

---

## Some Properties of Regularized Kernel Methods

Ernesto De Vito, Lorenzo Rosasco, Andrea Caponnetto,  
Michele Piana, Alessandro Verri

Technical Report

DISI

---

DISI, Università di Genova  
v. Dodecaneso 35, 16146 Genova, Italy

<http://www.disi.unige.it/>

## Abstract

In regularized kernel methods, the solution of a learning problem is found by minimizing functionals consisting of the sum of a data and a complexity term. In this paper we investigate some properties of a more general form of the above functionals in which the data term corresponds to the expected risk. First, we prove a quantitative version of the representer theorem holding for both regression and classification, for both differentiable and non differentiable loss functions, and for arbitrary offset terms. Second, we show that the case in which the offset space is non trivial corresponds to solving a standard problem of regularization in a Reproducing Kernel Hilbert Space in which the penalty term is given by a seminorm. Finally, we discuss the issues of existence and uniqueness of the solution. From the specialization of our analysis to the discrete setting it is immediate to establish a connection between the solution properties of sparsity and coefficient boundedness and some properties of the loss function. For the case of SVM for classification, we also obtain a complete characterization of the whole method in terms of the Khun-Tucker conditions with no need to introduce the dual formulation.

Keywords: Statistical learning, Reproducing kernel Hilbert spaces, Convex analysis, Representer theorem, Regularization theory

## 1. Introduction

The problem of learning from examples can be seen as the problem of estimating a unknown functional dependency given only a finite (possibly small) number of instances. The seminal work of Vapnik (Vapnik, 1988) shows that the key to effectively solve this problem is by controlling the complexity of the solution. In the context of statistical learning this leads to techniques known as *regularization networks* (Evgeniou et al., 2000) or *regularized kernel methods* (Vapnik, 1988, Schlkopf and Smola, 2002, Cristianini and Shawe Taylor, 2000) which solve a minimization problem of the form

$$\min_{f \in \mathcal{H}} \left\{ \frac{1}{\ell} \sum_{i=1}^{\ell} V(y_i, f(\mathbf{x}_i)) + \lambda \|f\|_{\mathcal{H}}^2 \right\}, \quad (1)$$

where  $V$  is the loss function,  $\mathcal{H}$  is the Hilbert space of the *hypotheses*,  $\lambda > 0$  is the regularization parameter and  $(\mathbf{x}_i, y_i)_{i=1}^{\ell}$  are the  $\ell$  pairs of examples. As shown by Evgeniou et al. (2000) the above minimization problem can also be seen as particular instance of Tikhonov Regularization (Tikhonov and Arsenin, 1977, Mukherjee et al., 2002) for a multivariate function approximation problem which is well known to be ill-posed (Poggio and Smale, 2003, Bertero et al., 1988, Evgeniou et al., 2000).

In this paper we study the following generalization of the above problem

$$\min_{(f,g) \in \mathcal{H} \times \mathcal{B}} \int_{X \times Y} V(y, f(\mathbf{x}) + g(\mathbf{x})) d\rho(\mathbf{x}, y) + \lambda \|f\|_{\mathcal{H}}^2, \quad (2)$$

that is, we consider the case of a generic measure  $\rho$  (what we call the *continuous setting*) and minimization takes place over  $\mathcal{H} \times \mathcal{B}$ , where  $\mathcal{H}$  and  $\mathcal{B}$  are Reproducing Kernel Hilbert Spaces (RKHS). As usual  $\mathcal{H}$  is the space of penalized functions whereas  $\mathcal{B}$  is the space of offset functions which often reduces to the set of constants.

Considering the continuous setting is meaningful for several reasons. First, it is useful in order to study the problem of the generalization properties of kernel methods (Steinwart, 2002a). In real problems an estimator  $f_S^\lambda$  can be found minimizing the discrete Tikhonov functional (see Eq. 1) associated to a given training set  $S$ . The generalization properties of  $f_S^\lambda$  can be quantified

by the discrepancy of its expected risk from the minimum obtainable risk that is,  $I[f_S^\lambda] - \inf I[f]$ . Following Cucker and Smale (2002) we can bound the previous quantity studying

$$I[f_S^\lambda] - I[f^\lambda] + I[f^\lambda] - \inf I[f]$$

where  $f^\lambda$  is the minimizer of the Tikhonov functional in the continuous setting. The above expression amounts to the typical splitting of the generalization error into sample error and approximation error (Niyogi and Girosi, 1999). Clearly, insight on the form of  $f^\lambda$  can be useful to obtain better bounds on both errors. Second, considering the continuous measure  $\rho$  corresponds intuitively to finding a stable solution to the learning problem in the case of infinite number of examples and, hence, gives information about the best we can do in the hypothesis space  $\mathcal{H} \times \mathcal{B}$  (Mukherjee et al., 2002). Third, we can treat both the empirical measure and the ideal (unknown) probability distribution describing the relation between the data  $\mathbf{x}$  and the labels  $y$ , in a unified framework.

The contribution of our work is threefold. First we provide a complete characterization of the explicit form of the solution of Problem (2), by exploiting a convexity assumption on the loss functions. Our result can be interpreted as a quantitative version of the representer theorem holding for both regression and classification and in which explicit care is taken of the offset space  $\mathcal{B}$ . Then, we discuss the role of the offset space  $\mathcal{B}$ . The starting point of our discussion is the obvious observation that the estimator given by the solution of Problem (2) is not the pair  $(f, g)$  but the sum  $f + g$ . In other words the natural hypothesis space is the sum  $\mathcal{H} + \mathcal{B}$  instead of the product  $\mathcal{H} \times \mathcal{B}$  (which is not even a space of functions from  $X$  to  $\mathbb{R}$ ). For arbitrary loss function we prove that Problem (2) is equivalent to a kernel method defined on  $\mathcal{H} + \mathcal{B}$ , which is a RKHS, with a penalty term given by a seminorm. Finally, for sake of completeness, we study the issues of the existence and uniqueness for Problem (2). When  $\mathcal{B}$  is not the empty set, both issues are not trivial. In particular, for  $\mathcal{B}$  equal to the set of constants, we prove existence under very reasonable conditions: for example, for classification, one needs at least two examples with different labels. About uniqueness we show that, for strictly convex loss functions, one has uniqueness if and only if the space  $\mathcal{B}$  is small enough to be separated by the measure  $\rho$ : for example, in the discrete setting, this last condition means that a function  $g \in \mathcal{B}$  is equal to 0 if and only if  $g(\mathbf{x}_i) = 0$  for all  $i$ . For the hinge loss function, which is convex but not strictly convex, we give an *ad hoc* condition in terms of number of support vectors of the two classes.

The plan of the paper is as follows. In section 2 we discuss our contributions with respect to previous works. In section 3 we introduce some basic concepts of learning theory and state the assumptions we make on the loss function  $V$  and hypothesis spaces  $\mathcal{H}$  and  $\mathcal{B}$ . In section 4 we study the form of the solution of Problem (2). In section 5 we discuss the theoretical meaning of the offset space  $\mathcal{B}$ . We discuss the problem of existence and uniqueness in section 6. In section 7 we apply our results to the discrete setting and focus on the case of Support Vector Machines. In the appendix we recall some notions from convex analysis in infinite dimensional spaces.

## 2. Putting our work in context

We now briefly discuss the relation between our results and the previous works on this subject. Results about the form of the solution of kernel methods are known in the literature as *representer theorems* (if  $\mathcal{B}$  is not trivial they are called *semiparametric representer theorems*).

The first result in this direction is due to Kimeldorf and Wahba (1970), see, also, Wahba (1990), for the quadratic loss function, but the structure of the proof holds for arbitrary loss function as shown by many authors, see, for example, Cox and O'Sullivan (1990) and, in the framework of statistical learning, Schölkopf et al. (2001) (in this last paper the penalty term can be any strictly increasing function of the norm). This kind of results shows that the minimizer of Eq. (1) can be written as

$$f(\mathbf{x}) = \sum_{i=1}^{\ell} \alpha_i K(\mathbf{x}, \mathbf{x}_i),$$

where  $K$  is the kernel. The above result holds for arbitrary loss function and for a large class of penalty terms. However, the form of the coefficients  $\alpha_i$  is unknown.

For the quadratic loss function, the form of the coefficients is well known in the context of inverse problem, see, for example, Tikhonov and Arsenin (1977), and reduces to solve a linear system of equations. For arbitrary differentiable functions, this problem was studied by Girosi (1998), Poggio and Girosi (1992), Wahba (1998) where the coefficients  $\alpha_i$  are solution of a system of algebraic equations.

This approach cannot be applied to hinge and  $\epsilon$ -insensitive loss function (Vapnik, 1988), since they are not differentiable: the form of the coefficients  $\alpha_i$  is recovered only through the usual dual Lagrangian formulation of the minimization problem, see, for example, Vapnik (1988), Cristianini and Shawe Taylor (2000).

Recently, Zhang (2001) and Steinwart (2002b) give a quantitative representer theorem in the classification setting that holds for arbitrary convex loss function, without using the dual problem. In these papers the form of the coefficients  $\alpha_i$  is given in terms of a closed equation involving the subgradient of the loss function. Moreover, they are able to extend the representer theorem to the continuous setting (a study of the solution of Tikhonov regularization in the continuous setting when the square loss is used can be found also in Cucker and Smale, 2002).

This paper, using similar techniques of Steinwart (2002b), extends the above result in the following points:

- our result holds both for regression and classification;
- we provide a general result that holds also when the offset term is considered. The presence of the offset space forces the coefficients  $\alpha_i$  to satisfy a system of linear equations;
- we do not assume that input space  $X$  and the outspace space  $Y$  are compact. In particular, for regression we can assume  $Y = \mathbb{R}$ ;
- we provide a simpler proof than the one of Steinwart (2002b) by using known results about integral convex functionals.

A discussion of the role of the offset terms can be found in Evgeniou et al. (2000) and in Poggio et al. (2001) when the space  $\mathcal{B}$  reduces to the set of constant functions. The results are close to our Theorem 5, but they are proved assuming that the unit constant is in the Mercer decomposition of the kernel and for the discrete setting, while our result holds true for offset term living in arbitrary Reproducing Kernel Hilbert Spaces.

The problem of the existence and uniqueness is discussed in Wahba (1998) for the discrete setting and with differentiable loss functions. For arbitrary  $\rho$  the papers by Steinwart (2002a,b)

study the existence for classification setting with offset space reduced to the constant functions. For the hinge loss and  $\epsilon$ -insensitive loss, the problem of uniqueness is treated in Burges and Crisp (2000, 2002). Their proof is based on the dual problem and on the Kuhn-Tucker conditions. Our results subsume the cited results as special cases, but are all obtained in the more general continuous setting. In particular our results on uniqueness of the SVM solution are similar to those in Burges and Crisp (2000, 2002) but do not make use of the dual formulation.

### 3. Notation and assumptions

In this section we first fix the notation and then state and comment the basic assumptions needed to derive the results described in the rest of the paper. We start with input and output spaces.

