dz
\]

\[
= \int_{\mathbb{R}} \left( \int_{\mathbb{R}} \left| e^{ivz} - 1 \right|^2 \sigma(|v|) |\mathcal{F}(u(\xi))|^2 \, d\xi \right) \, dv
\]

\[
= \int_{\mathbb{R}} \left( \int_{\mathbb{R}} \left| e^{ivz} - 1 \right|^2 \sigma(v/|\xi|) |\xi|^{2s} |\mathcal{F}(u(\xi))|^2 \, d\xi \right) \, dv.
\]

Here we use Plancherel’s theorem (see [5, Satz 4.16]) and the fact that translation in the time domain corresponds to modulation in the frequency domain (see [5, Lemma 4.5]) for the first and second equality, respectively. If we let

\[
\rho(\xi) := \int_{\mathbb{R}} \left| e^{ivz} - 1 \right|^2 \sigma(v/|\xi|) \, dv,
\]

then it follows that

\[
|u|_{\sigma,s}^2 = \int_{\mathbb{R}} \rho(\xi) |\xi|^{2s} |\mathcal{F}(u(\xi))|^2 \, d\xi.
\]

This shows that nonlocal energy seminorms behave similar as the fractional order Sobolev-Slobodeckij seminorm (see [17]). The only difference is an additional weighting term \( \rho \) depending on the frequency.

### 3. Linear case.

#### 3.1. Problem setting.

For \( \sigma \in W_{ad} \), we consider the lower level problems

\[
\min_{u \in H^s(\Omega)} \| Su - y_0 \|^2_{L^2(\Omega)} + |u|_{\sigma,s}^2, \quad (P(\sigma))
\]

where \( S \in \mathcal{L}(L^2(\Omega)) \), and \( y_0 \in L^2(\Omega) \) is a noisy measurement of the ground truth state. The solution set of the lower level problems is denoted by

\[
\mathcal{F} := \{ (\sigma, u) \in W_{ad} \times H^s(\Omega) \mid u \text{ solves } (P(\sigma)) \}.
\]

We address the following learning problem

\[
\min_{\sigma \in W_{ad}, u \in H^s(\Omega)} \frac{1}{2} \| u - u^\dagger \|^2_{L^2(\Omega)} + R(\sigma) \quad \text{subject to} \quad (\sigma, u) \in \mathcal{F}, \quad (BP)
\]

where \( u^\dagger \in L^2(\Omega) \) is the ground truth control and \( R : L^\infty((0,d)) \to [0, \infty) \) is a given regularization operator. We emphasize that when showing existence of solutions we specifically include the case \( R \equiv 0 \), i.e. we do not require additional regularization of the weights.
Example 3.1. As an example let $S$ be the solution operator to an elliptic PDE with Neumann boundary conditions. More precisely, for $(y, u) \in L^2(\Omega) \times L^2(\Omega)$ let

$$Su = y \quad \text{if and only if} \quad -\rho \Delta y + y = u \quad \text{and} \quad \frac{\partial y}{\partial n} = 0 \text{ on } \partial \Omega,$$

where $\rho$ is a positive constant. Clearly, $\text{rg } (S) \subset H^1(\Omega)$. As a straightforward consequence of the Lax-Milgram lemma and the standard Sobolev embedding, it follows that $S$ is a compact operator on $L^2(\Omega)$. Moreover, it is easy to see that $S$ is injective.

3.2. Preliminaries.

Proposition 2 (Uniform convexity). Let $S$ be injective on $\Pi_0(\Omega)$. Then there exist $c, C > 0$ such that for every $\sigma \in W_{ad}$ and every $u \in H^s(\Omega)$ we have

$$c ||u||^2_{H^s(\Omega)} \leq ||Su||^2_{L^2(\Omega)} + ||u||^2_{\sigma, s} \leq C ||u||^2_{H^s(\Omega)}.$$  \hspace{1cm} (14)

Proof. The second inequality follows from the continuity of $S$ on $L^2(\Omega)$ and Corollary 1. To prove the first inequality, first note that for all $\sigma \in W_{ad}$ it holds that

$$||u||^2_{\sigma, s} \leq ||u||_{\sigma, s} \quad \text{for all } u \in H^s(\Omega),$$

where $\sigma_{\text{min}} := \gamma_1 \chi_{(0, \delta)}$. Thus, it suffices to prove the first inequality for $\sigma = \sigma_{\text{min}}$. We begin by showing that there exists $c_1 > 0$ such that

$$c_1 ||u||^2_{L^2(\Omega)} \leq ||Su||^2_{L^2(\Omega)} + ||u||^2_{\sigma_{\text{min}}, s} \quad \text{for all } u \in H^s(\Omega).$$  \hspace{1cm} (15)

To do this, we argue by contradiction. If there is no $c_1 > 0$ such that (15) holds, then there exists a sequence $(u^n)$ in $H^s(\Omega)$ such that

$$||u^n||_{L^2(\Omega)} = 1, \quad \text{and} \quad ||Su^n||^2_{L^2(\Omega)} + ||u^n||^2_{\sigma_{\text{min}}, s} \leq 1/n \quad \text{for all } n \in \mathbb{N}.$$ 

Using Corollary 1, we deduce that $(u^n)$ is bounded in $H^s(\Omega)$. Since $H^s(\Omega)$ is reflexive, it follows that $(u^n)$ has an accumulation point with respect to the weak topology on $H^s(\Omega)$. Moreover, $H^s(\Omega)$ is compactly embedded in $L^2(\Omega)$, and consequently it follows that there is an accumulation point of $(u^n)$ with respect to the strong topology on $L^2(\Omega)$. Standard arguments show that

$$||u||_{L^2(\Omega)} = 1, \quad u \in \ker \cdot |\sigma_{\text{min}}, s = \Pi_0(\Omega) \quad \text{and} \quad Su = 0,$$

This contradicts the assumption that $S$ is injective on $\Pi_0(\Omega)$. Hence we have proven that there is $c_1 > 0$ such that (15) holds. Since by Theorem 2.2 we already know that

$$||u^n||^2_{H^s(\Omega)} \leq c_1 ||u^n||^2_{\sigma_{\text{min}}, s} + 4|\Omega|\delta^{-N-2s} ||u||^2_{L^2(\Omega)},$$

the claim follows by straightforward computations. \hfill $\square$

Proposition 3. Assume that $S$ is injective on $\Pi_0(\Omega)$. Then $(P(\sigma))$ has a unique solution.

Proof. Since $(P(\sigma))$ is a convex minimization problem, it follows that $u^* \in H^s(\Omega)$ solves $(P(\sigma))$ if and only if $u^*$ satisfies the first order optimality condition

$$\langle Su^*, S v \rangle_{L^2(\Omega)} + \langle u^*, v \rangle_{\sigma, s} = \langle y_\delta, Sv \rangle_{L^2(\Omega)} \quad \text{for all } v \in H^s(\Omega).$$  \hspace{1cm} (16)

Existence and uniqueness of solutions to (16) can be easily proven using the Lax-Milgram lemma (the required coercivity is a direct consequence of Proposition 2). \hfill $\square$
Optimality conditions for the lower level problem. To simplify notation, we define
\[ L^*(\sigma) : H^s(\Omega) \to H^s(\Omega)' \]
by
\[ |L^*(\sigma)u|v = \int_{\Omega \times \Omega} \frac{(u(x) - u(y))(v(x) - v(y))}{|x-y|^{n+2s}} \sigma(|x-y|) \, dx \, dy \]
for \((u, v) \in H^s(\Omega) \times H^s(\Omega)\).

Note that \(L^*(\sigma)\) is a bounded operator from \(H^s(\Omega)\) to \(H^s(\Omega)'\). Since \((P(\sigma))\) is convex, we obtain the following necessary and sufficient optimality condition for \((P(\sigma))\).

**Proposition 4.** An element \(u \in H^s(\Omega)\) solves \((P(\sigma))\) if and only if
\[ S^* Su - S^* y_b + L^*(\sigma)u = 0 \quad \text{in} \quad H^s(\Omega)' \]

**Remark 6.** An optimality system for \((P(\sigma))\) with \(S\) as in Example 3.1 can be obtained by standard Lagrangian methods. Indeed, let \(u \in H^s(\Omega)\) be a solution to \((P(\sigma))\) with \(S\) as in Example 3.1. Then there exists \(p \in H^1_0(\Omega)\) such that
\[ |L^*(\sigma)u|w + \int_\Omega p(x)w(x) \, dx = 0, \quad \text{for all} \quad w \in H^s(\Omega), \quad \text{(optimality)} \]
\[ \int_\Omega [y(x) - y_b(x)]v(x) + \nabla p(x) \cdot \nabla v(x) \, dx = 0, \quad \text{for all} \quad v \in H^1_0(\Omega), \quad \text{(adjoint eq.)} \]
\[ \int_\Omega \nabla y(x) \cdot \nabla z(x) - u(x)z(x) \, dx = 0, \quad \text{for all} \quad z \in H^1_0(\Omega). \quad \text{(state eq.)} \]

### 3.3. Existence of solutions
To prove that \((BP)\) has a solution, we apply the direct method of the calculus of variations. The crucial step in the proof is the argument proving that the feasible set is sequentially closed with respect to weak* convergence. Since the feasible set is defined by the lower level problem, this is related to stability of the lower level problem with respect to the weight function.

**Proposition 5.** Assume that \(S\) is injective on \(\Pi_0(\Omega)\) and that \(R\) is weak* sequentially lower semicontinuous. Then \((BP)\) has a solution.

**Proof.** First of all note that as a consequence of Proposition 3 the feasible set \(\mathcal{F}\) is nonempty. Thus, we can take a minimizing sequence for \((BP)\), i.e. a sequence \((\sigma^n, u^n) \in \mathcal{F}\) such that
\[ \lim_{n \to \infty} \frac{1}{2} \|u^n - u^\dagger\|^2_{L^2(\Omega)} + R(\sigma^n) = \inf_{(u, \sigma) \in \mathcal{F}_{ad}} \frac{1}{2} \|u - u^\dagger\|^2_{L^2(\Omega)} + R(\sigma). \]
It is easy to prove that \(W_{ad}\) is sequentially weak* compact. Consequently, \((\sigma^n)\) has a subsequence, again denoted by \((\sigma^n)\), such that \(\sigma^n \rightharpoonup^* \sigma^*\) in \(L^\infty((0, d)\)) for some \(\sigma \in W_{ad}\). Theorem 2.6 ensures that in this case \(\sigma^n(x, y) := \sigma^n(|x - y|)\) converges to \(\sigma^*(x, y) := \sigma^*\) in the weak* topology on \(L^\infty(\Omega \times \Omega)\). Using that
\[ \|Su^n - y_b\|^2_{L^2(\Omega)} + |u^n|_{\sigma^n, s}^2 \leq \|y_b\|^2_{L^2(\Omega)} \quad \text{for all} \quad n \in \mathbb{N}, \]
it follows from Proposition 2 that \((w^n)\) is bounded in \(H^s(\Omega)\). Since \(H^s(\Omega)\) is reflexive, this implies that \((w^n)\) has a subsequence, which we again denote by \((w^n)\), such that \(w^n \rightharpoonup u^*\) in \(H^s(\Omega)\) for some \(u^* \in H^s(\Omega)\). Note that if we set
\[ p^n := S^* Su^n - S^* y_b \quad \text{for all} \quad n \in \mathbb{N}, \]
then \((p^n)\) is a bounded sequence in \(L^2(\Omega)\), and thus it has a weakly converging subsequence, which we again denote by \((p^n)\), such that \(p^n \rightharpoonup p\) in \(L^2(\Omega)\). As a
consequence of Theorem 2.5 we now obtain $|u^*_n|^2_{\sigma^*_n,s} = \lim_{n \to \infty} |u^n|^2_{\sigma^n,s}$. From this, and using that $u^n$ solves the lower level problem $(P(\sigma))$, with $\sigma^n$ in place of $\sigma$, to justify the second inequality below, we deduce that

$$
\|Su^* - y_\delta\|^2_{L^2(\Omega)} + |u|^2_{\sigma^*_n,s} \leq \liminf_{n \to \infty} \|Su^n - y_\delta\|^2_{L^2(\Omega)} + |u^n|^2_{\sigma^n,s}
$$

$$
\leq \lim_{n \to \infty} \|Su - y_\delta\|^2_{L^2(\Omega)} + |u|^2_{\sigma^*_n,s} = \|Su - y_\delta\|^2_{L^2(\Omega)} + |u|^2_{\sigma^*_n,s}
$$

for all $u \in H^s(\Omega)$. Note that the last equality follows from the weak* convergence of $\sigma^*((x - y))$. This shows that $(\sigma^*, u^*) \in F$. Due to the weak* lower semi continuity of the involved functions we have

$$\frac{1}{2} \|u^* - u^\dagger\|^2_{L^2(\Omega)} + R(\sigma^*) \leq \liminf_{n \to \infty} \frac{1}{2} \|u^n - u^\dagger\|^2_{L^2(\Omega)} + R(\sigma^n).$$

Since $(\sigma^n, y^n, u^n)$ was chosen as a minimizing sequence of (BP), and $(\sigma^*, u^*) \in F$, this implies that

$$\frac{1}{2} \|u^* - u^\dagger\|^2_{L^2(\Omega)} + R(\sigma^*) = \inf_{(u,\sigma) \in F_{ad}} \frac{1}{2} \|u - u^\dagger\|^2_{L^2(\Omega)} + R(\sigma)$$

which shows that $(\sigma^*, u^*)$ is a solution to (BP).

