Dipartimento di Informatica e
Scienze dell'Informazione

Representer Theorem for Convex Loss Function

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

Technical Report **DISI**

DISI, Università di Genova
v. Dodecaneso 35, 16146 Genova, Italy
http://www.disi.unige.it/

Abstract

In this paper we use the subgradient technique to give a constructive proof of the Representer Theorem for learning algorithms derived from regularization. The proof holds for convex loss functions with no differentiability requirement. While the complete proof in the presence of a bias term requires some care, the case in which the penalty term is a norm is strikingly simple. The explicit form of the solution coefficients makes it clear the relation between the shape of the loss function and the solution properties, like sparsity and boundedness of the coefficients. In the case of Support Vector Machines, the proof obtains the Kuhn-Tucker conditions in the primal formulation with no need to go through the dual formulation.

1. Introduction

The problem of learning from examples can be seen as the problem of approximating a multivariate function from sparse data, which is well known to be ill posed (Poggio and Smale, 2003, Bertero et al., 1988). A classical way to restore well posedness is provided by regularization theory tools (Tikhonov and Larsenin, 1977). In the context of statistical learning this leads to learning algorithms looking for an estimator minimizing a functional of the form

$$J[f] = \frac{1}{\ell} \sum_{i=1}^{\ell} V(y_i, f(\mathbf{x}_i)) + \lambda \|f\|_{\mathcal{H}}^2 , \tag{1}$$

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 ℓ pairs of examples.

Results studing the form of the mimizer of (1) are known in the statistical learning literature as representer thorems. In this paper we study a compact method to derive the explicit form of the minimizer of $J[f]$ hence providing a new proof of the representer theorem.

If the loss function V is differentiable the study of the minimizer of $J[f]$ can be addressed by means of standard differentiation in functional spaces. This approach was studied in Girosi (1998), Poggio and Girosi (1992), Wahba (1998) where it was shown that the minimizer can be written as

$$f(\mathbf{x}) = \sum_{i=1}^{\ell} c_i K(\mathbf{x}, \mathbf{x}_i) \tag{2}$$

with coefficients c_i solution of a system of algebraic equations.

This approach cannot be applied to hinge loss function (Vapnik, 1988), since it is not differentiable. In Schölkopf et al. (2001), Wahba (1990, 1998) a generalized representer theorem was presented in order to cope with this case. However the form of the coefficients

c_i was recovered only through the usual dual Lagrangian formulation of the minimization problem (see Cristianini and Shawe Taylor, 2000, Vapnik, 1988).

Finally, Steinwart (2003) proves a representer theorem that holds for Lipschitz convex loss fuctions in the classification setting. His proof is based on infinite-dimensional convex analysis tools and gives the form of the coefficients c_i in terms of a closed equation involving the subgradient.

This paper, using similar techiques of Steinwart (2003), extends the above result in three points:

1. we remove the assumption that the loss function must be Lipschitz, without using *extension lemma*;

2. our result holds both for regression and classification;

3. we consider the semiparametric case where the functional (1) is replaced by.

$$J_0[f] = \frac{1}{\ell} \sum_{i=1}^{\ell} V(y_i, f(\mathbf{x}_i) + h(\mathbf{x}_i) + \lambda \|f\|_{\mathcal{H}}^2 , \tag{3}$$

   where h runs over a set of offset functions that are not penalized. In particular, we prove that the minimization of Eq. (3) is equivalent to the replacement of the penalty term $\|f\|_{\mathcal{H}}^2$ with a seminorm (we do not assume that the offset functions h are orthogonal to the penalized functions f).

The plan of the paper is as follows. We first recall the notions from convex analysis we need in our approach in section 2. In section 3 we state and prove our main result. In section 4 we discuss in detail the implication of our result for the square and the hinge loss in the classification setting. We assume the reader has some basic ideas about statistical learning theory (for a review see Vapnik, 1988, Evgeniou et al., 2000).

2. Some properties of convex functions

We briefly recall some properties of convex functions defined on a Hilbert space $\mathcal{H}$. For a detailed review see, for example (Ekeland and Turnbull, 1983).

A function $F : \mathcal{H} \to \mathbb{R}$ 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).