#### 3.1 Input and output spaces

As usual we denote with  $X$  and  $Y$  the input and output spaces respectively. We assume that  $X$  is a closed subset<sup>1</sup> of  $\mathbb{R}^d$  and  $Y$  is a closed subspace of  $\mathbb{R}$ . Unlike the standard setting  $X$  is not necessarily compact (therefore, in the case of subsets of  $\mathbb{R}^d$  this means we also include unbounded domains).

We let  $Z = X \times Y$  and endow it with a probability distribution  $\rho$  defined on the Borel  $\sigma$ -algebra of  $Z$ . In practice,  $\rho$  will be either the unknown distribution describing the relation between  $\mathbf{x}$  and  $y$  or the empirical measure

$$\rho_S = \frac{1}{\ell} \sum_{i=1}^{\ell} \delta_{(\mathbf{x}_i, y_i)},$$

associated with the *training set*  $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^{\ell}$  drawn i.i.d. with respect to  $\rho$ . We now deal with loss functions.

#### 3.2 Loss functions

The *loss function*  $V(y, f(\mathbf{x}))$  is the price we are willing to pay by using  $f(\mathbf{x})$  to predict the correct label  $y$ . We collect the mathematical assumptions on the loss function in the following definition and then comments on the purpose of each assumption.

**Definition 1** A loss function  $V$  is a map  $V : Y \times \mathbb{R} \rightarrow [0, +\infty[$  such that

1.  $\forall y \in Y, V(y, \cdot)$ , is a convex function on  $\mathbb{R}$ .
2.  $V$  is continuous on  $Y \times \mathbb{R}$
3. there are  $b \in [0, +\infty[$  and  $a : Y \rightarrow \mathbb{R}$  such that

$$V(y, w) \leq a(y) + b|w|^p \quad \forall w \in \mathbb{R}, y \in Y \quad (3)$$

$$\int_{X \times Y} |a(y)| d\rho(\mathbf{x}, y) < +\infty, \quad (4)$$

where  $p \geq 1$ .

---

1. In fact, this assumption could be replaced by the weaker hypothesis that  $X$  is a locally compact second countable topological space.

The convexity hypothesis is not restrictive, being satisfied by all the loss functions commonly in use. Moreover, it is powerful from a technical point of view: it allows for the use of subgradient techniques without assuming differentiability of  $V$  and makes it possible to use convex analysis tools in the study of existence and uniqueness of functional minimizers.

Assumption 2) aims at avoiding measurability problems in integration and it is always satisfied in applications. One could replace it by the weaker requirement that  $V$  is a Borel function.

Condition 3) is a technical hypothesis we need in order to use results from convex integral functional analysis. For example, it is satisfied in the following cases

1. if  $V$  is the square loss function,  $V(y, w) = (y - w)^2$ , and

$$\int_{X \times Y} y^2 d\rho(\mathbf{x}, y) < +\infty;$$

2. if  $V(y, \cdot)$  is Lipschitz on  $\mathbb{R}$  with a Lipschitz constant independent of  $y$  and

$$\int_{X \times Y} V(y, 0) d\rho(\mathbf{x}, y) < +\infty.$$

The only loss used in practice left out is the exponential loss. We now draw our attention to some functionals studied in statistical learning.

### 3.3 Learning functionals

The *expected risk* of a function  $f$  is defined as

$$I[f] = \int_{X \times Y} V(y, f(\mathbf{x})) d\rho(y, \mathbf{x}),$$

and can be seen as the average error obtained by the function  $f$ , where  $f$  is a possible solution of the learning problem and the probability measure  $\rho$  is unknown.

Given a training set  $S$ , a possible way to estimate  $I[f]$  is to evaluate the *empirical risk*

$$I_{\text{emp}}^S[f] = \frac{1}{\ell} \sum_{i=1}^{\ell} V(y_i, f(\mathbf{x}_i)).$$

The problem of learning is to find, given the training set  $S$ , an *estimator*  $f$  effectively predicting the label of a new point. This translates in finding a function  $f$  such that its expected risk is small with high probability.

A possible way to efficiently solve the learning problem is provided by *regularized kernel methods* which amounts to solving a problem of functional minimization as problem 1. A generalization of problem 1 to a continuous setting is provided by problem 2 in which the continuous measure  $\rho$  replaces the empirical measure  $\rho_S$  in the first term. In what follows we will refer to the functionals to be minimized in both 1 and 2 as *Tikhonov functionals* and to the solutions as the *regularized solutions*.

The second term of a Tikhonov functional is a *smoothness* or a *complexity* term measuring the norm of the function  $f$  in a suitable Hilbert space  $\mathcal{H}$ . The minimization takes place in the *hypothesis space*  $\mathcal{H} \times \mathcal{B}$ . We now collect the assumptions on the hypothesis space at the basis of our analysis.

### 3.4 Hypothesis space

The space of penalized functions  $\mathcal{H}$  is a RKHS on  $X$  with continuous kernel  $K$  and norm  $\|\cdot\|_{\mathcal{H}}$  and the space of offset functions  $\mathcal{B}$  is a RKHS on  $X$  with continuous kernel  $K^{\mathcal{B}}$  and norm  $\|\cdot\|_{\mathcal{B}}$ . However, notice that the product space  $\mathcal{H} \times \mathcal{B}$  is not a RKHS.

We recall that the RKHS  $\mathcal{H}$  is defined as the unique Hilbert space of continuous functions on  $X$  such that

$$f(\mathbf{x}) = \langle f, K_{\mathbf{x}} \rangle_{\mathcal{H}}, \quad (5)$$

where, for all  $\mathbf{x} \in X$ ,  $K_{\mathbf{x}}$  is the function on  $X$  defined by  $K_{\mathbf{x}}(\mathbf{s}) = K(\mathbf{x}, \mathbf{s})$ .

Moreover, we assume that the following inequalities hold

$$\int_{X \times Y} K(\mathbf{x}, \mathbf{x})^{\frac{p}{2}} d\rho(\mathbf{x}, y) < +\infty \quad (6)$$

$$\int_{X \times Y} K^{\mathcal{B}}(\mathbf{x}, \mathbf{x})^{\frac{p}{2}} d\rho(\mathbf{x}, y) < +\infty. \quad (7)$$

The inequalities (6), (7), always true if  $X$  is compact or the kernel is bounded, ensure that  $\mathcal{H}$  and  $\mathcal{B}$  can be considered as subspaces of  $L^p(Z, \rho)$ .

The fact that the hypothesis spaces  $\mathcal{H}$  and  $\mathcal{B}$  are RKHS with continuous kernel is a common assumption in learning theory, where, usually,  $\mathcal{B}$  is the one dimensional vector space of constant functions

$$\mathcal{B} = \{f : X \rightarrow \mathbb{R} \mid f(\mathbf{x}) = b \quad b \in \mathbb{R}\} = \mathbb{R}$$

with kernel  $K^{\mathcal{B}}$  simply given by  $K^{\mathcal{B}}(\mathbf{x}, \mathbf{s}) = 1$ . A possible mathematical justification is that reasonable estimators have to be continuous functions on  $X$ ; moreover, any tractable minimization problem has to be settled down in a Hilbert space. Now a result due to Schwartz (1964) shows that a Hilbert space (of functions on  $X$ ) is a subspace of  $\mathcal{C}(X)$  (with compact inclusion) if and only if it is a RKHS with a continuous kernel, so any reasonable and tractable hypothesis space has to be a RKHS with continuous kernel.

## 4. Explicit form of the regularized solution

In this section we determine the explicit form of the minimizer of the Tikhonov functional introduced in the previous section. We first state the main theorem and comment on the obtained result, then, we provide the mathematical proof.

### 4.1 Main theorem

**Theorem 2** *For the minimization problem*

$$\min_{(f,g) \in \mathcal{H} \times \mathcal{B}} \int_{X \times Y} V(y, f(\mathbf{x}) + g(\mathbf{x})) d\rho(\mathbf{x}, y) + \lambda \|f\|_{\mathcal{H}}^2. \quad (8)$$

*the following statements are equivalent:*

1. *the pair  $(f^\lambda, g^\lambda) \in \mathcal{H} \times \mathcal{B}$  is a solution;*

2. there are  $g^\lambda \in \mathcal{B}$  and  $f^\lambda \in \mathcal{H}$  such that

$$f^\lambda(\mathbf{s}) = -\frac{1}{2\lambda} \int_{X \times Y} K(\mathbf{s}, \mathbf{x}) \alpha(\mathbf{x}, y) d\rho(\mathbf{x}, y) \quad (9)$$

where  $\alpha \in L^q(Z, \rho)$ , with  $\frac{1}{q} + \frac{1}{p} = 1$ , satisfying

$$\alpha(\mathbf{x}, y) \in (\partial V)(y, f^\lambda(\mathbf{x}) + g^\lambda(\mathbf{x})) \quad \rho\text{-a.e.} \quad (10)$$

$$\int_{X \times Y} K^{\mathcal{B}}(\mathbf{s}, \mathbf{x}) \alpha(\mathbf{x}, y) d\rho(\mathbf{x}, y) = 0. \quad (11)$$

The proof of this theorem is given in the following subsection. A few important remarks are in order.

First, the theorem gives a general quantitative version of the representer theorem. The generality is obtained by considering the continuous setting which subsumes the discrete setting if the measure  $\rho$  is the empirical measure  $\rho_S$ . In this case, the integral reduces to a finite sum and we recover the well known result  $f^\lambda = \sum_{i=1}^{\ell} \alpha_i K_{\mathbf{x}_i}$ , where the  $\mathbf{x}_i$  form the training set. Moreover, the solution is quantitatively characterized since the coefficients  $\alpha$  are given by Eq. (10) involving the subgradient. For differentiable losses in the discrete setting, Eq. (10) reduces to

$$\alpha_i = V'(y_i, f^\lambda(\mathbf{x}_i) + g^\lambda(\mathbf{x}_i)),$$

where  $V'$  denotes the derivative with respect to the second variable (Girosi, 1998, Wahba, 1998).

Second, if  $\{\psi_i\}_{i=1}^m$  is a base for  $\mathcal{B}$ , the offset part of the solution can be written as  $g^\lambda = \sum_{i=1}^m d_i \psi_i$ , where the coefficients  $d_i$  are again constrained by Eq. (10). A discussion on how to solve explicitly Eq. (10) can be found in Wahba (1998). Furthermore, the presence of  $\mathcal{B}$  induces a system of linear constraints on the coefficients  $\alpha_i$  expressed by Eq. (11) that, for  $\mathcal{B} = \mathbb{R}$ , reduces to the well known condition

$$\sum_{i=1}^{\ell} \alpha_i = 0.$$

We stress that, unlike previous works, the above equation has been derived without introducing the dual formulation.

