Journal of Machine Learning Research 5 (2004) 1363–1390 Submitted 12/03; Revised 7/04; Published 10/04

Some Properties of Regularized Kernel Methods

Ernesto De Vito DEVITO@UNIMO.IT
Dipartimento di Matematica
Università di Modena
Modena, Italy and
INFN, Sezione di Genova
Genova, Italy

Lorenzo Rosasco ROSASCO@DISI.UNIGE.IT
Andrea Caponnetto CAPONNETTO@DISI.UNIGE.IT
DISI
Università di Genova,
Genova, Italy

Michele Piana PIANA@DIMA.UNIGE.IT
DIMA
Università di Genova,
Genova, Italy

Alessandro Verri VERRI@DISI.UNIGE.IT
DISI
Università di Genova,
Genova, Italy

Editor: Alexander J. Smola

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 Support Vector Machines 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 an 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; Cristianini and Shawe Taylor, 2000; Schölkopf and Smola, 2002). More precisely, given a training set $S = (\mathbf{x}_i, y_i)_{i=1}^{\ell}$ of ℓ pairs of examples, the estimator is defined as

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

where V is the loss function, $\mathcal{H}$ is the Hilbert space of the *hypotheses* and $\lambda > 0$ is the regularization parameter. 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 (Bertero et al., 1988; Evgeniou et al., 2000; Poggio and Smale, 2003).

In this paper we study the generalization of the above problem to the *continuous setting*, that is, given a probability distribution ρ defined on $X \times Y$ where X is the input space and Y is the output space, we study the properties of the estimator

$$(f^{\lambda}, g^{\lambda}) \in \operatorname*{argmin}_{(f,g) \in \mathcal{H} \times \mathcal{B}} \left\{ \int_{X \times Y} V(y, f(\mathbf{x}) + g(\mathbf{x})) d\rho(\mathbf{x}, y) + \lambda \| f \|_{\mathcal{H}}^2 \right\}, \tag{2}$$

where $\mathcal{H}$ and $\mathcal{B}$ are reproducing kernel Hilbert spaces (RKHS): $\mathcal{H}$ is the space of penalized functions and $\mathcal{B}$ is the offset space (Wahba, 1990).

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, 2002). To this purpose, one associates with each function $f : X \to \mathbb{R}$ its expected risk,

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

where ρ is the unknown probability distribution describing the relation between the input $\mathbf{x} \in X$ and the output $y \in Y$. Following Cucker and Smale (2002), for regularized kernel methods the discrepancy between the expected risk of the estimator, f_S^{λ}, and the minimum obtainable risk, $\inf_{f \in \mathcal{H}} I[f]$, can be decomposed as

$$I[f_S^{\lambda}] - \inf_{f \in \mathcal{H}} I[f] = \left(I[f_S^{\lambda}] - I[f^{\lambda}] \right) + \left(I[f^{\lambda}] - \inf_{f \in \mathcal{H}} I[f] \right),$$

where the first term represents the sample error and the second term the approximation error (Niyogi and Girosi, 1999). Clearly, insight on the form of f^{λ} can be useful to obtain better bounds on both errors. Second, considering the continuous measure ρ 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 in a unified framework.

The contribution of our work is threefold. First we provide a complete characterization of the explicit form of the estimator (f^λ, g^λ) given by Eq. (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 Problem (2) is not the *pair* (f^λ, g^λ) but the *sum* $f^\lambda + g^\lambda$. 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 ρ: 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) for the squared loss function (see also Wahba, 1990). However, the structure of the proof holds for arbitrary loss function as shown by many authors such as Cox and O'Sullivan (1990). In the framework of statistical learning, Schölkopf et al. (2001) give a proof of the representer theorem that holds for an arbitrary loss function and for any penalty term, being it a strictly increasing function of the norm. This kind of results shows that, if the $\mathcal{H}$ is a RKHS with kernel K, the estimator f_S^λ defined by Eq. (1) can be written as

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

The above result holds for arbitrary loss function and for a large class of penalty terms. However, the form of the coefficients α_i is unknown.

For the squared 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 Poggio and Girosi (1992); Girosi (1998); Wahba (1998) where the coefficients α_i are solution of a system of algebraic equations.

This approach cannot be applied to hinge and ε-insensitive loss function (Vapnik, 1988), since they are not differentiable: the form of the coefficients α_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) gives a quantitative representer theorem in the classification setting that holds for differentiable loss function and Steinwart (2003) extends this result for arbitrary convex loss function, without using the dual problem. In these papers the form of the coefficients α_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 techniques similar to those of Steinwart (2003), extends the above result in the following directions:

- 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 α_i to satisfy a system of linear equations;
- we do not assume that input space X and the output space Y are compact. In particular, for regression we can assume $Y = \mathbb{R}$;
- we provide a simpler proof than the one of Steinwart (2003) 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. (2002) when the space $\mathcal{B}$ reduces to the set of constant functions. The results are close to our Theorem 6, 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 RKHS.