The subgradient of F at point $v_0 \in \mathcal{H}$ is the subset of $\mathcal{H}$ given by

$$(\partial F)_{v_0} = \{w \in \mathcal{H} \,|\, F(v) \geq F(v_0) + \langle w, v - v_0 \rangle_{\mathcal{H}} \;\; \forall v \in \mathcal{H}\}. \tag{4}$$

If F is differentiable in v_0, the subgradient reduces to the usual gradient $F'(v_0)$ (Ekeland and Turnbull, 1983, Prop. III.2.8) and inequality (4) is the usual definition of convex function

$$F(v) \geq F(v_0) + \left\langle F'(v_0), v - v_0 \right\rangle.$$

If $\mathcal{H} = \mathbb{R}^2$, inequality (4) has a simple geometrical interpretation. A vector (w_1, w_2) is in the subgradient if and only if the plane

$$z = F(x_0, y_0) + w_1(x - x_0) + w_2(y - y_0)$$

is under the graph $z = F(x, y)$.

Remark 1 *If F is a convex function on $\mathbb{R}$, then it is continuous (Ekeland and Turnbull, 1983, Cor. III.1.2), left and right derivatives always exist with $F'_-(x) \leq F'_+(x)$ (Ekeland and Turnbull, 1983, Prop. III.2.7), and $(\partial F)_x = [F'_-(x), F'_+(x)]$.*

We need the following facts extending the linearity, extremality condition and chain rule properties of the gradient of differential functions.

Proposition 2 *The following facts hold:*

a) let F_1 and F_2 be continuous convex functions on $\mathcal{H}$ and $a, b \geq 0$, then $F = aF_1 + bF_2$ is convex and

$$(\partial F)_v = a(\partial F_1)_v + b(\partial F_2)_v;$$

b) F has a minimum point at v if and only if $0 \in (\partial F)_v$;

c) if F is defined on $\mathbb{R}$ and $w \in \mathcal{H}$, then the function on $\mathcal{H}$

$$v \to F(\langle v, w \rangle)$$

is convex, continuous and its subgradient at v_0 is given by

$$[F'_-(\langle v_0, w \rangle), F'_+(\langle v_0, w \rangle)]\; w.$$

Proof See Prop. III.2.9 and Cor. III.2.1 in Ekeland and Turnbull (1983) for item a), Prop. III.3.1 in Ekeland and Turnbull (1983) for item b) and Prop. III.2.12 in Ekeland and

Turnbull (1983) for item c). ■

Remark 3 *If F is a convex continuous function such that*

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

then F has a minimizer Ekeland and Turnbull (1983, Prop. II.4.6). If F is strictly convex, the minimizer is unique.

3. Main results

In this section we first fix the notation and then provide our proof of the representer theorem holding for arbitrary convex loss functions.

3.1 Background

We assume that the pair $(\mathbf{x}, y)$ is in $X \times Y$, where X is a subset of $\mathbb{R}^d$ and Y is a subset of $\mathbb{R}$ (for regression $Y = \mathbb{R}$, for binary classification $Y = \{-1, 1\}$). The training set of the ℓ-pairs of given examples will be denoted by $D = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_\ell, y_\ell))$.

Let the hypothesis space $\mathcal{H}$ be a Reproducing Kernel Hilbert Space (RKHS) with a continuous kernel $K : X \times X \to \mathbb{R}$ (Aronszajn, 1950). We recall that $\mathcal{H}$ is defined as the unique Hilbert space of continuous functions on X such that

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

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

We assume that the loss function $V(y, f(\mathbf{x}))$ is of the form

1. $V(y, f(\mathbf{x})) = V(yf(\mathbf{x}))$ for classification,
2. $V(y, f(\mathbf{x})) = V(y - f(\mathbf{x}))$ for regression,

where in both cases $V : \mathbb{R} \to [0, +\infty)$ is a *convex function of one variable*. We denote by $V'_{\pm}$ the left and right derivatives of V that always exist (see Remark 1 of Section 2).

Given a training set D, we can write the functional of Eq. (1) as

$$C \sum_{i=1}^{\ell} V(y_i, f(\mathbf{x}_i)) + \frac{1}{2} \|f\|^2_{\mathcal{H}}\,, \tag{6}$$

which is the standard form in the SVM setting. Here $C > 0$ controls the trade-off between the empirical error $\frac{1}{\ell} \sum_{i=1}^{\ell} V(y_i, f(\mathbf{x}_i))$ and the penalty term $\|f\|^2_{\mathcal{H}}$ and it is related to the

classical regularization parameter λ of Eq. (1) by the equality $C = \frac{1}{2\lambda\ell}$. We denote by f_D the minimizer[1] of the functional in (6).

