A theoretical framework for supervised learning from regions

Giorgio Gnecco

DIST–University of Genova
Via Opera Pia 13, 16145 Genova–Italy

Marco Gori

DII–University of Siena
Via Roma, 56, 53100 Siena–Italy

Stefano Melacci*

DII–University of Siena
Via Roma, 56, 53100 Siena–Italy

Marcello Sanguineti*

DIST–University of Genova
Via Opera Pia 13, 16145 Genova–Italy

Abstract

Supervised learning is investigated, when the data are represented not only by labeled points but also labeled regions of the input space. In the limit case, such regions degenerate to single points and the proposed approach changes back to the classical learning context. The adopted framework entails the minimization of a functional obtained by introducing a loss function that involves such regions. An additive regularization term is expressed via differential operators that model the smoothness properties of the desired input/output relationship. Representer theorems are given, proving that the optimization problem associated to learning from labeled regions has a unique solution, which takes on the form of a linear combination of kernel functions determined by the differential operators together with the regions themselves. As a relevant situation, the case of regions given by multi-dimensional intervals (i.e., "boxes") is investigated, which models prior knowledge expressed by logical propositions.

Keywords: supervised learning; kernel machines; propositional rules; variational calculus; infinite-dimensional optimization; representer theorems.

*Corresponding author
Email addresses: Giorgio.Gnecco@dist.unige.it (Giorgio Gnecco), marco@dii.unisi.it (Marco Gori), mela@dii.unisi.it (Stefano Melacci), marcello@dist.unige.it (Marcello Sanguineti)

1. Introduction

The classical supervised learning framework is based on a collection of ℓ labeled points, $\mathcal{L} = \{(x_\kappa, y_\kappa),\ \kappa = 1, \ldots, \ell\}$, where $x_\kappa \in \mathcal{X} \subseteq \mathbb{R}^d$ and $y_\kappa \in \{-1, 1\}$. We consider the situation in which supervised learning exploits not only labeled points but also $\ell_\mathcal{X}$ labeled regions $\mathcal{L}_\mathcal{X} = \{(\mathcal{X}_\kappa, y_\kappa),\ \kappa = 1, \ldots, \ell_\mathcal{X}\}$ of the input space, where $\mathcal{X}_\kappa \in 2^\mathcal{X}$ and $y_\kappa \in \{-1, 1\}$. In the limit case such regions degenerate to single points, so we focus on a fairly general context in which there is no distinction between the supervised entities and we deal with $\ell_t := \ell + \ell_\mathcal{X}$ labeled pairs. This framework and its potential impact in real-world applications has been investigated in different contexts (see [1] and the references therein).

A seminal work in this respect is [2], where it was proposed to embed labeled polyhedral sets into Support Vector Machines (SVMs). The corresponding model, called Knowledge-based SVM (KSVM), has been the subject of a number of various investigations [3, 4, 5, 6]. A particularly relevant situation corresponds to regions given by multi-dimensional intervals (i.e., "boxes") $\mathcal{X}_\kappa = \{x \in \mathbb{R}^d : \ x^i \in [a^i_\kappa, b^i_\kappa],\ i = 1, \ldots, d\}$, where $a_\kappa, b_\kappa \in \mathbb{R}^d$ collect the lower and upper bounds, respectively. The pair $(\mathcal{X}_\kappa, y_\kappa)$ formalizes the knowledge provided by a supervisor in terms of logical propositions of the form $\forall x \in \mathbb{R}^d,\ \bigwedge_{i=1}^d \big((x^i \geq a^i_\kappa) \wedge (x^i \leq b^i_\kappa)\big) \Rightarrow \ \text{class}(y_\kappa)\,.$

In [7], the problem of learning was extended by taking into account the supervision on multi-dimensional intervals of the input space, which model prior knowledge expressed by logical propositions. The effectiveness of such an approach was evaluated therein via simulations on real-world problems of medical diagnosis and image categorization. Taking the hint from the numerical experiments presented in [7], in this paper we give theoretical insights into the learning paradigm proposed therein.

We formulate the problem of learning via supervision on input regions by introducing a loss function that involves them and adopting the regularization framework proposed in [8]. Each region $\mathcal{X}_\kappa$ is associated with its characteristic function $1_{\mathcal{X}_\kappa}(x)$ and its normalized form $\hat{1}_{\mathcal{X}_\kappa}(x) := 1_{\mathcal{X}_\kappa}(x) / \int_\mathcal{X} 1_{\mathcal{X}_\kappa}(x)dx$ degenerates to the Dirac distribution $\delta(x - x_\kappa)$ in the case in which $\mathcal{X}_\kappa = \{x_\kappa\}$. We model the corresponding learning problem as the minimization, over an *infinite-dimensional space* (whose elements are the admissible solutions to the supervised learning task) of a functional, called *regularized functional risk*, that consists of two terms. The first term enforces closeness to the labeled data (regions and points), whereas the second one, called *regularization term*, expresses requirements on the global behavior of the desired input/output functional relationship. The trade-off between between such terms is achieved by a weight parameter, as typically done in Tikhonov's regularization [9].

We express the regularization term via differential operators, following the line of research proposed in [8]. The loss term results from the following two contributions: one from regions with non-null Lebesgue measure and another

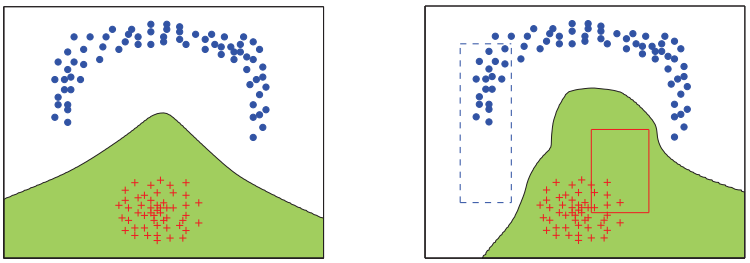
one that originates from points. As the minimization of such a functional entails a difficult infinite-dimensional problem, we also consider learning modeled as a more affordable variational task obtained by replacing the functional risk with an *average risk*. We show that in this case the infinite-dimensional optimization collapses to a finite-dimensional one. As ambient spaces we consider Sobolev spaces of orders guaranteeing that they are made up of continuous functions, in such a way that the learning functionals are well-defined when the regions degenerate to points.

Under the hypothesis that the Greeen's function of the regularization operator is the kernel of a Hilbert spaces of a special type, called *reproducing kernel Hilbert space*[1] (*RKHS*) [14], we prove new *representer theorems* (see, e.g., [15, p. 42], [16, 17, 18, 19]), showing that the minimization problem has a unique solution, which takes on the form of a linear combination of kernel functions determined by the differential operators together with the labeled regions. So, the solution to the regularized problem of learning from regions does not lead to the kernel expansion on the available data points and the kernel is no longer the Green's function of the associated operator, as instead it happens from classical results of this kind (see [20] and [21, p. 94]).

As a meaningful learning case, we investigate regions given by multidimensional intervals (i.e., "boxes"), which originates the *box kernels*. Figure 1 shows an example of box kernels in the case of a non-linear kernel-machine classifier. First, the classifier is trained on a 2-class data set of points, and the separating hyperplane is depicted in Figure 1 (a). Then, a supervision is given on two space regions bounded by multi-dimensional intervals; Figure 1 (b) shows the resulting separation boundary. The box kernel allows the classifier to embed knowledge on labeled space regions, whereas classical kernels are designed to operate merely on points.

The paper is organized as follows. In Section 2 we state the problem of learning from labeled regions and/or labeled points as the infinite-dimensional minimization of the functional risk. We investigate existence and uniqueness of its solution on the Sobolev space of functions that are square-integrable together with their partial derivatives up to a suitable order. Section 3 provides representer theorems for such a solution and considers the learning problem modeled via the minimization of the average risk. The particular case of regions given by multi-dimensional intervals (i.e., "boxes") is addressed in Section 4. In Section 5 we compare the two learning problems associated with the minimizations of the functional risk and the average risk, respectively. To make the paper self-contained, two appendices are provided on RKHSs and functionals. Preliminary results were presented in [7].

[1] Such spaces were introduced into applications closely related to learning by Parzen [10] and Wahba [11], and into learning theory by Cortes and Vapnik [12] and Girosi [13].



(a) (b)

Figure 1: (a) A two-class data set and the separating surface learned by a non-linear kernel-machine classifier. Examples of class 1 are represented by red crosses, whereas examples of class 2 are drawn with blue dots. (b) The data collection is augmented with two labeled space regions (bounded by the blue-dotted rectangle in the case of class 1, by the red-solid box for class 2). A kernel machine is trained using a box kernel, which allows the classifier to learn from the whole data collection (points and regions). The resulting class-separation boundary is shown.