Concerning the assumption on the weak* sequential lower semicontinuity of $R$ on $L^\infty((0, d))$ we observe that it is satisfied by all functions which are convex and lower semicontinuous on $L^p(\Omega)$ for some $1 \leq p < \infty$ (to verify this recall that $L^p((0, d))$ with $q = p/(p-1)$ is continuously embedded in $L^1((0, d))$ and apply [5, Korollar 6.28]).

3.4. **Optimality conditions.** In the following we derive optimality conditions for the bilevel problem. It is useful to recall that the optimality conditions in Proposition 4 are necessary and sufficient for $u \in H^s(\Omega)$ to be a solution to the lower level problem. Hence, we can simply replace the lower level problem constraint in (BP) by the equality constraint in Proposition 4 and then use standard techniques for constrained optimization problem to obtain optimality conditions for (BP). The Lagrange function $L: W_{ad} \times H^s(\Omega) \times H^s(\Omega) \to \mathbb{R}$ of the bilevel problem (with the lower level problem replaced by the corresponding optimality condition) is given by

$$(\sigma, u, q) \mapsto \frac{1}{2} \|u - u^\dagger\|^2_{L^2(\Omega)} + R(\sigma) + [L^*(\sigma)u]q + \langle Su - y_\delta, Sq \rangle_{L^2(\Omega)}.$$

**Proposition 6** (optimality system). If $(u, \sigma) \in F_{ad}$ is a solution to (BP), then there exists $q \in H^s(\Omega)$ such that

$$R'(\sigma)[w - \sigma] + [L^*(\omega - \sigma)u]q \geq 0 \quad \text{for all } \omega \in W_{ad},$$

$$u - u^\dagger + L^*(\sigma)q + S^*Sq = 0 \quad \text{in } H^s(\Omega)^\prime,$$

$$S^*Su - S^*y_\delta + L^*(\sigma)u = 0 \quad \text{in } H^s(\Omega)^\prime.$$ (constraint)

**Proof.** Since as a consequence of Proposition 2 the lower level problem is uniformly convex for all $\sigma \in W_{ad}$, this follows using the standard Lagrangian based approach. \qed
4. Nonlinear case. In this section we study how nonlocal regularization operators can be learned for the inverse problem of determining parameters in partial differential equations. In this context, the forward operator is often only defined implicitly as the solution operator to a PDE which depends on the sought-after parameters. We let the function describing the PDE be denoted by \( e: Y \times H^s(\Omega) \to Z \). Here \( Z \) is a general Hilbert space, and \( Y \) is a Hilbert space that is continuously embedded in \( L^2(\Omega) \). If for every \( u \in H^s(\Omega) \) there exists a unique \( y \in Y \) such that \( e(y, u) = 0 \), then the corresponding forward operator \( S: H^s(\Omega) \to Y \) is such that \( u \mapsto y \), where \( y \) satisfies \( e(y, u) = 0 \). In the following we prefer to state our hypotheses directly in terms of the function \( e \) in order to facilitate the use of the presented results in practice. A detailed discussion of a particular choice of \( e \), for which no constraints are required to obtain a well-posed forward problem, is provided in Section 4.3. The lower level problems are given by

\[
\min_{(y, u) \in Y \times H^s(\Omega)} \| y - y_\delta \|_{L^2(\Omega)}^2 + |u|_{\sigma,s}^2 \quad \text{s.t.} \quad e(y, u) = 0. \quad (P(\sigma))
\]

Here, \( y_\delta \in L^2(\Omega) \) is a noisy measurement of the ground truth state. We let

\[
F_{ad} := \{(y, u) \in Y \times H^s(\Omega) \mid e(y, u) = 0 \}
\]

denote the feasible set of the lower level problem. Moreover,

\[
\mathcal{F} := \{(\sigma, y, u) \in L^\infty((0, d)) \times Y \times H^s(\Omega) \mid \sigma \in W_{ad} \text{ and } (y, u) \text{ solves } (P(\sigma))\}
\]

denotes the solution set of the lower level problems. The learning problem is

\[
\min_{(\sigma, y, u) \in W_{ad} \times Y \times H^s(\Omega)} \frac{1}{2} \| u - u^l \|_{L^2(\Omega)}^2 + R(\sigma) \quad \text{subject to } (\sigma, y, u) \in \mathcal{F}. \quad (BP)
\]

As in the linear case, \( u^l \in L^2(\Omega) \) is the ground truth parameter and \( R: L^\infty((0, d)) \to [0, \infty) \) represents an additional regularization operator for the weight.

4.1. Existence of solutions.

**Definition 4.1 (Stability).** We say that \( \{P(\sigma)\}_{\sigma \in W_{ad}} \) is stable (resp. weak*-to-weak stable) if the following holds: \( P(\sigma) \) has a solution for every \( \sigma \in W_{ad} \), and for every sequence \( (\sigma^n) \) in \( W_{ad} \) such that \( \sigma^n \to \sigma \in L^\infty((0, d)) \) (resp. \( \sigma^n \to^* \sigma \) in \( L^\infty((0, d)) \)) for some \( \sigma \in W_{ad} \), it follows that every sequence of corresponding solutions to \( P(\sigma) \) for \( \sigma = \sigma^n \) has a strong (resp. weak) accumulation point, and every such accumulation point is a solution to \( P(\sigma) \).

**Theorem 4.2 (Existence of solutions).** Assume that \( P(\sigma) \) is weak*-to-weak stable and \( R \) is weak* sequentially lower semicontinuous. Then (BP) has a solution.

**Proof.** Since \( W_{ad} \) is clearly weak* sequentially compact, weak*-to-weak stability of \( P(\sigma) \) implies that \( \mathcal{F} \) is weak*-weak sequentially compact. The claim now follows from the well-known fact that a function which is sequentially lower semicontinuous with respect to some topology, attains a minimum on a nonempty set which is sequentially compact with respect to the same topology.

4.2. Optimality conditions. We derive necessary optimality conditions for the learning problem (BP). Here we essentially follow the discussion provided in [25, Section 5]. Throughout this section it is assumed that the function \( e \) describing the state constraint is well defined and at least once continuously F-differentiable on \( Y \times H^s(\Omega) \). We begin by recalling the Karush-Kuhn-Tucker conditions of the lower level problem.
Definition 4.3 (Karush-Kuhn-Tucker conditions). We say that \((y, u, \lambda) \in Y \times H^s(\Omega) \times Z'\) satisfies the Karush-Kuhn-Tucker (KKT) conditions of \((P(\sigma))\) if

\[
\begin{align*}
\mathcal{L}^s(\sigma)u + \lambda D_{\sigma}e(y, u) &= 0, \quad (17a) \\
\langle y - y_\delta, w \rangle_{L^2(\Omega)} + \langle \lambda D_{\sigma}e(y, u), w \rangle &= 0, \quad \text{for all } w \in Y, \quad (17b) \\
e(y, u) &= 0, \quad \text{in } Z. \quad (17c)
\end{align*}
\]

The KKT conditions constitute a system of first order necessary optimality conditions provided a suitable regularity assumption is met. More precisely, if \((y, u) \in Y \times H^1(\Omega)\) solves \((P(\sigma))\) and \(D_{\sigma}e(y, u) \in \mathcal{L}(Y, Z)\) is bijective, then there exists a unique \(\lambda \in Z'\) such that \((y, u, \lambda)\) satisfies the KKT conditions of \((P(\sigma))\). Here existence of \(\lambda\) follows from [31, Theorem 3.1] and proving uniqueness is straightforward. Next, we recall the Lagrange function \(L: W_{ad} \times Y \times H^s(\Omega) \times Z' \to \mathbb{R}\) of the lower level problem given by

\[
(y, u, \lambda) \mapsto \|y - y_\delta\|^2_{L^2(\Omega)} + |u|^2_{\sigma,\delta} + \lambda e(y, u)
\]

for \((y, u, \lambda) \in W_{ad} \times Y \times H^s(\Omega) \times Z'\), which enables us to write second order sufficient optimality conditions in a compact form. Second order sufficient optimality conditions have many important practical implications, see e.g. [9] and the references given therein. In particular, they are closely related to stability properties of the solution mapping, see [26, Chapter 2]. It is thus of no surprise that second order sufficient conditions of the lower level problem are important for the derivation of optimality conditions for the learning problem. From now on, we assume that \(e\) is at least twice continuously \(F\)-differentiable in an open neighbourhood of \(Y \times H^s(\Omega)\).

Definition 4.4 (second order sufficient optimality condition). We say that a point \((y, u) \in Y \times H^s(\Omega)\) satisfies the second order sufficient optimality condition of \((P(\sigma))\) if \(D_{\sigma}e(y, u) \in \mathcal{L}(Y, Z)\) is bijective and there exist \(\lambda \in Z'\) and \(\mu > 0\) such that \((y, u, \lambda)\) satisfies the KKT conditions, and

\[
D_{y, u}^2 L((\delta_y, \delta_u, (\Delta_y, \Delta_u)) \geq \mu \|(\delta_y, \delta_u)\|^2_{H^1(\Omega) \times H^s(\Omega)}
\]

for all \((\delta_y, \delta_u) \in \ker De(y, u)\).

The constraint that feasible points must be solutions to a lower level problem prevents the direct use of Lagrangian based approaches for obtaining optimality conditions. To overcome this issue, at least to some extend, one usually considers the KKT reformulation of bilevel optimization problem, in which the lower level problem is replaced by its KKT conditions, see e.g. [16, Section 5.5]. The KKT reformulation of the learning problem is given by

\[
\begin{align*}
\min_{\sigma} & \frac{1}{2} \|u - u\|^2_{L^2(\Omega)} + R(\sigma) \\
\text{subject to } & (\sigma, y, u, \lambda) \in W_{ad} \times Y \times H^s(\Omega) \times Z' \text{ satisfies (17a)-(17c)}.
\end{align*}
\]

In general, the constraints of \((BP^*)\) are easier to handle than the constraints of the original problem \((BP)\). For example, since there are no control constraints in the lower level problem, the constraints of the KKT reformulated problem consist only of equality and convex constraints. In order to use the KKT reformulation \((BP^*)\) to obtain optimality conditions for the learning problem \((BP)\), the relation between both problems needs to be investigated. Clearly, if the lower level problem is convex for every weight \(\sigma \in W_{ad}\), then both problems are equivalent. In general, this is not the case since points satisfying the KKT conditions of the lower level problem need not be solutions to the lower level problem. Note, however, that we are
only interested in (BP) to obtain optimality conditions for the learning problem. Consequently, for our purposes it is sufficient to know under which conditions a solution to the learning problem is guaranteed to be a local solution to (BP). This question is investigated in the following theorem. The theorem can be proven using the same arguments as in [25, Theorem 5.1] (replacing α by σ and \(W_{ad}\)), however, since in contrast to [25, Theorem 5.1] we do not include control constraints here, we give a simpler proof than the one of [25, Theorem 5.1].

**Theorem 4.5.** Let \((\sigma^*, y^*, u^*)\) be a solution to (BP) and assume that the following statements hold:

(A1) \(\{(P(\sigma))\}_{\sigma \in W_{ad}}\) is stable with respect to the weights,
(A2) \((y^*, u^*)\) satisfies the second order sufficient optimality condition of \((P(\sigma))\) for \(\sigma = \sigma^*\),
(A3) \((y^*, u^*)\) is the unique solution to \((P(\sigma))\) for \(\sigma = \sigma^*\).

Then there is a unique \(\lambda^* \in Y'\) such that \((\sigma^*, y^*, u^*, \lambda^*)\) is a local solution to (BP).

**Proof.** From the discussion following Theorem 4.3 it is clear that there exists a unique \(\lambda^* \in Y'\) such that \((\sigma^*, y^*, u^*, \lambda^*)\) satisfies (17a)-(17c). The proof is divided into three steps.