The problem of the existence and uniqueness is discussed in Wahba (1998) for the discrete setting and with differentiable loss functions. For arbitrary ρ the papers by Steinwart (2002, 2003) study the existence for the classification setting with offset space reduced to the constant functions. For the hinge loss and ε-insensitive loss, the problem of uniqueness is treated in Burges and Crisp (2000, 2003). 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 SVM solution are similar to those in Burges and Crisp (2000, 2003) 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 upon 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 locally compact second countable space (this assumption is satisfied for instance if X is a closed subset of $\mathbb{R}^d$) and Y is a closed subspace of $\mathbb{R}$.

We let $Z = X \times Y$ and endow it with a probability distribution ρ defined on the Borel σ-algebra of Z. We recall that, since ρ is a bounded measure and Z is second countable, ρ is a Radon measure. In practice, ρ 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 ρ. We now deal with loss functions.

3.2 Loss Functions

We collect the mathematical assumptions on the loss function in the following definition and we comment on the purpose of each assumption.

Definition 1 *Given $p \in [1, +\infty[$, a function $V : Y \times \mathbb{R} \to [0, +\infty[$ such that*

1. *for all $y \in Y$ the function $V(y, \cdot)$ is convex on $\mathbb{R}$;*
2. *the function V is measurable on $Y \times \mathbb{R}$;*
3. *there are $b \in [0, +\infty[$ and $a : Y \to [0, +\infty[$ such that*

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

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

is called a p-loss function with respect to ρ.

If the context is clear, V is simply called a loss function. 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. Finally, this requirement ensures stronger bounds for the sample error (Bartlett et al., 2002; Bartlett, 2003; Bartlett et al., 2003).

Assumption 2. is a minimal requirement for defining the expected risk and it is usually satisfied since loss functions commonly in use are continuous on Z.

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. for $p = 2$, 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. for $p = 1$, if $V(y,\cdot)$ is Lipschitz on $\mathbb{R}$ with a Lipschitz constant independent of y and

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

We now restrict our analysis to some functionals studied in statistical learning.

3.3 Learning Functionals

The *expected risk* of a measurable function $f : X \to \mathbb{R}$ 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 ρ is unknown.

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

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

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 ρ replaces the empirical measure ρ_S in the first term. In what follows we will refer to the functionals to be minimized in both Eq. (1) and Eq. (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

First of all, we recall the definition of reproducing kernel Hilbert space. A RKHS $\mathcal{H}$ on X with kernel $K : X \times X \to \mathbb{R}$ is defined as the unique Hilbert space of real valued functions on X such that, for all $f \in \mathcal{H}$,

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

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

Given a probability meausure ρ on Z and $p \in [1,+\infty[$, we say that the kernel K is p-bounded with respect to ρ if the function K is measurable on $X\times X$ and

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

Clearly the above condition depends only on the marginal distribution of ρ on X and ensures that $\mathcal{H}$ is a subspace of $L^p(Z,\rho)$ with continuous inclusion (see Lemma 4 in Section 4). This fact is

essential for proving our results. In particular, the p-boundedness of the kernel is fulfilled for all $p \in [1, +\infty[$ if X is compact and the kernel is continuous or if the kernel is measurable and bounded.

We are now ready to discuss the assumptions on the hypothesis space. We fix the probability measure ρ on Z and $p \in [1, +\infty[$ such that V is p-bounded with respect to ρ. We require that the space of penalized functions $\mathcal{H}$ and the space of offset functions $\mathcal{B}$ are RKHS on X such that the corresponding kernels K and $K^{\mathcal{B}}$ are p-bounded with respect to ρ. We denote the corresponding norms by $\|\cdot\|_{\mathcal{H}}$ and $\|\cdot\|_{\mathcal{B}}$. Finally, we notice that, in general, the product space $\mathcal{H} \times \mathcal{B}$ is not a RKHS.

In learning theory usually X is compact, K is continuous and $\mathcal{B}$ is the one dimensional vector space of constant functions

$$\mathcal{B} = \{ f : X \to \mathbb{R} \,|\, f(\mathbf{x}) = b,\ b \in \mathbb{R} \} = \mathbb{R}$$

with kernel $K^{\mathcal{B}}$ simply given by $K^{\mathcal{B}}(\mathbf{x}, \mathbf{s}) = 1$. Another example of offset space, which arises in approximation problems in RKHS on a bounded interval, is the space of splines of order n, whose corresponding kernel is continuous (Wahba, 1990). In both case the p-boundedness assumption is satisfied for all p. Our framework allows to treat arbitrary (possibly infinite-dimensional) offset spaces with the possibility to incorporate jumps in the offset term.

