Consistency of Empirical Risk Minimization for Unbounded Loss Functions

Marco Muselli[1] and Francesca Ruffino[2]

[1] Istituto di Elettronica e di Ingegneria dell'Informazione e delle Telecomunicazioni, Consiglio Nazionale delle Ricerche, Genova, Italy
marco.muselli@ieiit.cnr.it
[2] Dipartimento di Scienze dell'Informazione, Università di Milano, Milano, Italy
ruffino@dsi.unimi.it

Abstract. The theoretical framework of Statistical Learning Theory (SLT) for pattern recognition problems is extended to comprehend the situations where an infinite value of the loss function is employed to prevent misclassifications in specific regions with high reliability.
Sufficient conditions for ensuring the consistency of the Empirical Risk Minimization (ERM) criterion are then established and an explicit bound, in terms of the VC dimension of the class of decision functions employed to solve the problem, is derived.

1 Introduction

Pattern recognition problems deal with the important task of performing a binary classification of data pertaining to a given physical system by examining a finite collection of examples, usually called *training set*.

A variety of different methods have been proposed for solving pattern recognition problems; normally, the theoretical framework employed to establish the consistence of the followed approach is the one proposed by Vapnik & Chervonenkis more than thirty years ago [1–3] and currently referred to as Statistical Learning Theory (SLT).

In this framework the solution of any pattern recognition problem is shown to be equivalent to a proper functional optimization problem, where the (probability) measures involved are totally unknown and must be (implicitly) estimated through the examples contained in the training set. In particular, the functional to be minimized, called *expected risk*, is the expected value of a binary *loss function* that assumes value 1 in correspondence with a given input data, if a misclassification occurs.

The adoption of a binary loss function amounts to treat in the same manner all the examples in the training set; consequently, no a priori information is supposed to be available about the reliability of the data at hand. In fact, if this information would be accessible, a possible way of taking into account the highest confidence associated with a specific subset of the input space could be to increase the value of the loss function in that region.

In the limit case we could assign an infinite value of the loss function in correspondence with the data belonging to the region with high reliability, thus preventing any misclassification inside it. However, the adoption of this choice violates a basic requirement for the application of SLT, since the consistency of the Empirical Risk Minimization (ERM) criterion (usually adopted in pattern recognition techniques) is established only if the expected risk is always finite.

In this paper an extension of the theoretical framework of SLT is proposed to comprehend the case of pattern recognition problems where the loss function can assume an infinite value. In particular, it is shown that the finiteness of the VC dimension for the class of decision functions employed is still a sufficient condition for the consistency of the ERM criterion. An explicit upper bound for the error probability is provided, depending on the size of the available training set.

Due to space limitations, some proofs have been omitted.

2 The theoretical framework for pattern recognition problems

Consider a general pattern recognition problem, where vectors $x \in D \subset \mathbb{R}^d$ have to be assigned to one of two possible classes, associated with the values of a binary output y, coded by the integers -1 and $+1$. Every solution for the pattern recognition problem at hand is given by a binary function $\varphi : D \to \{-1, 1\}$, called *classifier* or *decision function*.

Usually, a sufficiently large set of classifiers $\Gamma = \{\varphi(x, \alpha), \alpha \in \Lambda\}$ is considered and the best decision function $\varphi(x, \alpha^*)$ that minimizes the expected risk

$$R(\alpha) = \int Q(z, \alpha) dF(z) \ , \qquad \alpha \in \Lambda$$

is selected. Here, $F(z)$ is the joint cumulative distribution function (c.d.f.) of $z = (x, y)$, whereas Q is called *loss function* and is given by

$$Q(z, \alpha) = |y - \varphi(x, \alpha)| = \begin{cases} 0 & \text{if } y = \varphi(x, \alpha) \\ 1 & \text{if } y \neq \varphi(x, \alpha) \end{cases} \tag{1}$$

However, when solving real world pattern recognition problems, usually we do not know the distribution function $F(z)$, but have only access to a training set S_l containing l samples (x_j, y_j), $j = 1, \ldots, l$, supposed to be obtained through l i.i.d. applications of F.