2. Learning from labeled regions and points

We formulate the problem of learning from labeled regions and/or labeled points in a unique framework, where each point corresponds to a singleton. Given a labeled set $\mathcal{X}_\kappa$, the characteristic function $1_{\mathcal{X}_\kappa}(x)$ associated with it is identically 1 when $x \in \mathcal{X}_\kappa$, otherwise it is identically 0. Denoting by $\mathrm{vol}(\mathcal{X}_\kappa) = \int_{\mathbb{R}^d} 1_{\mathcal{X}_\kappa}(x)dx$ the measure of the set, the normalized characteristic function is $\hat{1}_{\mathcal{X}_\kappa}(x) := 1_{\mathcal{X}_\kappa}(x)/\mathrm{vol}(\mathcal{X}_\kappa)$. When the region degenerates to a single point x_κ we denote by $\hat{1}_{\mathcal{X}_\kappa}(x)$ the Dirac delta $\delta(x - x_\kappa)$.

Let $w : \mathcal{X} \to \mathbb{R}^+$ be a continuous weight function (e.g., proportional to the probability density $p : \mathcal{X} \to \mathbb{R}^+$ of the inputs), $V : \mathbb{R}^2 \to \mathbb{R}^+$ a convex and differentiable loss function, $\lambda > 0$ a regularization parameter, and $P := (P^0, \ldots, P^{r-1})$ a vector of r finite-order differential operators of maximum order of derivation k and with constant coefficients, with formal adjoint $P^\star$ [22, 23]. Adopting the framework described in [8], we formulate the problem of learning from labeled regions as the minimization on a suitable class of functions $\mathcal{F}$ of the functional

$$R(f) := \sum_{\kappa \in \mathbb{N}_{\ell_t}} \int_{\mathbb{R}^d} V(y_\kappa, f(x)) \cdot w(x) \cdot \hat{1}_{\mathcal{X}_\kappa}(x)dx + \frac{\lambda}{2}\|Pf\|^2 , \tag{1}$$

where $\mathbb{N}_m$ denotes the set of the first m positive integers,

$$\|Pf\|^2 := (Pf, Pf) = (P^\star Pf, f) = (Lf, f) ,$$

$$(f, g) := \int_{\mathbb{R}^d} f(x) \cdot g(x)dx ,$$

and $L := P^\star P$ (which has $2k$ as its maximum order of derivation). We call $R(\cdot)$ in equation (1) the *regularized functional risk*. When all the regions degenerate to points, we get $\ell_t = \ell$, $\hat{1}_{\mathcal{X}_\kappa}(x) = \delta(x - x_k)$, and equation (1) becomes

$$R(f) := \sum_{\kappa \in \mathbb{N}_\ell} V(y_\kappa, f(x_\kappa)) \cdot w(x_\kappa) + \frac{\lambda}{2} \|Pf\|^2 , \tag{2}$$

which is the classical form of the regularized risk [24] when supervision is performed on labeled points.

We search for the minimizer f° in the Sobolev space $\mathcal{F} = \mathcal{W}^{k,2}$, i.e., the subset of $\mathcal{L}^2$ whose functions have square-integrable weak partial derivatives up to the order k. The loss term in equation (1) can be considered as resulting from the following two contributions: $\sum_{\kappa \in \mathbb{N}_{\ell_\mathcal{X}}} \int_{\mathbb{R}^d} V(y_\kappa, f(x)) \cdot w(x) \cdot \hat{1}_{\mathcal{X}_\kappa}(x) dx$, coming from a region with non-null Lebesgue measures $\mathrm{vol}(\mathcal{X}_\kappa)$, and $\sum_{\kappa \in \mathbb{N}_\ell} w(x_\kappa) \cdot V(y_\kappa, f(x_\kappa))$, which originates from points. Note also that $V(y_\kappa, f(x))$ measures the error for any $x \in \mathcal{X}_\kappa$ of the set with respect to the target y_κ (which is the same for all points in the same set $\mathcal{X}_\kappa$).

We make the following assumption. For the definition and the basic properties of *Reproducing Kernel Hilbert Spaces* (*RKHSs*) and their role in learning, see Appendix A.

Assumption 1. *For a positive integer $k > \frac{d}{2}$, let $(Pf, Pf)^{1/2}$ be a norm on $\mathcal{W}^{k,2}$, the Green's function $g(\cdot, \cdot)$ of L a (plain) kernel of a RKHS, and for each fixed $\zeta \in \mathcal{X}$ let $g(\cdot, \zeta) \in \mathcal{W}^{k,2}$.*

The requirement $k > \frac{d}{2}$ guarantees that the ambient space $\mathcal{F} = \mathcal{W}^{k,2}$ is made up of continuous functions (thanks to the Sobolev's Embedding Theorem [25, Chapter 4]), so the term $\int_{\mathbb{R}^d} V(y_\kappa, f(x)) \cdot w(x) \cdot \hat{1}_{\mathcal{X}_\kappa}(x) dx$ in the functional (1) is well-defined when $\mathcal{X}_\kappa$ degenerates to a single point. Moreover, it also guarantees the continuity of the functional. The other conditions of Assumption 1 shall be exploited in the proof of the next Theorem 1. For $\sigma > 0$, an example of a regularization operator L that satisfies Assumption 1 (for a suitable P) is $L = (\sigma^2 I - \nabla^2)^k$, which originates the *Sobolev spline kernel* [26]. There is no regularization operator L associated to a vector of differential operators P for linear and polynomial kernels of RKHSs. See [23] for other examples of operators P and L that satisfy Assumption 1.

3. Representer theorems

The following result guarantees existence and uniqueness of the minimizer to the functional (1) and provides an expression for such a minimizer. See Appendix B for some definitions and notations used in the proof.

Theorem 1 (Representer theorem for learning with the risk R). *Let $\mathcal{F} = \mathcal{W}^{k,2}$ and Assumption 1 be satisfied. Then the functional (1) has a unique global minimizer f°, which can be expressed as*

$$f^\circ(x) = -\frac{1}{\lambda} \sum_{\kappa \in \mathbb{N}_{\ell_t}} \int_{\mathbb{R}^d} g(x,\zeta) \cdot V_f' \left(y_\kappa, f^\circ(\zeta)\right) \cdot w(\zeta) \cdot \hat{1}_{\mathcal{X}_\kappa}(\zeta) d\zeta \,, \tag{3}$$

where $V_f'(y,f) := \partial V(y,f)/\partial f$ is the partial derivative of the loss function $V(y,f)$ with respect to its second argument.

PROOF. The assumptions that the loss function V is convex and continuous, that $k > \frac{d}{2}$ and that $(Pf, Pf)^{1/2}$ is a norm on $\mathcal{W}^{k,2}$ imply that the functional (1) is uniformly convex with modulus of convexity $\frac{\lambda}{2}\, t^2$. Then, the existence and the uniqueness of the global minimizer f° follow by the fact that (1) is uniformly convex and $\mathcal{W}^{k,2}$ is a Hilbert space.

We show that f° satisfies the Euler-Lagrange equation

$$Lf^\circ(x) = -\frac{1}{\lambda} \sum_{\kappa \in \mathbb{N}_{\ell_t}} V_f' \left(y_\kappa, f^\circ(x)\right) \cdot w(x) \cdot \hat{1}_{\mathcal{X}_\kappa}(x) \tag{4}$$

and that (4) has a unique solution. Let $\mathcal{C}_0^\infty \subset \mathcal{W}^{k,2}$ denote the set of compactly-supported, infinitely-differentiable, and real-valued functions on $\mathbb{R}^d$. For every $\alpha \in \mathbb{R}$ and $\varphi \in \mathcal{C}_0^\infty$, we have

$$\begin{aligned} R(f^\circ + \alpha\varphi) - R(f^\circ) &= \lambda \left(Pf, P(\alpha\varphi)\right) + \frac{\lambda}{2} \left(P(\alpha\varphi), P(\alpha\varphi)\right) \\ &+ \sum_{\kappa \in \mathbb{N}_{\ell_t}} \int_{\mathbb{R}^d} V(y_\kappa, (f^\circ + \alpha\varphi)(x)) \cdot w(x) \cdot \hat{1}_{\mathcal{X}_\kappa}(x) dx \\ &- \sum_{\kappa \in \mathbb{N}_{\ell_t}} \int_{\mathbb{R}^d} V(y_\kappa, f^\circ(x)) \cdot w(x) \cdot \hat{1}_{\mathcal{X}_\kappa}(x) dx \,. \end{aligned}$$

As the loss function V is differentiable, we get

$$\begin{aligned} R(f^\circ + \alpha\varphi) - R(f^\circ) &= \lambda\alpha \left(Pf^\circ, P\varphi\right) \\ &+ \sum_{\kappa \in \mathbb{N}_{\ell_t}} \int_{\mathbb{R}^d} V_f'(y_\kappa, f^\circ(x)) \cdot w(x) \cdot \hat{1}_{\mathcal{X}_\kappa}(x) \cdot \alpha\varphi(x) dx + o(\alpha) \,. \end{aligned}$$