Finally, the requirement that the hypothesis space is a RKHS is due to the fact that minimization of a convex functional in a Hilbert space is easier to treat than in an arbitrary Banach space since in the former case the subgradient of the functional is an element of the space itself. Moreover, in the proofs we use extensively the reproducing property given by Eq. (5).

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 *Let ρ be a probability measure on $X \times Y$ where X is a locally compact second countable space and Y is a closed subset of $\mathbb{R}$. Let V be a p-loss function with respect to ρ, $p \in [1, +\infty[$. Let $\mathcal{H}$ and $\mathcal{B}$ reproducing kernel Hilbert spaces such that the corresponding kernels K and $K^{\mathcal{B}}$ are p-bounded with respect to ρ. Define $q =]1, +\infty]$ such that $\frac{1}{q} + \frac{1}{p} = 1$.*

Let $\lambda > 0$ and $(f^\lambda, g^\lambda) \in \mathcal{H} \times \mathcal{B}$, then

$$(f^\lambda, g^\lambda) \in \operatorname*{argmin}_{(f,g) \in \mathcal{H} \times \mathcal{B}} \left\{ \int_{X \times Y} V(y, f(\mathbf{x}) + g(\mathbf{x})) d\rho(\mathbf{x}, y) + \lambda \|f\|_{\mathcal{H}}^2 \right\} \tag{7}$$

if and only if there is $\alpha \in L^q(Z, \rho)$ satisfying

$$\alpha(\mathbf{x}, y) \in (\partial V)(y, f^\lambda(\mathbf{x}) + g^\lambda(\mathbf{x})) \quad (x, y) \in X \times Y \text{ a.e.} \tag{8}$$

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

$$0 = \int_{X \times Y} K^{\mathcal{B}}(\mathbf{s}, \mathbf{x}) \alpha(\mathbf{x}, y) d\rho(\mathbf{x}, y) \quad \mathbf{s} \in X. \tag{10}$$

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 ρ is the empirical measure ρ_S. In this case, the integral reduces to a finite sum and we recover the well known result that $f_S^\lambda = \sum_{i=1}^\ell \alpha_i K_{\mathbf{x}_i}$, where the $\mathbf{x}_i$ form the training set. Moreover, the solution is quantitatively characterized since the coefficients α are given by Eq. (8) involving the subgradient. For differentiable loss functions in the discrete setting, Eq. (8) reduces to

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

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. (8). A discussion on how to solve explicitly Eq. (8) can be found in Wahba (1998). Furthermore, the presence of $\mathcal{B}$ induces a system of linear constraints on the coefficients α_i expressed by Eq. (10) 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 α is bounded. For the square loss, clearly $p = 2$, so that $q = 2$ and α is square-integrable. As shown by Steinwart (2003), for classification and compact X, one can again remove Assumption 3) of Definition 1 using the fact that a convex function is locally Lipschitz and the range of possible y is bounded.

The following corollary is the restatement of the representer theorem without offset space.

Corollary 3 *With the assumptions of Theorem* 2, *let* $f^\lambda \in \mathcal{H}$ *then*

$$f^\lambda \in \operatorname*{argmin}_{f \in \mathcal{H}} \left\{ \int_{X \times Y} V(y, f(\mathbf{x})) d\rho(\mathbf{x}, y) + \lambda \| f \|_{\mathcal{H}}^2 \right\}$$

if and only if there is $\alpha \in L^q(Z, \rho)$ *satisfying*

$$\begin{aligned} \alpha(\mathbf{x}, y) &\in (\partial V)(y, f^\lambda(\mathbf{x})) \qquad (x, y) \in X \times Y \text{ a.e.} \\ f^\lambda(\mathbf{s}) &= -\frac{1}{2\lambda} \int_{X \times Y} K(\mathbf{s}, \mathbf{x}) \alpha(\mathbf{x}, y) d\rho(\mathbf{x}, y) \quad \mathbf{s} \in X. \end{aligned}$$

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 $I[f+g]+\lambda\|f\|^2_{\mathcal{H}}$ is a convex map on $\mathcal{H}\times\mathcal{B}$, so (f^λ,g^λ) is a minimizer of the Tikhonov functional if and only if $(0,0)$ is in its subgradient, which is a subset of $\mathcal{H}\times\mathcal{B}$. Using linearity, the computation of the subgradient of the Tikhonov functional reduces to the computation of the subgradient of $I[f+g]$ and $\|f\|^2_{\mathcal{H}}$ 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 $L^p(Z,\rho)$ and then use a fundamental result of convex analysis to interchange the integral and the subgradient.