In this case we have not sufficient information to retrieve the minimum of the expected risk. A possible way to proceed is to apply the *Empirical Risk Minimization* (ERM) method, which suggests to calculate the function in Γ that minimizes the empirical risk, i.e. the risk computed on the training set.

$$R_{emp}(\alpha) = \frac{1}{l} \sum_{j=1}^{l} Q(z_j, \alpha) \tag{2}$$

It is then important to obtain necessary and sufficient conditions for the consistence of the ERM approach. Vapnik [3, page 82] has shown that a stronger definition of consistency allows to rule out trivial situations:

Definition 1. *The ERM method is* strictly consistent *for the set of functions* $\{Q(z,\alpha),\ \alpha \in \Lambda\}$ *and the probability distribution function* $F(z)$ *if for any nonempty subset* $\Lambda(c) = \{\alpha \in \Lambda : R(\alpha) \geq c\}$ *with* $c \in (-\infty, +\infty)$ *the following convergence holds*

$$\inf_{\alpha \in \Lambda(c)} R_{emp}(\alpha) \xrightarrow[l\to\infty]{\mathbf{P}} \inf_{\alpha \in \Lambda(c)} R(\alpha)$$

Necessary and sufficient conditions for strict consistency are provided by the following theorem [3, page 88].

Theorem 1. *If two real constants* a *and* A *can be found such that for every* $\alpha \in \Lambda$ *the inequalities* $a \leq R(\alpha) \leq A$ *hold, then the following two statements are equivalent:*

1. *The empirical risk minimization method is strictly consistent on the set of functions* $\{Q(z,\alpha),\ \alpha \in \Lambda\}$.
2. *The uniform one-sided convergence of the mean to their matematical expectation takes place over the set of functions* $\{Q(z,\alpha),\ \alpha \in \Lambda\}$, *i.e.*

$$\lim_{l\to\infty} \mathbf{P}\left\{\sup_{\alpha\in\Lambda}(R(\alpha) - R_{emp}(\alpha)) > \varepsilon\right\} = 0, \quad \textit{for all } \varepsilon > 0$$

Vapnik also gives an upper bound for the rate of convergence [3, page 130]:

$$\mathbf{P}\left\{\sup_{\alpha\in\Lambda}(R(\alpha) - R_{emp}(\alpha)) > \varepsilon\right\} \leq 4\exp\left\{\left(\frac{G^{\Lambda}(2l)}{l} - \left(\varepsilon - \frac{1}{l}\right)^2\right) l\right\} \qquad (3)$$

where $G^{\Lambda}(m)$ is the so called *Growth function*.

The quantity $\exp(G^{\Lambda}(m))$ represents the highest number of different classifications achievable by the functions in Γ on a sample of m points; note that $G^{\Lambda}(m)$ depends only on Λ and m. Furthermore it can be shown that the growth function assumes only two possible behaviors: linear for all values of m or linear for all $m \leq h$, where h is a positive integer called *VC dimension*, and logarithmic for $m > h$. This result allows to characterize completely the consistence of the ERM approach; in fact for any c.d.f. $F(z)$, a sufficient condition for the consistency of the ERM method is that the set Γ has a finite VC dimension.

3 A natural extension to unbounded loss functions

The theoretical framework described in the last section treats all the examples (x_j, y_j) of the training set in the same way; no information is supposed to be known about the confidence of the output value y_j assigned to the input vector x_j. On the other hand, if this kind of information is actually available, we can

properly modify the loss function Q to take into account the different reliability associated with each portion of the input space.

In the limit case, if we have high confidence in output values included in samples belonging to a given subset $C \subset Z$, we can assume that the loss function Q takes an infinite value in these points. Denote with

$$C^+ = \{x \in D : (x, +1) \in C\}\,, \qquad C^- = \{x \in D : (x, -1) \in C\}$$

the subsets of C with positive and negative label respectively, and with

$$D_\alpha^+ = \{x \in D : \varphi(x, \alpha) = +1\}\,, \qquad D_\alpha^- = \{x \in D : \varphi(x, \alpha) = -1\}$$

the partition of X in two regions made by the function $\varphi(x, \alpha) \in \Gamma$.

