Are Loss Functions All the Same?

L. Rosasco[*] E. De Vito[†] A. Caponnetto[‡] M. Piana[§]
A. Verri[¶]

September 30, 2003

Abstract

In this paper we investigate the impact of choosing different loss functions from the viewpoint of statistical learning theory. We introduce a convexity assumption - which is met by all loss functions commonly used in the literature, and study how the bound on the estimation error changes with the loss. We also derive a general result on the minimizer of the expected risk for a convex loss function in the case of classification. The main outcome of our analysis is that, for classification, the hinge loss appears to be the loss of choice. Other things being equal, the hinge loss leads to a convergence rate practically indistinguishable from the logistic loss rate and much better than the square loss rate. Furthermore, if the hypothesis space is sufficiently rich, the bounds obtained for the hinge loss are not loosened by the thresholding stage.

1 Introduction

A main problem of statistical learning theory is finding necessary and sufficient conditions for the consistency of the Empirical Risk Minimization principle. Traditionally, the role played by the loss is marginal and the choice of which loss to use for which

[*] INFM - DISI, Università di Genova, Via Dodecaneso 35, 16146 Genova (I)

[†] Dipartimento di Matematica, Università di Modena, Via Campi 213/B, 41100 Modena (I), and INFN, Sezione di Genova

[‡] DISI, Università di Genova, Via Dodecaneso 35, 16146 Genova (I)

[§] INFM - DIMA, Università di Genova, Via Dodecaneso 35, 16146 Genova (I)

[¶] INFM - DISI, Università di Genova, Via Dodecaneso 35, 16146 Genova (I)

problem is usually regarded as a computational issue (Vapnik, 1995; Vapnik, 1998; Alon et al., 1993; Cristianini and Shawe Taylor, 2000). The technical results are usually derived in a form which makes it difficult to evaluate the role played, if any, by different loss functions.

The aim of this paper is to study the impact of choosing a different loss function from a purely theoretical viewpoint. By introducing a convexity assumption – which is met by all loss functions commonly used in the literature, we show that different loss functions lead to different theoretical behaviors. Our contribution is twofold. First, we extend the framework introduced in Cucker and Smale (2002b), based on the square loss for regression, to a variety of loss functions for both regression and classification allowing for an effective comparison between the convergence rates achievable using different loss functions. Second, in the classification case, we show that for all convex loss functions the sign of the minimum of the expected risk coincides with the Bayes optimal solution. This can be interpreted as a consistency property supporting the meaningfulness of the convexity assumption at the basis of our study. This property is related to the problem of the Bayes consistency (Lugosi and Vayatis, 2003; Zhang, 2003).

The main outcome of our analysis is that, for classification, the hinge loss appears to be the loss of choice. Other things being equal, the hinge loss leads to a convergence rate which is practically indistinguishable from the logistic loss rate and much better than the square loss rate. Furthermore, the hinge loss is the only one for which, if the hypothesis space is sufficiently rich, the thresholding stage has little impact on the obtained bounds.

The plan of the paper is as follows. In Section 2 we fix the notation and discuss the mathematical conditions we require on loss functions. In Section 3, we generalize the result in Cucker and Smale (2002b) to convex loss functions. In Section 4 we discuss the convergence rates in terms of various loss functions and focus our attention on classification. In Section 5 we summarize the obtained results.

2 Preliminaries

In this section we fix the notation and then make explicit the mathematical properties required in the definition of loss functions which can be profitably used in statistical learning.