Finally, we discuss the role of assumption 3) in Definition 1. From the proof, it is apparent that this assumption is needed to ensure the continuity of the first term in the Tikhonov functional which in the discrete setting is trivially guaranteed. Therefore, for the discrete setting Theorem 2 holds for any convex loss function. In particular,  $L^q(Z, \rho_S) = \mathbb{R}^\ell$  and the condition  $\alpha \in L^q(Z, \rho_S)$  is always satisfied. Back to the continuous setting, if  $V(y, \cdot)$  is Lipschitz on  $\mathbb{R}$  with a Lipschitz constant independent of  $y$  and

$$\int_{X \times Y} V(y, 0) d\rho(\mathbf{x}, y) < +\infty,$$

one can choose  $p = 1$ , so that  $q = +\infty$  and condition  $\alpha \in L^\infty(Z, \rho)$  means that  $\alpha$  is bounded. For the square loss, clearly  $p = 2$ , so that  $q = 2$  and  $\alpha$  is square-integrable. As shown by Steinwart (2002b), for classification and compact  $X$ , one can again remove Assumption 3) using the fact that a convex function is locally Lipschitz.



## 4.2 Proof of the main theorem

Before giving the proof of the theorem we discuss the proof structure, which aside from some technicalities is very simple, and is based on two lemmas. The Tikhonov functional  $I[f + g] + \lambda \|f\|_{\mathcal{H}}^2$  is a convex map on  $\mathcal{H} \times \mathcal{B}$ , so  $(f^\lambda, g^\lambda)$  is a minimizer of the Tikhonov functional if and only if  $(0, 0)$  is in its subgradient. Using linearity, the computation of the subgradient of the Tikhonov functional reduces to the computation of the subgradient of  $I[f + g]$  and  $\|f\|_{\mathcal{H}}^2$  respectively. Since the latter functional is differentiable, the subgradient evaluation is straightforward. Some care is needed for the subgradient of the former. First, we rewrite it as an integral functional on  $L^2(Z, \rho)$  and then use a fundamental result of convex analysis to interchange the integral with the subgradient.

For sake of simplicity we give the proof for  $p = 2$  (to avoid the technicalities of dealing with Banach spaces). Thus, we have that Eqs. (3), (6), and (7) hold with  $p = 2$ .

Proof [of Theorem 2] Clearly,  $\lambda \|f\|_{\mathcal{H}}^2$  is continuous and, by Lemma 3, the functional  $I[f + g] = I_0(\mathcal{J}(f, g))$  is continuous and finite. So, from item 5 of Proposition 13 in Appendix, one has that

$$\partial \left( I[f + g] + \lambda \|f\|_{\mathcal{H}}^2 \right) = \partial(I[f + g]) + \lambda \partial(\|f\|_{\mathcal{H}}^2).$$

Now, the map

$$(f, g) \rightarrow \lambda \|f\|_{\mathcal{H}}^2$$

is differentiable with derivative  $(2\lambda f, 0)$  and, therefore, by item 1 of Proposition 13,

$$\partial(\|f\|_{\mathcal{H}}^2) = \{(2\lambda f, 0)\}. \quad (12)$$

The main difficulty is the evaluation of the subgradient of the map  $I[f + g]$  given in Lemma 4. Thus, we obtain that the elements of the subgradient of  $I[f + g]$  at  $(f, g)$  are of the form

$$\left( \int_{X \times Y} K(\mathbf{x}, \cdot) \alpha(\mathbf{x}, y) d\rho(\mathbf{x}, y), \int_{X \times Y} K^B(\mathbf{x}, \cdot) \alpha(\mathbf{x}, y) d\rho(\mathbf{x}, y) \right), \quad (13)$$

where  $\alpha \in L^2(Z, \rho)$  satisfies Eq. (10).

Now, by combining (12) and (13), we have that the elements of the subgradient of  $I[f + g] + \lambda \|f\|_{\mathcal{H}}^2$  at point  $(f, g)$  are of the form

$$\left( \int_{X \times Y} K(\mathbf{x}, \cdot) \alpha(\mathbf{x}, y) d\rho(\mathbf{x}, y) + 2\lambda f, \int_{X \times Y} K^B(\mathbf{x}, \cdot) \alpha(\mathbf{x}, y) d\rho(\mathbf{x}, y) \right), \quad (14)$$

where  $\alpha \in L^2(Z, \rho)$  again satisfies Eq. (10).

From item 3 of Proposition 13, we have that an element  $(f^\lambda, g^\lambda) \in \mathcal{H} \times \mathcal{B}$  is a minimizer of  $I[f + g] + \lambda \|f\|_{\mathcal{H}}^2$  if and only if  $(0, 0)$  belongs to the subgradient evaluated at  $(f^\lambda, g^\lambda)$ . Using Eq. (14), one has that

$$f^\lambda(\mathbf{s}) = -\frac{1}{2\lambda} \int_{X \times Y} \alpha(\mathbf{x}, y) K(\mathbf{x}, \mathbf{s}) d\rho(\mathbf{x}, y)$$

$$\int_{X \times Y} \alpha(\mathbf{x}, y) K^B(\mathbf{x}, \mathbf{s}) d\rho(\mathbf{x}, y) = 0.$$

where  $\alpha \in L^2(Z, \rho)$ , once more, satisfies Eq. (10). This ends the proof. ■

Before computing the subgradient of the map  $I[f + g]$  in Lemma 4, we need to extend the definition of expected risk on  $L^2(Z, \rho)$ . We thus have

$$I_0[u] = \int_{X \times Y} V(y, u(\mathbf{x}, y)) \, d\rho(\mathbf{x}, y) \quad u \in L^2(Z, \rho),$$

and rewrite  $I[f + g] = I_0(\mathcal{J}(f, g))$  where  $\mathcal{J} : \mathcal{H} \times \mathcal{B} \rightarrow L^2(Z, \rho)$  is the linear map

$$\mathcal{J}(f, g) = f + g,$$

(the function  $f + g$  is viewed in a natural way as a function on  $Z$ ).

The following lemma collects some technical facts on  $I_0$  and  $\mathcal{J}$ .

**Lemma 3** *In the notation of above,*

1. *the functional  $I_0 : L^2(Z, \rho) \rightarrow [0, +\infty[$  is well-defined and continuous.*
2. *the operator  $\mathcal{J} : \mathcal{H} \times \mathcal{B} \rightarrow L^2(Z, \rho)$  is well-defined and continuous.*

*Proof* Since the loss function  $V$  can be regarded as function on  $Z \times \mathbb{R}$ , that is,  $V(z, w) = V(y, w)$  where  $z = (\mathbf{x}, y)$ , one has that  $I_0[u]$  is the Nemitski functional associated with  $V$  (see Appendix), that is,

$$I_0[u] = \int_Z V(z, u(z)) \, d\rho(z).$$

We claim that  $I_0[u]$  is finite. Indeed, given  $u \in L^2(Z, \rho)$ , by Eq. (3),

$$\begin{aligned} \int_{X \times Y} V(y, u(z)) \, d\rho(\mathbf{x}, y) &\leq \int_{X \times Y} a(y) + b|u(z)|^2 \, d\rho(\mathbf{x}, y) \\ &\leq \int_{X \times Y} |a(y)| + b|u(z)|^2 \, d\rho(\mathbf{x}, y) < +\infty. \end{aligned}$$

The proof that  $I_0$  is continuous can be found in Proposition III.5.1 of Ekeland and Turnbull (1983).

In order to prove the second item, we let  $f \in \mathcal{H}$ . Then, by Eq. (5),

$$\begin{aligned} \int_{X \times Y} |f(\mathbf{x})|^2 \, d\rho(\mathbf{x}, y) &= \int_{X \times Y} |\langle f, K_{\mathbf{x}} \rangle_{\mathcal{H}}|^2 \, d\rho(\mathbf{x}, y) \\ &\leq \|f\|_{\mathcal{H}}^2 \int_{X \times Y} K(\mathbf{x}, \mathbf{x}) \, d\rho(\mathbf{x}, y) \\ &= C \|f\|_{\mathcal{H}}^2 < +\infty. \end{aligned}$$

where  $C = \int_{X \times Y} K(\mathbf{x}, \mathbf{x}) \, d\rho(\mathbf{x}, y)$  is finite due to Eq. (6). In particular, the function  $(\mathbf{x}, y) \mapsto f(\mathbf{x})$  is in  $L^2(Z, \rho)$  and  $\|f\|_{L^2} \leq \sqrt{C} \|f\|_{\mathcal{H}}$ . The same relation clearly holds for  $g \in \mathcal{B}$ . It follows that  $\mathcal{J}$  is well defined and

$$\|f + g\|_{L^2} \leq \sqrt{C} \|f\|_{\mathcal{H}} + \sqrt{C'} \|g\|_{\mathcal{B}},$$

that is,  $\mathcal{J}$  is continuous. ■

Finally, the following lemma computes the subgradient of  $I = I_0 \circ \mathcal{J}$ .

**Lemma 4** Let  $(f, g) \in \mathcal{H} \times \mathcal{B}$ ; the following two statements are equivalent

1.  $(\phi, \psi) \in \partial I_0(\mathcal{J}(f, g))$
2. there is  $\alpha \in L^2(Z, \rho)$  such that

$$\begin{aligned}\phi(\mathbf{s}) &= \int_{X \times Y} K(\mathbf{s}, \mathbf{x}) \alpha(\mathbf{x}, y) d\rho(\mathbf{x}, y) \\ \psi(\mathbf{s}) &= \int_{X \times Y} K^{\mathcal{B}}(\mathbf{s}, \mathbf{x}) \alpha(\mathbf{x}, y) d\rho(\mathbf{x}, y).\end{aligned}$$

where  $\alpha(\mathbf{x}, y) \in (\partial V)(y, f(\mathbf{x}) + g(\mathbf{x}))$  for  $\rho$ -almost all  $(\mathbf{x}, y)$ .

*Proof* Since  $I_0$  is finite and continuous in  $0 = \mathcal{J}(0)$ , by point 6 of Proposition 13, we know that

$$\partial(I_0 \circ \mathcal{J})(f, g) = \mathcal{J}^*(\partial I_0)(\mathcal{J}(f, g)), \quad (15)$$

where  $\mathcal{J}^* : L^2(Z, \rho) \rightarrow \mathcal{H} \times \mathcal{B}$  is the adjoint of  $\mathcal{J}$ .

First of all, we compute  $\partial I_0$ . Since  $I_0[0] < +\infty$ , we can apply Proposition 14 so that, given  $u \in L^2(Z, \rho)$ , then  $\alpha \in (\partial I_0)(u)$  if and only if  $\alpha \in L^2(Z, \rho)$  and

$$\alpha(z) \in (\partial V)(y, u(\mathbf{x}, y)),$$

for  $\rho$ -almost all  $(\mathbf{x}, y) \in X \times Y$ .