For every α with $|\alpha|$ sufficiently small we have $R(f^\circ + \alpha\varphi) - R(f^\circ) \geq 0$. Hence

$$\begin{aligned} \lim_{\alpha \to 0} \frac{R(f^\circ + \alpha\varphi) - R(f^\circ)}{\alpha} &= \lambda \left(Pf^\circ, P\varphi\right) \\ &+ \sum_{\kappa \in \mathbb{N}_{\ell_t}} \int_{\mathbb{R}^d} V_f'(y_\kappa, f^\circ(x)) \cdot w(x) \cdot \hat{1}_{\mathcal{X}_\kappa}(x) \cdot \varphi(x) dx \\ &= 0 \,. \end{aligned} \tag{5}$$

As $\varphi \in \mathcal{C}_0^\infty$, by the Green's formula (see, e.g., [27, Proposition 5.6.2]) we obtain

$$(Pf^\circ, P\varphi) = \int_{\mathbb{R}^d} (P^\star P f^\circ(x)) \cdot \varphi(x)\, dx = \int_{\mathbb{R}^d} (Lf^\circ(x)) \cdot \varphi(x)\, dx$$

and so

$$\int_{\mathbb{R}^d} \left(\lambda L f^\circ(x) + \sum_{\kappa \in \mathbb{N}_{\ell_t}} V_f'(y_\kappa, f^\circ(x)) \cdot w(x) \cdot \hat{1}_{\mathcal{X}_\kappa}(x) \right) \cdot \varphi(x)\, dx = 0\,. \quad (6)$$

As $\varphi \in \mathcal{C}_0^\infty$ is arbitrary and with a suitable topology $\mathcal{C}_0^\infty$ coincides with the space of test functions used in distribution theory [28], it follows that (6) is equivalent to (4), where equality between the two sides of the equation has to be interpreted in the sense of distributions on $\mathbb{R}^d$.

Now, we now show that any solution $\tilde{f}^\circ \in \mathcal{W}^{k,2}$ to the equation (4) coincides with f°. Indeed, for every $\alpha \in [0,1]$ one has

$$R(\tilde{f}^\circ + \alpha\varphi) \leq \alpha R(\tilde{f}^\circ + \varphi) + (1-\alpha) R(\tilde{f}^\circ)\,,$$

so

$$R(\tilde{f}^\circ + \varphi) \geq \frac{R(\tilde{f}^\circ + \alpha\varphi) - R(\tilde{f}^\circ)}{\alpha} + R(\tilde{f}^\circ)\,.$$

Then, taking the limit as $\alpha \to 0^+$ and exploiting (5), one obtains $R(\tilde{f}^\circ + \varphi) \geq R(\tilde{f}^\circ)$ for every $\varphi \in \mathcal{C}_0^\infty$. By the density of $\mathcal{C}_0^\infty$ in $\mathcal{W}^{k,2}$ [25, Corollary 3.23] and the continuity of the functional (1), it follows that $\tilde{f}^\circ$ is a global minimizer, so it coincides with f° by the above-mentioned uniqueness.

Let us now derive the solution to equation (4), from which we get (3). Recall that by the definition of the Green's function, one has

$$Lg(x,\zeta) = \delta(x-\zeta), \quad \text{for each } \zeta \in \mathbb{R}^d\,. \quad (7)$$

Note that the right-hand side of equation (4) can be written as

$$-\frac{1}{\lambda} \sum_{\kappa \in \mathbb{N}_{\ell_t}} \int_{\mathbb{R}^d} \delta(x-\zeta) \cdot V_f'\left(y_\kappa, f^\circ(\zeta)\right) \cdot w(\zeta) \cdot \hat{1}_{\mathcal{X}_\kappa}(\zeta) d\zeta\,. \quad (8)$$

Hence, it follows by (7) and (8) that the general solution of (4) is

$$f^\circ(x) = -\frac{1}{\lambda} \sum_{\kappa \in \mathbb{N}_{\ell_t}} \int_{\mathbb{R}^d} g(x,\zeta) \cdot V_f'\left(y_\kappa, f^\circ(\zeta)\right) \cdot w(\zeta) \cdot \hat{1}_{\mathcal{X}_\kappa}(\zeta) d\zeta + h(x)\,, \quad (9)$$

where h is a generic element of $\operatorname{Ker} L$ (the null space of the operator L, interpreted as an operator acting on functions in $\mathcal{W}^{k,2}$). Finally, we show that the assumption that $(Pf, Pf)^{1/2}$ is a norm on $\mathcal{W}^{k,2}$ implies $\operatorname{Ker} L = \{0\}$. Let

$f \in \operatorname{Ker} L$. Then $Lf = P^\star Pf = 0$, so $(P^\star Pf, \varphi) = (Pf, P\varphi) = 0$ for all $\varphi \in \mathcal{C}_0^\infty$. Taking the limit for $\varphi \to f$ in $\mathcal{W}^{k,2}$ and exploiting the continuity of the operator P on $\mathcal{W}^{k,2}$, we obtain $(Pf, Pf) = 0$. Then $f = 0$, since $(Pf, Pf)^{1/2}$ is a norm on $\mathcal{W}^{k,2}$. So, we conclude that $h = 0$ in (9) and the representation (3) holds.

Remark 1. Equation (3) is an integral equation in the unknown f°. In the particular case in which the loss function V is quadratic, (3) is a linear Fredholm integral equation of the second kind.

Remark 2. Note that Theorem 1 holds also when the loss function V is continuous but differentiable only on a portion of its domain, provided that the term $V\left(y_\kappa, f^\circ(x)\right)$ does not intersect the region of non-differentiability of V when x varies in the set $\mathcal{X}_\kappa$. An example of such a loss function is the linear hinge loss $V(y_\kappa, f(x)) = \max(0, 1 - y_\kappa f(x))$. In such a case, the assumption of empty intersection with the region of non-differentiability of V in the set $\mathcal{X}_\kappa$ can be equivalently stated as follows: inside $\mathcal{X}_\kappa$, one has

1. $y_\kappa f^\circ(x) > 1$, for all $x \in \mathcal{X}_\kappa$ or
2. $y_\kappa f^\circ(x) < 1$, for all $x \in \mathcal{X}_\kappa$.

Note that, by the continuity of the linear hinge loss and of f°, the assumption $y_\kappa f^\circ(x) \neq 1$ inside the set $\mathcal{X}_\kappa$ implies that either item 1 or 2 above holds.

Unfortunately, in general the minimizer f° cannot be easily computed by formula (3), as in the integral representation the terms $V_f'\left(y_\kappa, f^\circ(\zeta)\right)$ are not known a-priori. To simplify the problem, in the following we replace the functional (1) with

$$R_m(f) := \sum_{\kappa \in \mathbb{N}_{\ell_t}} V(y_\kappa, m_{\mathcal{X}_\kappa}(f)) + \frac{\lambda}{2}\|Pf\|^2, \tag{10}$$

where the term $m_{\mathcal{X}_\kappa}(f) := \int_{\mathbb{R}^d} f(x) \cdot w(x) \cdot \hat{1}_{\mathcal{X}_\kappa}(x) dx$ is the average weighted value of f over $\mathcal{X}_\kappa$ with respect to the weight function $w(x)$. We call $R_m(\cdot)$ *regularized average risk*. Of course, when $\mathcal{X}_\kappa = \{x_\kappa\}$ we get $m_{\mathcal{X}_\kappa}(f) = f(x_\kappa) \cdot w(x_\kappa)$.

The minimization of $R_m(\cdot)$ is a more affordable variational problem, as shown in the following theorem.

Theorem 2 (Representer theorem for learning with the risk R_m). *Let $\mathcal{F} = \mathcal{W}^{k,2}$ and Assumption 1 be satisfied. Then the functional (10) has a unique global minimizer f_m°, which can be expressed as*

$$f_m^\circ(x) = \sum_{\kappa \in \mathbb{N}_{\ell_t}} \alpha_\kappa \beta(\mathcal{X}_\kappa, x), \tag{11}$$

where $\beta(\mathcal{X}_\kappa, x) := \int_{\mathbb{R}^d} g(x, \zeta) \cdot w(\zeta) \cdot \hat{1}_{\mathcal{X}_\kappa}(\zeta) d\zeta$ and $\alpha_\kappa := -V_f'(y_\kappa, m_{\mathcal{X}_\kappa}(f_m^\circ))/\lambda$ are scalar values, $\kappa \in \mathbb{N}_{\ell_t}$.

PROOF. Proceeding as in the proof of Theorem 1, there exists a unique global minimizer f_m°. It satisfies the Euler-Lagrange equation

$$Lf_m^\circ(x) = -\frac{1}{\lambda} \sum_{\kappa \in \mathbb{N}_{\ell_t}} V_f' \left(y_\kappa, m_{\mathcal{X}_\kappa}(f)\right) \cdot w(x) \cdot \hat{1}_{\mathcal{X}_\kappa}(x)\,.$$

As $\operatorname{Ker} L = \{0\}$, its solution is

$$\begin{aligned} f_m^\circ(x) &= -\frac{1}{\lambda} \sum_{\kappa \in \mathbb{N}_{\ell_t}} \int_{\mathbb{R}^d} g(x,\zeta) \cdot V_f' \left(y_\kappa, m_{\mathcal{X}_\kappa}(f)\right) \cdot w(\zeta) \cdot \hat{1}_{\mathcal{X}_\kappa}(\zeta) d\zeta \\ &= \sum_{\kappa \in \mathbb{N}_{\ell_t}} \alpha_\kappa \beta(\mathcal{X}_\kappa, x)\,. \end{aligned}$$

Theorem 1 proves how the optimal solution is expressed in the case of the regularized functional risk R. Theorem 2 focuses on the regularized average risk R_m; it allows one to represent the solution by means of a finite numbers of coefficients. This makes the optimization of the learning problem affordable, opening the road to applications based on the described learning framework.