2.1 Notation

We denote with $p(\mathbf{x}, y)$ the density describing the probability distribution of the pair $(\mathbf{x}, y)$ with $\mathbf{x} \in X$ and $y \in Y$. The sets X and Y are compact subsets of $\mathrm{I\!R}^d$ and $\mathrm{I\!R}$ respectively. We let $Z = X \times Y$ and we recall that $p(\mathbf{x}, y)$ can be factorized in the form $p(\mathbf{x}, y) = p(y|\mathbf{x}) \cdot p(\mathbf{x})$ where $p(\mathbf{x})$ is the marginal distribution defined over X and $p(y|\mathbf{x})$ is the conditional distribution [1]. We write the expected risk for a function f as

$$I[f] = \int_Z V(f(\mathbf{x}), y) p(\mathbf{x}, y) d\mathbf{x} dy, \tag{1}$$

for some nonnegative valued function V, named *loss* function, the properties of which we discuss in the next subsection. The ideal estimator – or target function, denoted with $f_0 : X \rightarrow \mathrm{I\!R}$, is the minimizer of

$$\min_{f \in \mathcal{F}} I[f],$$

where $\mathcal{F}$ is the space of measurable functions for which $I[f]$ is well-defined. In practice f_0 cannot be found since the probability distribution $p(\mathbf{x}, y)$ is unknown. What is known is a training set D of examples $D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_\ell, y_\ell)\}$, obtained by drawing ℓ *i.i.d.* pairs in Z according to $p(\mathbf{x}, y)$. A natural estimate of the expected risk is given by the empirical risk

$$I_{emp}[f] = \frac{1}{\ell} \sum_{i=1}^{\ell} V(f(\mathbf{x}_i), y_i). \tag{2}$$

The minimizer f_D of

$$\min_{f \in \mathcal{H}} I_{emp}[f], \tag{3}$$

can be seen as a coarse approximation of f_0. In (3) the search is restricted to a function space $\mathcal{H}$, named *hypothesis space*, allowing for the effective computation of the solution. A central problem of statistical learning theory is to find conditions under which f_D mimics the behavior of f_0.

[1] The results obtained throughout the paper, however, hold for any probability measure on Z.

2.2 RKHS and Hypothesis Space

The approximation of f_0 from a finite set of data is an ill-posed problem (Girosi et al., 1995; Evgeniou et al., 2000). The treatment of the functional and numerical pathologies due to ill-posedness can be addressed by using regularization theory. The conceptual approach of regularization is to look for approximate solutions by setting appropriate smoothness constraints on the hypothesis space $\mathcal{H}$. Within this framework, Reproducing Kernel Hilbert Space (RKHS) (Aronszajn, 1950) provides a natural choice for $\mathcal{H}$ (Wahba, 1990; Girosi et al., 1995; Evgeniou et al., 2000). In what follows we briefly summarize the properties of RKHSs needed in the next section. A RKHS is a Hilbert space $\mathcal{H}$ characterized by a symmetric positive definite function $K(\mathbf{x}, \mathbf{s})$, named Mercer kernel (Aronszajn, 1950)

$$K : X \times X \to \mathbb{R},$$

such that $K(\cdot, \mathbf{x})$, for all $\mathbf{x} \in X$, and the following reproducing property holds

$$f(\mathbf{x}) = \langle f, K(\cdot, \mathbf{x}) \rangle_{\mathcal{H}}. \tag{4}$$

For each given $R > 0$, we consider as hypothesis space $\mathcal{H}_R$, the ball of radius R in the RKHS $\mathcal{H}$, or

$$\mathcal{H}_R = \{f \in \mathcal{H}, \|f\|_{\mathcal{H}} \leq R\}.$$

We assume that K is a continuous function on $X \times X$. It follows from Eq. (4) that $\mathcal{H}_R$ is a compact subset of $C(X)$ in the sup norm topology

$$\|f\|_\infty = \sup |f(x)|.$$

In particular, this implies that given $f \in \mathcal{H}_R$ we have

$$\|f\|_\infty \leq RC_K \quad \text{with} \quad C_K = \sup_{\mathbf{x} \in X} \sqrt{K(\mathbf{x}, \mathbf{x})}.$$

We conclude by observing that C_K is finite since X is compact.

2.3 Loss functions

In Eqs. (1) and (2) the loss function

$$V : \mathbb{R} \times Y \to [0, +\infty)$$

represents the price we are willing to pay by predicting $f(\mathbf{x})$ in place of y. The choice of the loss function is typically regarded as an empirical problem, the solution to which depends essentially upon computational issues.