We now compute the adjoint of  $\mathcal{J}$ . Let  $\alpha \in L^2(Z, \rho)$  and  $(\phi, \psi) = \mathcal{J}^* \alpha \in \mathcal{H} \times \mathcal{B}$ . Using the reproducing property of  $\mathcal{H}$  and the definition of  $\mathcal{J}^*$  we can write

$$\begin{aligned}\phi(\mathbf{s}) &= \langle \phi, K_{\mathbf{s}} \rangle_{\mathcal{H}} \\ &= \langle \mathcal{J}^* \alpha, (K_{\mathbf{s}}, 0) \rangle_{\mathcal{H} \times \mathcal{B}} = \langle \alpha, \mathcal{J}(K_{\mathbf{s}}, 0) \rangle_{L^2(Z, \rho)}.\end{aligned}$$

Writing the scalar product explicitly we then find

$$\phi(\mathbf{s}) = \int_{X \times Y} K(\mathbf{s}, \mathbf{x}) \alpha(\mathbf{x}, y) d\rho(\mathbf{x}, y).$$

Reasoning in the same way we find that

$$\psi(\mathbf{s}) = \int_{X \times Y} K^{\mathcal{B}}(\mathbf{s}, \mathbf{x}) \alpha(\mathbf{x}, y) d\rho(\mathbf{x}, y).$$

Replacing the above formulas in Eq. (15), we have the thesis. ■

## 5. Dealing with the offset space $\mathcal{B}$

In this section we deal with the offset term which often appears in regularized solutions. We first motivate our analysis, and then state and discuss our main result on this issue. The formal proof of the presented theorem ends the section.

## 5.1 Motivations

In the previous section we minimized a Tikhonov functional on the set  $\mathcal{H} \times \mathcal{B}$ , dealing explicitly with the possible presence of an offset term in the form of the solution. However, the fact that the set  $\mathcal{H} \times \mathcal{B}$  is not a RKHS makes it cumbersome to extend typical statistical learning results to the general setting in which the offset term is considered. For example a separate analysis, with and without the offset term, is needed for measuring the complexity of the hypothesis space or studying algorithm consistency.

The fact that the estimator is  $f^\lambda(\mathbf{x}) + g^\lambda(\mathbf{x})$  (for regression) or  $\text{sgn}(f^\lambda(\mathbf{x}) + g^\lambda(\mathbf{x}))$  (for classification) suggests to replace  $\mathcal{H} \times \mathcal{B}$  with the sum

$$\mathcal{S} = \mathcal{H} + \mathcal{B} = \{f + g \in \mathcal{C}(X) \mid f \in \mathcal{H}, g \in \mathcal{B}\}.$$

The hypothesis space  $\mathcal{S}$  is a space of functions on  $X$  and, in particular, a RKHS, the kernel being the sum of the kernels of  $\mathcal{H}$  and  $\mathcal{B}$ . In this section we show that the minimization of a Tikhonov functional on  $\mathcal{H} \times \mathcal{B}$  is essentially equivalent to the minimization of an appropriate functional on  $\mathcal{S}$ . This provides a rigorous derivation of the following facts.

1. The equivalent functional on  $\mathcal{S}$  is also a Tikhonov functional. The penalty term is a seminorm penalizing the functions in  $\mathcal{S}$  orthogonal to  $\mathcal{B}$  only.
2. The estimator given by the minimization of the Tikhonov functional on  $\mathcal{S}$  depends only on the kernel sum.

Moreover, since the hypothesis space  $\mathcal{S}$  is a RKHS, a number of classical results of learning theory follows without further effort.

## 5.2 Main theorem

We recall that the norm in  $\mathcal{S}$  is given by

$$\|f + g\|_{\mathcal{S}}^2 = \inf_{\substack{f' \in \mathcal{H}, g' \in \mathcal{B} \\ f+g=f'+g'}} \left( \|f'\|_{\mathcal{H}}^2 + \|g'\|_{\mathcal{B}}^2 \right). \quad (16)$$

and, with respect to this norm, that  $\mathcal{S}$  is a RKHS on  $X$  with kernel  $K + K^{\mathcal{B}}$  (Schwartz, 1964).

We are now ready to state the following result.

**Theorem 5** *Let  $Q$  be the orthogonal projection on the closed subspace of  $\mathcal{S}$*

$$\mathcal{S}_0 = \{s \in \mathcal{S} \mid \langle s, g \rangle_{\mathcal{S}} = 0 \quad \forall g \in \mathcal{B}\},$$

*that is the subset of functions orthogonal to  $\mathcal{B}$  w.r.t. the scalar product in  $\mathcal{S}$ . We have the following facts.*

1. *If  $(f^\lambda, g^\lambda) \in \mathcal{H} \times \mathcal{B}$  is a solution of the problem*

$$\min_{(f,g) \in \mathcal{H} \times \mathcal{B}} \{I[f + g] + \lambda \|f\|_{\mathcal{H}}^2\},$$

*then  $s^\lambda = f^\lambda + g^\lambda \in \mathcal{S}$  is a solution of the problem*

$$\min_{s \in \mathcal{S}} \{I[s] + \lambda \|Qs\|_{\mathcal{S}}^2\}$$

*and  $f^\lambda = Qs^\lambda$ .*

2. If  $s^\lambda \in \mathcal{S}$  is a solution of the problem

$$\min_{s \in \mathcal{S}} \{I[s] + \lambda \|Qs\|_{\mathcal{S}}^2\},$$

let  $f^\lambda = Qs^\lambda$  and  $g^\lambda = s^\lambda - Qs^\lambda$ , then

$$I[f^\lambda + g^\lambda] + \lambda \|f^\lambda\|_{\mathcal{H}}^2 = \inf_{(f,g) \in \mathcal{H} \times \mathcal{B}} \{I[f + g] + \lambda \|f\|_{\mathcal{H}}^2\}.$$

In particular, if  $g^\lambda \in \mathcal{B}$ , then  $(f^\lambda, g^\lambda) \in \mathcal{H} \times \mathcal{B}$  is a minimizer of  $I[f + g] + \lambda \|f\|_{\mathcal{H}}^2$ .

Before giving the proof in the following subsection we comment on this result.

First, notice that if  $\mathcal{H} \cap \mathcal{B} = \{0\}$  then  $\mathcal{S} = \mathcal{H} \times \mathcal{B}$  and

$$\|f + g\|_{\mathcal{S}}^2 = \|f\|_{\mathcal{H}}^2 + \|g\|_{\mathcal{B}}^2.$$

In this case the theorem is trivial. However, in the arbitrary case care is needed because there are functions in  $\mathcal{H}$  not orthogonal to  $\mathcal{B}$ . Moreover, the norm  $\|\cdot\|_{\mathcal{S}}$  restricted to  $\mathcal{H}$  and  $\mathcal{B}$  could be different from  $\|\cdot\|_{\mathcal{H}}$  and  $\|\cdot\|_{\mathcal{B}}$ : in particular, it could happen that  $(\mathcal{B}^\perp)^\perp \neq \mathcal{B}$ , where the orthogonality  $^\perp$  is meant with respect to the dot product in  $\mathcal{S}$ . This pathology is at the root of the fact that there are cases in which the problem

$$\min_{s \in \mathcal{S}} \{I[s] + \lambda \|Qs\|_{\mathcal{S}}^2\}$$

has a solution, whereas the functional  $I[f + g] + \lambda \|f\|_{\mathcal{H}}^2$  does not admit a minimizer on  $\mathcal{H} \times \mathcal{B}$  (see example below). In practice, since  $\mathcal{H} \cap \mathcal{B}$  in most applications is finite dimensional, this pathology does not occur and the minimization problem on  $\mathcal{H} \times \mathcal{B}$  is fully equivalent to the one on  $\mathcal{S}$ .

Second, the advantage of using the penalty term  $\|f\|_{\mathcal{H}}^2$  instead of  $\|Qs\|_{\mathcal{S}}^2$  is that one can solve the minimization problem without knowing the explicit form of the projection  $Q$ . Conversely, the space  $\mathcal{S}$  is the natural space to address theoretical issues.

Finally, we observe that since the proof does not depend on the convexity of the loss function, the theorem holds for arbitrary (positive) loss functions. However, if  $V$  satisfies the hypotheses of Definition 1, from Theorem 2 it follows that the minimizer  $s^\lambda$  of  $I[s] + \lambda \|Qs\|_{\mathcal{S}}^2$  is of the form

$$\begin{aligned} s^\lambda(\mathbf{s}) &= -\frac{1}{2\lambda} \int_{X \times Y} \alpha(\mathbf{x}, y) (K(\mathbf{x}, \mathbf{s}) + K^{\mathcal{B}}(\mathbf{x}, \mathbf{s})) d\rho(\mathbf{x}, y) + g^\lambda(\mathbf{s}) \\ &= -\frac{1}{2\lambda} \int_{X \times Y} \alpha(\mathbf{x}, y) K(\mathbf{x}, \mathbf{s}) d\rho(\mathbf{x}, y) + g^\lambda(\mathbf{s}) \end{aligned}$$

where  $g^\lambda \in \overline{\mathcal{B}}$  and  $\alpha \in L^q(Z, \rho)$  satisfies

$$\begin{aligned} \alpha(\mathbf{x}, y) &\in (\partial V)(y, s^\lambda(\mathbf{x})) \\ \int_{X \times Y} \alpha(\mathbf{x}, y) K^{\mathcal{B}}(\mathbf{x}, \mathbf{s}) &= 0. \end{aligned}$$

In particular, this implies that, given  $h \in \mathcal{B}$ , one can replace the kernel  $K$  with  $K(\mathbf{x}, \mathbf{s}) + h(\mathbf{x})h(\mathbf{s})$ , without changing the form of the minimizer  $s^\lambda$ . For example if  $\mathcal{B}$  is the set of constant functions, the two kernels  $K(\mathbf{x}, \mathbf{s}) = \mathbf{x} \cdot \mathbf{s}$  and  $K(\mathbf{x}, \mathbf{s}) = \mathbf{x} \cdot \mathbf{s} + 1$  are equivalent since both penalize the functions orthogonal to 1, that is the space of linear functions.

### 5.3 Proof

Before giving the proof of Theorem 5 we need to prove the following technical lemma. To this purpose we recall that  $\mathcal{S}_0$  was defined as

$$\mathcal{S}_0 = \{s \in \mathcal{S} \mid \langle s, g \rangle_{\mathcal{S}} = 0 \quad \forall g \in \mathcal{B}\},$$

and  $Q$  was the corresponding orthogonal projection from  $\mathcal{S}$  onto  $\mathcal{S}_0$ . Moreover we let  $\mathcal{H}_0$  be the closed subspace of  $\mathcal{H}$  given by

$$\mathcal{H}_0 = \{f \in \mathcal{H} \mid \langle f, h \rangle_{\mathcal{H}} = 0 \quad \forall h \in \mathcal{H} \cap \mathcal{B}\}$$

and  $P$  be the corresponding orthogonal projection from  $\mathcal{H}$  onto  $\mathcal{H}_0$ .

In order to prove the main theorem we need the following technical lemma that characterizes the space  $\mathcal{S}_0$ .

**Lemma 6** *Let  $s = f + g \in \mathcal{S}$  with  $f \in \mathcal{H}$  and  $g \in \mathcal{B}$ , then*

$$Qs = Pf \tag{17}$$

$$\|Qs\|_{\mathcal{S}} = \|Pf\|_{\mathcal{H}} \tag{18}$$

and there is a sequence  $(f_n, g_n) \in \mathcal{H} \times \mathcal{B}$  such that

$$\lim_{n \rightarrow \infty} \|Pf - f_n\|_{\mathcal{H}} = 0 \tag{19}$$

with  $f_n + g_n = s$ .