With this definition only the classifiers $\varphi(x, \alpha)$ such that both the intersections $D_\alpha^- \cap C^+$ and $D_\alpha^+ \cap C^-$ are empty can lead to a finite value of the expected risk. This condition can be viewed as a too strong constraint on the solution we are searching for. In fact, even if the measure of the subset

$$T_\alpha = (D_\alpha^- \cap C^+) \cup (D_\alpha^+ \cap C^-)$$

is negligible, the expected risk goes to infinity.

To relax this constraint we can accept as possible solutions also the decision functions $\varphi(x, \alpha)$ for which the measure of T_α is smaller than a prescribed tolerance $\tau > 0$. The corresponding value of the expected risk $R(\alpha)$ can be kept finite if the following loss function is employed:

$$Q_\tau(z, \alpha) = \begin{cases} Q'(z, \alpha) & \text{if } \mu(T_\alpha) \geq \tau \\ Q(z, \alpha) & \text{if } \mu(T_\alpha) < \tau \end{cases} \tag{4}$$

where

$$Q'(z, \alpha) = \begin{cases} 0 & \text{if } y = \varphi(x, \alpha) \\ 1 & \text{if } y \neq \varphi(x, \alpha) \text{ and } (x, y) \notin C \\ \infty & \text{if } y \neq \varphi(x, \alpha) \text{ and } (x, y) \in C \quad \text{(i.e. if } x \in T_\alpha) \end{cases} \tag{5}$$

Using these definitions, the expected and the empirical risk become respectively

$$R_\tau(\alpha) = \int Q_\tau(z, \alpha) dF(z), \qquad R_{\tau,emp}(\alpha) = \frac{1}{l} \sum_{j=1}^{l} Q_\tau(z_j, \alpha)$$

Now, we want to extend results on consistency of the ERM method to this new setting. To this aim a generalization of Vapnik's theory is required to include situations where the loss function assume values in the range $[0, \infty]$.

Denote with $\Lambda_\tau = \{\alpha \in \Lambda : \mu(T_\alpha) < \tau\}$ the subset of Λ including only parameters α which provide a finite loss function and with Λ_∞ the complement of Λ_τ in Λ. Note that if $\alpha \in \Lambda_\tau$, the expected risk $R_\tau(\alpha)$ assumes a finite value, while $R_\tau(\alpha) = \infty$ for all $\alpha \in \Lambda_\infty$.

It can be easily seen that the definition of strict consistency for ERM method can be directly generalized to the present case. Note that, according to the hypothesis of Theorem 1, we suppose that two real constants a and $A \in \mathbb{R}$ exist such that for every $c \leq a$, $\Lambda(c) = \Lambda(a)$ and for every $c \geq A$, $\Lambda(c) = \Lambda_\infty$. Then, we can consider only the real values $c \in [a, A]$ and the case $c = \infty$.

The following three lemmas provide specific results that are needed to generalize Theorem 1. Denote with $\Lambda_\tau(c) = \{\alpha \in \Lambda_\tau : R_\tau(\alpha) > c\}$ the subset of $\Lambda(c)$ containing the parameters which provide a finite expected risk. Note that, for all $c \in [a, A]$,

$$\Lambda(c) \setminus \Lambda_\tau(c) = \Lambda(\infty) = \Lambda_\infty \tag{6}$$

Lemma 1. *If*

$$\inf_{\alpha \in \Lambda_\infty} R_{\tau,emp}(\alpha) \xrightarrow[l\to\infty]{\mathbf{P}} \inf_{\alpha \in \Lambda_\infty} R_\tau(\alpha) \tag{7}$$

then

$$\lim_{l\to\infty} \mathbf{P}\left\{\left|\inf_{\alpha\in\Lambda(c)} R_{\tau,emp}(\alpha) - \inf_{\alpha\in\Lambda_\tau(c)} R_{\tau,emp}(\alpha)\right| > \varepsilon\right\} = 0 \tag{8}$$

for every $\varepsilon > 0$ and every $c \in [a, A]$.