Proof [of Theorem 2] Clearly, $\lambda\|f\|^2_{\mathcal{H}}$ is continuous and, by Lemma 4, the functional $I[f+g]$ is continuous and finite. So, from item 5 of Proposition 14, one has that

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

Now, the map

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

is differentiable with derivative $(2f,0)$ and, therefore, by item 1 of Proposition 14,

$$\partial(\|f\|^2_{\mathcal{H}})=\{(2f,0)\}. \tag{11}$$

The main difficulty is the evaluation of the subgradient of the map $I[f+g]$ given in Lemma 5. By means of this lemma we obtain that the elements of the subgradient of $I[f+g]$ at (f,g) are of the form

$$\left(\int_{X\times Y}K(\mathbf{x},\cdot)\alpha(\mathbf{x},y)d\rho(\mathbf{x},y),\int_{X\times Y}K^{\mathcal{B}}(\mathbf{x},\cdot)\alpha(\mathbf{x},y)d\rho(\mathbf{x},y)\right), \tag{12}$$

where $\alpha\in L^q(Z,\rho)$ satisfies

$$\alpha(\mathbf{x},y)\in(\partial V)(y,f(\mathbf{x})+g(\mathbf{x})) \tag{13}$$

for ρ-almost all $(x,y)\in X\times Y$. Now, by combining Eq. (11) and Eq. (12), we have that the elements of the subgradient of $I[f+g]+\lambda\|f\|^2_{\mathcal{H}}$ at point (f,g) are of the form

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

where $\alpha\in L^q(Z,\rho)$ satisfies Eq. (13).

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

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

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

where, by means of Eq. (13), $\alpha\in L^q(Z,\rho)$ satisfies Eq. (8). This ends the proof. ■

Before computing the subgradient of the map $I[f+g]$ in Lemma 5, we need to extend the definition of expected risk on $L^p(Z,\rho)$. First of all, we let

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

so that $I[f+g] = I_0(\mathcal{J}(f,g))$ where $\mathcal{J} : \mathcal{H}\times\mathcal{B} \to L^p(Z,\rho)$ 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 4 *With the above notations,*

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

$$I_0[u] = \int_Z V(z,u(z))\, d\rho(z) \quad u \in L^p(Z,\rho).$$

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

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

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),

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

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

$$\|f+g\|_{L^p} \leq \sqrt[p]{C}\|f\|_{\mathcal{H}} + \sqrt[p]{C'}\|g\|_{\mathcal{B}}.$$

Since J is linear, it follows that $\mathcal{J}$ is continuous. ■

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

Lemma 5 *With the above notations, let* $(f,g) \in \mathcal{H} \times \mathcal{B}$*, then* $(\phi,\psi) \in \partial(I_0 \circ \mathcal{J})(f,g)$ *if and only if there is* $\alpha \in L^q(Z,\rho)$ *such that*

$$\begin{aligned} \alpha(\mathbf{x},y) &\in (\partial V)(y, f(\mathbf{x})+g(\mathbf{x})) \quad (\mathbf{x},y) \in X \times Y \text{ a.e.} \\ \phi(\mathbf{s}) &= \int_{X\times Y} K(\mathbf{s},\mathbf{x})\alpha(\mathbf{x},y)d\rho(\mathbf{x},y) \quad \mathbf{s} \in X \\ \psi(\mathbf{s}) &= \int_{X\times Y} K^{\mathcal{B}}(\mathbf{s},\mathbf{x})\alpha(\mathbf{x},y)d\rho(\mathbf{x},y) \quad \mathbf{s} \in X. \end{aligned}$$

Proof Since I_0 is finite and continuous in $0 = \mathcal{J}(0)$, by point 6 of Proposition 14, we know that

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

where $\mathcal{J}^* : L^q(Z,\rho) \to \mathcal{H} \times \mathcal{B}$ is the adjoint of $\mathcal{J}$, that is,

$$\langle \mathcal{J}^*\alpha, (f,g)\rangle_{\mathcal{H}\times\mathcal{B}} = \int_{X\times Y} \alpha(\mathbf{x},y)\mathcal{J}(f,g)(x,y)\, d\rho(x,y).$$

First of all, we compute ∂I_0. Since $I_0[0] < +\infty$, we can apply Proposition 15 so that, given $u \in L^p(Z,\rho)$, then $\alpha \in (\partial I_0)(u)$ if and only if $\alpha \in L^q(Z,\rho)$ and

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

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

We now compute the adjoint of $\mathcal{J}$. Let $\alpha \in L^q(Z,\rho)$ 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, then state and discuss our main result on this issue. Finally, we give the proof of the results.

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. Typical examples in which offset spaces arise are Support Vector Machine algorithms (Vapnik, 1988), where the offset term is a constant accounting for the translation invariance of the separating hyperplane, and penalization methods (Wahba, 1990), where the offset space is the kernel space of the penalization operator.

However, the fact that the set $\mathcal{H} \times \mathcal{B}$ is not a RKHS (in fact, it is not even a function space) makes it cumbersome to extend of 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.