Different choices of the loss V give rise to different learning techniques. For example, for regression,

- square loss $V(y, f(\mathbf{x})) = (y - f(\mathbf{x}))^2 \rightarrow$ Regularization Network for Regression (RN),
- L^1-loss $V(y, f(\mathbf{x})) = |y - f(\mathbf{x})| \rightarrow$ Robust Statistics,
- ϵ-insensitive loss $V(y, f(\mathbf{x})) = |y - f(\mathbf{x})|_\epsilon \rightarrow$ Support Vector Machine Regression (SVMR),

and, for classification,

- square loss $V(y, f(\mathbf{x})) = (1 - yf(\mathbf{x}))^2 \rightarrow$ Regularization Network for Classification (RN),
- hinge loss $V(y, f(\mathbf{x})) = |1 - yf(\mathbf{x})|_+ \rightarrow$ Support Vector Machine Classification (SVMC),
- logistic loss $V(y, f(\mathbf{x})) = \log_2(1 + e^{-yf(\mathbf{x})}) \rightarrow$ Logistic Regression,
- exponential loss $V(y, f(\mathbf{x})) = e^{-yf(\mathbf{x})} \rightarrow$ AdaBoost.

3.2 Representer theorem

We are now ready to state and proof the so called Representer Theorem. Our result can be compared to Wahba (1990), Girosi (1998), Evgeniou et al. (2000) where the differentiable case is studied, with (Schölkopf et al., 2001) where a similar result is deduced without the convexity assumption but the explicit form of the coefficient is not given. Finally a recent paper of Steinwart, (Steinwart, 2003), treats the case of convex (non-differentiable) loss functions for classification.

Theorem 4 *The problem*

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

has a unique solution. Moreover the following two statement are equivalent:

1. *f_D is the minimizer*

1. see Theorem 4 below for existence and uniqueness.

2. f_D is of the form

$$\begin{aligned} \text{classification}: f_D &= \sum_{i=1}^{\ell} \alpha_i y_i K_{\mathbf{x}_i} \\ \text{regression}: f_D &= \sum_{i=1}^{\ell} \alpha_i K_{\mathbf{x}_i} \end{aligned} \tag{7}$$

where

$$\begin{aligned} \text{classification}: -CV'_+(y_i f_D(\mathbf{x}_i)) &\leq \alpha_i \leq -CV'_-(y_i f_D(\mathbf{x}_i)) \\ \text{regression}: CV'_-(y_i - f_D(\mathbf{x}_i)) &\leq \alpha_i \leq CV'_+(y_i - f_D(\mathbf{x}_i)) \end{aligned} \tag{8}$$

Proof For sake of simplicity, we develop our result in the context of binary classification. The extension of our analysis to regression is straightforward. First of all, we notice that, taking into account Eq. (5), the functional to minimize can be written as

$$J[f] = C \sum_{i=1}^{\ell} V(\langle f, g_i \rangle_{\mathcal{H}}) + \frac{1}{2}\|f\|^2_{\mathcal{H}}, \tag{9}$$

where $g_i = y_i K_{\mathbf{x}_i}$.

Since V is a convex function on $\mathbb{R}$, by Remark 1 in Section 2, V is continuous and, hence, the map

$$f \mapsto V(\langle f, g_i \rangle_{\mathcal{H}})$$

is convex and continuous. Since $f \mapsto \|f\|_{\mathcal{H}}$ is strictly convex and continuous, then $J[f]$ is strictly convex, continuous and

$$\lim_{\|f\|_{\mathcal{H}} \to +\infty} J[f] \geq \lim_{\|f\|_{\mathcal{H}} \to +\infty} \frac{1}{2}\|f\|^2_{\mathcal{H}} = +\infty.$$

By Remark 3 of Section 2, J has a unique minimizer.

We now show that 1) $\Leftrightarrow$ 2). Using properties a) and c) of Proposition 2 from Section 2) one has that

$$(\partial J)_f = \sum_{i=1}^{\ell} C \ \partial V(\langle f, g_i \rangle_{\mathcal{H}}) \ g_i + f.$$

Moreover, by Remark 1,

$$(\partial V)_{\langle f, g_i \rangle_{\mathcal{H}}} = [V'_-(\langle f, g_i \rangle_{\mathcal{H}}), V'_+(\langle f, g_i \rangle_{\mathcal{H}})].$$

