Are Loss Functions All the Same?

Lorenzo Rosasco
rosasco@disi.unige.it
INFM–DISI, Università di Genova, 16146 Genova, Italy

Ernesto De Vito
devito@unimo.it
Dipartimento di Matematica, Università di Modena, 41100 Modena, Italy, and
INFN, Sezione di Genova, 16146 Genova, Italy

Andrea Caponnetto
caponnetto@disi.unige.it
DISI, Università di Genova, 16146 Genova, Italy

Michele Piana
piana@dima.unige.it
INFM–DIMA, Università di Genova, 16146 Genova, Italy

Alessandro Verri
verri@disi.unige.it
INFM–DISI, Università di Genova, 16146 Genova, Italy

In this letter, 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 problem is usually regarded as a computational issue (Vapnik, 1995, 1998; Alon, Ben-David, Cesa-Bianchi, & Haussler, 1993; Cristianini & Shawe Taylor, 2000). The technical results are usually derived in a form that makes it difficult to evaluate the role, if any, played by different loss functions.

The aim of this letter 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 & Vayatis, in press; Zhang, in press).

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 that 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 letter 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 that 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 $\mathbb{R}^d$ and $\mathbb{R}$, 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

$$I[f] = \int_Z V(f(\mathbf{x}), y)p(\mathbf{x}, y)d\mathbf{x}dy, \tag{2.1}$$

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 $f_0 : X \to \mathbb{R}$, 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 independent and identically distributed pairs in Z according to $p(\mathbf{x}, y)$. A natural estimate of the expected risk is given by the empirical risk

$$I_{emp}[f] = \frac{1}{\ell} \sum_{i=1}^{\ell} V(f(\mathbf{x}_i), y_i). \tag{2.2}$$

The minimizer f_D of

$$\min_{f \in \mathcal{H}} I_{emp}[f] \tag{2.3}$$

can be seen as a coarse approximation of f_0. In equation 2.3, 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.

2.2 RKHS and Hypothesis Space. The approximation of f_0 from a finite set of data is an ill-posed problem (Girosi, Jones, & Poggio, 1995; Evgeniou, Pontil, & Poggio, 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.

---

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

An 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 $K(\cdot, \mathbf{x})$ belongs to $\mathcal{H}$ for all $\mathbf{x} \in X$, and the following reproducing property holds:

$$f(\mathbf{x}) = \langle f, K(\cdot, \mathbf{x}) \rangle_{\mathcal{H}}. \tag{2.4}$$

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 equation 2.4 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

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

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

2.3 Loss Functions. In equations 2.1 and 2.2, 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 on computational issues.

We now introduce a mathematical requirement that 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, that is, for every $M > 0$, there exists a constant $L_M > 0$ such that

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

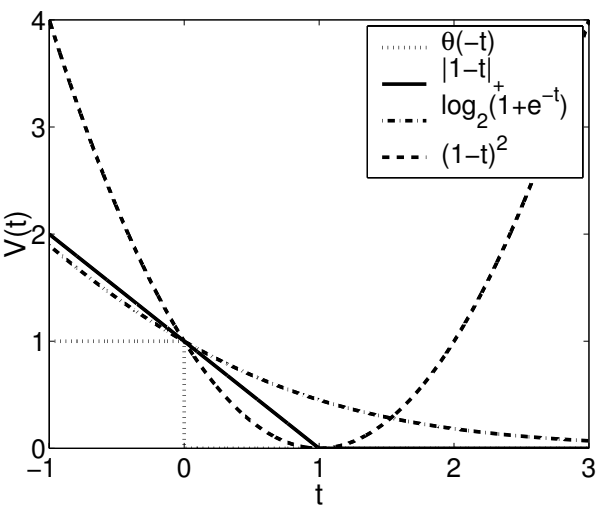for all $w_1, w_2 \in [-M, M]$ and for all $y \in Y$.

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

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 depend 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 ϵ-insensitive loss $V(w, y) = \max\{|w - y| - \epsilon, 0\} =: |w - y|_\epsilon$ 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

Table 1: Optimal Values of L_M and C_0 for a Number of Loss Functions for Regression and Classification.