In this section we show that under very weak conditions the presence of an offset term is equivalent to solving a standard regularization problem with a seminorm (Wahba, 1990).

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

$$\mathcal{S} = \mathcal{H} + \mathcal{B} = \{ f + g \,|\, f \in \mathcal{H},\ g \in \mathcal{B} \}.$$

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.

Finally, we notice that the norm of $\mathcal{B}$ (hence the kernel $K^{\mathcal{B}}$) plays no role in the functional

$$I[f+g] + \lambda \|f\|^2_{\mathcal{H}},$$

that is, all kernels, whose corresponding RKHS is $\mathcal{B}$ as a vector space, give rise to the same minimizers (f^λ, g^λ). This fact is confirmed by Eq. (18) below (see also Eq. (20)).

5.2 Main Theorem

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

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

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

We are now ready to state the following result.

Theorem 6 *Let Q be the orthogonal projection on the closed subspace of $\mathcal{S}$*

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

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*

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

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

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

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

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

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

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

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

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

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

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

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

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

has a solution, whereas the functional $I[f+g] + \lambda \|f\|^2_{\mathcal{H}}$ 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 $\|f\|^2_{\mathcal{H}}$ instead of $\|Qs\|^2_{\mathcal{S}}$ 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.

Third, 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^λ of $I[s]+\lambda\|Qs\|_{\mathcal{S}}^2$ is of the form

$$s^\lambda(\mathbf{s}) = \int_{X\times Y} \alpha(\mathbf{x},y)\left(K(\mathbf{x},\mathbf{s})+K^{\mathcal{B}}(\mathbf{x},\mathbf{s})\right) d\rho(\mathbf{x},y)+g^\lambda(\mathbf{s}) \tag{17}$$

$$= \int_{X\times Y} \alpha(\mathbf{x},y)K(\mathbf{x},\mathbf{s})\, d\rho(\mathbf{x},y)+g^\lambda(\mathbf{s}) \tag{18}$$

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

$$\alpha(\mathbf{x},y) \in (\partial V)(y,s^\lambda(\mathbf{x})) \tag{19}$$

$$\int_{X\times Y} \alpha(\mathbf{x},y)K^{\mathcal{B}}(\mathbf{x},\mathbf{s}) = 0. \tag{20}$$

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^λ. 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 6 we need to prove the following technical lemma. For this purpose we recall that $\mathcal{S}_0$ was defined as

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

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

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

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

$$Qs = Pf \tag{21}$$

$$\|Qs\|_{\mathcal{S}} = \|Pf\|_{\mathcal{H}} \tag{22}$$

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

$$\lim_{n\to\infty} \|Pf - f_n\|_{\mathcal{H}} = 0 \tag{23}$$

with $f_n+g_n = s$.

Equations (21) and (22) 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 (23) is a technical trick to overcome this pathology.

Proof [of Lemma 7] To give the proof of the lemma we need some preliminary facts. Let $\mathcal{K}$ be the closed subspace of $\mathcal{H} \times \mathcal{B}$

$$\mathcal{K} = \{(f,g) \in \mathcal{H} \times \mathcal{B} \,|\, (f,h)_{\mathcal{H}} = (g,h)_{\mathcal{B}} \; \forall h \in \mathcal{H} \cap \mathcal{B}\}.$$

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 $(f',g') \in \mathcal{H} \times \mathcal{B}$,

$$\left\langle s, f' + g' \right\rangle_{\mathcal{S}} = \left\langle f, f' \right\rangle_{\mathcal{H}} + \left\langle g, g' \right\rangle_{\mathcal{B}}. \tag{24}$$

From Eq. (16) one has that

$$\|f\|_{\mathcal{S}} \leq \|f\|_{\mathcal{H}} \quad f \in \mathcal{H} \tag{25}$$

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. (24), for all $g' \in \mathcal{B}$,

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

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

$$\|f\|_{\mathcal{S}}^2 = \langle f + 0, f + 0 \rangle_{\mathcal{S}} = \langle f, f \rangle_{\mathcal{H}} = \|f\|_{\mathcal{H}}^2. \tag{26}$$

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

$$\lim_{n \to \infty} \|h - h_n\|_{\mathcal{H}} = 0. \tag{27}$$

Since, by Eq. (25), $\|h - h_n\|_{\mathcal{S}} \leq \|h - h_n\|_{\mathcal{H}}$ and Q is continuous, it follows that $Qh = \lim_{n \to \infty} Qh_n = 0$, 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 (21) is proved. Equation (22) follows from Eq. (26). 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. (23) follows from Eq. (27). ■

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

Proof [Theorem 6] 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. (22)

$$I[s] + \lambda \|Qs\|_{\mathcal{S}}^2 = I[f + g] + \lambda \|Pf\|_{\mathcal{H}}^2 \tag{28}$$