In Section 5, we shall investigate the relationships between the two minimization problems associated with the functionals (1) and (10), resp.

Remark 3. Likewise (3), equation (11) is an integral equation in the unknown f_m°. For a quadratic loss function V, it is a Fredholm integral equation of the second kind.

If we separate the contributions coming from points and sets, the representation (11) can be written as

$$f_m^\circ(x) = \sum_{\kappa \in \mathbb{N}_\ell} \alpha_\kappa g(x_\kappa, x) + \sum_{\kappa \in \mathbb{N}_{\ell_\mathcal{X}}} \alpha_\kappa \beta(\mathcal{X}_\kappa, x).$$

For $\kappa_1, \kappa_2 \in \mathbb{N}_{\ell_t}$, let

$$\mathcal{K}(\mathcal{X}_{\kappa_1}, \mathcal{X}_{\kappa_2}) := \int_{\mathbb{R}^d} \beta(\mathcal{X}_{\kappa_1}, x) \cdot w(x) \cdot \hat{1}_{\mathcal{X}_{\kappa_2}}(x) dx. \tag{12}$$

The following proposition gives insights into the cases where either $\mathcal{X}_{\kappa_1}$ or $\mathcal{X}_{\kappa_2}$ degenerate into points. It extends [7, Proposition 1] to the case of possibly not uniform weight functions w (an interesting case is a mixture of Gaussians: $w(x) = \sum_{k=1}^{k_{\max}} \hat{\eta}_k \exp\left(-\frac{\|x-\hat{x}_k\|_2^2}{2\hat{\sigma}_k^2}\right)$, where $k_{\max} \in \mathbb{N}$, $\hat{\eta}_k \geq 0$, $\hat{x}_k \in \mathbb{R}^d$, and $\hat{\sigma}_k > 0$ are parameters). For a uniform weight function w, Proposition 1 reduces to [7, Proposition 1].

Proposition 1. *The following hold.*

i. $\mathcal{K}(\mathcal{X}_{\kappa_1}, \{x_{\kappa_2}\}) = w(x_2) \cdot \beta(\mathcal{X}_{\kappa_1}, x_{\kappa_2})\,;$

ii. $\mathcal{K}(\{x_{\kappa_1}\}, \{x_{\kappa_2}\}) = w(x_1) \cdot w(x_2) \cdot g(x_{\kappa_1}, x_{\kappa_2})\,.$

PROOF. (i) If $\mathcal{X}_{\kappa_2} = \{x_{\kappa_2}\}$ then $\hat{1}_{\mathcal{X}_{\kappa_2}}(x) = \delta(x - x_{\kappa_2})$. So, the statement follows by (12).
(ii) In this case $\hat{1}_{\mathcal{X}_{\kappa_1}}(x) = \delta(x - x_{\kappa_1})$. So, we conclude by the definition of $\beta(\mathcal{X}_{\kappa_1}, x)$ and the symmetry of $g(\cdot, \cdot)$.

Exploiting the definition (12) of the function $\mathcal{K}$ we can devise an efficient algorithmic scheme based on the collapsing to a finite dimension of the infinite-dimensional minimization of the functional $R_m(\cdot)$ (see (10)) over the Sobolev space $\mathcal{W}^{k,2}$. This is investigated in the next theorem, which holds under the same hypotheses of Theorem 2.

Theorem 3. *Let $\mathcal{F} = \mathcal{W}^{k,2}$ and Assumption 1 be satisfied. Then $R_m(f_m^\circ) = \hat{\mathcal{R}}_m(\alpha)$, where α is a vector of dimension ℓ_t and*

$$\begin{aligned} \hat{\mathcal{R}}_m(\alpha) &= \sum_{\kappa_1 \in \mathbb{N}_{\ell_t}} V(y_{\kappa_1}, \sum_{\kappa_2 \in \mathbb{N}_{\ell_t}} \alpha_{\kappa_2} \mathcal{K}(\mathcal{X}_{\kappa_2}, \mathcal{X}_{\kappa_1})) \\ &+ \frac{\lambda}{2} \sum_{\kappa_1, \kappa_2 \in \mathbb{N}_{\ell_t}} \alpha_{\kappa_1} \alpha_{\kappa_2} \mathcal{K}(\mathcal{X}_{\kappa_1}, \mathcal{X}_{\kappa_2}) \,. \end{aligned}$$

PROOF. By plugging the expression (11) of f_m° into (10) and exploiting the definition of the Green's function $g(\cdot, \cdot)$, we get

$$\begin{aligned} R_m(f_m^\circ) &= \sum_{\kappa_1 \in \mathbb{N}_{\ell_t}} V\left(y_{\kappa_1}, \int_{\mathbb{R}^d} \sum_{\kappa_2 \in \mathbb{N}_{\ell_t}} \alpha_{\kappa_2} \beta(\mathcal{X}_{\kappa_2}, x) \cdot w(x) \cdot \hat{1}_{\mathcal{X}_{\kappa_1}}(x) dx \right) \\ &\quad + \frac{\lambda}{2} \left(\sum_{\kappa_1 \in \mathbb{N}_{\ell_t}} \alpha_{\kappa_1} \beta(\mathcal{X}_{\kappa_1}, x), L\left(\sum_{\kappa_2 \in \mathbb{N}_{\ell_t}} \alpha_{\kappa_2} \beta(\mathcal{X}_{\kappa_2}, x) \right) \right) \\ &= \sum_{\kappa_1 \in \mathbb{N}_{\ell_t}} V\left(y_{\kappa_1}, \sum_{\kappa_2 \in \mathbb{N}_{\ell_t}} \alpha_{\kappa_2} \int_{\mathbb{R}^d} \beta(\mathcal{X}_{\kappa_2}, x) \cdot w(x) \cdot \hat{1}_{\mathcal{X}_{\kappa_1}}(x) dx \right) \\ &\quad + \frac{\lambda}{2} \sum_{\kappa_1, \kappa_2 \in \mathbb{N}_{\ell_t}} \alpha_{\kappa_1} \alpha_{\kappa_2} \left(\beta(\mathcal{X}_{\kappa_1}, x), w(x) \cdot \hat{1}_{\mathcal{X}_{\kappa_2}}(x) \right) . \end{aligned} \tag{13}$$

The statement follows by (12).

According to Theorems 2 and 3, in order to determine f_m° it is sufficient to minimize the function $\hat{R}_m(\alpha)$. Note that for a quadratic loss function V, $\hat{R}_m(\alpha)$ is quadratic, too.

4. Learning with box kernels

The function $\mathcal{K}(\cdot, \cdot)$ defined in (12) originates from the (plain) kernel $g(\cdot, \cdot)$; its arguments $\mathcal{X}_{\kappa_1}$ and $\mathcal{X}_{\kappa_2}$ can be space regions or points. When the regions

are multi-dimensional intervals (i.e., "boxes"), the function $\mathcal{K}(\cdot,\cdot)$ is called the *box kernel* associated with $g(\cdot,\cdot)$. Boxes formalize the type of knowledge that we introduced in Section 1.

The box kernel can be plugged in every existing kernel-based classifier, allowing it to process at the same time labeled points and labeled box regions with nonzero volume, without any modification to the learning algorithm. The function $\mathcal{K}(\cdot,\cdot)$ inherits a number of properties from the kernel $g(\cdot,\cdot)$.

Proposition 2. *Let $\mathbb{K} \in \mathbb{R}^{\ell_t,\ell_t}$ be the Gram matrix associated with the function $\mathcal{K}(\cdot,\cdot)$, i.e., the matrix with entries $\mathbb{K}_{\kappa_1,\kappa_2} := \mathcal{K}(\mathcal{X}_{\kappa_1},\mathcal{X}_{\kappa_2})$. If g is the kernel of a RKHS, then $\mathbb{K}$ is a positive semidefinite matrix.*

PROOF. For a box $\mathcal{X}_\kappa$, let $\mathrm{vol}(\mathcal{X}_\kappa) := \prod_{i=1}^{d} |a_\kappa^i - b_\kappa^i|$. We distinguish the following three cases.