**Step 1.** Define \(F(\sigma, y, u) := \|y - y_0\|^2_{L^2(\Omega)} + |u|^2_{H^s} \). It is straightforward to verify that \(F\) is infinitely many times F-differentiable on \(L^\infty((0, d)) \times Y \times H^s(\Omega)\) (since \(F\) is a continuous polynomial function in its variables). Assumption (A2) implies that for every \((y^*, u^*, \sigma\) \(Y' \times Z' \times Z\), the quadratic problem

\[
\begin{align*}
\min_{(\delta_y, \delta_u) \in Y \times H^s(\Omega)} & \quad D^2_{(y, u)} L(\sigma, y^*, u^*, \lambda^*)[(\delta_y, \delta_u), (\delta_y, \delta_u)] - y'(\delta_y) - u'(\delta_u) \\
\text{subject to} & \quad Dc(\sigma, y, u)(\delta_y, \delta_u) = z
\end{align*}
\]

has a unique solution \((\delta_y^*, \delta_u^*, \delta_\lambda^*)\), which is uniquely characterized by

\[
\begin{pmatrix}
F_{yy} + \sigma e_{yy} & F_{yu} + \sigma e_{yu} & e_y^* \\
F_{yu} + \sigma e_{yu} & F_{uu} + \sigma e_{uu} & e_u^*
\end{pmatrix}
\begin{pmatrix}
\delta_y^* \\
\delta_u^*
\end{pmatrix}
= \begin{pmatrix}
y' \\
u'
\end{pmatrix}.
\]

Here, we write \(F_{yy} = F_{yy}(\sigma^*, y^*, u^*)\), \(e_{yy} = e_{yy}(y^*, u^*)\) et cetera. This proves that the mapping

\[
(\delta_y, \delta_u, \delta_\lambda) \rightarrow \begin{pmatrix}
F_{yy} + \sigma e_{yy} & F_{yu} + \sigma e_{yu} & e_y^* \\
F_{yu} + \sigma e_{yu} & F_{uu} + \sigma e_{uu} & e_u^*
\end{pmatrix}
\begin{pmatrix}
\delta_y \\
\delta_u \\
\delta_\lambda
\end{pmatrix}
\]

is bijective as a mapping from from \(Y \times U \times Z'\) to \(Y' \times U' \times Z\). Since this mapping coincides with the derivative of the function describing the KKT conditions with respect to \((y, u, \lambda)\), it follows by the implicit function theorem [7, Proposition 4.7.1] that there exist neighborhoods \(V(\sigma^*)\) of \(\sigma^*\) and \(V(y^*, u^*, \lambda^*)\) of \((y^*, u^*, \lambda^*)\) and a continuously F-differentiable function \(\phi: V(\sigma^*) \rightarrow V(y^*, u^*, \lambda^*)\) such that for all \((y, u, \lambda) \in V(y^*, u^*, \lambda^*)\) we have

\[(y, u, \lambda) \text{ is a KKT point of } (P(\sigma)) \text{ if and only if } \phi(\sigma) = (y, u, \lambda).
\]

**Step 2.** We now claim that there exists an open subset \(U(\sigma^*)\) of \(V(\sigma^*)\) which contains \(\sigma^*\) such that for all \(\sigma \in U(\sigma^*) \cap W_{ad}\) and \((y(\sigma), u(\sigma), \lambda(\sigma)) = \phi(\sigma)\) the point \((y(\sigma), u(\sigma))\) solves the lower level problem \((P(\sigma))\). The proof is by contradiction. Indeed, if our claim was wrong, then there would be a sequence \((\sigma^n)\) in \(V(\sigma^*) \cap W_{ad}\) converging to \(\sigma^*\) in \(L^\infty((0, d))\) and a sequence \((y^n, u^n, \lambda^n)\) in \(Y \times U \times Z\) which
does not intersect $V(y^*, u^*, \lambda^*)$ such that $(y^n, u^n)$ solves the lower level problem $P(\sigma)$ with $\sigma = \sigma^n$ and $\lambda^n$ satisfies (17b) if $D_\sigma e(y^n, u^n)$ is bijective, and is chosen arbitrarily otherwise. Using (A1) and 4.5 it follows that $(y^n, u^n)$ must converge strongly in $H^1(\Omega) \times H^s(\Omega)$ to $(y, u)$. This also yields that $\lambda^n \to \lambda$ in $Z^s$, since the set $\{(y, u) \in Y \times H^s(\Omega) \mid D_\sigma e(y, u) \text{ is bijective}\}$ is open. This is a contradiction to the fact that we chose $(y^n, u^n, \lambda^n)$ such that it does not intersect $V(y^*, u^*, \lambda^*)$.

**Step 3.** Since we have proven that for all $(\sigma, y, u, \lambda) \in U(\sigma^*) \cap W_{ad} \times V(y^*, u^*, \lambda^*)$ the point $(y, u, \lambda)$ is a KKT point if and only if $(y, u)$ is a local solution to $P(\sigma)$, it follows that $(\sigma^*, y^*, u^*, \lambda^*)$ is a solution to (BP*) with the feasible set restricted to $U(\sigma^*) \cap W_{ad} \times V(y^*, u^*, \lambda^*)$. By definition this means that $(\sigma^*, y^*, u^*, \lambda^*)$ is a local solution to (BP*).}

Concerning assumption (A1), a sufficient condition is provided in the Appendix. As a direct consequence of the above theorem, we obtain the following: Any solution to (BP), for which the assumptions of Theorem 4.5 hold, satisfies the optimality conditions for a local solution to (BP*). Optimality conditions for (BP*) are derived in the following proposition.

**Proposition 7.** Let $(\sigma^*, y^*, u^*, \lambda^*)$ be a local solution to (BP*) with $(y^*, u^*)$ satisfying the second order sufficient optimality condition of (P(\sigma)) for $\sigma = \sigma^*$. Then there is a unique $(p^*, q^*, z^*) \in Y \times H^s(\Omega) \times Z^s$ such that

$$D_\sigma R(\sigma^*)[\sigma - \sigma^*] + [\mathcal{L}(\sigma - \sigma^*)u^*]q^* \geq 0, \quad \forall \sigma \in W_{ad},$$

(18a)

$$p^* + \lambda^* e_{uy}(y^*, u^*)p^* + \lambda^* e_{yu}(y^*, u^*)q^* + z^* e_y(y^*, u^*) = 0,$$

(18b)

$$u^* - u^1 + \lambda^* e_{uy}(y^*, u^*)p^* + \mathcal{L}(\sigma^*)q^* + \lambda^* e_{uu}(y^*, u^*)q^* + z^* e_u(y^*, u^*) = 0,$$

(18c)

$$e_y(y^*, u^*)p^* + e_u(y^*, u^*)q^* = 0.$$  

(18d)

Here we write $e_{uy}(y^*, u^*) := D_y^2 e(y^*, u^*)$, $e_{yu} := D_y e(y^*, u^*)$ and analogously for the other partial derivatives.

**Proof.** This can be proven using the same arguments as in [25, Lemma 5.1] (replacing $\alpha^*$ and the interval $[a, \bar{a}]$ by $\alpha^*$ and $W_{ad}$). For the convenience of the reader, we provide an outline of the proof. Let $E : L^\infty((0, d)) \times H^1(\Omega) \times H^s(\Omega) \times Z^s$ be the function describing the left-hand sides of (17a)-(17c). It is straightforward to verify that $E$ is continuously F-differentiable. In the proof of Theorem 4.5 in [14] we have shown that under the assumption that the second order sufficient condition holds the partial derivative $D_{(y,u)}E \in L(H^1(\Omega) \times H^s(\Omega) \times Z^s, Y' \times H^s(\Omega) \times Z)$ is bijective, which proves that the regularity condition in [26, Theorem 1.6 on p. 6] is satisfied. The exact conditions can then be obtained by straightforward computations. □

4.3. Estimation of the convection term. We consider the problem of estimating a vector valued convection term $b^i \in H^s(\Omega)^N$ in an elliptic PDE based on a noisy observation $y_N \in L^2(\Omega)$ of the ground truth state $y^l \in H^s_0(\Omega)$. The function $e : H^s_0(\Omega) \times L^q(\Omega)^N \to H^{-1}(\Omega)$ describing the PDE is in its weak form given by

$$e(y, b)w = \int_\Omega \nabla y \cdot \nabla w + b \cdot (w \nabla y) + cyw - f w \text{d}x,$$

(19)

for $(y, b, w) \in H^s_0(\Omega) \times L^q(\Omega)^N \times H^s_0(\Omega)$. Here $f \in L^2(\Omega)$ is a nonzero given source term, $c \in L^\infty(\Omega)$ is a given potential term, which is assumed to be nonnegative almost everywhere, and $1 < q < \infty$. We restrict ourselves to dimension $N \in \{1, 2, 3\}$. As we will see (Proposition 8), we need to require $q > 2$ if $N \in \{1, 2\}$ and
\(q \geq 3\) if \(N = 3\) to ensure that \(e\) is well-defined, and that the PDE \(e(y, b) = 0\) has a unique solution \(y \in H^0_\sigma(\Omega)\) for every \(b \in L^q(\Omega)^N\). Since we are interested in the case \(b \in H^s(\Omega)^N\), we often require that \(H^s(\Omega)\) is compactly embedded in \(L^q(\Omega)\) for \(q\) satisfying the above requirements. To achieve this, we frequently make the following assumption:

(B) If \(N \in \{1, 2\}\) then \(s \in (0, 1)\) and if \(N = 3\) then \(s \in (1/2, 1)\).

In particular if \((N, s)\) satisfies (B), there exist \(p, q > 1\) such that

\(1/p + 1/q = 1/2\), \(H^s(\Omega)\) is compactly embedded in \(L^q(\Omega)\), and \(H^0_\sigma(\Omega)\) is continuously embedded in \(L^p(\Omega)\).

We emphasize that making use of results from [8] we neither assume that \(\text{div } b = 0\) nor that \(b\) is small in the \(L^\infty(\Omega)^N\) norm. The lower level problem is given by

\[
\min_{(y, b) \in H^0_\sigma(\Omega) \times H^s(\Omega)^N} \|y - y_b\|_{L^2(\Omega)}^2 + |b|_{\sigma, s}^2 \text{ subject to } e(y, b) = 0. \quad (\text{P}_\text{adv}(\sigma))
\]

The definition of the nonlocal energy seminorm \(|\cdot|_{\sigma, s}\) is thereby extended to vector valued functions \(b = (b_1, \ldots, b_N)^T \in H^s(\Omega)^N\) by letting \(|b|_{\sigma, s}^2 := \sum_{i=1}^N |b_i|_{\sigma, s}^2\). The learning problem is

\[
\min_{(\sigma, y, b) \in W_{ad} \times H^0_\sigma(\Omega) \times H^s(\Omega)^N} \frac{1}{2} \|b - b^\dagger\|_{L^2(\Omega)}^2 + R(\sigma) \text{ subject to } (\sigma, y, b) \in \mathcal{F}. \quad (\text{BP}_\text{adv})
\]

We begin by verifying that \(e\) is well-defined and stable with respect to the state and convection term.

**Proposition 8.** Let \(e\) be as in (19) with \(q > 2\) if \(N \in \{1, 2\}\) and \(q \geq 3\) if \(N = 3\). Then

i) \(e\) is well-defined, infinitely many times \(F\)-differentiable, and (weak, strong)-to-weak sequentially continuous as a mapping from \(H^0_\sigma(\Omega) \times L^q(\Omega)^N\) to \(H^{-1}(\Omega)\),

ii) for every \(b \in L^q(\Omega)^N\) there is a unique \(y(b) \in H^0_\sigma(\Omega)\) such that \(e(y(b), b) = 0\),

iii) \(D_y e(y, b)\) is bijective for every \((y, b) \in H^0_\sigma(\Omega) \times L^q(\Omega)^N\),

iv) the mapping \(b \mapsto y(b)\) such that \(e(y(b), b) = 0\) is continuously \(F\)-differentiable as a mapping from \(L^q(\Omega)^N\) to \(H^0_\sigma(\Omega)\).

**Proof.**

i) Since the affine part of \(e\), which depends only on the state, is clearly well-defined, infinitely many times \(F\)-differentiable, and (weak, strong)-to-weak sequentially continuous, it remains to verify i) with \(e\) replaced by the bilinear part \(B: H^0_\sigma(\Omega) \times L^q(\Omega)^N \rightarrow H^{-1}(\Omega)\) of \(e\), which is given by

\[
B(y, b)w := \int_\Omega \sum_{i=1}^N b_i(x)w(x)\partial_i y(x) \, dx
\]

for every \((y, b, w) \in H^0_\sigma(\Omega) \times L^q(\Omega)^N \times H^0_\sigma(\Omega)\). Using classical Sobolev embeddings (see e.g. [1, Theorem 4.12]) and the assumption on \(q\), there exists \(1 < p < \infty\) with \(1/q + 1/p = 1/2\) such that \(H^0_\sigma(\Omega)\) is continuously embedded in \(L^p(\Omega)\). Applying Hölder’s inequality, we estimate

\[
|B(y, b)w| \leq \sum_{i=1}^N \|b_i\|_{L^q(\Omega)}\|\partial_i y\|_{L^2(\Omega)}\|w\|_{L^p(\Omega)} \leq C\|b\|_{L^q(\Omega)}\|y\|_{H^1(\Omega)}\|w\|_{H^1(\Omega)}.
\]
for a suitable constant $C > 0$. This proves that $B$ is well-defined and continuous. Since bilinear continuous functions are always infinitely many times F-differentiable and (weak, strong)-to-weak continuous, this concludes the proof of $iii$).

This follows from [8, Theorem 2.2].

iv) Using the first three assertions, the claim follows from the implicit function theorem (see e.g. [7, Theorem 4.7.1]).

\[ \square \]

Corollary 2. Let $e$ be as in (19) and let $(N, s)$ satisfy (B). Then

i) $e$ is infinitely many times continuously F-differentiable and (weak, weak)-to-weak sequentially continuous as a mapping from $H^s_0(\Omega) \times H^s(\Omega)^N$ to $H^{-1}(\Omega)$,

ii) for every $b \in H^s(\Omega)^N$ there is a unique $y(b) \in H^s_0(\Omega)$ such that $e(y(b), b) = 0$,

iii) $D_y e(y, b)$ is bijective for every $(y, b) \in H^s_0(\Omega) \times H^s(\Omega)^N$,

iv) the mapping $b \mapsto y(b)$ such that $e(y(b), b) = 0$ is continuously F-differentiable as a mapping from $H^s(\Omega)^N$ to $H^{-1}(\Omega)$,

v) the mapping $b \mapsto y(b)$ such that $e(y(b), b) = 0$ is Lipschitz continuous as a mapping from $H^s(\Omega)^N$ to $H^s_0(\Omega)$ on every bounded subset of $H^s(\Omega)^N$.