Let $(f_n, g_n) \in \mathcal{H} \times \mathcal{B}$ as in Lemma 7, 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

$$I[s]+\lambda\|Qs\|_{\mathcal{S}}^2=\lim_n\left(I[f_n+g_n]+\lambda\|f_n\|_{\mathcal{H}}^2\right). \tag{29}$$

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. (29) and the definition of minimizer, one has that, for all $s\in\mathcal{S}$,

$$I[s]+\lambda\|Qs\|_{\mathcal{S}}^2\geq I[f^\lambda+g^\lambda]+\lambda\left\|f^\lambda\right\|_{\mathcal{H}}^2. \tag{30}$$

In particular with the choice $s=s^\lambda$, by means of Eq. (22), one has that

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

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

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

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

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

$$I[s]+\lambda\|Qs\|_{\mathcal{S}}^2=I[f+g]+\lambda\|Pf\|_{\mathcal{H}}^2\leq I[f+g]+\lambda\|f\|_{\mathcal{H}}^2. \tag{31}$$

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. (31) and Eq. (22), it follows that

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

However, using Eq. (29) with $s=f^\lambda+g^\lambda$, one has that

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

So $I[f^\lambda+g^\lambda]+\lambda\left\|f^\lambda\right\|_{\mathcal{H}}^2$ 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^λ,g^λ) is a minimizer of $I[f+g]+\lambda\|f\|_{\mathcal{H}}^2$. ■

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 $\mathcal{H}=\ell_2=\{f=(f_n)_{n\in\mathbb{N}}\,|\,\sum_n f_n^2<+\infty\}$. The space ℓ_2 is a RKHS on $\mathbb{N}$ with respect to the kernel $K(n,m)=\delta_{n,m}$. Let $\mathcal{B}=\{f\in\ell_2\,|\,\sum_n n^2f_n^2<+\infty\}$ with the scalar product*

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

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 7, $Q = 0$.

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

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

then

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

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

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

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.

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 suitable conditions, fulfilled in the case of constant offsets. 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, 2003) 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 8 *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

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

if $\|f\|_{\mathcal{H}}$ goes to $+\infty$. From item 4 of Proposition 14 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 in two different settings.

The first proposition holds only for classification under the assumption that the loss function V goes to infinity when $yf(\mathbf{x})$ goes to $-\infty$ (see Condition 1 of Proposition 9 below). Similar results were obtained in Steinwart (2002). We let ν be the marginal measure on X associated with ρ and $\operatorname{supp}\nu$ its support.

Proposition 9 *Assume that the following conditions hold*

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

Then there is at least one solution of the problem

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

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

We observe that Assumption *2.* is satisfied if X is compact and K is continuous. Assumption *3.* has a very natural interpretation in the discrete setting where it simply amounts to have one example for each class. This condition is need since Assumption *1.* does not requires that V goes to $+\infty$ when $yf(\mathbf{x})$ goes to $+\infty$. Typical example of loss function satisfying Assumption *1.* is the hinge loss.

The second result holds both for regression and classification, but it requires the loss function going to infinity when $f(\mathbf{x})$ goes to $\pm\infty$, uniformly in y (compare Assumption *1.* of Proposition 10 and Assumption *1.* of Proposition 9).

Proposition 10 *Assume that the following conditions hold*

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

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

We observe that for classification with symmetric loss functions, as the squared loss function, this proposition gives a sharper result than Proposition 9.

We now prove Proposition 9 and omit the proof of Proposition 10 since it is essentially the same.
Proof [of Proposition 9] The idea of the proof is to show that the functional we have to minimize goes to $+\infty$ when $\|s\|_{\mathcal{S}}$ goes to $+\infty$. With 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

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

since $R \geq 2(1+C)\sqrt{\frac{M}{\lambda}}$. If $\|Qs\| \leq \frac{R}{2(1+C)}$, 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 $\mathbf{x} \in \operatorname{supp}\nu$

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

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

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

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

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

Since the functional is continuous, from item 4 of Proposition 14 the existence of the minimizer follows. ■

6.2 Uniqueness

The first proposition completely characterizes uniqueness for strictly convex functions.

Proposition 11 *Let s^λ 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 $g(\mathbf{x}) = 0$ for ν-almost all $x \in X$.*

Let us comment on this proposition before providing the proof. We recall that a solution s^λ 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^λ (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^λ: in the discrete setting 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 11]

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

$$\left\| (1-t)Qs^\lambda + tQs' \right\|_{\mathcal{S}}^2 < (1-t) \left\| Qs^\lambda \right\|_{\mathcal{S}}^2 + t \left\| Qs' \right\|_{\mathcal{S}}^2.$$

   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

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

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

2. Let $s' = s^\lambda + g$ with g as in item *1*. 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 measurable set

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

   By contradiction, let us assume that $\nu(U) > 0$ 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,

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

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

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 $\nu(U)=0$, that is, $g(\mathbf{x})=0$ for ν-almost all $\mathbf{x}\in X$.