1. $\mathrm{vol}(\mathcal{X}_{\kappa_1}) > 0$ and $\mathrm{vol}(\mathcal{X}_{\kappa_2}) > 0$. Since g is the kernel of a RKHS, there exists a feature mapping $\bar{\phi}$ to some inner-product feature space $\mathcal{F}$ such that $\forall x,\zeta \in \mathcal{X}$ one has $g(x,\zeta) = <\bar{\phi}(x),\bar{\phi}(\zeta)>_{\mathcal{F}}$. By the definition (12) we get

$$\begin{aligned}
\mathcal{K}(\mathcal{X}_{\kappa_1},\mathcal{X}_{\kappa_2}) &= \int_{\mathbb{R}^d}\left(\int_{\mathbb{R}^d} g(x,\zeta)\cdot w(\zeta)\cdot \hat{1}_{\mathcal{X}_{\kappa_1}}(\zeta)d\zeta\right)\cdot w(x)\cdot \hat{1}_{\mathcal{X}_{\kappa_2}}(x)dx \\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} <\bar{\phi}(x),\bar{\phi}(\zeta)>_{\mathcal{F}} w(\zeta)\cdot \hat{1}_{\mathcal{X}_{\kappa_1}}(\zeta)\cdot w(x)\cdot \hat{1}_{\mathcal{X}_{\kappa_2}}(x)d\zeta dx \\
&= <\int_{\mathbb{R}^d}\bar{\phi}(\zeta)w(\zeta)\cdot \hat{1}_{\mathcal{X}_{\kappa_1}}(\zeta)d\zeta, \int_{\mathbb{R}^d}\bar{\phi}(x)w(x)\cdot \hat{1}_{\mathcal{X}_{\kappa_2}}(x)dx>_{\mathcal{F}} \\
&= <\bar{\Phi}(\mathcal{X}_{\kappa_1}),\bar{\Phi}(\mathcal{X}_{\kappa_2})>_{\mathcal{F}}, \qquad (14)
\end{aligned}$$

   where $\mathcal{Z} \in 2^{\mathcal{X}}$, $\bar{\Phi}(\mathcal{Z}) := \int_{\mathcal{Z}} \phi(x)w(x)\cdot \hat{1}_{\mathcal{Z}}(x)dx$, and the last equalities in (14) follow by the definition of Bochner's integral [25, Chapter 7].

2. $\mathrm{vol}(\mathcal{X}_{\kappa_1}) > 0$ and $\mathcal{X}_{\kappa_2} = \{x_{\kappa_2}\}$. By the same arguments as above, we get

$$\begin{aligned}
\mathcal{K}(\mathcal{X}_{\kappa_1},\{x_{\kappa_2}\}) &= \int_{\mathbb{R}^d}\left(\int_{\mathbb{R}^d} g(x,\zeta)\cdot w(\zeta)\cdot \hat{1}_{\mathcal{X}_{\kappa_1}}(\zeta)d\zeta\right)\cdot w(x)\cdot \delta(x-x_{\kappa_2})dx \\
&= w(x_{\kappa_2})\int_{\mathbb{R}^d} g(x_{\kappa_2},\zeta)\cdot w(\zeta)\cdot \hat{1}_{\mathcal{X}_{\kappa_1}}(\zeta)d\zeta \\
&= <w(x_{\kappa_2})\bar{\phi}(x_{\kappa_2}), \int_{\mathbb{R}^d}\bar{\phi}(\zeta)w(\zeta)\cdot \hat{1}_{\mathcal{X}_{\kappa_1}}(\zeta)d\zeta>_{\mathcal{F}} \\
&= <w(x_{\kappa_2})\bar{\phi}(x_{\kappa_2}),\bar{\Phi}(\mathcal{X}_{\kappa_1})>_{\mathcal{F}}, \qquad (15)
\end{aligned}$$

   where $w(x_{\kappa_2})\bar{\phi}(x_z)$ is the degenerate case of $\bar{\Phi}(\mathcal{Z})$ (in which $\mathcal{Z}$ degenerates into the point x_z).

3. $\mathcal{X}_{\kappa_1} = \{x_{\kappa_1}\}$ and $\mathcal{X}_{\kappa_2} = \{x_{\kappa_2}\}$. We immediately get $\mathcal{K}(\mathcal{X}_{\kappa_1},\mathcal{X}_{\kappa_2}) = w(x_{\kappa_1})\cdot w(x_{\kappa_2})\cdot g(x_{\kappa_1},x_{\kappa_2}) =< w(x_{\kappa_1})\bar{\phi}(x_{\kappa_1}), w(x_{\kappa_2})\bar{\phi}(x_{\kappa_2})>_{\mathcal{F}}$.

The statement follows by constructing the Gram matrices corresponding to the cases 1, 2, and 3, resp.

Proposition 2 can be extended to regions with arbitrary shapes. However, from a practical point of view, the multi-dimensional integrals that define $\mathcal{K}(\cdot,\cdot)$ (hence the Gram matrix $\mathbb{K}$) are particularly simple to evaluate for regions with a certain shapes and for some plain kernels $g(\cdot,\cdot)$. This was shown in [7] in the case of box-shaped regions and Gaussian plain kernels.

5. Functional risk versus average risk: A comparison

We now compare the two learning problems associated with the minimizations of the functional and average risks $R(\cdot)$ and $R_m(\cdot)$, resp., on the Sobolev space $\mathcal{W}^{k,2}$.

As described in the following proposition, a first relationship occurs when the loss function is the linear hinge loss $V(y_\kappa, f(x)) = \max(0, 1 - y_\kappa f(x))$, the weight function satisfies $w(x) \geq 1$ inside each set $\mathcal{X}_\kappa$, and the assumption made in Remark 2 holds.

Proposition 3. *If V is the linear hinge loss function, $w(x) \geq 1$ and $y_\kappa f^\circ(x) \neq 1$ inside each set $\mathcal{X}_\kappa$, then one has $R(f^\circ) = R_m(f^\circ)$.*

Proof. It follows from the assumption $y_\kappa f^\circ(x) \neq 1$ inside each set $\mathcal{X}_\kappa$, and the continuity of this loss function that either

$$y_\kappa f^\circ(x) > 1, \text{for all}\, x \in \mathcal{X}_\kappa$$

or

$$y_\kappa f^\circ(x) < 1, \text{for all}\, x \in \mathcal{X}_\kappa$$

holds. This, combined with the assumption on the weight function and the definition of $m_{\mathcal{X}_\kappa}(f)$, shows that, for each set $\mathcal{X}_\kappa$, one has

$$\begin{aligned} & \int_{\mathbb{R}^d} V(y_\kappa, f^\circ(x)) \cdot w(x) \cdot \hat{1}_{\mathcal{X}_\kappa}(x) dx \\ = & \int_{\mathbb{R}^d} \max(0, 1 - y_\kappa f^\circ(x)) \cdot w(x) \cdot \hat{1}_{\mathcal{X}_\kappa}(x) dx \\ = & \max\left(0, 1 - y_\kappa \int_{\mathbb{R}^d} f^\circ(x) \cdot w(x) \cdot \hat{1}_{\mathcal{X}_\kappa}(x) dx\right) \\ = & V(y_\kappa, m_{\mathcal{X}_\kappa}(f^\circ)), \end{aligned} \tag{16}$$

so $R(f^\circ) = R_m(f^\circ)$.

Note that the equality

$$\int_{\mathbb{R}^d} V(y_\kappa, f^\circ(x)) \cdot w(x) \cdot \hat{1}_{\mathcal{X}_\kappa}(x) dx = V(y_\kappa, m_{\mathcal{X}_\kappa}(f^\circ))$$

obtained in the proof of Proposition 3 holds also for the linear loss $V(y_\kappa, f(x)) = -y_\kappa f(x)$ (even without the assumptions $w(x) \geq 1$ and $y_\kappa f^\circ(x) \neq 1$ inside each set $\mathcal{X}_\kappa$), so also for this loss function one has $R(f^\circ) = R_m(f^\circ)$.

A second relationship between $R(\cdot)$ and $R_m(\cdot)$ is provided by the following proposition, which holds for a general convex and differentiable loss function V and requires that, for all $\kappa \in \mathbb{N}_{\ell_t}$, $w(x) \cdot \hat{1}_{\mathcal{X}_\kappa}(x)$ is a probability density on $\mathcal{X}_\kappa$ (this happens, e.g., for the uniform weight function $w(x) \equiv 1$).

Proposition 4. *For $\kappa \in \mathbb{N}_{\ell_t}$, let $\int_{\mathbb{R}^d} w(x) \cdot \hat{1}_{\mathcal{X}_\kappa}(x)dx = 1$. Then for every $f \in \mathcal{W}^{k,2}$ one has $R(f) \geq R_m(f)$.*

Proof. For every $f \in \mathcal{W}^{k,2}$, the risks $R(f)$ and $R_m(f)$ differ by the two terms

$$\sum_{\kappa \in \mathbb{N}_{\ell_t}} \int_{\mathbb{R}^d} V(y_\kappa, f(x)) \cdot w(x) \cdot \hat{1}_{\mathcal{X}_\kappa}(x)dx$$

and

$$\sum_{\kappa \in \mathbb{N}_{\ell_t}} V(y_\kappa, m_{\mathcal{X}_\kappa}(f)) = \sum_{\kappa \in \mathbb{N}_{\ell_t}} V(y_\kappa, \int_{\mathbb{R}^d} f(x) \cdot w(x) \cdot \hat{1}_{\mathcal{X}_\kappa}(x)dx)\,.$$

By applying Jensen's inequality[2] to the probability density $w(x) \cdot \hat{1}_{\mathcal{X}_\kappa}(x)$, for every $\kappa \in \mathbb{N}_{\ell_t}$ we get

$$\int_{\mathbb{R}^d} V(y_\kappa, f(x)) \cdot w(x) \cdot \hat{1}_{\mathcal{X}_\kappa}(x)dx \geq V\big(y_\kappa, \int_{\mathbb{R}^d} f(x) \cdot w(x) \cdot \hat{1}_{\mathcal{X}_\kappa}(x)dx\big)\,.$$

The next proposition shows that, for λ "large enough" and "sufficiently small" diameters of the sets $\mathcal{X}_\kappa$, the values of the two risks $R(\cdot)$ and $R_m(\cdot)$ at optimality are very similar. The proposition requires a slightly larger minimum value of k with respect to the previous ones.