Equations (17) and (18) show that  $\mathcal{S}_0$  and  $\mathcal{H}_0$  are the same Hilbert space and, in particular,  $Qs \in \mathcal{H}$ . However, in general, it could happen that  $s - Qs \notin \mathcal{B}$ . Equation (19) is a technical trick to overcome this pathology.

Proof [of Lemma 6] To give the proof of the lemma we need some preliminary facts. Let  $\mathcal{K}$  be the closed subspace of  $\mathcal{H} \times \mathcal{B}$

$$\mathcal{K} = \{(f, g) \in \mathcal{H} \times \mathcal{B} \mid \langle f, h \rangle_{\mathcal{H}} = \langle g, h \rangle_{\mathcal{B}} \quad \forall h \in \mathcal{H} \cap \mathcal{B}\}.$$

It is known (Schwartz, 1964) that, given  $s \in \mathcal{S}$ , there is a unique  $(f, g) \in \mathcal{K}$  such that  $s = f + g$ . Moreover for all  $(f', g') \in \mathcal{H} \times \mathcal{B}$ ,

$$\langle s, f' + g' \rangle_{\mathcal{S}} = \langle f, f' \rangle_{\mathcal{H}} + \langle g, g' \rangle_{\mathcal{B}}. \tag{20}$$

From Eq. (16) one has that

$$\|f\|_{\mathcal{S}} \leq \|f\|_{\mathcal{H}} \quad f \in \mathcal{H} \tag{21}$$

First of all we claim that  $\mathcal{H}_0 \subset \mathcal{S}_0$ . Clearly, if  $f \in \mathcal{H}_0$ , then  $(f, 0) \in \mathcal{K}$  and, by Eq. (20), for all  $g' \in \mathcal{B}$ ,

$$\langle f + 0, 0 + g' \rangle_{\mathcal{S}} = \langle f, 0 \rangle_{\mathcal{H}} + \langle 0, g' \rangle_{\mathcal{B}} = 0,$$

that is  $f \in \mathcal{S}_0$ . This shows the claim. Moreover,

$$\|f\|_{\mathcal{S}}^2 = \langle f + 0, f + 0 \rangle_{\mathcal{S}} = \langle f, f \rangle_{\mathcal{H}} = \|f\|_{\mathcal{H}}^2. \tag{22}$$

Let  $s = f + g$  with  $f \in \mathcal{H}$  and  $g \in \mathcal{B}$ . Clearly,  $f = Pf + h$  where  $h \in \mathcal{H}_0^\perp = ((\mathcal{H} \cap \mathcal{B})^\perp)^\perp = \mathcal{H} \bar{\cap} \mathcal{B}$  (here  $^\perp$  denotes the orthogonal complement with respect to the scalar product of  $\mathcal{H}$ ). It follows that there is a sequence  $h_n \in \mathcal{H} \cap \mathcal{B}$  such that

$$\lim_{n \rightarrow \infty} \|h - h_n\|_{\mathcal{H}} = 0. \quad (23)$$

Since, by Eq. (21),  $\|h - h_n\|_{\mathcal{S}} \leq \|h - h_n\|_{\mathcal{H}}$  and  $Q$  is continuous, it follows that  $Qh = \lim_{n \rightarrow \infty} Qh_n = 0$ , since  $Qh_n = 0$ . The statements of the theorem easily follow from the above facts. Indeed

$$Qs = Q(Pf + h + g) = QPf = Pf,$$

since  $Pf \in \mathcal{H}_0 \subset \mathcal{S}_0$ , and Equation (17) is proved. Equation (18) follows from Eq. (22). Finally let now  $f_n = Pf + h - h_n$  and  $g_n = g + h_n$ . Clearly,  $f_n + g_n = f + g = s$ ,  $f_n \in \mathcal{H}$  and  $g_n \in \mathcal{B}$  and moreover Eq. (19) follows from Eq. (23).  $\blacksquare$

We are now ready to prove the main theorem of this section.

Proof [Theorem 5] First of all we note the following facts. Let  $f \in \mathcal{H}$ ,  $g \in \mathcal{B}$  and  $s = f + g \in \mathcal{S}$ . By Eq. (18)

$$I[s] + \lambda \|Qs\|_{\mathcal{S}}^2 = I[f + g] + \lambda \|Pf\|_{\mathcal{H}}^2 \quad (24)$$

Let  $(f_n, g_n) \in \mathcal{H} \times \mathcal{B}$  as in Lemma 6, then

$$I[f + g] + \lambda \|Pf\|_{\mathcal{H}}^2 = \lim_n \left( I[f_n + g_n] + \lambda \|f_n\|_{\mathcal{H}}^2 \right).$$

From the above equalities it follows that

$$I[s] + \lambda \|Qs\|_{\mathcal{S}}^2 = \lim_n \left( I[f_n + g_n] + \lambda \|f_n\|_{\mathcal{H}}^2 \right) \quad (25)$$

We can now prove the first part of the theorem. Assume that  $(f^\lambda, g^\lambda) \in \mathcal{H} \times \mathcal{B}$  is a minimizer of  $I[f + g] + \lambda \|f\|_{\mathcal{H}}^2$  and let  $s^\lambda = f^\lambda + g^\lambda$ . From Eq. (25) and the definition of minimizer, one has that, for all  $s \in \mathcal{S}$ ,

$$I[s] + \lambda \|Qs\|_{\mathcal{S}}^2 \geq I[f^\lambda + g^\lambda] + \lambda \|f^\lambda\|_{\mathcal{H}}^2. \quad (26)$$

In particular with the choice  $s = s^\lambda$ , by means of Eq. (18), one has that

$$\|Qs\|_{\mathcal{S}} = \|Pf^\lambda\|_{\mathcal{H}} \geq \|f^\lambda\|_{\mathcal{H}},$$

and, hence, that  $Qs^\lambda = Pf^\lambda = f^\lambda$ . Therefore, it follows that

$$I[s] + \lambda \|Qs\|_{\mathcal{S}}^2 \geq I[s^\lambda] + \lambda \|Qs^\lambda\|_{\mathcal{S}}^2,$$

that is,  $s^\lambda$  is a minimizer of  $I[s] + \lambda \|Qs\|_{\mathcal{S}}^2$ .

Before proving the second part of the theorem we note that the following inequality follows as a simple consequence of the definition of projection.

$$I[s] + \lambda \|Qs\|_{\mathcal{S}}^2 = I[f + g] + \lambda \|Pf\|_{\mathcal{H}}^2 \leq I[f + g] + \lambda \|f\|_{\mathcal{H}}^2. \quad (27)$$

Assume now that  $s^\lambda \in \mathcal{S}$  is a minimizer of  $I[s] + \lambda \|Qs\|_{\mathcal{S}}^2$ . Let  $f^\lambda = Qs^\lambda$  and  $g^\lambda = s - f^\lambda$ , then, by Eq. (27) and Eq. (18), it follows that

$$I[f^\lambda + g^\lambda] + \lambda \|f^\lambda\|_{\mathcal{H}}^2 \leq \inf_{(f,g) \in \mathcal{H} \times \mathcal{B}} \{I[f + g] + \lambda \|f\|_{\mathcal{H}}^2\}.$$

However, using Eq. (25) with  $s = f^\lambda + g^\lambda$ , one has that

$$I[f^\lambda + g^\lambda] + \lambda \|f^\lambda\|_{\mathcal{H}}^2 \geq \inf_{(f,g) \in \mathcal{H} \times \mathcal{B}} \{I[f + g] + \lambda \|f\|_{\mathcal{H}}^2\}.$$

So  $I[f^\lambda + g^\lambda] + \lambda \|f^\lambda\|_{\mathcal{H}}^2$  is the infimum of  $I[f + g] + \lambda \|f\|_{\mathcal{H}}^2$  on  $\mathcal{H} \times \mathcal{B}$ . Clearly, if  $g^\lambda \in \mathcal{B}$ , it follows that  $(f^\lambda, g^\lambda)$  is a minimizer of  $I[f + g] + \lambda \|f\|_{\mathcal{H}}^2$ .  $\blacksquare$

#### 5.4 A counterexample

The following example shows that in some pathological framework the minimization on  $\mathcal{H} \times \mathcal{B}$  is not equivalent to the one on  $\mathcal{S} = \mathcal{H} + \mathcal{B}$ .

**Example 1** Let  $\mathcal{H} = \ell_2 = \{f = (f_n)_{n \in \mathbb{N}} \mid \sum_n f_n^2 < +\infty\}$ . The space  $\ell_2$  is a RKHS on  $\mathbb{N}$  with respect to the kernel  $K(n, m) = \delta_{n,m}$ . Let  $\mathcal{B} = \{f \in \ell_2 \mid \sum_n n^2 f_n^2 < +\infty\}$  with the scalar product

$$\langle f, g \rangle_{\mathcal{B}} = \sum_n n^2 f_n g_n.$$

The space  $\mathcal{B}$  is a RKHS with respect to the kernel  $K^{\mathcal{B}}(n, m) = \frac{1}{n^2} \delta_{n,m}$ .

Clearly,  $\mathcal{B} \subset \mathcal{H}$ , so that  $\mathcal{H} \cap \mathcal{B} = \mathcal{B}$ , which is not closed in  $\mathcal{H}$ . Since  $\mathcal{B}$  is dense in  $\mathcal{H}$ ,  $P = 0$  and, by Lemma 6,  $Q = 0$ .

Let  $V$  be the quadratic loss function and choose  $h = (h_n)_{n \in \mathbb{N}} \in \mathcal{H}$  such that  $h \notin \mathcal{B}$ . Let  $\rho(n, y) = \delta(y - h_n)$  so that

$$I[s] = \|s - h\|_{\mathcal{H}}^2,$$

then

$$I[s] + \lambda \|Qs\|_{\mathcal{S}}^2 = \|s - h\|_{\mathcal{H}}^2,$$

and the minimizer is  $s^\lambda = h$ . Moreover, by our theorem, one has that

$$\inf_{f \in \mathcal{H}, g \in \mathcal{B}} \{I[f + g] + \lambda \|f\|_{\mathcal{H}}^2\} = I[s^\lambda] + \lambda \|Qs^\lambda\|_{\mathcal{S}}^2 = 0.$$

If  $(f^\lambda, g^\lambda) \in \mathcal{H} \times \mathcal{B}$  were a minimizer, then  $f^\lambda = 0$  and, hence,  $g^\lambda = h$ , but this is impossible since  $h \notin \mathcal{B}$ .



## 6. Existence and uniqueness

We now discuss existence and uniqueness of the regularized solution in  $\mathcal{S}$ . Before stating and proving the main results we summarize our findings and show that if the offset space is empty both existence and uniqueness are easily obtained.

Our analysis extends existence to all cases of interest under some weak assumptions on the kernel and the loss function for both regression and classification. For classification the assumption of having at least two examples with different labels is also needed.