We now introduce a mathematical requirement a function needs to satisfy in order to be naturally thought of as a loss function.

We first notice that the loss function is always a true function of only one variable t, with $t = w - y$ for regression and $t = wy$ for classification. The basic assumption we make is that the mapping

$$t \to V(t)$$

is convex for all $t \in \mathbb{R}$. This convexity hypothesis has two technical implications (Rockafellar, 1970).

1. A loss function is a Lipschitz function, *i.e.* for every $M > 0$ there exists a constant $L_M > 0$ such that

   $$|V(w_1, y) - V(w_2, y)| \leq L_M |w_1 - w_2|$$

   for all $w_1, w_2 \in [-M, M]$ and for all $y \in Y$.

2. There exists a constant C_0 such that, $\forall y \in Y$,

   $$V(0, y) \leq C_0 \ .$$

The explicit values of L_M and C_0 depends on the specific form of the loss function. In what follows we consider

- the square loss $V(w, y) = (w - y)^2$,
- the absolute value loss $V(w, y) = |w - y|$, and
- the $\epsilon-$insensitive loss $V(w, y) = \max\{|w - y| - \epsilon, 0\} =: |w - y|_\epsilon$

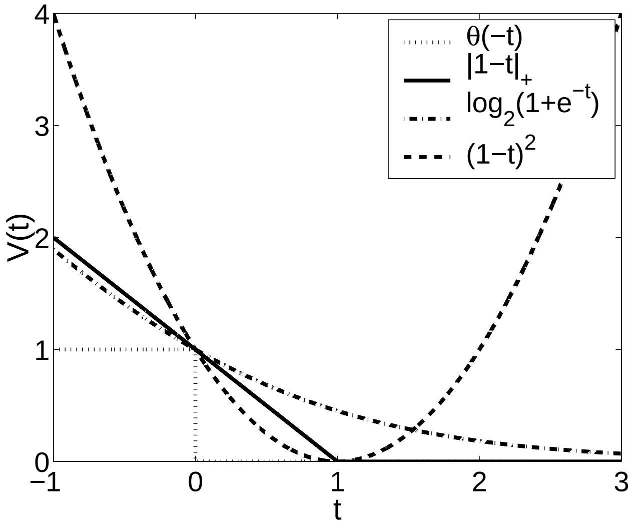


Figure 1: Various loss functions used in classification. Here $t = yf(\mathbf{x})$.

for regression, and

- the square loss $V(w, y) = (w - y)^2 = (1 - wy)^2$,
- the hinge loss $V(w, y) = \max\{1 - wy, 0\} =: |1 - wy|_+$, and
- the logistic loss $V(w, y) = (\ln 2)^{-1} \ln(1 + e^{-wy})$

for classification (see Figure 1). The constant in the logistic loss ensures that all losses for classification equal 1 for $w = 0$. The values of L_M and C_0 for the various loss functions are summarized in Table 1. We observe that for regression the value of δ in C_0 depends on the interval $[a, b]$ in $\mathbb{R}$ and is defined as $\delta = \max\{|a|, |b|\}$. For classification, instead, $C_0 = 1$ for all loss functions.

Notice that the $0 - 1$ loss, the natural loss function for binary classification, does not satisfy the convexity assumption. In practice, this does not constitute a limitation since the $0 - 1$ loss leads to intractable optimization problems.

Table 1: Optimal values of L_M and C_0 for a number of loss functions for regression (regr) and classification (class). In regression ($Y = [a, b]$) $\delta = \max\{|a|, |b|\}$.

problem	loss	L_M	C_0
regr	quad	$2M + \delta$	δ^2
regr	abs val	1	δ
regr	ϵ-insensitive	1	δ
class	quad	$2M + 2$	1
class	hinge	1	1
class	logistic	$(\ln 2)^{-1} e^M/(1 + e^M)$	1

3 Estimation error bounds for convex loss functions

It is well known that by introducing an hypothesis space $\mathcal{H}_R$, the *generalization error* $I[f_D] - I[f_0]$, can be written as

$$I[f_D] - I[f_0] = (I[f_D] - I[f_R]) + (I[f_R] - I[f_0]) \tag{5}$$

with f_R defined as the minimizer of $\min_{f \in \mathcal{H}_R}\{I[f]\}$.