Applying property b) of Proposition 2 (again from Section 2), one has that f_D is the minimizer of J if and only if $0 \in (\partial J)_{f_D}$, that is, if and only if there exist

$$a_i \in [V'_-(\langle f, g_i \rangle_{\mathcal{H}}), V'_+(\langle f, g_i \rangle_{\mathcal{H}})] \tag{10}$$

such that

$$0 = C \sum_{i=1}^{\ell} a_i g_i + f \quad .$$

Equation (7) follows defining $\alpha_i = -Ca_i$ and Eq. (8) is a restatement of formula (10). ■

3.3 Semiparametric representer theorem

In this section, we extend the representer theorem in order to deal to the case in which a bias term appears in the solution. To do this we study the general case of learning algorithms in which the penalty term do not penalize a class $\mathcal{B}$ of *bias* functions. In this case the representer theorem is sometimes called *semiparametric* (Schölkopf et al., 2001),

We assume that $\mathcal{B}$ is RKHS with a continuous kernel $K^{\mathcal{B}}$ and we study the algorithm defined by the following minimization problem

$$\min_{f_1 \in \mathcal{H}, f_2 \in \mathcal{B}} \{C \sum_{i=1}^{\ell} V\left(y_i, f_1(\mathbf{x}_i) + f_2(\mathbf{x}_i)\right) + \frac{1}{2} \|f_1\|^2_{\mathcal{H}}\}. \tag{11}$$

The hypothesis space of the algorithm is now given by

$$\mathcal{S} = \mathcal{H} + \mathcal{B} = \{f_1 + f_2 \,|\, f_1 \in \mathcal{H},\ f_2 \in \mathcal{B}\} \quad .$$

The space $\mathcal{S}$ is a reproducing kernel Hilbert space with kernel

$$K^{\mathcal{S}}(\mathbf{x}, \mathbf{s}) = K(\mathbf{x}, \mathbf{s}) + K^{\mathcal{B}}(\mathbf{x}, \mathbf{s}),$$

the spaces $\mathcal{H}$ and $\mathcal{B}$ are closed subspace of $\mathcal{S}$ and $\mathcal{S}$ has a natural norm defined in the following way. Let $\mathcal{I} = \mathcal{H} \cap \mathcal{B}$, if $g = f_1 + f_2 \in \mathcal{S}$, with $f_1 \in \mathcal{H}$, $f_2 \in \mathcal{B}$, then

$$\|g\|^2_{\mathcal{S}} = \|f_1 + f_2\|^2_{\mathcal{S}} = \inf_{h \in \mathcal{I}} (\ \|f_1 + h\|^2_{\mathcal{H}} + \|f_2 - h\|^2_{\mathcal{B}}). \tag{12}$$

see (Aronszajn, 1950).

Remark 5 *If $\mathcal{H}$ and $\mathcal{B}$ have a null intersection the definition in (12) reduces to*

$$\|g\|_{\mathcal{S}}^2 = \|f_1\|_{\mathcal{H}}^2 + \|f_2\|_{\mathcal{B}}^2 .$$

Now we denote by P be the orthogonal projection from $\mathcal{S}$ onto $\mathcal{B}$ and we let $Q = I - P$. We show now that Problem (11) is equivalent to a regularized algorithm in $\mathcal{S}$ where the penalty term is a seminorm, (see Wahba, 1990, Poggio et al., 2001). Precisely,

Lemma 6 *Let P be the orthogonal projection from $\mathcal{S}$ onto $\mathcal{B}$ and $Q = I - P$, then*

$$\inf_{\substack{f_1 \in \mathcal{H} \\ f_2 \in \mathcal{B}}} \{C \sum_{i=1}^{\ell} V\left(y_i, f_1(\mathbf{x}_i) + f_2(\mathbf{x}_i)\right) + \frac{1}{2} \|f_1\|_{\mathcal{H}}^2\} = \inf_{g \in \mathcal{S}} \{C \sum_{i=1}^{\ell} V\left(y_i, g(\mathbf{x}_i)\right) + \frac{1}{2} \|Qg\|_{\mathcal{S}}^2\}. \quad (13)$$

Proof In order to prove Eq. (13), we first show that the following equalities hold

$$\|Qg\|_{\mathcal{S}}^2 = \|Qf_1\|_{\mathcal{S}}^2 = \inf_{h \in \mathcal{I}} \|f_1 + h\|_{\mathcal{H}}^2 , \quad (14)$$

where $g = f_1 + f_2 \in \mathcal{S}$ with $f_1 \in \mathcal{H}$, $f_2 \in \mathcal{B}$.