Proof. If $N \in \{1, 2\}$, then for all $s \in (0, 1)$ there is $q > 2$ such that $H^s(\Omega)$ is compactly embedded in $L^q(\Omega)$. If $N = 3$ and $s \in (1/2, 1)$, then there is $q \geq 3$ such that $H^s(\Omega)$ is compactly embedded in $L^q(\Omega)$. Combining this observation with Proposition 8, the first four assertions follow easily. To prove $v)$, it suffices to prove that the mapping is Lipschitz continuous on every ball $B_r(\hat{b}) := \{ b \in H^s(\Omega)^N \mid \|b - \hat{b}\|_{H^s(\Omega)^N} \leq r \}$ with radius $r > 0$ and center $\hat{b} \in H^s(\Omega)^N$. Observe that due to $iv)$ and the compactness of the embedding of $H^s(\Omega)^N$ into $L^q(\Omega)^N$ we have $M := \sup_{b \in B_r(\hat{b})} \|Dy(b)\|_{L^q(\Omega)^N, H^s(\Omega)^N}$ is finite. Moreover, it is easy to see that

$$\|Dy(b)\|_{L^q(\Omega)^N, H^s(\Omega)^N} \leq C_{s, q} \|Dy(b)\|_{L^q(\Omega)^N, H^s(\Omega)^N} \leq C_{s, q} M$$

for all $b \in B_r(\hat{b})$:

where $C_{s, q}$ denotes the embedding constant of $H^s(\Omega)$ into $L^q(\Omega)$. It now follows from [7, Theorem 3.3.2] that $b \mapsto y(b)$ is Lipschitz continuous on $B_r(\hat{b})$ (with Lipschitz constant bounded by $C_{s, q} M$).

The following technical result is needed for the existence proof in Proposition 9.

Lemma 4.6. Let $y \in H^1_0(\Omega)$ and $v \in \mathbb{R}^N \setminus \{0\}$ be such that $v \cdot \nabla y = 0$ a.e. on $\Omega$. Then $y = 0$.

Proof. Let $\tilde{y}$ denote the zero extension of $y$ to the complement of $\Omega$ in $\mathbb{R}^N$. It is well-known, see e.g. [1, Lemma 3.27 on p. 71], that $\tilde{y} \in H^1(\mathbb{R}^N)$ and

$$\nabla \tilde{y}(x) = 0 \quad \text{a.e. on } \mathbb{R}^N \setminus \Omega.$$

Consequently, we have $v \cdot \nabla \tilde{y}(x) = 0$ a.e. on $\mathbb{R}^N$. Let $(\rho^n)$ in $C^\infty(\mathbb{R}^N)$ be a sequence of mollifiers as defined in [6, p. 109]. Define the sequence $(w^n)$ by $w^n := \rho^n \ast \tilde{y}$ for every $n \in \mathbb{N}$. It follows from [6, Proposition 4.20 on p. 107 and Lemma 9.1 on p. 266] that $w^n \in C^\infty(\mathbb{R}^N)$ and

$$v \cdot \nabla w^n = \rho^n \ast (v \cdot \nabla \tilde{y}) = 0.$$

Moreover, $w^n$ has compact support, since $\rho^n$ and $\tilde{y}$ have compact support (see [6, Proposition 4.18 on p. 106]). Consequently, for arbitrary $x \in \mathbb{R}^N$ there exists $\alpha > 0$
such that \( x + \alpha v \notin \text{supp}(w^n) \). We have
\[
0 = \int_0^\alpha v \cdot \nabla w^n(x + sv) \, ds = w^n(x + \alpha v) - w^n(x) = -w^n(x).
\]
Since \( x \in \mathbb{R}^N \) was arbitrary, this proves that \( w^n \) is zero on \( \mathbb{R}^N \). Since \( (w^n) \) also converges to \( \tilde{y} \) in \( L^2(\mathbb{R}^N) \) as \( n \to \infty \) (see [6, Theorem 4.22 on p. 109]) this implies that \( y \) is zero.

**Proposition 9.** Let \((N,s) \) satisfy \((B)\). Then \((P_{\text{adv}}(\sigma))\) has a solution if and only if there exists \((y,b) \in F_{\text{adv}}\) such that
\[
\|y - y_s\|_{L^2(\Omega)}^2 + |b|_{\sigma,s}^2 \leq \|y_s\|_{L^2(\Omega)}^2.
\]

**Proof.** To prove that the existence of \((y,b) \in F_{\text{adv}}\) satisfying \((20)\) is necessary for existence of solutions, it suffices to show that
\[
\inf_{(y,b) \in F_{\text{adv}}} \|y - y_s\|_{L^2(\Omega)}^2 + |b|_{\sigma,s}^2 \leq \|y_s\|_{L^2(\Omega)}^2.
\]

To do this, let \((b^n)\) be a sequence of constant functions in \( H^s(\Omega)^N \) such that \( \|b^n\|_{H^s(\Omega)} \) diverges to \( \infty \) as \( n \to \infty \). Interpreting \((b^n)\) as a sequence of vectors in \( \mathbb{R}^N \), we have \( |b^n| \to \infty \). Using that \( \text{div} b^n = 0 \) for every \( n \in \mathbb{N} \), it is straightforward to prove that \((y^n)\) is bounded in \( H^1(\Omega) \). Let \( v^n = b^n/|b^n| \). Then there exists subsequences of \((y^n)\) and \((v^n)\), again denoted by \((y^n)\) and \((v^n)\), such that
\[
y^n \to y \quad \text{in} \quad H^1(\Omega) \quad \text{and} \quad v^n \to v \quad \text{in} \quad \mathbb{R}^N.
\]

Let \( \phi \in D(\Omega) \) be arbitrary. Testing \( e(y^n,b^n) \) with \( w^n = \phi/|b^n| \) yields
\[
\int_{\Omega} v \cdot (\phi \nabla y) \, dx = \lim_{n \to \infty} \int_{\Omega} v^n \cdot (\phi \nabla y^n) \, dx = \lim_{n \to \infty} -\frac{1}{|b^n|} \int_{\Omega} \nabla y^n \cdot \nabla \phi + cy^n \phi - f \phi \, dx = 0.
\]

Consequently, by the fundamental lemma of the calculus of variations it follows that \( v \cdot \nabla y = 0 \) a.e. on \( \Omega \). By Theorem 4.6, we deduce that \( y = 0 \). It follows that
\[
\inf_{(y,b) \in F_{\text{adv}}} \|y - y_s\|_{L^2(\Omega)}^2 + |b|_{\sigma,s}^2 \leq \lim_{n \to \infty} \|y^n - y_s\|_{L^2(\Omega)}^2 = \|y_s\|_{L^2(\Omega)}^2.
\]

where we use that for every \( n \in \mathbb{N} \) we have \( |b^n|_{\sigma,s} = 0 \) (since \( b^n \) is constant). This proves \((21)\), which in turn implies that the existence of \((y,b) \in F_{\text{adv}}\) satisfying \((20)\) is necessary for the existence of solutions to \((P_{\text{adv}}(\sigma))\).

We now prove that the existence of \((y,b) \in F_{\text{adv}}\) satisfying \((20)\) is also sufficient to guarantee existence of solutions. It follows from Corollary 2 ii) that the feasible set \( F_{\text{adv}} \) is nonempty. Consequently, we can take a minimizing sequence \((y^n,b^n)\) in \( F_{\text{adv}} \) to \((P_{\text{adv}}(\sigma))\). We divide the proof into three steps.

**Step 1.** In the first step we prove that \((y^n)\) is bounded in \( H^1(\Omega) \). Since \((y^n,b^n)\) is a minimizing sequence to \((P_{\text{adv}}(\sigma))\), using the cost functional in \((P_{\text{adv}}(\sigma))\), it can be easily derived that
\[
(|b^n|_{\sigma,s}) \quad \text{and} \quad (\|y^n\|_{L^2(\Omega)}) \quad \text{are bounded.} \tag{24}
\]

For every \( n \in \mathbb{N} \) we can write \( b^n \) as the sum of a constant function and a function with mean value zero, i.e. \( b^n = b^n_0 + b^n_1 \) in \( \Pi^0(\Omega) + \Pi^0(\Omega)^\perp \). It then follows from \((24)\), the fact that \( |b^n|_{\sigma,s} = |b^n_0|_{\sigma,s} \) for all \( n \in \mathbb{N} \), Theorem 2.4, and Corollary 1
that \((b_2^n)\) is bounded in \(H^s(\Omega)^N\). Using the chain rule and integration by parts we moreover have
\[
\int_\Omega b_1^n \cdot (y^n \nabla y^n) \, dx = \frac{1}{2} \int_\Omega b_1^n \cdot (\nabla(y^n)^2) \, dx = \frac{1}{2} \int_{\partial\Omega} y^2(b_1^n \cdot \nu) \, d\mathcal{N} - 0, \quad (25)
\]
since \(b_1^n\) is constant and \(y \in H^1_0(\Omega)\). Using \((25)\), testing \(e(y^n, b^n) = 0\) with \(y^n\) yields
\[
\int_{\Omega} |\nabla y^n|^2 \, dx = - \int_{\Omega} (b_2^n \cdot (y^n \nabla y^n) + y^2 - y f y^n) \, dx.
\]
Applying Poincaré’s and Hölder’s inequality, and using that \((y^n)\) is bounded in \(L^2(\Omega)\), we deduce that
\[
c\|y^n\|_{H^1(\Omega)}^2 \leq \int_{\Omega} |\nabla y^n|^2 \, dx = - \int_{\Omega} (b_2^n \cdot (y^n \nabla y^n) + cy^2 - y f y^n) \, dx
\leq \|b_2^n\|_{L^p(\Omega)} \|y^n\|_{L^p(\Omega)} \|y^n\|_{H^1(\Omega)} + C, \quad (26)
\]
where \(c, C > 0\) are suitably chosen constants and \(1 < p, q < \infty\) are such that \(1/p + 1/q = 1/2\), \(H^s(\Omega)\) is compactly embedded in \(L^q(\Omega)\), and \(H^1_0(\Omega)\) is compactly embedded in \(L^p(\Omega)\). If \((y^n)\) is bounded in \(L^p(\Omega)\), then it follows from \((26)\) that \(\|y^n\|_{H^1(\Omega)}\) is bounded and the first step is finished. If \((y^n)\) is not bounded in \(L^p(\Omega)\), it has a subsequence, again denoted by \((y^n)\), such that \(\|y^n\|_{L^p(\Omega)}\) tends to \(\infty\) as \(n \to \infty\). If we define \(z^n := \frac{y^n}{\|y^n\|_{L^p(\Omega)}}\), then \(\|z^n\|_{L^p(\Omega)} = 1\) and
\[
c\|z^n\|_{H^1(\Omega)}^2 \leq \|b_2^n\|_{L^q(\Omega)} \|z^n\|_{L^q(\Omega)} \|z^n\|_{H^1(\Omega)} + D^n,
\]
where \(D^n \to 0\) as \(n \to \infty\). Dividing this inequality by \(\|z^n\|_{H^1(\Omega)}\), it follows that \((z^n)\) is bounded in \(H^1(\Omega)\). Consequently, there exists a subsequence again denoted by \((z^n)\) such that \(z^n \to z \in H^1_0(\Omega)\). Since \(H^1_0(\Omega)\) is compactly embedded in \(L^p(\Omega)\) we have \(z^n \to z\) in \(L^p(\Omega)\). It follows that \(\|z\|_{L^p(\Omega)} = 1\). However since \((y^n)\) is bounded in \(L^2(\Omega)\) we also have
\[
\|z^n\|_{L^2(\Omega)} = \|y^n\|_{L^2(\Omega)} \|y^n\|_{L^p(\Omega)}^{-1} \to 0 \quad \text{as} \quad n \to \infty,
\]
which implies that \(z = 0\). This is a contradiction, and consequently \((y^n)\) must be bounded in \(H^1(\Omega)\). Hence, \((y^n)\) has a subsequence, again denoted by \((y^n)\) such that \(y^n \to y\) in \(H^1_0(\Omega)\) for some \(y \in H^1_0(\Omega)\).

**Step 2.** We now prove that \((b_2^n)\) has a bounded subsequence. Here we interpret the sequence of constant functions \((b_1^n)\) as a sequence of vectors in \(\mathbb{R}^N\). We argue by contradiction and assume that there is a subsequence, again denoted \((b_2^n)\), such that \(b_2^n \neq 0\) for all \(n \in \mathbb{N}\) and \(|b_2^n| \to \infty\) as \(n \to \infty\). If we let \(v^n := b_2^n/|b_2^n|\), then \((v^n)\) has a subsequence converging to some \(v \in \mathbb{R}^N\). Let \(\phi \in \mathcal{D}(\Omega)\) be arbitrary. Testing \(e(y^n, b^n)\) with \(w^n := \frac{\phi}{|b_2^n|}\) we deduce
\[
\int_{\Omega} v \cdot (\phi \nabla y) = \lim_{n \to \infty} \int_{\Omega} v^n \cdot (\phi \nabla y^n)
\leq \lim_{n \to \infty} - |b_2^n| \int_{\Omega} \nabla y^n \cdot \nabla \phi + b_2^n \cdot (\phi \nabla y^n) + cy^n w - f w \, dx = 0.
\]
Consequently, by the fundamental lemma of the calculus of variations it follows that \(v \cdot \nabla y = 0\) a.e. in \(\Omega\). By Theorem 4.6, we deduce that \(y = 0\). Consequently, we must have
\[
\inf_{(y,b) \in \mathbb{F}_{ad}} \|y - y_s\|^2 + \|b\|^2_{\sigma,s} = \lim_{n \to \infty} \|y^n - y_s\|^2_{L^2(\Omega)} + \|b^n\|^2_{\sigma,s} \geq \|y_s\|^2_{L^2(\Omega)}. \quad (27)
\]
Now if (27) really holds, then existence of solutions follows from (20) and the proof is finished. If (27) does not hold, then our assumption must have been wrong and consequently (\(b_1\)) must be bounded.