The first term in the r.h.s of (5) is the sample or estimation error, whereas the second term – which does not depend on the data – is the approximation error. In this section we provide a bound on the estimation error for all loss functions through a rather straightforward extension of Theorem C in (Cucker and Smale, 2002b). We let $N(\epsilon)$ be the covering number of $\mathcal{H}_R$ (which is well defined because $\mathcal{H}_R$ is a compact subset of $C(X)$) and start by proving the following sufficient condition for uniform convergence from which the derivation of the probabilistic bound on the estimation error will be trivially obtained.

Lemma: *Let $M = C_K R$ and $B = L_M M + C_0$. For all $\epsilon > 0$,*

$$Prob\{D \in Z^\ell | \sup_{f \in \mathcal{H}_R} |I[f] - I_{emp}[f]| \leq \epsilon\} \geq 1 - 2N(\frac{\epsilon}{4L_M})e^{-\frac{\ell\epsilon^2}{8B^2}} \quad . \tag{6}$$

Proof. Since $\mathcal{H}_R$ is compact, both M and B are finite. We start by writing

$$\mathcal{L}_D[f] = I[f] - I_{emp}[f].$$

Given $f_1, f_2 \in \mathcal{H}_R$, for the Lipschitz property we have that

$$|\mathcal{L}_D[f_1] - \mathcal{L}_D[f_2]| \leq |I[f_1] - I[f_2]| + |I_{emp}[f_1] - I_{emp}[f_2]| \leq 2L_M \|f_1 - f_2\|_\infty \,. \qquad (7)$$

The mean value μ of the random variable on Z defined as $\xi(\mathbf{x}, y) = V(f(\mathbf{x}), y)$, is

$$\mu := \int_Z V(f(\mathbf{x}), y)\; dp(\mathbf{x}, y) = I[f],$$

and, since $0 \leq \mu, \xi \leq B$, we have that $|\xi(\mathbf{x}, y) - \mu| \leq B$, for all $\mathbf{x} \in X$, $y \in Y$.

Given $f \in \mathcal{H}$, we denote with

$$A_f = \{D \in Z^\ell \,|\, |\mathcal{L}_D[f]| \geq \epsilon\}$$

the collection of training sets for which convergence in probability of $I_{emp}[f]$ to $I[f]$ with high confidence is not attained. By Hoeffding inequality (Cucker and Smale, 2002b) we have that,

$$p^\ell(A_f) \leq 2e^{-\frac{\ell\epsilon^2}{2B^2}}.$$

If $m = N(\frac{\epsilon}{2L_M})$, by definition of covering number, there exist m functions $f_1, \ldots, f_m \in \mathcal{H}_R$ such that the m balls of radius $\frac{\epsilon}{2L_M}$, $\mathcal{B}(f_i, \frac{\epsilon}{2L_M})$ cover $\mathcal{H}_R$, or $\mathcal{H}_R \subset \cup_{i=1}^m \mathcal{B}(f_i, \frac{\epsilon}{2L_M})$. Equivalently, for all $f \in \mathcal{H}_R$, there exists some $i \in \{1, ..., m\}$ such that $f \in B(f_i, \frac{\epsilon}{2L_M})$, or

$$\|f - f_i\|_\infty \leq \frac{\epsilon}{2L_M}. \qquad (8)$$

If we now define $A = \cup_{i=1}^m A_{f_i}$, we then have

$$p^\ell(A) \leq \sum_{i=1}^m p^\ell(A_{f_i}) \leq m2e^{-\frac{\ell\epsilon^2}{2B^2}}.$$

Thus, for all $D \notin A$ we have that $|\mathcal{L}_D[f_i]| \leq \epsilon$ and, by combining Eqs. (7) and (8), we obtain

$$|\mathcal{L}_D[f] - \mathcal{L}_D[f_i]| \leq \epsilon.$$

Therefore, for all $f \in \mathcal{H}_R$ and $D \notin A$ we have that

$$|\mathcal{L}_D[f]| \leq 2\epsilon.$$

The thesis follows replacing ϵ with $\frac{\epsilon}{2}$. QED.