■

Two important examples of convex loss functions which are not strictly convex are the hinge and the ε-insensitive loss. The next proposition deals with the hinge loss though a similar result can be also derived for the ε-insensitive loss, see Burges and Crisp (2000). 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 12 *Let $Y=\{\pm 1\}$, $V(y,w)=|1-yw|_+$ and $\mathcal{B}=\mathbb{R}$. Let s^λ 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

$$I_+=\{i\,|\,y_i=1,\,s^\lambda(\mathbf{x})<1\} \qquad I_-=\{i\,|\,y_i=-1,\,s^\lambda(\mathbf{x})>-1\}$$
$$B_+=\{i\,|\,y_i=1,\,s^\lambda(\mathbf{x}_1)=1\} \qquad B_-=\{i\,|\,y_i=-1,\,s^\lambda(\mathbf{x}_1)=-1\}.$$

The solution is unique if and only if

$$\#I_+\neq\#I_-+\#B_- \tag{32}$$

and

$$\#I_-\neq\#I_++\#B_+, \tag{33}$$

where # denotes set cardinality.

Proof Assume that s' is another solution. From item *1* of proposition 11, we have that $Qs^\lambda=Qs'$ and $s'=s^\lambda+b$. Since both functions are minimizers, one concludes that

$$\sum_{i=1}^{\ell}|1-y_is^\lambda(\mathbf{x}_i)|_+=\sum_{i=1}^{\ell}|1-y_is'(\mathbf{x}_i)|_+ \tag{34}$$

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 (34) 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. (32) holds).
2. if $b < 0$, then $\sum_{i \in I_+ \cup I_-} y_i = \sum_{B_+} y_i$ (that is, Eq. (33) holds).

Now, if neither Eq. (32) nor Eq. (33) holds, then $b = 0$ and s^λ is unique. Conversely, assume for example that Eq. (32) 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. (32) holds;
2. for the solution with $b = b_+$, Eq. (33) holds;
3. for the solution with $b_- < b < b_+$, both Eqs. (32) and (33) 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 ρ_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 14 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 13 *Let $\mathcal{S} = \mathcal{H} + \mathbb{R}$ and Q the projection on*

$$\{s \in \mathcal{S} \mid \langle s, 1\rangle_{\mathcal{S}} = 0\}.$$

Given $\lambda > 0$, let $f^\lambda \in \mathcal{H}$ and $b^\lambda \in \mathbb{R}$ and define $s^\lambda = f^\lambda + b^\lambda \in \mathcal{S}$, then

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

if and only if

$$\begin{aligned} s^\lambda &\in \operatorname*{argmin}_{s\in\mathcal{S}} \left\{ \frac{1}{\ell}\sum_i V(y_i, s(\mathbf{x}_i)) + \lambda \|Qs\|^2_{\mathcal{H}} \right\} \\ f^\lambda &= Qs^\lambda \end{aligned}$$

if and only if there are $\alpha_1, \ldots, \alpha_\ell \in \mathbb{R}$ such that

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

We notice two facts. First, α_i can be zero only if $0 \in (\partial V)(y_i, f^\lambda(\mathbf{x}_i)+b^\lambda)$ – that is, 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 (2003). Second if V_- and V_+ are bounded by a constant $M > 0$, one has that $|\alpha_i| \le 2\lambda\ell M$ – that is, a sufficient conditions 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

$$[V'_-(y,w), V'_+(y,w)] = \begin{cases} -y & \text{for } yw < 1 \\ [\min\{-y,0\}, \max\{0,-y\}] & \text{for } yw = 1 \\ 0 & \text{for } yw > 1 \end{cases}. \tag{35}$$

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 α_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 $(\alpha_1, \ldots, \alpha_\ell, b^\lambda)$ solves the following algebraic system of inequalities

$$\begin{aligned} 0 \le \alpha_i \le C \quad &\text{if} \quad y_i \left(\sum_{j=1}^{\ell} \alpha_j y_j K(\mathbf{x}_i, \mathbf{x}_j) + b^\lambda \right) = 1 \\ \alpha_i = 0 \quad &\text{if} \quad y_i \left(\sum_{j=1}^{\ell} \alpha_j y_j K(\mathbf{x}_i, \mathbf{x}_j) + b^\lambda \right) > 1 \\ \alpha_i = C \quad &\text{if} \quad y_i \left(\sum_{j=1}^{\ell} \alpha_j y_j K(\mathbf{x}_i, \mathbf{x}_j) + b^\lambda \right) < 1 \\ \sum_i \alpha_i y_i \quad &= \quad 0 \end{aligned} \tag{36}$$

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

Looking at Eqs.(35-36), it is immediate to see that the box constraints ($0 \le \alpha_i \le C$) 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).