**Step 3.** In the first two steps we have established that either existence of solutions holds trivially or \((y^*, b^*)\) is bounded in \(H_0^1(\Omega) \times H^*({\Omega})^N\). In the second case, the sequence \((y^n, b^n)\) has a weak accumulation point \((y, b)\) in \(H_0^1(\Omega) \times H^*({\Omega})^N\). The (weak,weak)-to-weak sequential continuity established in Corollary 2 implies that \(e(y, b) = 0\). Since the cost functional in \((P_{\text{adv}}(\sigma))\) is sequentially weakly lower semicontinuous, it follows that \((y, b)\) is a solution to \((\text{BP}_{\text{adv}})\). This finishes the proof. \(\square\)

**Remark 7.** Observe that (20) is a necessary and sufficient condition for existence of solutions to \((P_{\text{adv}}(\sigma))\); hence, it can not be improved in general. In some (extreme) cases (20) may be violated, take e.g. \(y_\delta = 0\) and \(f \neq 0\) (the latter condition implies that \(y = 0\) is not feasible). To illustrate (20) we discuss a particular case, where we assume that the \(L^2\) norm of the error contribution to the data is small compared to the \(L^2\) norm of the ground truth state, and that the same is true for the nonlocal energy seminorm of the ground truth convection coefficient. More precisely, let \((y^*, b^*)\) denote the ground truth solution, which satisfies \(e(y^*, b^*) = 0\). Let \(y_\delta = y^* + z\), where \(z \in L^2(\Omega)\) represents the error. Since the squared \(L^2\) norm is convex, we have

\[
\|y_\delta\|^2_{L^2(\Omega)} \geq \frac{1}{2} \|y^*\|^2_{L^2(\Omega)} - \|z\|^2_{L^2(\Omega)}.
\]

Since \((y^*, u^*) \in F_{\text{ad}}\), we infer that a sufficient condition for (20) to be satisfied is

\[
\|z\|^2_{L^2(\Omega)} + \|b^\dagger\|_{\sigma, s} \leq \frac{1}{2} \|y^*\|^2_{L^2(\Omega)} - \|z\|^2_{L^2(\Omega)}.
\]

Now if the squared \(L^2\) norm of the error is proportional to the squared \(L^2\) norm of the ground truth state with a small proportionality constant, i.e. \(\|z\|^2_{L^2(\Omega)} = \epsilon \|y^*\|^2_{L^2(\Omega)}\) for \(\epsilon > 0\) small, then (28) is equivalent to

\[
\|b^\dagger\|_{\sigma, s} \leq \left(\frac{1}{2} - 2\epsilon\right) \|y^*\|^2_{L^2(\Omega)}.
\]

Thus, we are left with a condition on the unknown ground truth convection coefficient. It is trivially satisfied if \(b\) is constant. Moreover, if \(\sigma\) is constant, then \(\|b^\dagger\|^2_{\sigma, s} \equiv \sigma \|b^\dagger\|^2_{H^*(\Omega)}\), where \(\sigma\) can be considered as a regularization parameter which is typically small.

Under stronger assumptions than the one in Proposition 9 it can be shown that \(\{P_{\text{adv}}(\sigma)\}_{\sigma \in W_{\text{ad}}}\) is weak*-to-weak stable, see Proposition 14, which would then allow us to apply the general existence result in Theorem 4.2 to argue that \((\text{BP}_{\text{adv}})\) has a solution. In fact, using the specific structure of \((\text{BP}_{\text{adv}})\) and the fact that weak*-to-weak stability as in Theorem 4.1 need only hold for a minimizing sequence to \((\text{BP}_{\text{adv}})\) we can show existence of solutions using as only assumption that the feasible set of the learning problem is nonempty, which Proposition 9 is in turn equivalent to the assumption that there exists \(\sigma \in W_{\text{ad}}\) such that (20) is satisfied for some \((y, b) \in F_{\text{ad}}\).

**Proposition 10** (Existence of solutions to the learning problem). Let \((N, s)\) satisfy (B) and assume that \(R\) is weak* sequentially lower semicontinuous on \(L^\infty(\{0, d\})\).
Then (BP_{adv}) has a solution if and only if there exist \( \sigma \in W_{ad} \) and \((y, b) \in F_{ad}\) such that
\[
\|y - y_\delta\|_{L^2(\Omega)}^2 + |b|_{\sigma, s}^2 \leq \|y_\delta\|_{L^2(\Omega)}^2.
\] (29)

**Proof.** As a straightforward consequence of Proposition 9 we obtain that the feasible set \( \mathcal{F} \) of (BP_{adv}) is nonempty if and only if there exist \( \sigma \in W_{ad} \) and \((y, b) \in F_{ad}\) satisfying (29). This proves that the condition above is necessary for the existence of solutions. To prove that it is also sufficient, note that if the above condition holds, then the feasible set \( \mathcal{F} \) is nonempty. Consequently, we can take a minimizing sequence \((\sigma^n, y^n, b^n)\) in \( \mathcal{F} \) to (BP_{adv}). It follows from the minimizing sequence property that \((b^n)\) is bounded in \(L^2(\Omega)\) and from the feasibility that \((b^n|_{\sigma^n, s})\) is bounded. In combination with Corollary 1, this yields that \((b^n)\) is bounded in \(H^s(\Omega)^N\). By Corollary 2 v), the boundedness of \((b^n)\) in \(H^s(\Omega)^N\) in turn implies that \((y^n)\) is bounded in \(H^1_0(\Omega)\). Since \(W_{ad}\) is weak* sequentially compact, it follows that \((\sigma^n, y^n, b^n)\), has a subsequence, again denoted by \((\sigma^n, y^n, b^n)\), such that \((\sigma^n, y^n, b^n)\) converges to \((\sigma, y, b) \in W_{ad} \times H^1_0(\Omega) \times H^s(\Omega)^N\) in the (weak*, weak, weak) sense in \(L^\infty((0,\delta]) \times H^1_0(\Omega) \times H^s(\Omega)^N\). As a consequence of Corollary 2 i) moreover we have \((y, u) \in F_{ad}\). Our next aim is to prove that \((y, b)\) solves the lower level problem with \(\sigma\). Using the KKT conditions in Theorem 4.3, we deduce that there exists a sequence \((\lambda^n)\) in \(H^1_0(\Omega)\) such that for every \(n \in \mathbb{N}\) and every \(1 \leq i \leq N\) we have
\[
[L^n(\sigma^n)b^n_i]v + \int_\Omega v(x)\lambda^n(x)\partial_i y^n(x) \, dx = 0, \quad \text{for all } v \in H^s(\Omega),
\]
\[
\int_\Omega (y^n - y_\delta)w + \nabla \lambda^n \cdot \nabla w + b^n \cdot (\lambda^n \nabla w) + c\lambda^n w \, dx = 0, \quad \text{for all } w \in H^1_0(\Omega).
\]
Arguing similarly as in Corollary 2 v) one can prove that \((\lambda^n)\) is bounded in \(H^1_0(\Omega)\). Let \(p^n_i(x) = \lambda^n(x)\partial_i y^n(x)\). We now distinguish between two cases.

1. If \(N \in \{1, 2\}\), then \((p^n_i)\) is bounded in \(L^{2-\delta}(\Omega)\) for arbitrary \(0 < \delta < 1\).
2. If \(N = 3\) and \(s \in (1/2, 1)\), then \((p^n_i)\) is bounded in \(L^{3/2}(\Omega)\) and \(H^s(\Omega)^N\) is compactly embedded in \(L^3(\Omega)\).

In either case it follows from Theorem 2.5 that
\[
\lim_{n \to \infty} |b^n_i|_{\sigma^n, s} = |b_i|_{\sigma, s} \quad \text{for every } 1 \leq i \leq N.
\] (30)

This implies that
\[
\|y - y_\delta\|_{L^2(\Omega)}^2 + |b|_{\sigma, s}^2 = \lim_{n \to \infty} \|y^n - y_\delta\|_{L^2(\Omega)}^2 + |b^n|_{\sigma^n, s}^2 \leq \lim_{n \to \infty} \|w - y_\delta\|_{L^2(\Omega)}^2 + |v|_{\sigma^n, s}^2 + \|w - y_\delta\|_{L^2(\Omega)}^2 + |v|_{\sigma, s}^2,
\]
for all \((w, v) \in F_{ad}\). This shows that \((\sigma, y, b) \in \mathcal{F}\). Since by assumption the cost functional is weak*-weak-weak sequentially lower semi continuous, it follows that \((\sigma, y, b)\) solves (BP_{adv}), which finishes the proof. \(\square\)

### 4.3.1. Second order sufficient optimality conditions.

In view of Theorem 4.5, it is important to know under which circumstances a computed KKT point of the lower level problem satisfies the second order sufficient optimality conditions stated in Theorem 4.4. Although a complete answer to this question is beyond the scope of this paper, we aim to provide some insight into the problem. First, we prove that second order sufficient optimality conditions are satisfied if and only if the minimum of a certain quadratic cost function over a convex set (which is proven to be attained) is strictly larger than \(-1\). Since this quadratic cost function is not guaranteed to be
Proposition 11. Suppose \((N, s)\) satisfies \((B)\). Let \((y, b, \lambda)\) denote a KKT point of \((P_{\text{adv}}(\sigma))\). Then the problem

\[
\min_{(\delta_y, \delta_b) \in H^1_0(\Omega) \times H^s_0(\Omega)} \int_{\Omega} \delta_y \cdot (\lambda \nabla \delta_y) \, dx
\]

subject to \(\|\delta_y\|_{L^2(\Omega)}^2 + |\delta_b|_{\sigma, s}^2 \leq 1\) and \(D e(y, b)(\delta_y, \delta_b) = 0\)

has a solution. Moreover, \((y, b, \lambda)\) satisfies the second order sufficient optimality conditions if and only if the minimal function value in (31) is strictly larger than \(-1\).

Proof. Since \((N, s)\) satisfies \((B)\), we can choose \(p, q > 1\) satisfying \((C)\). Now we divide the proof into two parts.

Part 1. First we assume that (31) has a solution and prove that in this case the second order sufficient optimality conditions are equivalent to the minimal function value in (31) being strictly larger than \(-1\). From Proposition 8 we know that the mapping \((y, b) \mapsto e(y, b)\) is continuously differentiable as a mapping from \(H^1_0(\Omega) \times L^q(\Omega)\) to \(H^{-1}(\Omega)\) and that \(D_y e(y, b) \in \mathcal{L}(H^1_0(\Omega), H^{-1}(\Omega))\) is bijective. In combination with the assumption that \(H^s(\Omega)\) is compactly embedded in \(L^q(\Omega)\), this implies that the operator \(S: H^s(\Omega) \rightarrow H^1_0(\Omega)\), \(\delta_b \mapsto \delta_y\), where \(\delta_y\) is the unique element in \(H^1_0(\Omega)\) such that \((\delta_b, \delta_y) \in \ker D e(y, b)\), is well-defined, injective, and compact. Proposition 2 now yields that \(\sqrt{\|\delta_y\|_{L^2(\Omega)}^2 + |\delta_b|_{\sigma, s}^2}\) is a norm on \(\ker D e(y, b)\), which is equivalent to the original norm \(\sqrt{\|\delta_y\|_{L^2(\Omega)}^2 + |\delta_b|_{H^s(\Omega)}^2}\). It follows that \((y, b, \lambda)\) satisfies the second order sufficient optimality conditions if and only if there is \(\mu > 0\) such that

\[
D^2_{(y, b)} L(\sigma, y, b, \lambda)[(\delta_y, \delta_b), (\delta_y, \delta_b)] \geq \mu \left(\|\delta_y\|_{L^2(\Omega)}^2 + |\delta_b|_{\sigma, s}^2\right)
\]

(32)

for all \((\delta_y, \delta_b) \in \ker D e(y, b)\). After some computations we obtain that

\[
D^2_{(y, b)} L(\sigma, y, b, \lambda)[(\delta_y, \delta_b), (\delta_y, \delta_b)] = \|\delta_y\|_{L^2(\Omega)}^2 + |\delta_b|_{\sigma, s}^2 + \int_{\Omega} \delta_b \cdot (\lambda \nabla \delta_y) \, dx.
\]

Since the mapping \(\delta_b \mapsto \delta_y\) such that \((\delta_b, \delta_y) \in \ker D e(y, b)\) is linear, it follows that (32) is equivalent to

\[
\int_{\Omega} \delta_b \cdot (\lambda \nabla \delta_y) \, dx \geq \mu - 1 \quad \forall (\delta_b, \delta_y) \in \ker D e(y, b): \|\delta_y\|_{L^2(\Omega)}^2 + |\delta_b|_{\sigma, s}^2 = 1,
\]

(33)

which immediately yields that if the minimum in (31) is attained and strictly larger than \(-1\), then the second order sufficient optimality conditions are satisfied. Conversely, assume that the second order sufficient optimality conditions are satisfied. Observe that the case that the cost functional in (31) is nonnegative on the feasible set is trivial. If the minimum in (31) is negative, then it is attained in \((\delta_y, \delta_b)\) such
that \( \|y\|_{L^2(\Omega)}^2 + |b|_{\sigma,s}^2 = 1 \) (which can be proven by a straightforward rescaling argument). Consequently, since (32) and (33) are equivalent, if the second order sufficient optimality conditions are satisfied, then the minimal functional value in (31) is strictly larger than \(-1\).