Proposition 5. *Let $k \geq \frac{d+1}{2}$, f° and f_m° be the minimum points over $\mathcal{W}^{k,2}$ of $R(\cdot)$ and $R_m(\cdot)$, resp., and for every $\kappa \in \mathbb{N}_{\ell_t}$ let $\int_{\mathbb{R}^d} w(x) \cdot \hat{1}_{\mathcal{X}_\kappa}(x)dx = 1$. Then there exist two constants $C_1, C_2 > 0$ such that*

$$i. \quad |R(f^\circ) - R_m(f^\circ)| \leq \sum_{\kappa \in \mathbb{N}_{\ell_t}} \operatorname{diam}(\mathcal{X}_\kappa) \cdot L_\kappa(\lambda) \cdot C_2 \frac{2R(0)}{\lambda}\,; \tag{17}$$

$$ii. \quad |R(f_m^\circ) - R_m(f_m^\circ)| \leq \sum_{\kappa \in \mathbb{N}_{\ell_t}} \operatorname{diam}(\mathcal{X}_\kappa) \cdot L_\kappa(\lambda) \cdot C_2 \frac{2R(0)}{\lambda}\,,$$

where $L_\kappa(\lambda) := \sup_{\delta \in \left[-C_1 \frac{2R(0)}{\lambda}, C_1 \frac{2R(0)}{\lambda}\right]} \left|V'_f(y_\kappa, \delta)\right|$.

[2] Jensen's inquality states that for every convex function ϕ and every random variable z, one has $\mathbb{E}\{\phi(z)\} \geq \phi(\mathbb{E}\{z\})$.

PROOF. (i) We first evaluate the risk $R(f)$ for $f = 0$. By the definition of $R(\cdot)$ and the optimality of f°, we get

$$\frac{\lambda}{2}\|Pf^\circ\|^2 \leq R(f^\circ) \leq R(0)\,,$$

then $\|Pf^\circ\| \leq \frac{2R(0)}{\lambda}$. As $k > \frac{d+1}{2}$, by the Sobolev Embedding Theorem with $k > \frac{d+1}{2}$ it follows that f° is of class $\mathcal{C}^1$ and there exist two constants $C_1, C_2 > 0$ such that

$$\sup_{x\in\mathbb{R}^d} |f^\circ(x)| \leq C_1\|Pf^\circ\| \leq C_1\frac{2R(0)}{\lambda} \tag{18}$$

and

$$\sup_{x\in\mathbb{R}^d} \|\nabla f^\circ(x)\| \leq C_2\|Pf^\circ\| \leq C_2\frac{2R(0)}{\lambda}\,. \tag{19}$$

By (18) we have

$$\sup_{x\in\mathcal{X}_\kappa} \left|V_f'(y_\kappa, f^\circ(x))\right| \leq \sup_{\delta\in\left[-C_1\frac{2R(0)}{\lambda}, C_1\frac{2R(0)}{\lambda}\right]} \left|V_f'(y_\kappa,\delta)\right| = L_\kappa(\lambda)\,.$$

By (19) and the definition of $L_\kappa(\lambda)$, we obtain

$$\sup_{x,y\in\mathcal{X}_\kappa} |V(y_\kappa, f^\circ(x)) - V(y_\kappa, f^\circ(y))| \leq \operatorname{diam}(\mathcal{X}_\kappa)\cdot L_\kappa(\lambda)\cdot C_2\frac{2R(0)}{\lambda}\,. \tag{20}$$

As $w(x)\cdot \hat{1}_{\mathcal{X}_\kappa}(x)$ is a probability density on $\mathcal{X}_\kappa$, we have

$$\min_{x\in\mathcal{X}_\kappa} V(y_\kappa, f^\circ(x)) \leq V(y_\kappa, m_{\mathcal{X}_\kappa}(f^\circ)) \leq \max_{x\in\mathcal{X}_\kappa} V(y_\kappa, f^\circ(x))\,. \tag{21}$$

Hence, (20) and (21) provide

$$\begin{aligned} &\left|V(y_\kappa, m_{\mathcal{X}_\kappa}(f^\circ)) - \int_{\mathbb{R}^d} V(y_\kappa, f^\circ(x))\cdot w(x)\cdot \hat{1}_{\mathcal{X}_\kappa}(x)dx\right| \\ \leq\ & \operatorname{diam}(\mathcal{X}_\kappa)\cdot L_\kappa(\lambda)\cdot C_2\frac{2R(0)}{\lambda}\,. \end{aligned} \tag{22}$$

Finally, (17) follows by the estimate (22) on each set $\mathcal{X}_\kappa$.

(ii) is proved likewise item i, observing that $R(0) = R_m(0)$.

Remark 4. The function $L_\kappa(\lambda)$ is nonincreasing and for a quadratic loss function $V(y_\kappa, f(x)) = \frac{1}{2}\left(y_\kappa - f(x)\right)^2$ we have $L_\kappa(\lambda) \leq |y_k| + C_1\frac{2R(0)}{\lambda}$. In such a case, formula (17) becomes

$$|R(f^\circ) - R_m(f^\circ)| \leq \sum_{\kappa\in\mathbb{N}_{\ell_t}} \operatorname{diam}(\mathcal{X}_\kappa)\cdot\left(|y_k| + C_1\frac{2R(0)}{\lambda}\right)\cdot C_2\frac{2R(0)}{\lambda}\,.$$

Let

$$K(\lambda) := \sum_{\kappa \in \mathbb{N}_{\ell_t}} \operatorname{diam}(\mathcal{X}_\kappa) \cdot L_\kappa(\lambda) \cdot C_2 \frac{2R(0)}{\lambda} .$$

The following proposition provides a relationship between f° and f_m°, showing that their difference is small when λ is "large enough" and the diameters of the sets $\mathcal{X}_\kappa$ are "small enough".

Proposition 6. *Let $k \geq \frac{d+1}{2}$, f° and f_m° be the minimum points over $\mathcal{W}^{k,2}$ of $R(\cdot)$ and $R_m(\cdot)$, resp., and for every $\kappa \in \mathbb{N}_{\ell_t}$ let $\int_{\mathbb{R}^d} w(x) \cdot \hat{1}_{\mathcal{X}_\kappa}(x) dx = 1$. Then there exist a constant $C_1 > 0$ (the same as in Proposition 5) such that*

$$\sup_{x \in \mathbb{R}^d} |f^\circ(x) - f_m^\circ(x)| \leq C_1 \, \|P(f^\circ - f_m^\circ)\| \leq C_1 \sqrt{\frac{4K(\lambda)}{\lambda}} . \tag{23}$$

Proof. By Proposition 5 and the optimality of f° and f_m°, we have

$$\begin{aligned} |R(f_m^\circ) - R(f^\circ)| &= R(f_m^\circ) - R(f^\circ) \\ &\leq R(f_m^\circ) - R_m(f_m^\circ) + R_m(f_m^\circ) - R_m(f^\circ) + R_m(f^\circ) - R(f^\circ) \\ &\leq |R(f_m^\circ) - R_m(f_m^\circ)| + R_m(f_m^\circ) - R_m(f^\circ) + |R_m(f^\circ) - R(f^\circ)| \\ &\leq 2K(\lambda) . \end{aligned} \tag{24}$$

As the functional $R(\cdot)$ is uniformly convex with modulus of convexity $\frac{\lambda}{2}t^2$, by (B.1) and the optimality of f° we get

$$|R(f_m^\circ) - R(f^\circ)| \geq \frac{\lambda}{2} \|P(f_m^\circ - f^\circ)\|^2 . \tag{25}$$

As $k > \frac{d+1}{2}$, by the Sobolev Embedding Theorem with $k > \frac{d+1}{2}$ it follows that f° is of class $\mathcal{C}^1$ and there exist a constant $C_1 > 0$ such that

$$\sup_{x \in \mathbb{R}^d} |f^\circ(x)| \leq C_1 \|P f^\circ\| \leq C_1 \frac{2R(0)}{\lambda} \tag{26}$$

The statement (23) follows by combining equations (25), (26), and (24).