The above Lemma can be compared to the classic result in the book of Vapnik (1998) (see Chapter 3 and 5 therein) where a different notion of covering number that depends on the given sample is considered. The relation between these two complexity measures of hypothesis space has been investigated by some authors (Zhou, 2002; Pontil, 2003). In particular, from the results in Pontil (2003) the generalization of our proof to the case of data dependent covering number does not seem straightforward.

We are now in a position to generalize Theorem C in Cucker and Smale (2002b) and obtain the probabilistic bound by observing that for a fixed η the confidence term in Eq. (6) can be solved for ϵ.

Theorem *Given* $0 < \eta < 1$, $\ell \in \mathbb{N}$ and $R > 0$, *with probability at least* $1 - \eta$,

$$I[f_D] \leq I_{emp}[f_D] + \epsilon(\eta, \ell, R) \quad \text{and}$$

$$|I[f_D] - I[f_R]| \leq 2\epsilon(\eta, \ell, R)$$

with $\lim_{\ell \to \infty} \epsilon(\eta, \ell, R) = 0$.

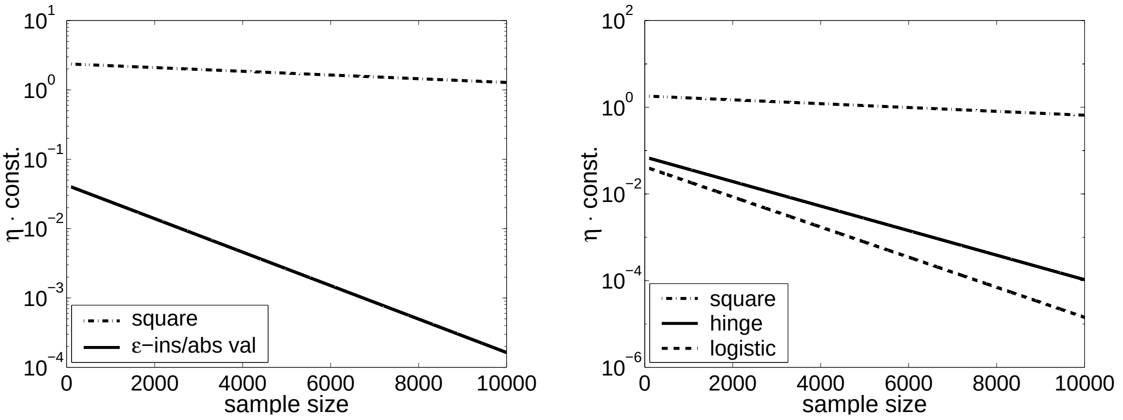
Proof. The first inequality is evident from the above lemma and the definition of ϵ. The second one follows observing that, with probability at least $1 - \eta$,

$$I[f_D] \leq I_{emp}[f_D] + \epsilon(\eta, \ell, R) \leq I_{emp}[f_R] + \epsilon(\eta, \ell, R) \leq I[f_R] + 2\epsilon(\eta, \ell, R),$$

and, by definition of f_R, $I[f_R] \leq I[f_D]$. That $\lim_{\ell \to \infty} \epsilon(\eta, \ell, R) = 0$ follows elementarily by inverting $\eta = \eta(\epsilon)$ with respect to ϵ.

4 Statistical properties of loss functions

We now move on to study some statistical properties of various loss functions.



(a)

(b)

Figure 2: Semilogarithmic plots of the convergence rates of various loss functions for regression (a) and classification (b). In both cases, $R = 1.5$, $\epsilon = 0.2$ and the dimensionality of the input-space used to estimate the covering number is 10. For regression we set $\delta = 1.5$.