**Part 2.** Since \( p, q > 1 \) were chosen such that \((C)\) holds, by Hölder’s inequality the mapping \((\delta_y, \delta_b, \lambda) \mapsto \int_\Omega \delta_b \cdot (\lambda \nabla \delta_y) \, dx\) is continuous on \( H^1_0(\Omega) \times H^*(\Omega) \times H^1_0(\Omega) \).

From the first step we deduce that the feasible set is weakly sequentially compact in \( L^2(\Omega) \times H^*(\Omega) \). Existence of solutions to (31) follows with classical arguments using compactness of the mapping \( \delta_b \mapsto \delta_y \) from \( H^*(\Omega) \) to \( H^1_0(\Omega) \) and the fact that continuous bilinear forms are (weak-strong)-to-weak sequentially continuous. \(\square\)

**Remark 8.** Since the minimum in (31) is always attained, the proof of Proposition 11 reveals that the second order sufficient optimality conditions for \((P_{\text{adv}}(\sigma))\) are equivalent to the requirement that

\[
D^2_{(y,b)}(\sigma, y, b, \lambda)(\delta_y, \delta_b, (\delta_y, \delta_b)) > 0
\tag{34}
\]

for all \((\delta_y, \delta_b) \in \ker De(y, b) \setminus \{(0, 0)\} \).

**Proposition 12.** Let \((N, s)\) satisfy \((B)\), and let \( p, q > 1 \) be as in \((C)\). Let \((y, b, \lambda)\) be a KKT point of \((P_{\text{adv}}(\sigma))\). Then there exists a constant \( C > 0 \) (depending on \( b \)) such that if \( \|\lambda\|_{L^p(\Omega)} \leq C/\|\nabla y\|_{L^2(\Omega)} \), then \((y, b, \lambda)\) satisfies the second order sufficient optimality conditions.

**Proof.** By Hölder’s inequality, we have

\[
\int \delta_b \cdot (\lambda \nabla \delta_y) \geq -\|\delta_b\|_{L^1(\Omega)}\|\delta_y\|_{H^1(\Omega)}\|\lambda\|_{L^p(\Omega)}.
\]

Now let \( C_{\sigma,q,s} > 0 \) be such that

\[
\|\delta_b\|_{L^2(\Omega)} \leq C_{\sigma,q,s}\left(\|\delta_y\|_{L^2(\Omega)}^2 + |\delta_b|_{\sigma,s}^2\right)^{1/2}
\]

and \( C(b) > 0 \) be such that for all \( g \in H^{-1}(\Omega) \) and \( w \in H^1(\Omega) \) such that

\[-\Delta w + b \cdot \nabla w = g\]

we have \( \|\nabla w\|_{L^2(\Omega)} \leq C(b)\|g\|_{H^{-1}(\Omega)} \). It follows that

\[
\|\delta_b\|_{L^2(\Omega)} \leq C(b)\|\nabla \delta_y\|_{H^{-1}(\Omega)}
\]

\[
\leq C(b)C_{\sigma,q,s}C_{1,p}\|\nabla \delta_y\|_{L^2(\Omega)}\left(\|\delta_y\|_{L^2(\Omega)}^2 + |\delta_b|_{\sigma,s}^2\right)^{1/2},
\tag{35}
\]

where \( C_{1,p} \) is the embedding constant of the embedding \( H^1_0(\Omega) \) into \( L^p(\Omega) \). Thereby the bound on \( \|\delta_b \cdot \nabla y\|_{H^{-1}(\Omega)} \) can be obtained using Hölder’s inequality, by which for all \( p \in H^1_0(\Omega) \) with \( \|p\|_{H^1(\Omega)} = 1 \) we have

\[
\int_\Omega \delta_b \cdot (p \nabla y) \, dx \leq \|\delta_b\|_{L^1(\Omega)}\|p\|_{L^p(\Omega)}\|\nabla y\|_{L^2(\Omega)}
\]

\[
\leq C_{\sigma,q,s}C_{1,p}\|\nabla \delta_y\|_{L^2(\Omega)}\left(\|\delta_y\|_{L^2(\Omega)}^2 + |\delta_b|_{\sigma,s}^2\right)^{1/2}.
\]

We deduce that

\[
\int \delta_b \cdot (\lambda \nabla \delta_y) \geq -C(b)C_{\sigma,q,s}C_{1,p}\|\nabla \delta_y\|_{L^2(\Omega)}\|\lambda\|_{L^p(\Omega)}\left(\|\delta_y\|_{L^2(\Omega)}^2 + |\delta_b|_{\sigma,s}^2\right).
\]

It is now evident that if

\[
\|\lambda\|_{L^p(\Omega)} < (C(b)C_{\sigma,q,s}C_{1,p}\|\nabla \delta_y\|_{L^2(\Omega)})^{-1},
\]
then the optimal value is strictly larger than $-1$, which yields the claim. \hfill \square

Using similar ideas as in Corollary 2 v) one can prove that for bounded subsets $B$ of $H^s(\Omega)$ we have $\sup_{b \in B} C(b) < \infty$, where $C(b)$ is the Lipschitz constant in the proof of Proposition 12.

4.3.2. Uniqueness of solutions. A similar condition as in Proposition 12 can also be given for uniqueness of solutions. The discussion is restricted to the case $c = 0$.

**Proposition 13.** Let $(N, s)$ satisfy (B), and let $p, q > 1$ be such that $1/p + 1/q = 1/2$ and $H^s(\Omega)$ is continuously embedded in $L^q(\Omega)$. Assume that there exists $(y, b) \in F_{ad}$ such that (29) holds with strict inequality. Let $(\bar{y}, \bar{b}, \bar{\lambda})$ be a KKT point of $(P_{adv}(\sigma))$. Then there exists a constant $C > 0$ such that

$$
\|\bar{\lambda}\|_{L^p(\Omega)} < C,
$$

then $(\bar{y}, \bar{b})$ is the unique solution to $(P_{adv}(\sigma))$.

**Proof.** A look at the proof of Proposition 9 reveals that under our assumption all solutions to $(P_{adv}(\sigma))$ must be contained in a norm ball with finite radius in $H^1_0(\Omega) \times H^s(\Omega)$. In the following we denote this ball by $B$. The proof is heavily based on the observation that

$$
eq \equiv$$

$$
\|y - \bar{y}\|_{L^2(\Omega)}^2 + \|b - \bar{b}\|_{H^s(\Omega)}^2 \geq c\|b - \bar{b}\|_{L^2(\Omega)}^2.
$$

(37)

Theorem 2.2 yields that suffices to show that there exists $\kappa > 0$ such that for all $(y, b) \in F_{ad} \cap B$ we have

$$
\|y - \bar{y}\|_{L^2(\Omega)}^2 + \|b - \bar{b}\|_{H^s(\Omega)}^2 \geq \kappa\|b - \bar{b}\|_{L^2(\Omega)}^2.
$$

(38)

We begin by proving that there exist $\kappa_1 > 0$ and $M > 0$ such that (38) with $\kappa = \kappa_1$ holds for all $(y, b) \in F_{ad} \cap B$ with $\|b - \bar{b}\|_{L^2(\Omega)} < M$. To do this, we argue by contradiction: If this is false, then there is a sequence $(y^n, b^n)$ in $(F_{ad} \cap B) \setminus \{(\bar{y}, \bar{b})\}$ with $\|b^n - \bar{b}\|_{L^2(\Omega)} \to 0$ as $n \to \infty$ and a nullsequence $(\kappa^n)$ in $(0, \infty)$ such that

$$
\|y^n - \bar{y}\|_{L^2(\Omega)}^2 + \|b^n - \bar{b}\|_{H^s(\Omega)}^2 < \kappa^n\|b^n - \bar{b}\|_{L^2(\Omega)}^2.
$$

(39)

It follows that $\|b^n - \bar{b}\|_{H^s(\Omega)} \to 0$ as $n \to \infty$. We define a sequence $(\delta^n)$ in $H^s(\Omega)^N$ by setting $\delta^n := (b^n - \bar{b})/\|b^n - \bar{b}\|_{L^2(\Omega)}$ for every $n \in \mathbb{N}$. By definition we have $\|\delta^n\|_{L^2(\Omega)} = 1$. Moreover, it follows from (39) that $|\delta^n|_{\sigma,s} \to 0$ as $n \to \infty$. Hence $(\delta^n)$ has a subsequence, which converges weakly in $H^s(\Omega)$ and strongly in $L^2(\Omega)$ to
a constant and nonzero element $\delta_y$ of $H^s(\Omega)$. Additionally, for the sequence $(\delta_y^n)$ defined by $\delta_y^n := (y^n - \bar{y})/\|b^n - \bar{b}\|_{L^2(\Omega)}$ we have
\[
\|\delta_y^n\|_{H^1(\Omega)} \leq \frac{C^2 \left( \|b^n - \bar{b}\|_{L^2(\Omega)} + \|b^n - \bar{b}\|_{H^s(\Omega)}^2 \right)}{\|b^n - \bar{b}\|_{L^2(\Omega)}^2} \leq C^2(1 + \kappa^n/c),
\]
where $C > 0$ is the Lipschitz constant of the solution operator $b \mapsto y(b)$ on $B$ with respect to the norm $\| \cdot \|_{H^s}$ on $H^s(\Omega)$. Consequently, $(\delta_y^n)$ has a weak cluster point $\delta_y$ in $H^1_0(\Omega)$. Let $\phi \in \mathcal{D}(\Omega)$ be arbitrary and let $\phi^n := \phi/\|b^n - \bar{b}\|_{L^2(\Omega)}$. Testing (36) for $(y, b) = (y^n, b^n)$ with $\phi^n$ we obtain
\[
\int_{\Omega} \nabla \delta_y^n \cdot \nabla \phi + (b \cdot \nabla \delta_y^n)\phi + (\delta_y^n \cdot \nabla \bar{y})\phi + [\delta_y^n \cdot \nabla (y^n - y)]\phi \, dx = 0. \tag{40}
\]
Taking the limit $n \to \infty$ in (40) and (39) we deduce that $(\delta_y, \delta_b) \in \ker D_e(\bar{y}, \bar{b})$ and $\|\delta_y\|_{L^2(\Omega)} = 0$. This is a contradiction since the mapping $\delta_b \mapsto \delta_y$ such that $(\delta_y, \delta_b)$ is injective on the space of constant functions.

It remains to show that there exists $\kappa_2 > 0$ such that (38) with $\kappa = \kappa_2$ holds for all $(y, b) \in F_{ad} \cap B$ with $\|b - \bar{b}\|_{L^2(\Omega)} \geq M$. For this, it suffices to prove that
\[
\min_{(y, b) \in F_{ad} \cap B : \|b - \bar{b}\|_{L^2(\Omega)} \geq M} \|y - \bar{y}\|_{L^2(\Omega)} + \|b - \bar{b}\|_{H^s(\Omega)}^2 \tag{41}
\]
is a solution, and that the optimal function value $\bar{\kappa}$ in (41) is strictly larger than zero, since then we have
\[
\|y - \bar{y}\|_{L^2(\Omega)} + \|b - \bar{b}\|_{H^s(\Omega)}^2 \geq \bar{\kappa} \geq \|y - \bar{y}\|_{L^2(\Omega)} + \|b - \bar{b}\|_{H^s(\Omega)}^2.
\]
for all $(y, b) \in F_{ad} \cap B$ with $\|b - \bar{b}\|_{L^2(\Omega)} \geq M$, where $r := \sup\{\|b - \bar{b}\|_{L^2(\Omega)} : (y, b) \in F_{ad} \cap B\}$. Proving existence of solutions to (41) is straightforward since
\[
\{(y, b) \in F_{ad} \cap B : \|b - \bar{b}\|_{L^2(\Omega)} \geq M\}
\]
is weakly sequentially compact; a fact which can be easily proven using that $H^s(\Omega)$ is compactly embedded in $L^2(\Omega)$ and Corollary 2. It now follows from the first step that the optimal function value is positive, as claimed above.