For a quadratic loss function $V(y_\kappa, f(x)) = \frac{1}{2}(y_\kappa, f(x))^2$, the difference $f^\circ(x) - f_m^\circ(x)$ can be evaluated directly by comparing the solutions to the two linear Fredholm integral equations of the second kind obtained by plugging $V(y_\kappa, f(x)) = \frac{1}{2}(y_\kappa, f(x))^2$ into (3) and (11). The corresponding equations are

$$\begin{aligned} f^\circ(x) &+ \frac{1}{\lambda} \sum_{\kappa \in \mathbb{N}_{\ell_t}} \int_{\mathbb{R}^d} g(x, \zeta) \cdot w(\zeta) \cdot \hat{1}_{\mathcal{X}_\kappa}(\zeta) \cdot f^\circ(\zeta) d\zeta \\ &= \frac{1}{\lambda} \sum_{\kappa \in \mathbb{N}_{\ell_t}} y_\kappa \int_{\mathbb{R}^d} g(x, \zeta) \cdot w(\zeta) \cdot \hat{1}_{\mathcal{X}_\kappa}(\zeta) d\zeta \end{aligned}$$

and

$$\begin{aligned} f_m^\circ(x) &+ \frac{1}{\lambda}\sum_{\kappa\in\mathbb{N}_{\ell_t}}\int_{\mathbb{R}^d}\left(\int_{\mathbb{R}^d} g(x,\xi)\cdot w(\xi)\cdot \hat{1}_{\mathcal{X}_\kappa}(\xi)d\xi\right)\cdot w(\zeta)\cdot \hat{1}_{\mathcal{X}_\kappa}(\zeta)\cdot f_m^\circ(\zeta)d\zeta \\ &= \frac{1}{\lambda}\sum_{\kappa\in\mathbb{N}_{\ell_t}} y_\kappa \int_{\mathbb{R}^d} g(x,\zeta)\cdot w(\zeta)\cdot \hat{1}_{\mathcal{X}_\kappa}(\zeta)d\zeta\,, \end{aligned}$$

resp. Such equations can be written in the forms

$$f^\circ(x) + \gamma\int_{\mathbb{R}^d} k(x,\zeta)\cdot f^\circ(\zeta)d\zeta = r(x) \tag{27}$$

and

$$f_m^\circ(x) + \gamma\int_{\mathbb{R}^d} k_m(x,\zeta)\cdot f_m^\circ(\zeta)d\zeta = r(x)\,, \tag{28}$$

resp., where

$$\gamma := \frac{1}{\lambda}\,,$$

$$k(x,\zeta) := \sum_{\kappa\in\mathbb{N}_{\ell_t}} g(x,\zeta)\cdot w(\zeta)\cdot \hat{1}_{\mathcal{X}_\kappa}(\zeta)$$

and

$$k_m(x,\zeta) := \sum_{\kappa\in\mathbb{N}_{\ell_t}}\left(\int_{\mathbb{R}^d} g(x,\xi)\cdot w(\xi)\cdot \hat{1}_{\mathcal{X}_\kappa}(\xi)d\xi\right)\cdot w(\zeta)\cdot \hat{1}_{\mathcal{X}_\kappa}(\zeta)$$

are their respective kernels (in the sense of Fredholm integral equations), and

$$r(x) := \frac{1}{\lambda}\sum_{\kappa\in\mathbb{N}_{\ell_t}} y_\kappa \int_{\mathbb{R}^d} g(x,\zeta)\cdot w(\zeta)\cdot \hat{1}_{\mathcal{X}_\kappa}(\zeta)d\zeta$$

is known.

Let us assume that there exists $\varepsilon > 0$ such that

$$\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} |k(x,\zeta) - k_m(x,\zeta)|^2\, dxd\zeta \le \varepsilon\,.$$

(By the definitions of $k(x,\zeta)$ and $k_m(x,\zeta)$, we expect ε decreases when decreasing the diameters of the sets $\mathcal{X}_\kappa$). We estimate the difference between f° and f_m°. By subtracting equation (28) from (27), we get

$$
\begin{aligned}
& f^\circ(x) - f_m^\circ(x) \\
= \; & -\gamma \left(\int_{\mathbb{R}^d} k(x,\zeta) \cdot f^\circ(\zeta) d\zeta - \int_{\mathbb{R}^d} k_m(x,\zeta) \cdot f_m^\circ(\zeta) d\zeta \right) \\
= \; & -\gamma \left(\int_{\mathbb{R}^d} (k(x,\zeta) - k_m(x,\zeta)) \cdot f^\circ(\zeta) d\zeta + \int_{\mathbb{R}^d} k_m(x,\zeta) \cdot (f^\circ(\zeta) - f_m^\circ(\zeta)) d\zeta \right) .
\end{aligned}
$$

As $\left\| \int_{\mathbb{R}^d} a(\cdot,\zeta) b(\zeta) d\zeta \right\| \leq \sqrt{\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |a(x,\zeta)|^2 dx d\zeta} \cdot \|b\|$, in terms of the $\mathcal{L}^2$-norm we obtain

$$
\|f^\circ - f_m^\circ\| \leq \gamma \left(\sqrt{\varepsilon} \cdot \|f^\circ\| + \sqrt{\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |k_m(x,\zeta)|^2 dx d\zeta} \cdot \|f^\circ - f_m^\circ\| \right) .
$$

Then, for $\gamma \cdot \sqrt{\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |k_m(x,\zeta)|^2 dx d\zeta} < 1$ and $f^\circ \neq 0$, we get

$$
\frac{\|f^\circ - f_m^\circ\|}{\|f^\circ\|} \leq \frac{\gamma \cdot \sqrt{\varepsilon}}{1 - \gamma \cdot \sqrt{\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |k_m(x,\zeta)|^2 dx d\zeta}} . \tag{29}
$$

Finally, recalling that $\gamma = \frac{1}{\lambda}$ and ε is nonincreasing when the diameters of the sets $\mathcal{X}_\kappa$ decrease, it follows by (29) that the relative error $\frac{\|f^\circ - f_m^\circ\|}{\|f^\circ\|}$ decreases when increasing λ and/or decreasing the diameters.

6. Discussion

We have developed a unified variational formulation for the class of learning problems introduced in [2], which incorporate both supervised points and supervised regions. Our approach takes the hint from the inspirational framework proposed in [8], where desired smoothness properties of the input/output relationship that provides the solution to the learning problem are modeled via differential operators. We have provided new representer theorems for the optimal solutions of the associated learning problems.

The basic ingredients of our approach are: (i) loss functions involving labeled regions (either with non-null Lebesgue measure or degenerating into points), (ii) regularization based on diffferential operators, (iii) Green's functions (of such operators) that are kernels of RKHSs, and (iv) Sobolev spaces of suitable orders.

As a particular case arising when prior knowledge is expressed by logical propositions, we have investigated the situation in which the labeled regions are multi-dimensional intervals (i.e., "boxes"). In such a context, we have shown that the solution to learning from labeled boxes/points is based on a novel class of kernels, obtained by joining a classical kernel with the collection of supervised boxes (which can degenerate to points).

Using the techniques described in [24], some results that we have obtained for finite-order differential operators can be extended to the case of infinite-order

differential operators and, more generally, to pseudo-differential ones. When $\sigma > 0$, this holds for the infinite-order differential operator $L = \sum_{n=0}^{\infty} \frac{(-1)^n \sigma^{2n}}{n! 2^n} \nabla^{2n}$ (where ∇^2 denotes the Laplace operator), whose Green's function is the Gaussian $g(x, \zeta) = \exp\left(-\frac{\|x-\zeta\|_2^2}{2\sigma^2}\right)$ (where $\|\cdot\|_2$ denotes the l_2-norm), to which refer the numerical results presented in [7].

Appendix A. Reproducing Kernel Hilbert Spaces (RKHSs)

A Reproducing Kernel Hilbert Space (RKHS) is a Hilbert space $\mathcal{H}_K(\Omega)$ formed by functions defined on a nonempty set Ω such that for every $u \in \Omega$ the evaluation functional Υ_u, defined for any $f \in \mathcal{H}_K(\Omega)$ as $\Upsilon_u(f) = f(u)$, is bounded. RKHSs were formally defined by Aronszajn [14] but their theory employs work by Schönberg [29], as well as many classical results on kernels and positive definite functions. We consider real RKHSs.

RKHSs can be characterized in terms of *kernels*, which are *symmetric positive semidefinite* functions $K : \Omega \times \Omega \to \mathbb{R}$, i.e., functions satisfying for all positive integers m, all $(w_1, \ldots, w_m) \in \mathbb{R}^m$, and all $(u_1, \ldots, u_m) \in \Omega^m$,

$$\sum_{i,j=1}^{m} w_i \, w_j \, K(u_i, u_j) \geq 0.$$

By the Riesz Representation Theorem [30, p. 200], for every $u \in \Omega$ there exists a unique element $K_u \in \mathcal{H}_K(\Omega)$, called the *representer* of u, such that for every $f \in \mathcal{H}$ one has

$$\Upsilon_u(f) = \langle f, K_u \rangle_{\mathcal{H}_K(\Omega)} \tag{A.1}$$

(*reproducing property*), where $\langle \cdot, \cdot \rangle_{\mathcal{H}_K(\Omega)}$ denotes the inner product in $\mathcal{H}_K(\Omega)$. It is easy to check that the function $K : \Omega \times \Omega$ defined for all $u, v \in \Omega$ as $K(u, v) = \langle u, v \rangle$ (where $\langle \cdot, \cdot \rangle$ is any inner product on $\mathbb{R}^d$) is a kernel.