4.1 Comparing convergence rates

Using Eq. (6) we first compare the convergence rates of the various loss functions. This is made possible since the constants appearing in the bounds depend explicitly on the choice of the loss function. For the sake of simplicity we assume $C_K = 1$ throughout.

For regression we have that the absolute value and the ϵ-insensitive loss functions have the same confidence, i.e.,

$$2N\left(\frac{\epsilon}{4}\right)\exp\left(-\frac{\ell\epsilon^2}{8(R+\delta)^2}\right) \tag{9}$$

from which we see that the radius, $\epsilon/4$, does not decrease when R increases, unlike the case of the square loss in which the confidence is

$$2N\left(\frac{\epsilon}{4(2R+\delta)}\right)\exp\left(-\frac{\ell\epsilon^2}{8(R(2R+\delta)+\delta^2)^2}\right). \tag{10}$$

Notice that for the square loss the convergence rate is also much slower given the different leading power of the R and δ factors in the denominator of the exponential arguments of (9) and (10). In Figure (2a) we compare the dependence of the estimated confidence η on the sample size ℓ for the square and the ϵ-insensitive loss for some fixed values of the various parameters (see the legend for details). The covering number has been estimated from the upper bounds found in Zhou (2002) for the Gaussian kernel. Clearly, to a steeper slope corresponds a better convergence rate.

Qualitatively, the behavior of the square loss does not change moving from regression to classification. For the hinge loss, instead, the confidence reads

$$2N\left(\frac{\epsilon}{4}\right)\exp\left(-\frac{\ell\epsilon^2}{8(R+1)^2}\right).$$

Here again, the covering number does not depend on R and the convergence rate is much better than for the square loss. The overall behavior of the logistic loss

$$2N\left(\frac{\epsilon}{4(\ln 2)^{-1}e^R/(1+e^R)}\right)\exp\left(-\frac{\ell\epsilon^2}{8(R((\ln 2)^{-1}e^R/(e^R+1))+1)^2}\right)$$

is very similar to the hinge case. This agrees with the intuition that these two losses have similar shape (see Figure 1). The behavior of the convergence rates for these three loss functions is depicted in Figure (2b) (again the covering number has been estimated using the upper bounds found in Zhou (2002) for the case of Gaussian kernel and to a steeper slope corresponds a better convergence rate). We conclude this section pointing out that this analysis is made possible by the fact that, unlike previous work, mathematical properties of the loss function have been incorporated directly into the bounds.

4.2 Bounds for classification

We now focus our attention to the case of classification. We start by showing that the convexity assumption ensures that the thresholded minimizer of the expected risk equals the Bayes optimal solution independently of the loss function. We then find that the hinge loss is the one for which the obtained bounds are tighter.

The natural restriction to indicator functions for classifications corresponds to considering the $0-1$ loss. Due to the intractability of the optimization problems posed by this loss, real valued loss functions must then be used (effectively solving a regression problem) and classification is obtained by thresholding the output.

We recall that in this case the best solution f_b for a binary classification problem is provided by the Bayes rule defined, for $p(1|\mathbf{x}) \neq p(-1|\mathbf{x})$, as

$$f_b(\mathbf{x}) = \begin{cases} 1 & \text{if } p(1|\mathbf{x}) > p(-1|\mathbf{x}) \\ -1 & \text{if } p(1|\mathbf{x}) < p(-1|\mathbf{x}). \end{cases}$$

We now prove the following fact relating the Bayes optimal solution to the real valued minimizer of the expected risk for a convex loss.