**Step 3.** Using (36), for every $(y, b) \in F_{ad} \cap B$ we compute
\[
\begin{align*}
\|\lambda e_y(\bar{y}, \bar{b})|y - \bar{y}| + e_b(\bar{y}, \bar{b})|b - \bar{b}| & = | - \lambda e_y|y - \bar{y}, b - \bar{b}| |y - \bar{y}|_{L^2(\Omega)} + | - \lambda e_y|y - \bar{y}, b - \bar{b}| |b - \bar{b}|_{H^s(\Omega)} |b - \bar{b}|_{H^s(\Omega)} |b - \bar{b}|_{H^s(\Omega)},
\end{align*}
\]
where $C_{s,q}$ is the embedding constant of the embedding of $H^s(\Omega)$ into $L^2(\Omega)$ and $D > 0$ is the Lipschitz constant of the solution operator $b \mapsto y(b)$ on $B$ with respect to the norm $\| \cdot \|_{H^1(\Omega)}$. Let $F(y, b) := \|y - y\|_{L^2(\Omega)} + \|b\|_{H^s(\Omega)}$ for every $(y, b) \in F_{ad}$. Moreover, let $c$ be the corresponding constant in (37). We have
\[
F(y, b) + F(\bar{y}, \bar{b}) = F_y(\bar{y}, \bar{b})(y - \bar{y}) + F_b(\bar{y}, \bar{b})(b - \bar{b}) + \|y - \bar{y}\|_{L^2(\Omega)} + \|b - \bar{b}\|_{H^s(\Omega)}^2 \\
\geq (c - C_{s,q}D)|\lambda|_{L^p(\Omega)} \|b - \bar{b}\|_{H^s(\Omega)}^2,
\]
Here we used that $(\bar{y}, \bar{b}, \bar{\lambda})$ is a KKT point for the second equality and the estimates in (38) and (42) for the last inequality. It follows that if
\[
\|\lambda\|_{L^p(\Omega)} < c/(C_{s,q}D),
\]
then $(\bar{y}, \bar{b})$ is the unique solution to $(P_{\text{adv}}(\sigma))$. \qed
5. **Numerical experiments.** In this section we present results of numerical experiments where we solve the learning problem for the linear forward problem from Example 3.1. Here we let $\Omega = (0, 1)$ and $\rho = 0.1$. As a regularization operator for the weights, we consider the particular choice

$$R(\sigma) = \beta \int_0^d \sigma \, dx + \alpha |\sigma|_{L^2((0,d))}^2$$

for $\sigma \in W_{ad}$, where $\alpha, \beta \geq 0$. The obtained results are compared to results obtained for choosing the optimal regularization parameter $\nu$ in (1) by solving a similar learning problem.

5.1. **Data.** We let $\omega = (\omega_1, \ldots, \omega_m)^T$ be an $m$-dimensional random variable following a uniform distribution on $[0,1]^m$. We distinguish between two cases.

(A) In the first case, we let $m = 3$ and

$$u^\dagger(x, \omega) = \sin(20 \omega_1 x) + \omega_3 \cos(40 \omega_2 x) \quad \text{for } x \in (0, 1).$$

(B) In the second case, we let $m = 3$ and

$$u^\dagger(x, \omega) = 3 \omega_3 \cos(6 \pi x + 10 \omega_1) + 2 \omega_2 \quad \text{for } x \in (0, 1).$$

To create data for training and validation, we take samples $\omega_i$ from $\omega$ and let $u^\dagger, i = u^\dagger(\cdot, \omega_i)$. We discretize the problem using linear Lagrange elements for equidistant grid points $0 = x_1 < \cdots < x_{N_E} = 1$, where $N_E = 128$. The corresponding (discrete) ground truth state $y^\dagger, i$ is computed by solving the discretized forward problem. Noisy data measurements are generated by point wise setting $y^\delta_i(x_j) = y^\dagger, i(x_j) + \epsilon \xi_{i,j}$, where $\xi_{i,j} \in \mathbb{R}$ are samples drawn from a normally distributed random variable with mean 0 and standard deviation 1, and $\epsilon$ is the noise level. In order to discretize the weights, we use piecewise constant FEM. We let $(u_1, \ldots, u_{N_E})$ and $(\sigma_1, \ldots, \sigma_{N_E+1})$ denote a basis for the control and the weight FEM spaces, respectively. The integrals

$$\mathcal{L}_s(\sigma_k) u_i u_j = \iint_{\Omega \times \Omega} \frac{(u_i(x)-u_i(y))(u_j(x)-u_j(y))}{|x-y|^{1+2s}} \sigma_k(|x-y|) \, dx \, dy$$

are computed analytically using symbolic integration.

5.2. **Applied methods.** Recall that the lower level problem has a unique solution for every regularization weight $\sigma \in W_{ad}$. Using this, we define the reduced cost functional $F: W_{ad} \to \mathbb{R}$ by

$$F(\sigma) := \frac{1}{2} \|u(\sigma) - u^\dagger\|_{L^2(\Omega)}^2 + R(\sigma), \quad \text{for every } \sigma \in W_{ad},$$

where $u(\sigma)$ is the unique solution to the lower level problem with weight $\sigma$. The learning problem (BP) can then be written as follows

$$\min_{\sigma \in W_{ad}} F(\sigma) \quad \text{subject to } \sigma_{\min} \leq \sigma \leq \sigma_{\max} \quad (43)$$

where $\sigma_{\min} \equiv \gamma_1 \chi_{[0,\delta]}$ and $\sigma_{\max} \equiv \gamma_2$. A necessary optimality condition for $\sigma^*$ to be a solution of (43) is

$$\langle F'(\sigma^*), \sigma - \sigma^* \rangle \geq 0 \quad \text{for all } \sigma \in W_{ad}, \quad (44)$$
If $F' (\sigma^*)$ has a Riesz representative $\nabla F (\sigma^*)$ in $L^2 (\Omega)$, then (44) is equivalent to

$$\sigma^* = P_{W_{ad}} (\sigma^* - c \nabla F (\sigma^*)),$$

for arbitrary $c > 0$, where $P_{W_{ad}}$ is the $L^2$-minimal projection on $W_{ad}$. We define

$$\Phi (\sigma) := \sigma - P_{W_{ad}} (\sigma - c \nabla F (\sigma)),$$

which can be interpreted pointwise almost everywhere on $(0, d)$ as

$$\Phi (\sigma) (x) = \sigma (x) - \max [\sigma_{\min} (x), \min [\sigma_{\max} (x), \sigma (x) - c \nabla F (\sigma) (x)]] .$$

In order to solve the reduced learning problem we use a non-linear primal-dual active set method provided in [27]. To solve the unconstrained problems on the inactive set we use a globalized quasi-Newton method accompanied by an Armijo line search (compare [30, algorithm 11.5 on p 60]).

Strictly speaking the convergence analysis provided in [27] does not apply to our setting. In practice the algorithm performed satisfactorily.

5.3. Results. We tested the algorithm in MATLAB for various choices of $s$. We create $N_{\text{train}}$ training and $N_{\text{val}}$ validation data vectors. The training set is divided into $N_{\text{batch}}$ training batches. Each training batch then consists of batchsize = $N_{\text{train}} / N_{\text{batch}}$ training vectors. For $1 \leq i \leq N_{\text{batches}}$ an optimal regularization weight $\sigma^{*,i}$ is computed for the $i$-th batch by solving the associated learning problem. Subsequently, the optimal weights are tested on the validation set. Thus for each validation vector $(y^1, u^1, y^2)$ and each optimal weight $\sigma^{*,i}$ we compute a solution $u_{\sigma^{*,i}}$ to the corresponding lower level problem. We then compute the validation error given by $\|u_{\sigma^{*,i}} - u^1\|_{L^2 (\Omega)}$. The average validation error is obtained by averaging the validation error over all validation vectors and weights $\sigma^{*,i}$. The same training and validation procedure was used to learn the optimal regularization parameter $\nu^*$ for regularization with i) the fractional order Sobolev seminorm (corresponding to a weight $\sigma \equiv 1$), ii) the $L^2 (\Omega)$ norm, and iii) the $H^1 (\Omega)$ seminorm $\|\nabla \cdot\|_{L^2 (\Omega)}$.

In the experiments, we always rescaled the nonlocal energy seminorm using the optimal regularization parameter $\nu^*$ for the fractional order Sobolev seminorm, i.e. we replaced $| \cdot |_{2, \sigma}^2$ by $\nu^* | \cdot |_{2, \sigma}^2$ in the lower level problem. This enabled us to use $\sigma \equiv 1$ as an initial guess for the weight. The following parameters, which were used in all experiments, as well as the values given for $\beta$ later on, are all given relative to this rescaling: $\alpha = 10^{-4}$, $\gamma_1 = 0.1$, $\gamma_2 = 10^{15}$, and $\delta = 1/8$ (see Assumption 2.1). 1 % additive noise (i.e. $\epsilon = 0.01$) was used to create the noisy measurements. The obtained training and validation errors for different batchsizes are provided in Tables 1 and 2. We notice the following behaviour:

1. In all tested cases, both the training and the validation error for the optimal weight $\sigma$ are smaller than the training and validation error for the optimal regularization parameter $\nu$ (see Tables 1 and 2).
2. Overall, the benefits of being able to choose a distance dependent weight $\sigma$ over choosing only a scalar regularization parameter were less pronounced for larger values of $s$ than for smaller ones (compare Table 1a with Table 1b and Table 2a with Table 2b).
3. Note that for $s = 0.1$ the optimal weight in case (B) has distinct peaks around $1/3$, $2/3$, and close to 1 (see Figure 1). This can be explained by the fact that functions created as in (B) are always periodic with a period 1/3.
4. In case (B) the influence of the weight was much larger compared to case (A). The validation error was significantly decreased for $s = 0.1$ (see Table 2a). We attribute this to the fact that in case (B) both the training and validation functions were periodic with the same period. This constitutes a case where in our opinion the impact of being able to choose a nonlocal weight is clearly visible.

5. Large batch sizes improve estimates for $\nu$ as well as $\sigma$. In fact, the results for smaller batch sizes (even after taking several smaller batches involving the same amount of training data in total) can not reach the results obtained for one batch consisting of the total training set (compare the rows).

6. While the $L^2$ norm was outperformed by all other tested regularization operators, the performances of the $H^1(\Omega)$ seminorm, the $H^s(\Omega)$ seminorm with $s = 0.9$, and the nonlocal energy seminorm with $s = 0.9$ were almost identical. In general, whether the additional computational effort when using fractional order regularization is justified, depends on the structure of the data. It should be noted that the significant improvement reported in case (B) is not surprising. In fact, case (B) was intentionally designed to provide an example where one would expect that being able to choose a distance dependent weight improves the reconstruction quality.

**Acknowledgments.** G. H. gratefully acknowledges support by the International Research Training Group IGDK 1754 “Optimization and Numerical Analysis for Partial Differential Equations with Nonsmooth Structures”, funded by the German Research Council (DFG) and the Austrian Science Fund (FWF):[W 1244-N18].

| Table 1. Average training and validation error for optimal regularization parameter $\nu^*$ (second and third column) and optimal weight $\sigma^*$ (fourth and fifth column) in case (A) using $\beta = 0$. The training and validation set both consisted of 512 data vectors. |
| --- |
| (a) $s = 0.1$ |
| batchsize | train error reg | val error reg | train error weight | val error weight |
| 8 | $1.81 \times 10^{-2}$ | $2.19 \times 10^{-2}$ | $1.42 \times 10^{-2}$ | $2.04 \times 10^{-2}$ |
| 64 | $1.88 \times 10^{-2}$ | $2.11 \times 10^{-2}$ | $1.61 \times 10^{-2}$ | $1.86 \times 10^{-2}$ |
| 512 | $1.90 \times 10^{-2}$ | $2.09 \times 10^{-2}$ | $1.65 \times 10^{-2}$ | $1.82 \times 10^{-2}$ |
| (b) $s = 0.9$ |
| batchsize | train error reg | val error reg | train error weight | val error weight |
| 8 | $1.53 \times 10^{-2}$ | $1.98 \times 10^{-2}$ | $1.51 \times 10^{-2}$ | $1.97 \times 10^{-2}$ |
| 64 | $1.63 \times 10^{-2}$ | $1.85 \times 10^{-2}$ | $1.61 \times 10^{-2}$ | $1.84 \times 10^{-2}$ |
| 512 | $1.65 \times 10^{-2}$ | $1.82 \times 10^{-2}$ | $1.64 \times 10^{-2}$ | $1.80 \times 10^{-2}$ |
| (c) $L^2$ and $H^1$ regularization with optimal $\nu^*$ |
| batchsize | train error $L^2$ | val error $L^2$ | train error $H^1$ | val error $H^1$ |
| 8 | $1.93 \times 10^{-2}$ | $2.30 \times 10^{-2}$ | $1.51 \times 10^{-2}$ | $1.97 \times 10^{-2}$ |
| 64 | $1.99 \times 10^{-2}$ | $2.22 \times 10^{-2}$ | $1.61 \times 10^{-2}$ | $1.84 \times 10^{-2}$ |
| 512 | $2.01 \times 10^{-2}$ | $2.20 \times 10^{-2}$ | $1.64 \times 10^{-2}$ | $1.80 \times 10^{-2}$ |
Table 2. Average training and validation error for optimal regularization parameter $\nu^*$ (second and third column) and optimal weight $\sigma^*$ (fourth and fifth column) in case (B) using $\beta = 0$. The training and validation set both consisted of 512 data vectors.