Proof. If (8) would not be valid, then, by using (6) we obtain for every $\varepsilon > 0$

$$\lim_{l\to\infty} \mathbf{P}\left\{\left|\inf_{\alpha\in\Lambda(c)} R_{\tau,emp}(\alpha) - \inf_{\alpha\in\Lambda_\infty} R_{\tau,emp}(\alpha)\right| > \varepsilon\right\} = 0 \tag{9}$$

and it can be easily shown that (7) leads to

$$\inf_{\alpha \in \Lambda(c)} R_{\tau,emp}(\alpha) \xrightarrow[l\to\infty]{\mathbf{P}} \inf_{\alpha \in \Lambda_\infty} R_\tau(\alpha) = \infty$$

This is not possible since $R_{\tau,emp}(\alpha) \in \mathbb{R}$ for all $\alpha \in \Lambda(c)$ with $c \in [a, A]$. □

Lemma 2. *Under hypothesis (7) the following two statements are equivalent for all $c \in [a, A]$:*

$$\lim_{l\to\infty} \mathbf{P}\left\{\left|\inf_{\alpha\in\Lambda(c)} R_\tau(\alpha) - \inf_{\alpha\in\Lambda(c)} R_{\tau,emp}(\alpha)\right| > \varepsilon\right\} = 0 \text{ for every } \varepsilon > 0 \tag{10}$$

$$\lim_{l\to\infty} \mathbf{P}\left\{\left|\inf_{\alpha\in\Lambda_\tau(c)} R_\tau(\alpha) - \inf_{\alpha\in\Lambda_\tau(c)} R_{\tau,emp}(\alpha)\right| > \varepsilon\right\} = 0 \text{ for every } \varepsilon > 0 \tag{11}$$

Proof. At first we can note that for all $c \in [a, A]$

$$\inf_{\alpha\in\Lambda(c)} R_\tau(\alpha) = \inf_{\alpha\in\Lambda_\tau(c)} R_\tau(\alpha) \tag{12}$$

since

$$\inf_{\alpha\in\Lambda_\infty} R_\tau(\alpha) = \infty$$

Now, let us prove that (10) implies (11); we have

$$\begin{aligned}
&\lim_{l\to\infty} \mathbf{P}\left\{\left|\inf_{\alpha\in\Lambda_\tau(c)} R_\tau(\alpha) - \inf_{\alpha\in\Lambda_\tau(c)} R_{\tau,emp}(\alpha)\right| > \varepsilon\right\} \\
&\le \lim_{l\to\infty} \mathbf{P}\left\{\left|\inf_{\alpha\in\Lambda_\tau(c)} R_\tau(\alpha) - \inf_{\alpha\in\Lambda(c)} R_{\tau,emp}(\alpha)\right| > \frac{\varepsilon}{2}\right\} \\
&+ \lim_{l\to\infty} \mathbf{P}\left\{\left|\inf_{\alpha\in\Lambda(c)} R_{\tau,emp}(\alpha) - \inf_{\alpha\in\Lambda_\tau(c)} R_{\tau,emp}(\alpha)\right| > \frac{\varepsilon}{2}\right\}
\end{aligned}$$

Due to (12) and (10) the first term at the right hand side vanishes; for the last term it is sufficient to apply Lemma 1.

To verify that (11) implies (10) we employ Lemma 1 to obtain that

$$\inf_{\alpha\in\Lambda(c)} R_{\tau,emp}(\alpha) \xrightarrow[l\to\infty]{\mathbf{P}} \inf_{\alpha\in\Lambda_\tau(c)} R_\tau(\alpha)$$

from which (10) follows after the application of (12). □

Lemma 3. *The following equality holds for every $\varepsilon > 0$:*

$$\mathbf{P}\left\{\left|\inf_{\alpha\in\Lambda_\infty} R_\tau(\alpha) - \inf_{\alpha\in\Lambda_\infty} R_{\tau,emp}(\alpha)\right| > \varepsilon\right\} = \mathbf{P}\left\{\sup_{\alpha\in\Lambda_\infty} (R_\tau(\alpha) - R_{\tau,emp}(\alpha)) > \varepsilon\right\}$$

Using previous lemmas we can prove the following two results which generalize Theorem 1 and the upper bound for the rate of convergence (3).