Problem	Loss	L_M	C_0
Regression	Quadratic	$2M + \delta$	δ^2
Regression	Absolute Value	1	δ
Regression	ϵ-insensitive	1	δ
Classification	Quadratic	$2M + 2$	1
Classification	Hinge	1	1
Classification	Logistic	$(\ln 2)^{-1} e^M/(1 + e^M)$	1

Note: In regression ($Y = [a, b]$) $\delta = \max\{|a|, |b|\}$.

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.

3 Estimation Error Bounds for Convex Loss Functions

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

$$I[f_D] - I[f_0] = (I[f_D] - I[f_R]) + (I[f_R] - I[f_0]) \tag{3.1}$$

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

The first term on the right-hand side of equation 3.1 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 1. *Let $M = C_K R$ and $B = L_M M + C_0$. For all $\epsilon > 0$,*

$$Prob\{D \in Z^\ell | \sup_{f \in \mathcal{H}_R} |I[f] - I_{emp}[f]| \le \epsilon\} \ge 1 - 2N\left(\frac{\epsilon}{4L_M}\right)e^{-\frac{\ell\epsilon^2}{8B^2}}. \tag{3.2}$$

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

$$|\mathcal{L}_D[f_1] - \mathcal{L}_D[f_2]| \le |I[f_1] - I[f_2]| + |I_{emp}[f_1] - I_{emp}[f_2]|$$
$$\le 2L_M \|f_1 - f_2\|_\infty. \tag{3.3}$$

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

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

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 & 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 $i \in \{1, \ldots, m\}$ such that $f \in B(f_i, \frac{\epsilon}{2L_M})$, or

$$\|f - f_i\|_\infty \leq \frac{\epsilon}{2L_M}. \tag{3.4}$$

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

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

Thus, for all $D \notin A$, we have that $|\mathcal{L}_D[f_i]| \leq \epsilon$ and, by combining equations 3.3 and 3.4, 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 $\frac{\epsilon}{2}$.

Lemma 1 can be compared to the classic result in Vapnik (1998, Chaps. 3 and 5) 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 Zhou (2002); and Pontil (in press). In particular, from the results in Pontil (in press), the generalization of our proof to the case of data-dependent covering numbers 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 equation 3.2 can be solved for ϵ.

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

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

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

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

Proof. The first inequality is evident from lemma 1 and the definition of ϵ. The second 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.

4.1 Comparing Convergence Rates. Using equation 3.2, we first compare the convergence rates of the various loss functions. This is possible because the constants appearing in the bounds depend explicitly on the choice of the loss function. For 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,

$$2N\left(\frac{\epsilon}{4}\right) \exp\left(-\frac{\ell\epsilon^2}{8(R + \delta)^2}\right), \tag{4.1}$$

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

$$2N\left(\frac{\epsilon}{4(2R + \delta)}\right) \exp\left(-\frac{\ell\epsilon^2}{8(R(2R + \delta) + \delta^2)^2}\right). \tag{4.2}$$

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 equations 4.1 and 4.2. 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, a steeper slope corresponds to a better convergence rate.
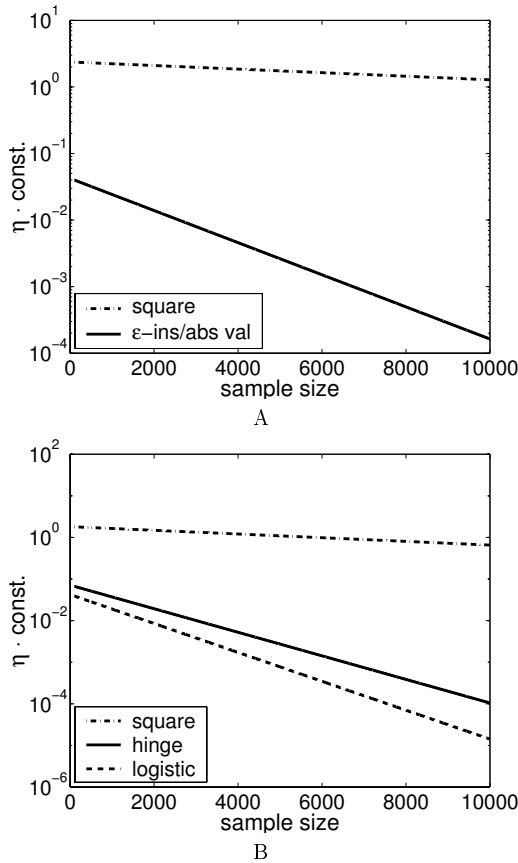
Figure 2: Semilogarithmic plots of the convergence rates of various loss functions for regression (A) and classification (B). In both cases, $R = 1.5$, $\epsilon = 0.2$, and the dimensionality of the input space used to estimate the covering number is 10. For regression, we set $\delta = 1.5$.