(\text{A}) $s = 0.1$

| batchsize | train error reg | val error reg | train error weight | val error weight |
|------------|-----------------|---------------|-------------------|-----------------|
| 8          | $2.03 \times 10^{-2}$ | $2.12 \times 10^{-2}$ | $1.02 \times 10^{-2}$ | $1.09 \times 10^{-2}$ |
| 64         | $2.05 \times 10^{-2}$ | $2.09 \times 10^{-2}$ | $1.02 \times 10^{-2}$ | $1.04 \times 10^{-2}$ |
| 512        | $2.06 \times 10^{-2}$ | $2.08 \times 10^{-2}$ | $1.14 \times 10^{-2}$ | $1.15 \times 10^{-2}$ |

(\text{B}) $s = 0.5$

| batchsize | train error reg | val error reg | train error weight | val error weight |
|------------|-----------------|---------------|-------------------|-----------------|
| 8          | $1.37 \times 10^{-2}$ | $1.44 \times 10^{-2}$ | $1.33 \times 10^{-2}$ | $1.41 \times 10^{-2}$ |
| 64         | $1.38 \times 10^{-2}$ | $1.42 \times 10^{-2}$ | $1.35 \times 10^{-2}$ | $1.38 \times 10^{-2}$ |
| 512        | $1.39 \times 10^{-2}$ | $1.41 \times 10^{-2}$ | $1.35 \times 10^{-2}$ | $1.38 \times 10^{-2}$ |

(\text{C}) $L^2$ and $H^1$ regularization with optimal $\nu^*$

| batchsize | train error $L^2$ | val error $L^2$ | train error $H^1$ | val error $H^1$ |
|------------|-------------------|-----------------|-------------------|-----------------|
| 8          | $2.48 \times 10^{-2}$ | $2.60 \times 10^{-2}$ | $1.33 \times 10^{-2}$ | $1.40 \times 10^{-2}$ |
| 64         | $2.51 \times 10^{-2}$ | $2.56 \times 10^{-2}$ | $1.35 \times 10^{-2}$ | $1.38 \times 10^{-2}$ |
| 512        | $2.52 \times 10^{-2}$ | $2.55 \times 10^{-2}$ | $1.35 \times 10^{-2}$ | $1.38 \times 10^{-2}$ |

K. K. gratefully acknowledges partial support by the ERC advanced grant 668998 (OCLOC) under the EU’s H2020 research program.

Appendix A. Stability.

Proposition 14 (Stability). Assume that there exists $(y, b) \in F_{ad}$ with $b$ constant such that $\|y - y_s\|_{L^2(\Omega)}^2 < \|y_s\|_{L^2(\Omega)}^2$. Then $\{P_{adv}(\sigma)\}_{\sigma \in W_{ad}}$ is both weak*--to--weak stable and stable.

Proof. We begin by proving weak*--to--weak stability. For this purpose, let $(\sigma^n)$ be a sequence in $W_{ad}$ and $\sigma \in W_{ad}$ be such that $\sigma^n \rightharpoonup^* \sigma$ in $L^\infty((0, d))$ as $n \to \infty$. Let $(y^n, b^n)$ be a sequence in $H^1_0(\Omega) \times H^s(\Omega)$ such that, for every $n \in \mathbb{N}$, $(y^n, b^n)$ is a solution to $P_{adv}(\sigma)$ for $\sigma = \sigma^n$. We need to show that $(y^n, b^n)$ has a weak accumulation point in $H^1_0(\Omega) \times H^s(\Omega)$, and that every weak accumulation point of $(y^n, b^n)$ is a solution to $P_{adv}(\sigma)$. We first prove that $(y^n, b^n)$ is bounded in $H^1_0(\Omega) \times H^s(\Omega)$. For this purpose observe that

$$\|y^n - y_s\|_{L^2(\Omega)}^2 + \|b^n\|_{s, \sigma}^2 \leq \|y - y_s\|_{L^2(\Omega)}^2 < \|y_s\|_{L^2(\Omega)}^2$$

for every $n \in \mathbb{N}$. It follows that there is $\kappa > 0$ such that for every $n \in \mathbb{N}$ we have

$$\|y^n - y_s\|_{L^2(\Omega)}^2 + \|b^n\|_{s, \sigma}^2 < \|y_s\|_{L^2(\Omega)}^2 - \kappa.$$  \hfill (45)

This implies that $(\|b^n\|_{s, \sigma})$ is bounded. In order to prove that $(\|b^n\|_{L^2(\Omega)})$ is bounded, we argue by contradiction. Indeed, if $(\|b^n\|_{L^2(\Omega)})$ is not bounded, arguing as in the proof of Proposition 9 we obtain that, up to subsequences, $y^n \rightharpoonup 0$ in $L^2(\Omega)$. This implies that there exists $N \in \mathbb{N}$ such that for all $n \geq N$ we have

$$\|y_s\|_{L^2(\Omega)}^2 - \kappa \leq \|y_s\|_{L^2(\Omega)}^2 - \kappa + \|b^n\|_{s, \sigma}^2 \leq \|y^n - y_s\|_{L^2(\Omega)}^2 + \|b^n\|_{s, \sigma}^2$$
This contradicts (45). Consequently, \((b^n)\) must be bounded in \(H^s(\Omega)\). It follows that \((b^n)\) has a weakly converging subsequence in \(H^s(\Omega)\), for simplicity again denoted by \((b^n)\). By Proposition 8 we deduce that \(y^n \rightarrow y\) in \(H^1_0(\Omega)\). Now for every \(n \in \mathbb{N}\) and every \((w, v) \in F_{ad}\) we have
\[
\|y^n - y\|^2_{L^2(\Omega)} + |b^n|_{\sigma^n,s}^2 \leq \|w - y\|^2_{L^2(\Omega)} + |v|_{\sigma^n,s}^2,
\]
since \((y^n, b^n)\) solves the lower level problem \(P_{ad}(\sigma)\) for \(\sigma = \sigma^n\). Using the KKT conditions in Theorem 4.3, we deduce that there exists a sequence \((\lambda^n)\) in \(H^1_0(\Omega)\) such that for every \(n \in \mathbb{N}\) and every \(1 \leq i \leq N\)
\[
\int_{\Omega} (\mathcal{L} \sigma^n b^n_i)v + \int_{\Omega} v(x)\lambda^n(x)\partial_i y^n(x) \, dx = 0, \quad \text{for all } v \in H^s(\Omega),
\]
\[
\int_{\Omega} (y^n - y\delta)w + \nabla \lambda^n \cdot \nabla w + b^n \cdot (\lambda^n \nabla w) + c\lambda^n w \, dx = 0, \quad \text{for all } w \in H^1_0(\Omega).
\]
Arguing similarly as in Corollary 2 v) one can prove that \((\lambda^n)\) is bounded in \(H^1_0(\Omega)\).

Let \(p^n_i(x) := \lambda^n(x)\partial_i y^n(x)\). We now distinguish between two cases.
1. If $N \in \{1, 2\}$, then $(p^n_i)$ is bounded in $L^{2-\delta}(\Omega)$ for arbitrary $0 < \delta < 1$.
2. If $N = 3$ and $s \in (1/2, 1)$, then $(p^n_i)$ is bounded in $L^{3/2}(\Omega)$ and $H^s(\Omega)$ is compactly embedded in $L^3(\Omega)$.

In either case it follows from Theorem 2.5 that

$$\lim_{n \to \infty} |b^n_i|_{\sigma^n,s} = |b_i|_{\sigma,s} \quad \text{for every } 1 \leq i \leq N. \quad (46)$$
This implies that
\[
\|y - y_\delta\|_{L^2(\Omega)}^2 + |b|_{\sigma,s}^2 = \lim_{n \to \infty} \|y^n - y_\delta\|_{L^2(\Omega)}^2 + |b^n|_{\sigma,s}^2 \\
\leq \lim_{n \to \infty} \|w - y_\delta\|_{L^2(\Omega)}^2 + |v|_{\sigma,s}^2 = \|w - y_\delta\|_{L^2(\Omega)}^2 + |v|_{\sigma,s}^2,
\]
for all \((w, v) \in F_{\text{adv}}\). This shows that \((y, b)\) solves \(F_{\text{adv}}(\sigma)\), which proves weak*-to-weak stability. If additionally \(\sigma^n \to \sigma\) in \(L^\infty((0, d))\), then \(\lim_{n \to \infty} |b^n|_{\sigma,s}^2 = |b|_{\sigma,s}^2\), which follows from the estimate
\[
|b^n_i|_{\sigma,s}^2 - |b_i|_{\sigma,s}^2 = |b^n_i|_{\sigma,s}^2 - |b^n_i|_{\sigma^n,s}^2 + |b^n_i|_{\sigma^n,s}^2 - |b_i|_{\sigma,s}^2 \\
\leq |b^n_i - b_i|_{H^s(\Omega)} \|\sigma^n - \sigma\|_{L^\infty((0, d))} + |b^n_i|_{\sigma^n,s}^2 - |b_i|_{\sigma,s}^2.
\]
Since \(|\cdot|_{\sigma,s}\) is a uniformly convex norm, weak convergence of \(|b^n_i|_{\sigma,s}\) to \(b_i\) in \(H^s(\Omega)\) and \(\lim_{n \to \infty} |b^n_i|_{\sigma,s}^2 = |b|_{\sigma,s}^2\) are sufficient to conclude that \(b^n \to b\) in \(H^s(\Omega)\), see [6, Proposition 3.32]. Since by Corollary 2 the solution operator \(b \mapsto y(b)\) of the state equation is continuous as mapping from \(H^s(\Omega)\) to \(H^1_0(\Omega)\), this finishes the proof.

\[
\text{REFERENCES}
\]

[1] R. A. Adams and J. J. F. Fournier, Sobolev Spaces, Pure and Applied Mathematics (Amsterdam), 140, Elsevier/Academic Press, Amsterdam, 2003.
[2] B. Alali, K. Liu and M. Gunzburger, A generalized nonlocal vector calculus, Z. Angew. Math. Phys., 66 (2015), 2807–2828.
[3] H. Antil, E. Otárola and A. J. Salgado, Optimization with respect to order in a fractional diffusion model: Analysis, approximation and algorithmic aspects, J. Sci. Comput., 77 (2018), 204–224.
[4] H. Antil and C. N. Rautenberg, Sobolev spaces with non-Muckenhoupt weights, fractional elliptic operators, and applications, SIAM J. Math. Anal., 51 (2019), 2479–2503.
[5] K. Bredies and D. Lorenz, Mathematische Bildverarbeitung, Vieweg+Teubner Verlag, 2011.
[6] H. Brezis, Functional Analysis, Sobolev Spaces and Partial Differential Equations, Universitext, Springer, New York, 2011.
[7] H. Cartan, Differential Calculus, Hermann, Paris; Houghton Mifflin Co., Boston, Mass., 1971.
[8] E. Casas, M. Mateos and A. Rösch, Analysis of control problems of nonmonotone semilinear elliptic equations, ESAIM: COCV, 26 (2020).
[9] E. Casas and F. Tröltzsch, Second order optimality conditions and their role in PDE control, Jahresber. Dtsch. Math.-Ver., 117 (2015), 3–44.
[10] J. Chung, M. Chung and D. P. O’Leary, Designing optimal spectral filters for inverse problems, SIAM J. Sci. Comput., 33 (2011), 3132–3152.
[11] J. Chung and M. I. Español, Learning regularization parameters for general-form Tikhonov, Inverse Problems, 33 (2017), 21pp.
[12] M. D’Elia, J. C. De Los Reyes and A. Miniguano Trujillo, Bilevel parameter optimization for nonlocal image denoising models, preprint, arXiv:1912.02347v3.
[13] M. D’Elia and M. Gunzburger, The fractional Laplacian operator on bounded domains as a special case of the nonlocal diffusion operator, Comput. Math. Appl., 66 (2013), 1245–1260.
[14] 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 (2016), 464–500.
[15] J. C. De Los Reyes and C.-B. Schönlieb, Image denoising: Learning the noise model via nonsmooth PDE-constrained optimization, Inverse Probl. Imaging, 7 (2013), 1183–1214.
[16] S. Dempe, Foundations of Bilevel Programming, Nonconvex Optimization and its Applications, 61, Kluwer Academic Publishers, Dordrecht, 2002.
[17] E. Di Nezza, G. Palatucci and E. Valdinoci, Hitchhiker’s guide to the fractional Sobolev spaces, Bull. Sci. Math., 136 (2012), 521–573.
[18] Q. Du, M. Gunzburger, R. B. Lehoucq and K. Zhou, Analysis and approximation of nonlocal diffusion problems with volume constraints, SIAM Rev., 54 (2012), 667–696.
[19] L. C. Evans and R. F. Gariepy, Measure Theory and Fine Properties of Functions, Studies in Advanced Mathematics, CRC Press, Boca Raton, FL, 1992.
[20] G. B. Folland, *Real Analysis. Modern Techniques and Their Applications*, Pure and Applied Mathematics, John Wiley & Sons, Inc., New York, 1999.

[21] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Monographs and Studies in Mathematics, 24, Pitman (Advanced Publishing Program), Boston, MA, 1985.

[22] M. Gunzburger and R. B. Lehoucq, A nonlocal vector calculus with application to nonlocal boundary value problems, *Multiscale Model. Simul.*, 8 (2010), 1581–1598.

[23] E. Haber and L. Tenorio, Learning regularization functionals–A supervised training approach, *Inverse Problems*, 19 (2003), 611–626.

[24] M. Hintermüller and C. N. Rautenberg, Optimal selection of the regularization function in a weighted total variation model. Part I: Modelling and theory, *J. Math. Imaging Vision*, 59 (2017), 498–514.

[25] G. Holler, K. Kunisch and R. C. Barnard, A bilevel approach for parameter learning in inverse problems, *Inverse Problems*, 34 (2018), 28pp.

[26] K. Ito and K. Kunisch, *Lagrange Multiplier Approach to Variational Problems and Applications*, Advances in Design and Control, 15, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2008.

[27] K. Ito and K. Kunisch, The primal-dual active set method for nonlinear optimal control problems with bilateral constraints, *SIAM J. Control Optim.*, 43 (2004), 357–376.

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

[29] 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*, Lecture Notes in Comput. Sci., 9087, Springer, Cham, 2015, 654–665.

[30] M. Ulbrich and S. Ulbrich, *Nichtlineare Optimierung*, Mathematik Kompakt, Birkhäuser, Basel, 2012.

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

Received February 2020; revised September 2020.

E-mail address: gernot.holler@uni-graz.at
E-mail address: karl.kunisch@uni-graz.at