Theorem 2. *The following two statements are equivalent:*

1. *The ERM method is strictly consistent on the set of functions $\{Q_\tau(z,\alpha), \alpha \in \Lambda\}$.*
2. *For every $\varepsilon > 0$*

$$\lim_{l\to\infty} \mathbf{P}\left\{\sup_{\alpha\in\Lambda}(R_\tau(\alpha) - R_{\tau,emp}(\alpha)) > \varepsilon\right\} = 0 \tag{13}$$

Proof. Since

$$\begin{aligned}
&\lim_{l\to\infty} \mathbf{P}\left\{\sup_{\alpha\in\Lambda} (R_\tau(\alpha) - R_{\tau,emp}(\alpha)) > \varepsilon\right\} \\
&\le \lim_{l\to\infty} \mathbf{P}\left\{\sup_{\alpha\in\Lambda_\infty} (R_\tau(\alpha) - R_{\tau,emp}(\alpha)) > \varepsilon\right\} \\
&+ \lim_{l\to\infty} \mathbf{P}\left\{\sup_{\alpha\in\Lambda_\tau} (R_\tau(\alpha) - R_{\tau,emp}(\alpha)) > \varepsilon\right\}
\end{aligned} \tag{14}$$

to obtain that 1 implies 2 it is sufficient to prove that the two terms at the right hand side of (14) vanish for every $\varepsilon > 0$.

For the first term we can apply Lemma 3 by noting that, when $c = \infty$, the definition of strict consistency gives

$$\lim_{l\to\infty} \mathbf{P}\left\{\left|\inf_{\alpha\in\Lambda_\infty} R_\tau(\alpha) - \inf_{\alpha\in\Lambda_\infty} R_{\tau,emp}(\alpha)\right| > \varepsilon\right\} = 0 \text{ for every } \varepsilon > 0 \qquad (15)$$

For the second term we can use Lemma 2, thus obtaining for $c \in [a, A]$ that

$$\lim_{l\to\infty} \mathbf{P}\left\{\left|\inf_{\alpha\in\Lambda(c)} R_\tau(\alpha) - \inf_{\alpha\in\Lambda(c)} R_{\tau,emp}(\alpha)\right| > \varepsilon\right\} = 0 \text{ for every } \varepsilon > 0$$

is equivalent to

$$\lim_{l\to\infty} \mathbf{P}\left\{\left|\inf_{\alpha\in\Lambda_\tau(c)} R_\tau(\alpha) - \inf_{\alpha\in\Lambda_\tau(c)} R_{\tau,emp}(\alpha)\right| > \varepsilon\right\} = 0 \text{ for every } \varepsilon > 0$$

Now, when $\alpha \in \Lambda_\tau(c)$, we have $Q_\tau(z, \alpha) = Q(z, \alpha)$; then, Theorem 1 can be employed to ensure that

$$\lim_{l\to\infty} \mathbf{P}\left\{\sup_{\alpha\in\Lambda_\tau} (R_\tau(\alpha) - R_{\tau,emp}(\alpha)) > \varepsilon\right\} = 0 \text{ for every } \varepsilon > 0$$

To verify that 2 implies 1, we note that (13) implies

$$\lim_{l\to\infty} \mathbf{P}\left\{\sup_{\alpha\in\Lambda_\infty} (R_\tau(\alpha) - R_{\tau,emp}(\alpha)) > \varepsilon\right\} = 0$$

and

$$\lim_{l\to\infty} \mathbf{P}\left\{\sup_{\alpha\in\Lambda_\tau} (R_\tau(\alpha) - R_{\tau,emp}(\alpha)) > \varepsilon\right\} = 0$$