Uniqueness depends critically on the convexity assumption. For strictly convex functions we prove that the solution is unique if and only if the offset space satisfies certain conditions (always met for offset spaces consisting of constant functions). For loss functions which are not strictly convex we limit our attention to the hinge loss and show that the solution is unique unless some particular conditions on the number and location of the support vectors are met. In Burges and Crisp (2000, 2002) similar results were obtained considering the dual formulation of the minimization problem.

If the offset space is empty, strict convexity and coerciveness of the penalty term trivially imply both existence and uniqueness. Indeed, we have the following proposition.

**Proposition 7** *Given  $\lambda > 0$ , there exists a unique solution of the problem*

$$\min_{f \in \mathcal{H}} \left( I[f] + \lambda \|f\|_{\mathcal{H}}^2 \right).$$

*Proof* The function  $\left( I[f] + \lambda \|f\|_{\mathcal{H}}^2 \right)$  is strictly convex and continuous. Moreover

$$I[f] + \lambda \|f\|_{\mathcal{H}}^2 \geq \lambda \|f\|_{\mathcal{H}}^2 \rightarrow +\infty$$

if  $\|f\|_{\mathcal{H}}$  goes to  $+\infty$ . From item 4 of Proposition 13 both existence and uniqueness follow. ■

### 6.1 Existence

We now consider existence. If  $\mathcal{B}$  is not trivial, there are no general results (see Wahba, 1990, for a discussion on this subject). However, if  $\mathcal{B}$  is the set of constant functions, we derive existence of the solution under very reasonable conditions for both classification and regression.

The first proposition considers the classification case.

**Proposition 8** *Assume that the following conditions hold*

1.  $\lim_{w \rightarrow -\infty} V(1, w) = +\infty$  and  $\lim_{w \rightarrow +\infty} V(-1, w) = +\infty$
2. there is  $C > 0$  such that  $\sqrt{K(\mathbf{x}, \mathbf{x})} \leq C$  for all  $\mathbf{x} \in \text{supp } \nu$
3.  $\rho(X \times \{1\}) > 0$  and  $\rho(X \times \{-1\}) > 0$

*Then there is at least one solution of the problem*

$$\min_{s \in \mathcal{S}} \left( I[s] + \lambda \|Q_S\|_{\mathcal{S}}^2 \right),$$

where  $\mathcal{S} = \mathcal{H} + \mathbb{R}$ .

For regression we have a similar proposition.

**Proposition 9** *Assume that the following conditions hold*

1.  $\lim_{w \rightarrow \pm\infty} (\inf_{y \in Y} V(y, w)) = +\infty$ .
2. *there is  $C > 0$  such that  $\sqrt{K(\mathbf{x}, \mathbf{x})} \leq C$  for all  $\mathbf{x} \in \text{supp } \nu$ .*

*Then there is at least one solution of the problem*

$$\min_{s \in \mathcal{S}} \left( I[s] + \lambda \|Qs\|_{\mathcal{S}}^2 \right),$$

where  $\mathcal{S} = \mathcal{H} + \mathbb{R}$ .

Before proving proposition 8 (the proof of proposition 9, being essentially the same, is omitted), two comments are in order.

First, for both propositions condition 1 is true if  $Y$  is bounded for all loss functions, while condition 2 is always satisfied if  $X$  or  $K$  are bounded. Second the last condition for classification has a very natural interpretation in the discrete setting where it simply amounts to have one example for each class.

Proof [of Proposition 8] The idea of the proof is to show that the functional we have to minimize goes to  $+\infty$  when  $\|s\|_{\mathcal{S}}$  goes to  $+\infty$ . To this aim, let

$$\alpha = \min\{\rho(X \times \{1\}), \rho(X \times \{-1\})\}.$$

By assumption 3,  $\alpha > 0$ . For a fixed  $M > 0$ , we are looking for  $R > 0$  such that for all  $s \in \mathcal{S}$  with  $\|s\|_{\mathcal{S}} \geq R$ ,

$$I[s] + \lambda \|Qs\|_{\mathcal{S}}^2 \geq M.$$

Due to assumption 1, there is  $r > 0$  such that, for all  $w \leq -r$ ,  $V(1, w) \geq \frac{M}{\alpha}$  and, for all  $w \geq r$ ,  $V(-1, w) \geq \frac{M}{\alpha}$ . We now let  $R = \max\{2(1+C)\sqrt{\frac{M}{\lambda}}, 2r\}$  and choose  $s \in \mathcal{S}$  with  $\|s\|_{\mathcal{S}} \geq R$ . If  $\|Qs\|_{\mathcal{S}} = \|Qs\|_{\mathcal{H}} \geq \frac{R}{2(1+C)}$ , then

$$\begin{aligned} I[s] + \lambda \|Qs\|_{\mathcal{S}}^2 &\geq \lambda \|Qs\|_{\mathcal{S}}^2 \\ &\geq \lambda \left( \frac{R}{2(1+C)} \right)^2 \\ &\geq M, \end{aligned}$$

since  $R \geq 2(1+C)\sqrt{\frac{M}{\lambda}}$ . If  $\|Qs\|_{\mathcal{S}} \leq \frac{R}{2(1+C)}$ , let  $b = s - Qs \in \mathbb{R}$ , then

$$\begin{aligned} |b| &= \|s - Qs\|_{\mathcal{S}} \\ &\geq \|s\|_{\mathcal{S}} - \|Qs\|_{\mathcal{S}} \\ &\geq R - \frac{R}{2(1+C)} = R \frac{2C+1}{2C+2} \end{aligned}$$

Assume, for example, that  $b > 0$ . For all  $\mathbf{x} \in \text{supp } \nu$

$$\begin{aligned}
s(\mathbf{x}) &= \langle Qs, K_{\mathbf{x}} \rangle_{\mathcal{H}} + b \\
&\geq b - \|Qs\|_{\mathcal{H}} \|K_{\mathbf{x}}\|_{\mathcal{H}} \\
&\geq R \frac{2C+1}{2C+2} - \frac{R}{2(1+C)} C \\
&\geq R \frac{C+1}{2C+2} \\
&= \frac{R}{2} \geq r,
\end{aligned}$$

since  $R \geq \frac{r}{2}$ . By definition of  $r$ , one has that for all  $\mathbf{x} \in \text{supp } \nu$

$$V(-1, s(\mathbf{x})) \geq \frac{M}{\alpha}.$$

Integrating both sides, we find

$$\int_{X \times \{-1\}} V(-1, s(\mathbf{x})) d\rho(\mathbf{x}, -1) \geq \frac{M}{\alpha} \rho(X \times \{-1\}) \geq M$$

from which it follows that

$$I[s] + \lambda \|Qs\|_{\mathcal{S}}^2 \geq M.$$

The same proof holds when  $b < 0$  replacing the integration on  $X \times \{-1\}$  with the integration on  $X \times \{1\}$ . Since  $M$  is arbitrary, we have that

$$I[s] + \lambda \|Qs\|_{\mathcal{S}}^2 \geq \lambda \|Qs\|_{\mathcal{S}}^2 \rightarrow +\infty.$$

Since the functional is continuous, from item 4 of Proposition 13 the existence of the minimizer follows. ■

## 6.2 Uniqueness

The first proposition completely characterizes uniqueness for strictly convex functions.

**Proposition 10** *Let  $s^\lambda$  be a solution of the problem*

$$\min_{s \in \mathcal{S}} \left( I[s] + \lambda \|Qs\|_{\mathcal{S}}^2 \right).$$

1. *If  $s'$  is another solution, then  $Qs' = Qs^\lambda$ .*
2. *If  $V(y, \cdot)$  is strictly convex for all  $y \in Y$  then all the minimizers are of the form  $s^\lambda + g$ , with  $g \in \mathcal{S}$  such that  $Qg = 0$  and, for all  $\mathbf{x} \in \text{supp } \nu$ ,  $g(\mathbf{x}) = 0$ .*

Let us comment on this proposition before providing the proof. We recall that a solution  $s^\lambda$  is the sum of two terms:  $f^\lambda = Qs^\lambda$  which is orthogonal to  $\mathcal{B}$  and  $g^\lambda = s^\lambda - f^\lambda$ . The uniqueness of  $f^\lambda$  (item 1) is due to the strict convexity of the penalty term. Item 2 states the general conditions

that should be satisfied by offset functions to obtain uniqueness on  $s^\lambda$ : in the discrete setting  $\text{supp}\nu$  reduces to the set of the input points  $\{\mathbf{x}_1, \dots, \mathbf{x}_\ell\}$  and one has uniqueness if and only if the condition  $g(\mathbf{x}_i) = 0$  for all  $i$  implies that  $g$  is equal to zero. Clearly, if  $\mathcal{B}$  is the space of constant functions uniqueness is ensured. We now give the proof of the proposition.  
 Proof [of Proposition 10]

1. Let  $s'$  another minimizer and assume that  $Qs^\lambda \neq Qs'$ . Then, by the strict convexity of  $\|\cdot\|_S^2$ , one has that, for all  $t \in ]0, 1[$ ,

$$\left\| (1-t)Qs^\lambda + tQs' \right\|_S^2 < (1-t) \left\| Qs^\lambda \right\|_S^2 + t \left\| Qs' \right\|_S^2.$$

Since  $I[s]$  is convex, one has that

$$I[(1-t)s^\lambda + ts'] \leq (1-t)I[s^\lambda] + tI[s'].$$

From the above two inequalities we find

$$\begin{aligned} I[(1-t)s^\lambda + ts'] + \lambda \left\| Q \left( (1-t)s^\lambda + ts' \right) \right\|_S^2 & < (1-t) \left( I[s^\lambda] + \lambda \left\| Qs^\lambda \right\|_S^2 \right) + t \left( I[s'] + \lambda \left\| Qs' \right\|_S^2 \right) \\ & = \min_{s \in \mathcal{S}} \left( I[s] + \lambda \left\| Qs \right\|_S^2 \right). \end{aligned}$$

Since this is impossible, it follows that  $Qs^\lambda = Qs'$ .