Fact: Assume that the loss function $V(w, y) = V(wy)$ is convex and that it is decreasing in a neighborhood of 0. If $f_0(\mathbf{x}) \neq 0$, then

$$f_b(\mathbf{x}) = \operatorname{sgn}(f_0(\mathbf{x})).$$

Proof. We recall that, since V is convex, V admits left and right derivative in 0 and, since it is decreasing, $V'_-(0) \leq V'_+(0) < 0$. Observe that

$$I[f] = \int_X \left(p(1|\mathbf{x})V(f(\mathbf{x}) + (1 - p(1|\mathbf{x}))V(-f(\mathbf{x}))\right) p(\mathbf{x})d\mathbf{x},$$

with $p(1|\mathbf{x}) = p(y = 1|\mathbf{x})$, we have $f_0(\mathbf{x}) = \operatorname{argmin}_{w\in\mathbb{R}} \psi(w)$, where

$$\psi(w) \quad = \quad p(1|\mathbf{x})V(w) + (1 - p(1|\mathbf{x}))V(-w)$$

(we assume existence and uniqueness to avoid pathological cases).

Assume, for example, that $p(1|\mathbf{x}) > \frac{1}{2}$. Then,

$$\begin{aligned} \psi'_-(0) &= p(1|\mathbf{x})V'_-(0) - (1 - p(1|\mathbf{x}))V'_+(0) \\ &\leq p(1|\mathbf{x})V'_+(0) - (1 - p(1|\mathbf{x}))V'_+(0) \\ &= (2p(1|\mathbf{x}) - 1)V'_+(0) \leq 0, \end{aligned}$$

Since ψ is also a convex function in w, this implies that for all $w \leq 0$

$$\psi(w) \geq \psi(0) + \psi'_-(0)w \geq \psi(0),$$

so that the minimum point w^* of $\psi(w)$ is such that $w^* \geq 0$. Since $f_0(\mathbf{x}) = w^*$, it follows that if $f_0(\mathbf{x}) \neq 0$

$$\operatorname{sgn} f_0(\mathbf{x}) = \operatorname{sgn}(2p(1|\mathbf{x}) - 1) = f_b(\mathbf{x}).$$

This ends the proof.

Remark: The technical condition $f_0(\mathbf{x}) \neq 0$ is always met by all loss functions considered in this paper and in practical applications and is equivalent to require the differentiability of V in the origin[2].

The above fact ensures that in the presence of infinite data all loss functions used in practice, though only rough approximations of the $0-1$ loss, lead to consistent results. Therefore, our result can be interpreted as a consistency property shared by all convex loss functions.

It can be shown that for the hinge loss (Lin et al., 2003)

$$I[f_0] = I[f_b]. \tag{11}$$

By directly computing f_0 for different loss functions (see Hastie *et al.* (2001), pp. 381, for example) it is easy to prove that this result does not hold for the other loss functions used in this paper.

We now use this result to show that the hinge loss has a further advantage on the other loss functions. In the case of finite data, we are interested in bounding

$$I[\mathrm{sgn}(f_D)] - I[f_b], \tag{12}$$

but we can only produce bounds of the type

$$I[f_D] - I[f_R] \leq 2\epsilon(\eta, \ell, R).$$

We observe that for all loss functions

$$I[\mathrm{sgn}(f_D)] \leq I[f_D] \tag{13}$$

see Figure (1). Now, if the hypothesis space is rich enough to contain f_0, i.e. when the approximation error can be neglected, we have $f_R = f_0$.

For the hinge loss, using Eqs. (11) and (13) and the theorem, we obtain that for

[2] Consider the case $p(1|\mathbf{x}) > \frac{1}{2}$. Computing the right derivative of ψ in 0, $\psi'_+(0)$, and observing that $\psi'_+(0) \geq 0$ for $p(1|\mathbf{x}) \in (\frac{1}{2}, \frac{V'_-(0)}{V'_-(0)+V'_+(0)})$, it follows that this interval is empty if and only if $V'_-(0) = V'_+(0)$. For more details see Rosasco *et al.* (2003).

$0 < \eta < 1$ and $R > 0$ with probability at least $1 - \eta$

$$0 \leq I[\mathrm{sgn}(f_D)] - I[f_b] \leq I[f_D] - I[f_0] \leq 2\epsilon(\eta, \ell, R).$$

We stress that the simple derivation of the above bound follows naturally from the special property of the hinge loss expressed in Eq. (11). For other loss functions similar results can be derived through a more complex analysis (Lugosi and Vayatis, 2003; Zhang, 2003).