Since $Q = I - P$ by definition of projection we have

$$\begin{aligned} \|Qg\|_{\mathcal{S}}^2 &= \|g - Pg\|_{\mathcal{S}}^2 \\ &= \inf_{f \in \mathcal{B}} \|g - f\|_{\mathcal{S}}^2 \\ &= \inf_{f \in \mathcal{B}} \|f_1 + (f_2 - f)\|_{\mathcal{S}}^2 . \end{aligned}$$

Then by Eq. (12)

$$\begin{aligned} \|Qg\|_{\mathcal{S}}^2 &= \inf_{f \in \mathcal{B}} \left(\inf_{h \in \mathcal{I}} (\|f_1 + h\|_{\mathcal{H}}^2 + \|f_2 - f - h\|_{\mathcal{B}}^2) \right) \\ &= \inf_{f \in \mathcal{B}, h \in \mathcal{I}} \left(\|f_1 + h\|_{\mathcal{H}}^2 + \|f_2 - f - h\|_{\mathcal{B}}^2 \right) . \end{aligned}$$

Clearly the second term attains 0 for the choice $f = f_2 - h$, so the claim follows.

We now prove Eq. (13). We first note that the following equality holds $\forall h \in \mathcal{I}$

$$\inf_{f_1 \in \mathcal{H}} (\|f_1\|_{\mathcal{H}}^2) = \inf_{f_1 \in \mathcal{H}} (\|f_1 + h\|_{\mathcal{H}}^2) \quad (15)$$

Now let $J_{emp}[f] = C\sum_{i=1}^{\ell} V(y_i, f(\mathbf{x}_i))$, from Eq. (15) we have that

$$\inf_{f_1\in\mathcal{H}, f_2\in\mathcal{B}}(J_{emp}[f_1+f_2] + \frac{1}{2}\|f_1\|^2_{\mathcal{H}}) = \inf_{\substack{f_1\in\mathcal{H}h\in\mathcal{I}\\ f_2\in\mathcal{B}}}\inf(J_{emp}[(f_1+f_2] + \frac{1}{2}\|f_1+h\|^2_{\mathcal{H}})$$

Finally using Eq. (14) it follows that

$$\begin{aligned}\inf_{f_1\in\mathcal{H}, f_2\in\mathcal{B}}(J_{emp}[f_1+f_2] + \frac{1}{2}\|f_1\|^2_{\mathcal{H}}) &= \inf_{\substack{f_1\in\mathcal{H}\\ f_2\in\mathcal{B}}}(J_{emp}[f_1+f_2] + \frac{1}{2}\|Qf_1\|^2_{\mathcal{S}})\\ &= \inf_{g\in\mathcal{S}}(J_{emp}[g] + \frac{1}{2}\|Qg\|^2_{\mathcal{S}}).\end{aligned}$$

■

We can now state the following semiparametric representer theorem. Let $(\phi_n)_{n=1}^m$ be a basis for $\mathcal{B}$.

Theorem 7 *The following two statements are equivalent:*

1. *g_D is the minimizer of*

$$\min_{g\in\mathcal{S}}\{C\sum_{i=1}^{\ell} V(y_i, g(\mathbf{x}_i)) + \frac{1}{2}\|Qg\|^2_{\mathcal{S}}\}.$$

2. *$g_D = f_1^* + f_2^*$, where*

$$\text{classification}: f_1^* = \sum_{i=1}^{\ell}\alpha_i y_i K_{\mathbf{x}_i} \tag{16}$$

$$\begin{aligned}\text{regression}: f_1^* &= \sum_{i=1}^{\ell}\alpha_i K_{\mathbf{x}_i}\\ f_2^* &= \sum_{n=1}^{m}\beta_n\phi_n \in \mathcal{B}\end{aligned} \tag{17}$$

and

$$\begin{aligned}\text{classification}: -CV'_+(y_i f_D(\mathbf{x}_i)) &\le \alpha_i \le -CV'_-(y_i f_D(\mathbf{x}_i)) \qquad (18)\\ \text{regression}: CV'_-(y_i - f_D(\mathbf{x}_i)) &\le \alpha_i \le CV'_+(y_i - f_D(\mathbf{x}_i))\end{aligned}$$

$$\sum_{i=1}^{\ell}\alpha_i y_i\phi_n(\mathbf{x}_i) = 0 \qquad \forall n \tag{19}$$