2. Let  $s' = s^\lambda + g$  with  $g$  as in item 2. By straightforward computation we have that  $s'$  is a minimizer. It is left to show that the minimizers are only the functions written in the above form. From item 1 we have that  $Qg = 0$ . Let  $U$  be the open set

$$U = \{\mathbf{x} \in X \mid g(\mathbf{x}) \neq 0\} = \{\mathbf{x} \in X \mid s'(\mathbf{x}) \neq s^\lambda(\mathbf{x})\}.$$

By contradiction, let us assume that  $U \cap \text{supp}\nu \neq \emptyset$ , then  $\nu(U) > 0$  and, hence,  $\rho(U \times Y) > 0$ . Fix  $t \in ]0, 1[$ . since  $V(y, \cdot)$  is strictly convex, for all  $(\mathbf{x}, y) \in U \times Y$ , one has that

$$V(y, (1-t)s^\lambda(\mathbf{x}) + ts'(\mathbf{x})) < (1-t)V(y, s^\lambda(\mathbf{x})) + tV(y, s'(\mathbf{x})).$$

Therefore, by integration,

$$\begin{aligned} \int_{U \times Y} V(y, (1-t)s^\lambda(\mathbf{x}) + ts'(\mathbf{x})) d\rho(\mathbf{x}, y) & < \\ & < (1-t) \int_{U \times Y} V(y, s^\lambda(\mathbf{x})) d\rho(\mathbf{x}, y) + t \int_{U \times Y} V(y, s'(\mathbf{x})) d\rho(\mathbf{x}, y). \end{aligned}$$

On the complement of  $U \times Y$ , we have  $V(y, s^\lambda(\mathbf{x})) = V(y, s'(\mathbf{x}))$ , so that

$$I[(1-t)s^\lambda + ts'] < (1-t)I[s^\lambda] + tI[s'].$$

By the same line of reasoning of item 1, one finds a contradiction. It follows that  $U \cap \text{supp}\nu = \emptyset$ , which is equivalent to the thesis.

■

Two important examples of convex loss functions which are not strictly convex are the hinge and the  $\epsilon$ -insensitive loss. The next proposition deals with the hinge loss though a similar result can be also derived for the  $\epsilon$ -insensitive loss. For the sake of simplicity we develop our result in the discrete setting for the case of constant offset functions. In this case uniqueness of the solution is expressed as a condition on the number of support vectors of the two classes. Similar but a little bit more involved conditions can be found considering the continuous setting.

**Proposition 11** *Let  $Y = \{\pm 1\}$ ,  $V(y, w) = |1 - yw|_+$  and  $\mathcal{B} = \mathbb{R}$ . Let  $s^\lambda$  be a solution of*

$$\min_{s \in \mathcal{S}} \left( \frac{1}{\ell} \sum_{i=1}^{\ell} V(y_i, s(\mathbf{x}_i)) + \lambda \|Qs\|_{\mathcal{S}}^2 \right),$$

and define

$$\begin{aligned} I_+ &= \{i \mid y_i = 1, s^\lambda(\mathbf{x}) < 1\} & I_- &= \{i \mid y_i = -1, s^\lambda(\mathbf{x}) > -1\} \\ B_+ &= \{i \mid y_i = 1, s^\lambda(\mathbf{x}_1) = 1\} & B_- &= \{i \mid y_i = -1, s^\lambda(\mathbf{x}_1) = -1\}. \end{aligned}$$

The solution is not unique if and only if

$$\#I_+ = \#I_- + \#B_- \tag{28}$$

or

$$\#I_- = \#I_+ + \#B_+, \tag{29}$$

where  $\#$  denotes set cardinality.

*Proof* Assume that  $s'$  is another solution. From item 1 of proposition 10, we have that  $Qs^\lambda = Qs'$  and  $s' = s^\lambda + b$ . Since both functions are minimizers, one concludes that

$$\sum_{i=1}^{\ell} |1 - y_i s^\lambda(\mathbf{x}_i)|_+ = \sum_{i=1}^{\ell} |1 - y_i s'(\mathbf{x}_i)|_+ \tag{30}$$

We notice that if  $yw_1 < 1$  and  $yw_2 > 1$ , then

$$V(y, (1-t)w_1 + tw_2) < (1-t)V(y, w_1) + tV(y, w_2).$$

Reasoning as in the proof of the previous proposition, one has that, for all  $i \in I_+ \cup I_-$ ,

$$y_i s'(\mathbf{x}_i) \leq 1$$

and, for all  $i \notin (I_+ \cup I_- \cup B_+ \cup B_-)$

$$y_i s'(\mathbf{x}_i) \geq 1.$$

Using the above two equations, it follows that equality (30) becomes

$$\sum_{i \in I_+ \cup I_-} (1 - y_i s^\lambda(\mathbf{x}_i)) = \sum_{i \in I_+ \cup I_-} (1 - y_i s'(\mathbf{x}_i)) + \sum_{i \in B_+ \cup B_-} |-by_i|_+,$$

(if the index set is empty, we let the corresponding sum be equal to 0). The above equation is equivalent to

$$\sum_{i \in I_+ \cup I_-} by_i = \sum_{i \in B_+ \cup B_-} |-by_i|_+,$$

that has a not trivial solution if and only if both the following conditions are true

1. if  $b > 0$ , then  $\sum_{i \in I_+ \cup I_-} y_i = -\sum_{B_-} y_i$  (that is, Eq. (28) holds).
2. if  $b < 0$ , then  $\sum_{i \in I_+ \cup I_-} y_i = \sum_{B_+} y_i$  (that is, Eq. (29) holds).

Now, if neither Eq. (28) nor Eq. (29) holds, then  $b = 0$  and  $s^\lambda$  is unique. Conversely, assume for example that Eq. (28) holds. It is simple to check that there is  $b > 0$  such that for all  $i \in I_+ \cup I_-$ ,

$$y_i(s^\lambda(\mathbf{x}_i) + b) \leq 1$$

and, for all  $i \notin (I_+ \cup I_- \cup B_+ \cup B_-)$

$$y_i(s^\lambda(\mathbf{x}_i) + b) \geq 1.$$

Finally, by direct computation one has that

$$I[s^\lambda] = I[s^\lambda + b].$$

■

If the solution is not unique, the solution family is parameterized as  $s^\lambda + b$ , where  $b$  runs in a closed, not necessarily bounded interval. However, if there is at least one example for each class,  $b$  lies in the bounded interval  $[b_-, b_+]$  and one can easily show that

1. for the solution with  $b = b_-$ , Eq. (28) holds;
2. for the solution with  $b = b_+$ , Eq. (29) holds;
3. for the solution with  $b_- < b < b_+$ , both Eqs. (28) and (29) hold, from which it follows that  $\#I_+ = \#I_-$  and  $\#B_+ = \#B_- = 0$ .

## 7. Discrete Tikhonov Regularization

We now specialize our results to the case in which the probability measure is the empirical distribution  $\rho_S$  and  $\mathcal{B}$  is the space of constant functions ( $K^{\mathcal{B}} = 1$ ) and discuss in detail Support Vector Machines for classification.

We start by recalling that, from item 2 of Proposition 13 it follows that the left and right derivatives of  $V(y, \cdot)$  always exist and

$$(\partial V)(y, w) = [V'_-(y, w), V'_+(y, w)].$$

**Corollary 12** *Let  $\mathcal{S}$  be the RKHS with kernel  $K + 1$ . The following three conditions are equivalent.*

1. The pair  $(f^\lambda, b^\lambda) \in \mathcal{H} \times \mathbb{R}$  is a solution of

$$\min_{f \in \mathcal{H}, b \in \mathbb{R}} \left( \frac{1}{\ell} \sum_i V(y_i, f(\mathbf{x}_i) + b) + \lambda \|f\|_{\mathcal{H}}^2 \right)$$

2. The function  $s^\lambda = f^\lambda + b^\lambda$  is a solution of

$$\min_{s \in \mathcal{S}} \left( \frac{1}{\ell} \sum_i V(y_i, s(\mathbf{x}_i)) + \lambda \|Qs\|_{\mathcal{H}}^2 \right)$$

where  $Q$  is the projection on the functions that are orthogonal to 1 and  $f^\lambda = Qs^\lambda$ .

3. There are  $\alpha_1, \dots, \alpha_\ell \in \mathbb{R}$  such that

$$f^\lambda = \sum_{i=1}^{\ell} \alpha_i K_{\mathbf{x}_i} = \sum_{i=1}^{\ell} \alpha_i (K_{\mathbf{x}_i} + 1).$$

where

$$\begin{aligned} \frac{-1}{2\lambda\ell} V'_+(y_i, f^\lambda(\mathbf{x}_i) + b^\lambda) &\leq \alpha_i \leq \frac{-1}{2\lambda\ell} V'_-(y_i, f^\lambda(\mathbf{x}_i) + b^\lambda) \\ \sum_{i=1}^{\ell} \alpha_i &= 0 \end{aligned}$$

We notice two facts. First,  $\alpha_i$  is zero if and only if  $0 \in (\partial V)(y_i, f^\lambda(\mathbf{x}_i) + b^\lambda)$  – that is, if and only if  $f^\lambda(\mathbf{x}_i) + b^\lambda$  is a minimizer of  $V(y_i, \cdot)$ . Therefore, a necessary condition for obtaining sparsity is a *plateaux* in the loss function. A quantitative discussion on this topic can be found in Steinwart (2002b). Second if  $V'_-$  and  $V'_+$  are bounded by a constant  $M > 0$ , one has that  $|\alpha_i| \leq (2\lambda\ell)^{-1} M$  – that is, a sufficient condition for box constraints on the coefficients.

In the rest of this section we consider Support Vector Machines for classification showing that through our analysis the solution is completely characterized in the primal formulation.

A simple calculation for the hinge loss shows that

$$[V'_-(y, w), V'_+(y, w)] = \begin{cases} -y & \text{for } yw < 1 \\ [\min\{-y, 0\}, \max\{0, -y\}] & \text{for } yw = 1 \\ 0 & \text{for } yw > 1 \end{cases} . \quad (31)$$

To be consistent with the notation used in the literature, we let  $C = \frac{1}{2\lambda\ell}$  and factorize the labels  $y_i$  from the coefficients  $\alpha_i$ . Then, according to the above corollary, the solution of the SVM algorithm is given by

$$s^\lambda = \sum_{i=1}^{\ell} \alpha_i y_i K_{\mathbf{x}_i} + b^\lambda$$

where the set  $(\alpha_1, \dots, \alpha_\ell, b^\lambda)$  solves the following algebraic system of inequalities

$$\begin{aligned}
0 \leq \alpha_i \leq C & \quad \text{if} \quad y_i \left( \sum_{j=1}^{\ell} \alpha_j y_j K(\mathbf{x}_i, \mathbf{x}_j) + b^\lambda \right) = 1 \\
\alpha_i = 0 & \quad \text{if} \quad y_i \left( \sum_{j=1}^{\ell} \alpha_j y_j K(\mathbf{x}_i, \mathbf{x}_j) + b^\lambda \right) > 1 \\
\alpha_i = C & \quad \text{if} \quad y_i \left( \sum_{j=1}^{\ell} \alpha_j y_j K(\mathbf{x}_i, \mathbf{x}_j) + b^\lambda \right) < 1 \\
\sum_i \alpha_i y_i & = 0
\end{aligned} \tag{32}$$

Interestingly, the above inequalities, which fully characterize the support vectors associated to the solution, are usually obtained as the Kuhn-Tucker conditions of the dual QP optimization problem (Vapnik, 1988).

Looking at Eqs.(31-32), it is immediate to see that the box constraints  $(0 \leq \alpha_i \leq C)$  are due to the linearity of  $V(yf(\mathbf{x}))$  for  $yf(\mathbf{x}) < 1$ , whereas sparsity  $(\alpha_i = 0)$  follows from the constancy of  $V(yf(\mathbf{x}))$  for  $yf(\mathbf{x}) > 1$ .