5 Conclusion

In this paper we consider a probabilistic bound on the estimation error based on covering numbers and depending explicitly on the form of the loss function for both regression and classification problems. Our analysis makes explicit an implicit convexity assumption met by all loss functions used in the literature. Unlike previous results, constants related to the behavior of different loss functions are directly incorporated in the bound. This allows us to analyze the role played by the choice of the loss function in statistical learning: we conclude that the built-in statistical robustness of loss functions like the hinge or the logistic loss for classification and the ϵ-insensitive loss for regression leads to better convergence rates than the classic square loss. It remains to be seen whether the same conclusions on the convergence rates can be drawn using different bounds.

Furthermore, for classification, we derived in a simple way results relating the classification problem to the regression problem that is actually solved in the case of real valued loss functions. In particular we pointed out that only for the hinge loss the solution of the regression problem with infinite data returns the Bayes rule. Using this fact the bound on the generalization error for the hinge can be written ignoring the thresholding stage.

Finally, we observe that our results are found considering the regularization setting of the Ivanov type - that is, empirical risk minimization in balls of radius R in the RKHS $\mathcal{H}$. Many kernel methods consider a functional of the form

$$I_{emp}[f] + \lambda \|f\|^2_{\mathcal{H}}$$

that can be seen as the Tikhonov version of the above regularization problem. The question arises of whether, or not, the results presented in this paper can be generalized to the Tikhonov setting. For the square loss a positive answer is given in Cucker and Smale (2002a), where the proofs heavily rely on the special properties of the square loss. Current work focuses on extending this result to the wider class of loss functions considered in this paper.

We thank Stefano Tassano for useful discussions. L. Rosasco is supported by an INFM fellowship. A. Caponnetto is supported by a PRIN fellowship within the project "Inverse problems in medical imaging", n. 2002013422. This research has been partially funded by the INFM Project MAIA, the FIRB Project ASTA, and by the EU Project KerMIT.

References

Alon, N., Ben-David, S., Cesa-Bianchi, N., and Haussler, D. (1993). Scale-sensitive dimensions,uniform convergence and learnability. In *Proc. of the* 34*th Annual IEEE Conf. on Foundations of Computer Science*, pages 292–301.

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

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

Cucker, F. and Smale, S. (2002a). Best choices for regularization parameters in learning theory: on the bias-variance problem. *Foundation of Computational Mathematics*, 2:413–428.

Cucker, F. and Smale, S. (2002b). On the mathematical foundation of learning. *Bull. A.M.S.*, 39:1–49.

Evgeniou, T., Pontil, M., and Poggio, T. (2000). Regularization networks and support vector machines. *Adv. Comp. Math.*, 13:1–50.

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

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer, New York.

Lin, Y. Wahba, G., Zhang, H., and Lee, Y. (2003). Statistical properties and adaptive tuning of support vector machines. *Machine Learning*, 48:115–136.

Lugosi, G. and Vayatis, N. (2003). On the bayes risk consistency of regularized boosting methods. *Annals of Statistics.* (to appear).

Pontil, M. (2003). A note on different covering numbers in learning theory. *J. Complexity.* (to appear).

Rockafellar, R. (1970). *Convex Analysis.* Princeton University Press, Princeton, N.J.

Rosasco, L., De Vito, E., Caponnetto, A., Piana, M., and Verri (2003). Notes on the use of different loss functions. Technical Report DISI-TR-03-07, Department of Computer Science, University of Genova, Italy.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory.* Springer, New York.

Vapnik, V. (1998). *Statistical Learning Theory.* Wiley, New York.

Wahba, G. (1990). *Splines Models for Observational Data*, volume 59 of *Series in Applied Mathematics.* SIAM.

Zhang, T. (2003). Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics.* (to appear).

Zhou, D. (2002). The covering number in learning theory. *J. Complexity*, 18:739–767.