Proof Again, for sake of simplicity, we proof our theorem for classification, but the result can be easily generalized to the case of regression. We first note that the functional to minimize can be written as

$$J_0[g] = C \sum_{i=1}^{\ell} V(\langle f, g_i \rangle_{\mathcal{H}}) + \frac{1}{2} \|Qg\|_{\mathcal{S}}^2 .$$

where $g_i = y_i K_{\mathbf{x}_i}$. Clearly, J_0 is convex so we can mimic the proof of Theorem 4. Taking into account that the gradient of the function $g \mapsto \|Qg\|_{\mathcal{S}}^2$ is Qg, one has that g_D is the unique minimizer of J_0 if and only if $\mathrm{Qg}_D = \sum_{i=1}^{\ell} \alpha_i y_i K_{\mathbf{x}_i}^{\mathcal{S}}$
$-CV'_+(y_i g_D(\mathbf{x}_i)) \leq \alpha_i \leq -CV'_-(y_i g_D(\mathbf{x}_i))$.(19)

We now proof 1) $\Rightarrow$ 2). Assume that g_D is a minimizer, then the above relations hold and we have to proof that conditions (3.3) and (3.3) imply 2).
Let $f_1^* = Qg_D$ and $f_2^* = g_D - f_1^* = Pg_D$. Clearly $f_2^* \in \mathcal{B}$ and, since $(\phi_n)_{n=1}^m$ is a basis for $\mathcal{B}$, Eq. (17) holds.
Now from Eq. (3.3), one has that

$$f_1^* = \sum_{i=1}^{\ell} \alpha_i y_i K_{\mathbf{x}_i}^{\mathcal{S}} \tag{20}$$

where α_i satisfy conditions (3.3).
Using this last equation we can easily prove that Eq. (19) holds. Let $n = 1, \ldots, m$ and observe that, since ϕ_n is in $\mathcal{B}$,

$$\begin{aligned} 0 &= \langle f_1^*, \phi_n \rangle_{\mathcal{S}} \\ &= \sum_{i=1}^{\ell} \alpha_i y_i \left\langle K_{\mathbf{x}_i}^{\mathcal{S}}, \phi_n \right\rangle_{\mathcal{S}} \\ &= \sum_{i=1}^{\ell} \alpha_i y_i \phi_n(\mathbf{x}_i). \end{aligned}$$

Finally we have to show that Eq. (16) with the condition (18) holds. Due to Eq. (20) and the fact that $K_{\mathbf{x}_i}^{\mathcal{S}} = K_{\mathbf{x}_i} + K_{\mathbf{x}_i}^{\mathcal{B}}$, it is enough to prove that $\sum_{i=1}^{\ell} \alpha_i y_i K_{\mathbf{x}_i}^{\mathcal{B}} = 0$. From Eq. (19) we obtain that

$$\begin{aligned} \sum_{i=1}^{\ell} \alpha_i y_i K_{\mathbf{x}_i}^{\mathcal{B}} &= \sum_{i=1}^{\ell} \alpha_i y_i \sum_{n=1}^{m} \left\langle K_{\mathbf{x}_i}^{\mathcal{B}}, \phi_n \right\rangle_{\mathcal{B}} \phi_n \\ &= \sum_{n=1}^{m} \phi_n \left(\sum_{i=1}^{\ell} \alpha_i y_i \phi_n(\mathbf{x}_i) \right) \end{aligned}$$

$$= \quad 0,$$

and in particular, we have that $f_1^* \in \mathcal{H}$ and inequalities (18) are a restatement of conditions (3.3).

The second part of the proof is to show that 2) $\Rightarrow$ 1). Now, let $g_D = f_1^* + f_2^*$ from Eq. (17) we know that since $f_2^* \in \mathcal{B}$, $Qg_D = Qf_1^*$ and, using to Eq. (19) and reasoning as above, one has that

$$\sum_{i=1}^{\ell} \alpha_i y_i K_{\mathbf{x}_i}^{\mathcal{B}} = 0.$$

In particular, $f_1^* = \sum_{i=1}^{\ell} \alpha_i y_i K_{\mathbf{x}_i}^{\mathcal{S}}$. In a similar way one check that $Qf_1^* = f_1$, so that Qg_D satisfies Eq. (3.3) with conditions (3.3), that is g_D is the minimizer. ■

In most of the applications, as the Regularization Networks and Support Vector Machines, the space of bias functions $\mathcal{B}$ reduces to the set of constant functions

$$\mathcal{B} = \{f_2(\mathbf{x}) = b\,\mathbf{1} \mid b \in \mathbb{R}\}.$$

In the following, we apply our result to this particular case (for a discussion of the role of b see Vapnik, 1988, Evgeniou et al., 2000, Poggio et al., 2001).