Then Lemma 3 ensure that

$$\lim_{l\to\infty} \mathbf{P}\left\{\left|\inf_{\alpha\in\Lambda_\infty} R_\tau(\alpha) - \inf_{\alpha\in\Lambda_\infty} R_{\tau,emp}(\alpha)\right| > \varepsilon\right\} = 0 \text{ for every } \varepsilon > 0 \qquad (16)$$

whereas the application of Theorem 1 yields

$$\lim_{l\to\infty} \mathbf{P}\left\{\left|\inf_{\alpha\in\Lambda_\tau(c)} R_\tau(\alpha) - \inf_{\alpha\in\Lambda_\tau(c)} R_{\tau,emp}(\alpha)\right| > \varepsilon\right\} = 0 \text{ for every } \varepsilon > 0$$

By using Lemma 2 we obtain therefore

$$\lim_{l\to\infty} \mathbf{P}\left\{\left|\inf_{\alpha\in\Lambda(c)} R_\tau(\alpha) - \inf_{\alpha\in\Lambda(c)} R_{\tau,emp}(\alpha)\right| > \varepsilon\right\} = 0 \qquad (17)$$

for every $\varepsilon > 0$ and every $c \in [a, A]$. □

Theorem 3. *The following inequality holds*

$$\mathbf{P}\left\{\sup_{\alpha\in\Lambda}\left(\int Q_\tau(z,\alpha)dF(z) - \frac{1}{l}\sum_{j=1}^{l} Q_\tau(z_j,\alpha)\right) > \varepsilon\right\}$$
$$\leq 4\exp\left\{\left(\frac{G^{\Lambda_\tau}(2l)}{l} - \left(\varepsilon - \frac{1}{l}\right)^2\right) l\right\} + 4\exp\left\{\left(\frac{G^{\Lambda_\infty}(2l)}{l} - \left(\tau - \frac{2}{l}\right)^2\right) l\right\} \qquad (18)$$

4 A more practical choice for the empirical risk

Unfortunately, in real-world applications the measure μ on the input space D is unknown and only the training set is available. In these cases the empirical risk $R_{\tau,emp}(\alpha)$, which depends on $\mu(T_\alpha)$, cannot be calculated. Thus we have to use a different form of the empirical risk that allows a direct evaluation while ensuring the convergence in probability to $\inf_{\alpha\in\Lambda} R_\tau(\alpha)$ when l increases indefinitely. In this way the replacement does not prejudice the consistency of the ERM method.

A possible choice is the following

$$R'_{emp}(\alpha) = \frac{1}{l}\sum_{j=1}^{l} Q'(z_j, \alpha)$$

where $Q'(z, \alpha)$ is defined in (5).

We can prove that, under mild conditions, this form of the empirical risk shares the same convergence properties of $R_{\tau,emp}(\alpha)$.

If $\Lambda_0 = \{\alpha \in \Lambda_\tau : \mu(T_\alpha) = 0\}$, the corresponding classifiers $\varphi(x, \alpha)$, with $\alpha \in \Lambda_0$ do not misclassify any point of the certainty region C. Then $\Lambda_{0,\tau} = \Lambda_\tau \backslash \Lambda_0$ includes the values of α for which $0 < \mu(T_\alpha) < \tau$.

The following corollary establishes the convergence properties of $R'_{emp}(\alpha)$.

Corollary 1. *If*

$$\inf_{\alpha\in\Lambda_0} R_\tau(\alpha) \leq \inf_{\alpha\in\Lambda_{0,\tau}} R_\tau(\alpha) \tag{19}$$

then

$$\inf_{\alpha\in\Lambda} R'_{emp}(\alpha) \xrightarrow[l\to\infty]{\mathbf{P}} \inf_{\alpha\in\Lambda} R_\tau(\alpha) \tag{20}$$

Furthermore, it can be easily proved that the rate of convergence of $R'_{emp}(\alpha)$ to $R_\tau(\alpha)$ can be upper bounded by the right hand side of (18).

References

1. V. N. Vapnik and A. Ya. Chervonenkis On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities. *Theory of Probability and Its Applications* **16** (1971) 264–280.
2. V. N. Vapnik *Estimation of Dependences Based on Empirical Data.* (1982) New York: Springer-Verlag.
3. V. N. Vapnik *Statistical Learning Theory.* (1998) New York: John Wiley & Sons.