On the other hand, every kernel $K : \Omega \times \Omega \to \mathbb{R}$ generates an RKHS $\mathcal{H}_K(\Omega)$ that is the completion of the linear span of the set $\{K_u : u \in \Omega\}$, with the inner product defined as $\langle K_u, K_v \rangle_{\mathcal{H}_K(\Omega)} = K(u, v)$ and the induced norm $\|\cdot\|_{\mathcal{H}_K(\Omega)}$ (see, e.g., [14] and [31, p. 81]).

By the reproducing property (A.1) and the Cauchy-Schwartz inequality, for every $f \in \mathcal{H}_K(\Omega)$ and every $u \in \Omega$ we have $|f(u)| = |\langle f, K_u \rangle_{\mathcal{H}_K(\Omega)}| \leq \|f\|_{\mathcal{H}_K(\Omega)} \sqrt{K(u, u)} \leq s_{K,\Omega} \, \|f\|_{\mathcal{H}_K(\Omega)}$, where $s_{K,\Omega} = \sup_{u \in \Omega} \sqrt{K(u, u)}$. Thus for every kernel K, we have

$$\sup_{u \in \Omega} |f(u)| \leq s_{K,\Omega} \|f\|_{\mathcal{H}_K(\Omega)} \, .$$

A paradigmatic example of a kernel on $\mathbb{R}^d$ is the *Gaussian kernel* $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, defined as $K(u, v) = \exp(-\|u - v\|^2)$. Other examples of kernels are $K(u, v) = \exp(-\|u - v\|)$, $K(u, v) = \langle u, v \rangle^p$ (*homogeneous polynomial* of

degree p), $K(u,v) = (1 + \langle u, v\rangle)^p$ (*inhomogeneous polynomial* of degree p), and $K(u,v) = (a^2 + \|u - v\|^2)^{-\alpha}$, with $\alpha > 0$ [15, p. 38].

For the role of kernels in learning theory see, e.g., [15] and [21].

Appendix B. Functionals on normed spaces

A functional $\Phi : F \to \mathbb{R}$ (where F is a convex subset of the normed linear space $\mathcal{F}$) is called *uniformly convex* iff there exists a non-negative function $\delta : [0, +\infty) \to [0, +\infty)$ such that $\delta(0) = 0$, $\delta(t) > 0$ for all $t > 0$, and for all $h, g \in F$ and all $\lambda \in [0, 1]$, one has

$$\Phi(\lambda h + (1-\lambda)g) \leq \lambda\Phi(h) + (1-\lambda)\Phi(g) - \lambda(1-\lambda)\delta(\|h - g\|_{\mathcal{F}})\,.$$

Any such function δ is called a *modulus of convexity* of Φ [32]. It is easy to show (see, e.g., [33, Proposition 2.1]) that, for $c > 0$, the functional $\Phi(f) = c\,\|f\|_{\mathcal{F}}^2$ is uniformly convex with modulus of convexity $\delta(t) = c\,t^2$. The sum of a convex functional Φ_1 and of a uniformly convex one Φ_2 is uniformly convex, and has the same modulus of convexity as Φ_2 (see, e.g., [33, Proposition 2.1]). A uniformly convex functional on a convex subset of a Hilbert space admits a unique minimum point (see, e.g., [34, p. 10]). A useful property of uniform convexity is that $f^\circ \in \operatorname{argmin}_{f\in F} \Phi(f)$ implies the lower bound

$$|\Phi(f^\circ) - \Phi(f)| \geq \delta\left(\|f^\circ - f\|_{\mathcal{F}}\right) \tag{B.1}$$

for any $f \in F$ (see, e.g., [33, Proposition 2.1]).

[1] F. Lauer, G. Bloch, Incorporating prior knowledge in support vector machines for classification: A review, Neurocomputing 71 (2008) 1578–1594.

[2] G. Fung, O. Mangasarian, J. Shavlik, Knowledge-based support vector machine classifiers, in: Advances in Neural Information Processing Systems 14, MIT Press, 2002, pp. 537–544.

[3] G. Fung, O. Mangasarian, J. Shavlik, Knowledge-based nonlinear kernel classifiers, in: Conference on Learning Theory, 2003, pp. 102–112.

[4] Q. Le, A. Smola, T. Gártner, Simpler knowledge-based support vector machines, in: Proceedings of ICML, 2006, pp. 521–528.

[5] O. Mangasarian, E. Wild, Nonlinear knowledge-based classification, IEEE Trans. on Neural Networks 19 (2008) 1826–1832.

[6] O. Mangasarian, E. Wild, G. Fung, Proximal knowledge-based classification, Statistical Analysis and Data Mining 1 (2009) 215–222.

[7] S. Melacci, M. Gori, Learning with box kernels, in: Proceedings of the International Conference on Neural Information Processing (ICONIP 2011), 2011, pp. 519–528.

[8] T. Poggio, F. Girosi, A theory of networks for approximation and learning-MIT A.I. Memo No. 1140, C. B. I. P. Paper No. 3.

[9] A. N. Tikhonov, Solutions of incorrectly formulated problems and the regularization method, Soviet Math. Dokl. 4 (1963) 1035–1038.

[10] E. Parzen, An approach to time series analysis, Annals of Mathematical Statistics 32 (1961) 951–989.

[11] G. Wahba, Spline Models for Observational Data, SIAM, Philadelphia, PA, 1990.

[12] C. Cortes, V. Vapnik, Support vector networks, Machine Learning 20 (1995) 1–25.

[13] F. Girosi, An equivalence between sparse approximation and support vector machines, Neural Computation 10 (1998) 1455–1480.

[14] N. Aronszajn, Theory of reproducing kernels, Trans. of the American Mathematical Society 68 (1950) 337–404.

[15] F. Cucker, S. Smale, On the mathematical foundations of learning, Bulletin of AMS 39 (2001) 1–49.

[16] F. Girosi, M. Jones, T. Poggio, Regularization theory and neural networks architectures, Neural Computation 7 (1995) 219–269.

[17] V. Kůrková, M. Sanguineti, Learning with generalization capability by kernel methods of bounded complexity, J. of Complexity 21 (2005) 350–367.

[18] G. Gnecco, M. Sanguineti, The weight-decay technique in learning from data: An optimization point of view, Computational Management Science 6 (2009) 53–79.

[19] G. Gnecco, M. Sanguineti, Regularization techniques and suboptimal solutions to optimization problemsin learning from data, Neural Computation 22 (2010) 793–829.

[20] A. J. Smola, B. Schölkopf, From regularization operators to support vector kernels, in: Advances in Neural Information Processing Systems 10, MIT Press, 1998, pp. 343–349.

[21] B. Schölkopf, A. J. Smola, Learning with Kernels, MIT, Cambridge, MA, 2002.

[22] M. E. Taylor, Pseudo-Differential Operators, Princeton University Press, 1981.

[23] G. Gnecco, M. Gori, M. Sanguineti, Learning with boundary conditions, Technical Report, University of Genoa (2011).

[24] Z. Chen, S. Haykin, On different facets of regularization theory, Neural Computation 14 (2002) 2791–2846.

[25] R. A. Adams, J. F. Fournier, Sobolev spaces, 2nd Edition, Academic Press, 2003.

[26] G. E. Fasshauer, Q. Ye, Reproducing kernels of generalized Sobolev spaces via a Green function approach with differential operators, IIT Technical Report (2010).

[27] H. Attouch, G. Buttazzo, G. Michaille, Variational Analysis in Sobolev and BV Spaces. Applications to PDEs and Optimization, SIAM, Philadelphia, PA, 2006.

[28] L. Schwartz, Théorie des Distributions, Herrman, Paris, 1978.

[29] I. J. Schönberg, Metric spaces and completely monotone functions, Annals of Mathematics 39 (1938) 811–841.

[30] A. Friedman, Foundations of Modern Analysis, Dover, New York, 1982.

[31] C. Berg, J. P. R. Christensen, P. Ressel, Harmonic Analysis on Semigroups, Springer, New York, 1984.

[32] E. S. Levitin, B. T. Polyak, Convergence of minimizing sequences in conditional extremum problems, Dokl. Akad. Nauk SSSR 168 (5) (1966) 764–767.

[33] V. Kůrková, M. Sanguineti, Error estimates for approximate optimization by the extended Ritz method, SIAM J. on Optimization 15 (2005) 461–487.

[34] A. L. Dontchev, Perturbations, Approximations and Sensitivity Analysis of Optimal Control Systems, Lecture Notes in Control and Information Sciences, 52, Springer, Berlin Heidelberg, 1983.