Acknowledgments

We thank the anonymous referees for suggestions leading to an improved version of the paper. 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 ASTAA and the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778.

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 Banach space and $\mathcal{H}^*$ its dual. A function $F : \mathcal{H} \to \mathbb{R} \cup +\infty$ 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

$$\partial F(v_0) = \{w \in \mathcal{H}^* \,|\, F(v) \geq F(v_0) + \langle w, v-v_0 \rangle \,,\ \forall v \in \mathcal{H}\}. \tag{37}$$

where $\langle \cdot, \cdot \rangle$ is the pairing between $\mathcal{H}^*$ and $\mathcal{H}$. 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 14 *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*

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

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

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

   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.2.12

■

We now recall the definition of *Nemitski* functional, adapted to our framework (Ekeland and Turnbull, 1983, p.143). Let Z be a locally compact second countable space, ρ be a finite measure on Z, and $W : Z \times \mathbb{R} \to [0, +\infty[$ be a measurable function on $Z \times \mathbb{R}$ such that $W(z, \cdot)$ is convex for all $z \in Z$ (since $W(z, \cdot)$ is convex on $\mathbb{R}$, it is continuous).

Let $p \in [1, +\infty[$ and $L^p(Z, \rho)$ be the Banach space of measurable functions $u : Z \to \mathbb{R}$ such that $\int_Z |u(z)|^p \, d\rho(z)$ is finite.

The *Nemitski* functional associated with W is defined as the map $I_0 : L^p(X, \nu) \to [0, +\infty[\cup \{+\infty\}$ given by

$$I_0[u] = \int_Z W(z, u(z)) d\rho(z). \tag{38}$$

Next proposition provides us with a straightforward method to study the subgradient (∂I_0). Let $q \in]1, +\infty]$ such that $\frac{1}{p} + \frac{1}{q} = 1$.

Proposition 15 *Assume that there is an element* $u_0 \in L^p(Z, \nu)$ *such that* $\sup_{z \in Z} |u_0(z)| < +\infty$ *and* $I_0[u_0] < +\infty$. *Given* $u \in L^p(Z, \rho)$

$$(\partial I_0)(u) = \{w \in L^q(Z, \rho) \mid \ w(z) \in (\partial W)(z, u(z)) \ \rho - \text{a.e.}\}. \tag{39}$$

Proof See the proof of Prop. III.5.3 of Ekeland and Turnbull (1983). The proof is for Z interval of $\mathbb{R}$, but can be easily extended to arbitrary Z, compare with Ekeland and Teman (1974). ■

References

Peter L. Bartlett. Prediction algorithms: complexity, concentration and convexity. In *Proceedings of the 13th IFAC Symposium on System Identification*, pages 1507–1517, 2003.

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. (submitted), 2002.

Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. Technical Report 638, Department of Statistics, U. C. Berkeley, 2003.

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. Uniqueness of the SVM solution. In *Proceedings of the Twelfth Conference on Neural Information Processing Systems*, pages 223–229, Cambridge, MA, 2000. MIT Press.

C. Burges and D. Crisp. Uniqueness theorems for kernel methods. *Neurocomputing*, 55:187–220, 2003.

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 Computationals Mathematics*, 2:413–428, 2002.

I. Ekeland and R. Teman. *Analyse convexe et problémes 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 stochastic processes and smoothing 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, P. Niyogi, T. Poggio, and R. Rifkin. Statistical learning: Stability is sufficient for generalization and necessary and sufficient for consistency for empirical risk minimization. C.B.C.L. Paper 223, Massachusetts Institute of Technology - Artificial Intelligence Laboratory, December 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, New Jersey, 1992.

T. Poggio, S. Mukherjee, R. Rifkin, A. Rakhlin, and A. Verri. B. In J. Winkler and M. Niranjan, editors, *Uncertainty in Geometric Computations*, pages 131–141. Kluwer Academic Publishers, 2002.

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, R. Herbrich, and A. J. Smola. A generalized representer theorem. In D. Helmbold and B. Williamson, editors, *Neural Networks and Computational Learning Theory*, number 81, pages 416–426. Springer, Berlin, Germany, 2001.

B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002. URL `http://www.learning-with-kernels.org`.

L. Schwartz. Sous-espaces hilbertiens d'espaces vectorials topologiques and noyaux associès. *Journal d'Analyse Mathematique*, 13:115–256, 1964.

I. Steinwart. Consistency of support vector machines and other regularized kernel machines. *submitted to IEEE Transactions on Information Theory*, 2002.

I. Steinwart. Sparseness of support vector machines. *Journal of Machine Learning Research*, 4: 1071–1105, 2003.

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. Len, T. K. Dietterich and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 357–363. MIT Press, 2001.