First of all we notice that $\mathcal{B}$ is a RKHS with respect to the trivial Mercer kernel $K^{\mathcal{B}}(\mathbf{x}, \mathbf{s}) = 1$. We now describe $\mathcal{S}$. There are two possibilities:

1. if $\mathbf{1} \in \mathcal{H}$, then $\mathcal{I} = \mathcal{B}$, $\mathcal{S} = \mathcal{H}$ and $Q(\mathcal{S}) = \{f \in \mathcal{H} \mid \langle f, \mathbf{1} \rangle_{\mathcal{H}} = 0\}$;

2. if $\mathbf{1} \notin \mathcal{H}$, then $\mathcal{I} = \{0\}$, $\mathcal{S} = \mathcal{H} \oplus \mathcal{B}$ and $Q(\mathcal{S}) = \mathcal{H}$.

In Theorem 7, we choose $\phi_1 = \mathbf{1}$, so the bias function f_2^* reduces to b^* and Eq. (19) is simply given by

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

and

$$f_1^* = \sum_{i=1}^{\ell} \alpha_i y_i K_{\mathbf{x}_i} = \sum_{i=1}^{\ell} \alpha_i y_i (K_{\mathbf{x}_i} + 1).$$

In particular, it follows that regularizing algorithms with bias give the same classifier using either the kernel $K(\mathbf{x}, \mathbf{s}) = \mathbf{x} \cdot \mathbf{s}$ or the kernel $K(\mathbf{x}, \mathbf{s}) = \mathbf{x} \cdot \mathbf{s} + 1$ (see Evgeniou et al., 2000, Poggio et al., 2001).

The above result holds without assuming that $\mathbf{1}$ is an eigenfunction of the spectral decomposition (Aronszajn, 1950) of

$$K(\mathbf{x}, \mathbf{s}) = \sum_n \lambda_n \phi_n(\mathbf{x}) \phi_n(\mathbf{s}), \tag{21}$$

(compare with Poggio et al., 2001).

4. Examples

We can now consider the implications of our results as one focuses on a specific loss function. If the loss function is differentiable the subgradient reduces to the usual derivative while if we consider the hinge loss we gain some important insight on the sparsity and box constraint properties of the solution.

4.1 Differentiable loss

The square, the logistic and the Ada Boost loss are all differentiable and the subgradient reduces to the usual derivative. For example if we consider the square loss we have that

$$V'(yg(\mathbf{x})) = -2(1 - yg(\mathbf{x})) = -2y(y - g(\mathbf{x})).$$

Applying Theorem 7, the estimator given by the RN algorithm has the form

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

where

$$\begin{aligned} \alpha_i &= 2C y_i (y_i - g_D(\mathbf{x}_i)) \\ &= 2C y_i (y_i - \sum_{j=1}^{\ell} \alpha_j y_j K(\mathbf{x}_i, \mathbf{x}_j) - b^*) \qquad (22) \\ \sum_i \alpha_i y_i &= 0 \qquad (23) \end{aligned}$$

The set (22) of ℓ equations gives rise to a system of linear equations for α_i parametrized by b^*, which is fixed by Eq. (23). Clearly $\alpha_i = 0$ if and only if $y_i = g_D(\mathbf{x}_i)$, that is, only if the estimator gives the correct label to the data $\mathbf{x}_i$.

Remark 8 *Both AdaBoost and Logistic Regression algorithms have no sparsity at all since for them* $V'(yg(\mathbf{x})) < 0$.

4.2 Hinge loss

For the hinge loss

$$[V'_-(yf(\mathbf{x})), V'_+(yf(\mathbf{x}))] = \begin{cases} -1 & yf(\mathbf{x}) < 1 \\ [-1, 0] & yf(\mathbf{x}) = 1 \\ 0 & yf(\mathbf{x}) > 1 \end{cases} . \tag{24}$$

According to Theorem 7, the SVMC algorithm needs to find

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

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

$$\begin{aligned} 0 \leq \alpha_i \leq C \quad & \text{if} \quad y_i g_D(\mathbf{x}_i) = 1 \\ \alpha_i = 0 \quad & \text{if} \quad y_i g_D(\mathbf{x}_i) > 1 \\ \alpha_i = C \quad & \text{if} \quad y_i g_D(\mathbf{x}_i) < 1 \\ \sum_i \alpha_i y_i \quad & = \quad 0 \end{aligned} \tag{25}$$

The above inequalities are usually obtained as the Kuhn-Tucker conditions of a QP optimization problem (Vapnik, 1988).