Qualitatively, the behavior of the square loss does not change moving from regression to classification. For the hinge loss, 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 a similar shape (see Figure 1). The behavior of the conver-
gence 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 a steeper slope corresponds
to a better convergence rate). We conclude this section by pointing out that
this analysis is made possible by the fact that unlike previous work, mathe-
matical properties of the loss function have been incorporated directly into
the bounds.

4.2 Bounds for Classification. We now focus on the case of classifica-
tion. We start by showing that the convexity assumption ensures that the
thresholded minimizer of the expected risk equals the Bayes optimal solu-
tion independent 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 corre-
sponds to considering the $0 - 1$ loss. Due to the intractability of the opti-
mization problems posed by this loss, real valued loss functions must then
be used (effectively solving a regression problem), and classification is ob-
tained 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

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

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

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

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

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

(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 $f_0(\mathbf{x}) \neq 0$,

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

Remark. The technical condition $f_0(\mathbf{x}) \neq 0$ is always met by all loss functions considered in this letter and in practical applications and is equivalent to requiring 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, Wahba, Zhang, & Lee, 2003),

$$I[f_0] = I[f_b]. \tag{4.3}$$

By directly computing f_0 for different loss functions (see Hastie, Tibshirani, & Friedman, 2001, p. 381, for example), it is easy to prove that this result does not hold for the other loss functions used in this letter.

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

$$I[\mathrm{sgn}(f_D)] - I[f_b], \tag{4.4}$$

---

[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, De Vito, Caponnetto, Piana, and Verri (2003).

but we can produce bounds only of the type

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

We observe that for all loss functions,

$$I[\text{sgn}(f_D)] \leq I[f_D] \tag{4.5}$$

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

For the hinge loss, using equations 4.3 and 4.5, and the theorem, we obtain that for $0 < \eta < 1$ and $R > 0$ with probability at least $1 - \eta$:

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

We stress that the simple derivation of the above bound follows naturally from the special property of the hinge loss expressed in equation 4.3. For other loss functions similar results can be derived through a more complex analysis (Lugosi & Vayatis, in press; Zhang, in press).

5 Conclusion

In this article, 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 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

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

that can be seen as the Tikhonov version of the above regularization problem. The question arises of whether the results presented in this article 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 article.

Acknowledgments

References

Alon, N., Ben-David, S., Cesa-Bianchi, N., & Haussler, D. (1993). Scale-sensitive dimensions, uniform convergence and learnability. In *Proceedings of the* 34*th Annual IEEE Conference on Foundations of Computer Science* (pp. 292–301). New York: IEEE.

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

Cristianini, N., & Shawe Taylor, J. (2000). *An introduction to support vector machines*. Cambridge: Cambridge University Press.

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

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

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

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

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning*. New York: Springer-Verlag.

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

Lugosi, G., & Vayatis, N. (in press). On the Bayes risk consistency of regularized boosting methods. *Annals of Statistics*.

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

Rockafellar, R. (1970). *Convex analysis*. Princeton, NJ: Princeton University Press.

Rosasco, L., De Vito, E., Caponnetto, A., Piana, M., & Verri. (2003). *Notes on the use of different loss functions* (Tech. Rep. No. DISI-TR-03-07). Genova: Department of Computer Science, University of Genova, Italy.

Vapnik, V. (1995). *The nature of statistical learning theory*. New York: Springer-Verlag.

Vapnik, V. (1998). *Statistical learning theory*. New York: Wiley.

Wahba, G. (1990). *Splines models for observational data*. Philadelphia: SIAM.

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

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