## 8. Conclusion

In this paper we study some properties of learning functionals derived from Tikhonov regularization. We develop our analysis in a continuous setting and use tools from convex analysis in infinite dimensional spaces to quantitatively characterize the explicit form of the regularized solution for both regression and classification. We also address the case with and without the offset term within the same unifying framework. We show that the presence of an offset term is equivalent to solving a standard problem of regularization in a Reproducing Kernel Hilbert Space in which the penalty term is given by a seminorm. Finally, we discuss issues of existence and uniqueness of the solution and specialize our results to the discrete setting.

Current work aims at extending these results to vector-valued functions (Micchelli and Pontil, 2003) and exploring possible use of offset functions to incorporate invariances (Girosi and Chan, 1995).

## Acknowledgment

L. Rosasco is supported by an INFM fellowship. A. Caponnetto is supported by a PRIN fellowship within the project ‘‘Inverse problems in medical imaging’’, n. 2002013422. This research has been partially funded by the INFM Project MAIA, the FIRB Project ASTA, and the EU Project KerMIT.

## Appendix A. Convex functions in infinite dimensional spaces

The proof of Theorem 2 is based on the properties of convex functions defined on infinite dimensional spaces. In particular, we use the notion of subgradient that extends the notion of



derivative to convex non-differentiable functions. In this appendix we collect the results we need. For details see the book Ekeland and Turnbull (1983) and also Ekeland and Teman (1974).

Let  $\mathcal{H}$  be a Hilbert space. A function  $F : \mathcal{H} \rightarrow \mathbb{R} \cup +\infty$  is *convex* if

$$F(tv + (1-t)w) \leq tF(v) + (1-t)F(w),$$

for all  $v, w \in \mathcal{H}$  and  $t \in [0, 1]$  (if the strict inequality holds for  $t \in (0, 1)$ ,  $F$  is called *strictly convex*).

Let  $v_0 \in \mathcal{H}$  such that  $F(v_0) < +\infty$ . The *subgradient* of  $F$  at point  $v_0 \in \mathcal{H}$  is the subset of  $\mathcal{H}$  given by

$$\partial F(v_0) = \{w \in \mathcal{H} \mid F(v) \geq F(v_0) + \langle w, v - v_0 \rangle_{\mathcal{H}} \quad \forall v \in \mathcal{H}\}. \quad (33)$$

If  $F(v) = +\infty$ , we let  $\partial F(v_0) = \emptyset$ .

In the following proposition we summarize the main properties of the subgradient we need.

**Proposition 13** *The following facts hold:*

1. If  $F$  is differentiable at  $v_0$ , the subgradient reduces to the usual gradient  $F'(v_0)$ .
2. If  $F$  is defined on  $\mathbb{R}$  and  $F(v_0) < +\infty$ , then  $F$  admits left and right derivative and

$$\partial F(v_0) = [F'_-(v_0), F'_+(v_0)].$$

3. Assume that  $F \neq +\infty$ . A point  $v_0$  is a minimizer of  $F$  if and only if  $0 \in \partial F(v_0)$ .

4. If  $F$  is continuous and

$$\lim_{\|v\|_{\mathcal{H}} \rightarrow \infty} F(v) = +\infty.$$

then  $F$  has a minimizer. If  $F$  is strictly convex, the minimizer is unique.

5. Let  $G$  be another convex function on  $\mathcal{H}$ . Assume that there is  $v_0 \in \mathcal{H}$  such that  $F$  and  $G$  are continuous and finite at  $v_0$ . Let  $a, b \geq 0$ , then  $aF + bG$  is convex and, for all  $v \in \mathcal{H}$ ,

$$\partial(aF + bG)(v) = a(\partial F)(v) + b(\partial G)(v).$$

6. Let  $\mathcal{H}'$  be another Hilbert space and  $\mathcal{J}$  be a continuous linear operator from  $\mathcal{H}'$  into  $\mathcal{H}$ . Assume that there is  $v'_0 \in \mathcal{H}'$  such that  $F$  is continuous and finite at  $\mathcal{J}v'_0$ . For all  $v' \in \mathcal{H}'$

$$\partial(F \circ \mathcal{J})(v') = \mathcal{J}^*(\partial F)(\mathcal{J}v'),$$

where  $\mathcal{J}^* : \mathcal{H} \rightarrow \mathcal{H}'$  is the adjoint of  $\mathcal{J}$  defined by

$$\langle v', \mathcal{J}^*v \rangle_{\mathcal{H}'} = \langle \mathcal{J}v', v \rangle_{\mathcal{H}}.$$

for all  $v \in \mathcal{H}$  and  $v' \in \mathcal{H}'$ .

**Proof** We simply give the references to the book of Ekeland and Turnbull (1983).

1. Prop. III.2.8
2. Prop. III.2.7

3. It is a simple consequence of Prop. III.3.1
4. It is a simple consequence of Prop. II.4.6.
5. Prop. III.2.13
6. Prop. III.2.8

■

We now recall the definition of *Nemitski* functional, adapted to our framework. Let  $\rho$  be a finite measure on  $Z$  (a closed subset of  $\mathbb{R}^n$ ) and  $W : Z \times \mathbb{R} \rightarrow [0, +\infty[$  be a continuous function on  $\mathbb{R} \times Z$  such that  $W(z, \cdot)$  is convex for all  $z \in Z$ .

Let  $L^2(Z, \rho)$  be the Hilbert space of square-integrable functions with respect to  $\rho$ . The *Nemitski* functional associated to  $W$  is defined as the map  $I_0 : L^2(Z, \rho) \rightarrow [0, +\infty]$  given by

$$I_0[u] = \int_Z W(z, u(z)) d\rho(z). \quad (34)$$

Next proposition provides us with a straightforward method to study the subgradient  $(\partial I_0)$ .

**Proposition 14** *Assume that there is an element  $u_0 \in L^2(Z, \rho)$  such that  $\sup_{z \in Z} |u_0(z)| < +\infty$  and  $I_0[u_0] < +\infty$ . Given  $u \in L^2(Z, \rho)$*

$$(\partial I_0)(u) = \{w \in L^2(Z, \rho) \mid w(z) \in (\partial W)(z, u(z)) \text{ } \rho - \text{a.e.}\}. \quad (35)$$

*Proof* See the proof of Prop. III.5.3 of Ekeland and Turnbull (1983) for  $p = 2$  observing that the fact that  $V$  is continuous on  $Z$  assures that  $V$  is a Borel function and that  $V(z, \cdot)$  is lower semicontinuous. The proof is for  $Z$  interval of  $\mathbb{R}$ , but can be easily extended to arbitrary  $Z$ , compare with Ekeland and Teman (1974). ■

## References

- M. Bertero, T. Poggio, and V. Torre. Ill-posed problems in early vision. *Proceedings of the IEEE*, 76:869–889, 1988.
- C. Burges and D. Crisp. The uniqueness of the svm solution. In *Proceedings of the Twelfth Conference on Neural Information Processing Systems*, pages 244–250, Cambridge, MA, 2000. MIT Press.
- C. Burges and D. Crisp. Uniqueness theorems for kernel methods. Technical Report MSR-TR-2002-11, Microsoft Research, 2002.
- D. Cox and F. O’Sullivan. Asymptotic analysis of penalized likelihood and related estimators. *Ann. Stat.*, 18:1676–1695, 1990.
- N. Cristianini and J. Shawe Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, Cambridge, UK, 2000.

- F. Cucker and S. Smale. Best choices for regularization parameters in learning theory: on the bias-variance problem. *Foundations of Computational Mathematics*, 2:413–428, 2002.
- I. Ekeland and R. Teman. *Analyse convexe et problèmes variationnels*. Gauthier-Villards, Paris, 1974.
- I. Ekeland and T. Turnbull. *Infinite-dimensional Optimization and Convexity*. Chicago Lectures in Mathematics. The University of Chicago Press, Chicago, 1983.
- T. Evgeniou, Pontil M., and T. Poggio. Regularization networks and support vector machines. *Adv. Comp. Math.*, 13:1–50, 2000.
- F. Girosi. An equivalence between sparse approximation and support vector machines. *Neural Computation*, 10:1455–1480, 1998.
- F. Girosi and N. Chan. Prior knowledge and the creation of virtual examples for rbf networks. In *Proceedings of the IEEE-SP Workshop on Neural Networks Signal Processing*, pages 201–210, Cambridge, MA, 1995. IEEE Signal Processing Society.
- G. Kimeldorf and G. Wahba. A correspondence between bayesian estimation of gaussian processes and gacsmoothing by splines. *Ann. Math. Stat.*, 41:495–502, 1970.
- C.A. Micchelli and M. Pontil. On learning vector-valued functions. *Research Note RN/03/08, Dept of Computer Science, UCL*, 2003.
- S. Mukherjee, T. Niyogi, P. and Poggio, and R. Rifkin. Statistical learning: Loo stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *CBCL Paper 223, Massachusetts Institute of Technology, Cambridge, MA*, 2002.
- P. Niyogi and F. Girosi. Generalization bounds for function approximation from scattered data. *Adv. Comp. Math.*, 10:51–80, 1999.
- T. Poggio and F. Girosi. A theory of networks for approximation and learning. In C. Lau, editor, *Foundation of Neural Networks*, pages 91–106. IEEE Press, Piscataway, N.J., 1992.
- T. Poggio, Mukherjee, Rifkin S., A. R., Rakhlin, and A. Verri. b. Technical Report AI-Moemo 2001-011, MIT Artificial Intelligence Laboratory, 2001.
- T. Poggio and S. Smale. The mathematics of learning: Dealing with data. *Notices of the American Mathematical Society (AMS)*, 50:537–544, 2003.
- B. Schölkopf and A.J. Smola. Learning with kernels. Technical Report 644, Cambridge, MA, 2002.
- B. Schölkopf, R. Herbrich, A. Smola, and R.C. Williamson. A generalized representer theorem. In *Proceedings of 14th COLT, Lectures Notes in Artificial Intelligence 2111*, pages 416–426. Springer-Verlag, 2001.
- L. Schwartz. Sous-espaces hilbertiens d’espaces vectoriels topologiques and noyaux associés. *Journal d’Analyse Mathématique*, 13:115–256, 1964.

- I. Steinwart. Consistency of support vector machines and other regularized kernel machines. *submitted to IEEE Transactions on Information Theory*, 2002a.
- T. Steinwart. Sparseness of support vector machines. *submitted to IEEE Transactions on Information Theory*, 2002b.
- A.N. Tikhonov and V.Y. Arsenin. *Solutions of Ill Posed Problems*. W. H. Winston, Washington, D.C., 1977.
- V. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1988.
- G. Wahba. *Splines Models for Observational Data*, volume 59 of *Series in Applied Mathematics*. SIAM, 1990.
- G. Wahba. Support vector machines, reproducing kernel hilbert spaces and randomized gacv. Technical Report 984, Department of Statistics, University of Wisconsin, 1998.
- T. Zhang. Convergence of large margin separable linear classification. In T.G. Leen, T.K. Dietterich and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 357–363. MIT Press, 2001.