Looking at Eqs.(24-25), it is immediate to establish a link between the form of the loss and the solution properties. The box constraints ($0 \leq \alpha_i \leq C$) are due to the fact that $V(yf(\mathbf{x}))$ has an asymptote for $yf(\mathbf{x}) \to -\infty$, whereas sparsity ($\alpha_i = 0$) follows from $V(yf(\mathbf{x}))$ being constant for $yf(\mathbf{x}) > 1$.

Moreover, from inequalities (25) one deduces easily the geometrical interpretation of SVM in terms of the *margin*, defined as

$$\{\mathbf{x} \in X \mid -1 \leq g_D(\mathbf{x}) \leq 1\}.$$

Indeed, if a point $\mathbf{x}_i$ is well classified and outside the margin, that is $y_i g_D(\mathbf{x}_i) > 1$, then $\alpha_i = 0$ and the example $(\mathbf{x}_i, y_i)$ can be removed by D without changing the estimator g_D. Conversely, if $(\mathbf{x}_i, y_i)$ is misclassified or inside the margin, that is $y_i g_D(\mathbf{x}_i) < 1$, then $\alpha_i = C$ and $\mathbf{x}_i$ is a boundary support vector. For the points on the boundary of the margin and well classified, that is $y_i g_D(\mathbf{x}_i) = 1$, $0 \leq \alpha_i \leq C$.

Finally, we can rewrite inequalities (25) in the feature space, dropping out the dependence on C. Let

$$\begin{aligned} h_i = \sqrt{C}\, y_i K^{\mathcal{S}}_{\mathbf{x}_i} \qquad & h_0 = \mathbf{1} \\ \beta_i = \frac{\alpha_i}{C} \qquad & \beta_0 = \frac{b^*}{\sqrt{C}} \\ \varphi_D = \frac{g_D}{\sqrt{C}} \quad = \quad & \sum_{i=0}^{\ell} \beta_i h_i, \end{aligned}$$

then g_D is a minimizer if and only if

$$\begin{aligned} \beta_i = 0 \quad \text{if} \quad & \langle \varphi_D, h_i \rangle_{\mathcal{S}} > 1 \\ 0 \leq \beta_i \leq 1 \quad \text{if} \quad & \langle \varphi_D, h_i \rangle_{\mathcal{S}} = 1 \\ \beta_i = 1 \quad \text{if} \quad & \langle \varphi_D, h_i \rangle_{\mathcal{S}} < 1 \\ \beta_0 \quad = \quad & \frac{\langle \varphi_D, h_0 \rangle_{\mathcal{S}}}{\langle h_0, h_0 \rangle_{\mathcal{S}}} \quad . \end{aligned}$$

5. Conclusion

In this paper we present a new proof of the representer theorem that holds for any convex loss function both for the parametric and semiparametric case and it is based on the notion of subgradient. Applying our result to the hinge loss function, which gives rise to the Support Vector Machines for classification, we deduce a set of algebraic inequalities that are equivalent to the usual Kuhn-Tucker conditions. From these inequalities a simple interpretation of the sparsity property of the solution can be given.

s

References

N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 686:337–404, 1950.

M. Bertero, T. Poggio, and V. Torre. Ill-posed problems in early vision. *Proceedings of the IEEE*, 76:869–889, 1988.

N. Cristianini and J. Shawe Taylor. *An Introduction to Support Vector Machines.* Cambridge University Press, Cambridge, UK, 2000.

I. Ekeland and T. Turnbull. *Infinite-dimensional optimization and convexity.* Chicago Lectures Notes in Mathematics. The University of Chicago Press, 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.

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, R. Herbrich, A. Smola, and R.C. Williamson. A generalized representer theorem. In *Procceedings of 14th COLT*, Lectures Notes in Artificial Intelligence 2111, pages 416–426. Springer-Verlag, 2001.

I. Steinwart. Sparseness of support vector machines. *Journal of Machine Learning Research*, 2003. accepted.

A.N. Tikhonov and V.Y. Larsenin. *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 lernel hilbert spaces and randomized gacv. Technical Report 984, Department of Statistics, University of Wisconsin, 1998.