Dipartimento di Informatica e
Scienze dell'Informazione

Thesis Proposal

Regularization Approaches to Learning Theory

Lorenzo Rosasco

Advisor: Prof. Alessandro Verri

Dipartimento di Informatica e Scienze dell'Informazione DISI, Università di Genova
v. Dodecaneso 35, 16146 Genova, Italy http://www.disi.unige.it/

Contents

Abstract

Statistical Learning Theory proved to be theoretically well founded and provided effective learning algorithms appliable in many different fields. Nevertheless a number of very important theoretical issues still need to be fully explained. In particualr it seems that the a strong connection between Regularization Theory for Ill Posed Inverse Problem (RT) and Learning Theory (LT) has not been completely exploited, yet. The aim of our study is twofold. First, to study the theoretical properties of Tikhonov regularization algorithm in the context of Learning Theory, emphasizing the possible relations between RT and LT. Second, to use this connection to develop new selection criteria for regularization parameters and to design novel and effective learning algorithms derived from regularization techniques.

1. Motivations and Aims of the Thesis

In recent years considerable progress has been made in the understanding of problems of learning and generalization. Learning in this context basically means the ability to perform well on new data after recovering a model on the basis of given data. Such problems arise in many different areas and are becoming increasingly important and crucial towards many applications such as bioinformatics, multimedia, computer vision, internet search and information retrieval, datamining and textmining, finance and several others. Often the dimensionality of the input spaces in these novel applications is huge as in the analysis of microarray data, where expression level of thousands of genes need to be analyzed, given only a limited number of experiments. Without performing dimensionality reduction, the classical statistical paradigms show fundamental shortcomings at this point. A modern approach to this kind of problems is provided by Statistical Learning Theory and learning from examples paradigm.

In this context the main problem of learning was posed by Vapnik as the problem of the consistency and generalization of Empirical Risk Minimization (ERM). This last method selects an estimator minimizing the error on the given examples. Its consistency refers to the capability of recovering the best solution as the number of examples goes to infinity while its generalization refers to the property of performing well on new inputs. It turned out (Vapnik, 1988, Alon et al., 1997) that consistency and generalization properties of ERM could be completely characterized by controlling the complexity of the solution. Hence to obtain profitable algorithms one should modify ERM so to be able to control the solution complexity. A class of learning algorithm inspired by this last observation are the so called Regularization Networks (RN) (Evgeniou et al., 2000) and in particular Support Vector Machines (SVM) (Vapnik, 1988, Cristianini and Shawe Taylor, 2000) which proved to be extremely effective in many different application field. Despite the very good practical results, the theoretical properties of such algorithms is far from being completely investigated. This is mostly due to the fact that for a long time they have been simply considered as a particular instance of Structural Risk Minimization principle (SRM)

(Vapnik, 1988). It is possible to see that RN, though clearly inspired from SRM, cannot be mathematically derived from it. The idea of using Regularization Theory to tackle this problem comes from the fact that the form of the functional minimized by RN closely resembles the functional which is usually minimized in Tikhonov Regularization for ill-posed inverse problems. Nonetheless conceptual differences between inverse problems and learning problems make the formalization of the above intuition not straightforward. Clearly, drawing a bridge between LT and RT would provide a number of very useful insights both from the theoretical and the practical point of view. In particular one of the main goals would be to define new effective selection rules for the regularization parameter whose choice is known to be one of the central issue both in LT and RT.

The rest of the document is divided as follows: first we introduce the main concepts of Learning Theory and fix the mathematical setting, second we present some classical results about the ERM algorithm, third we state some preliminary results about the mathematical properties of Regularization Networks, and finally we briefly sketch how we plan to schedule the future work.

2. Introduction on Statistical Learning Theory

In this section we introduce the main concepts of Learning Theory. Since will be dealing with theoretical issues, we spend some words to describe the mathematical framework in which the whole theory is developed.

2.1 Input and output spaces

We denote with X and Y the input and output spaces respectively. We assume that X is a compact [1] subset of $\mathbb{R}^d$ and Y is a compact subspace of $\mathbb{R}$. The output space is usually $\mathbb{R}$, for *regression problems*, or simply $\{-1, 1\}$ for *binary classification problems.* We let $Z = X \times Y$ and endow it with a probability distribution ρ defined on the Borel σ-algebra of Z. The probability measure ρ under very mild assumptions decomposes as $\rho(\mathbf{x}, y) = \rho(y|\mathbf{x})\nu(\mathbf{x})$, where $\rho(y|\mathbf{x})$ and $\nu(\mathbf{x})$ are the conditional and marginal probabilities respectively. The probabilistic setting in which the whole theory is developed is very general and allows to model different situations. Clearly different outputs y could correspond to the same input $\mathbf{x}$, for example it could happen that

- the underlying process is deterministic, but there is **noise** in the measurement of y;
- the underlying process is **not deterministic**;

[1] The requirement on X and Y to be compact set can be sometimes replaced by the weaker requirement of being simply closed subsets (De Vito et al., 2003). This can be useful to be able to deal with unbounded domains.

- the underlying process is deterministic, but only **incomplete** information is available.

In practice the probability measure is known only through a finite set of *examples* $S = (\mathbf{x}_1, y_1), \dots, (\mathbf{x}_\ell, y_\ell)$, called *training set*, drawn *i.i.d.* according to the unknown probability ρ. In this context the aim of the learning procedure is to find, given a training set S, an estimator $f : X \to Y$ such that

$$f(\mathbf{x}) = y,$$

that is a function that, given a new point $\mathbf{x}$, can "effectively" predict the associated label y. Later in the section we will formalize the above statement, here we just give the following two comments. First and most important, the estimator f should correctly predict not only the labels of the initial set of data, but especially the label of new points $\mathbf{x}$. If so, we say that the solution *generalizes*. Second, we note that, since there is not a unique output associated to an input point $\mathbf{x}$, we expect the estimator f to commit errors, the best solution being the one giving the minimum number of errors.

From the last comment it should be apparent that we need to define some measure of the error made by a certain estimator f; this leads naturally to the concept of loss function.

2.2 Loss functions

The *loss function* $V(y, f(\mathbf{x}))$ is the price we are willing to pay by using $f(\mathbf{x})$ to predict the correct label y. Probably the most typical example of such a function is the square loss

$$(y - f(\mathbf{x}))^2.$$

For classification problems a very natural choice is the so called $1 - 0$ loss

$$|-yf(\mathbf{x})|_*$$

where $|\tau|_* = 1$ if $\tau > 0$ and $|\tau|_* = 0$ otherwise; which is easy to show that simply counts the number of misclassification errors made by the function f. In table 2.2 we report some other common loss functions. The following definition collect the main mathematical assumptions on the loss function.

Definition 1 (Loss Function) *A loss function is a map $V : Y \times \mathbb{R} \to [0; +\infty]$ such that*

1. *$V(y, \omega) > 0$ for all $y \in Y$ and $w \in \mathbb{R}$,*

2. *V is continuous on $Y \times \mathbb{R}$,*

3. *$\forall\, y \in Y$, $V(y, \cdot)$ is a convex function on $\mathbb{R}$.*

We note that it is sometimes useful to look at the loss function as a map $\ell : Z \to \mathbb{R}$, where $\ell(\mathbf{z}) = V(f, \mathbf{z}) = V(y, f(\mathbf{x}))$. In the following section we show how to measure the average error made by a certain estimator f, given our definition of loss function.

problem	loss	
regr	square	$(y - f(\mathbf{x}))^2$
regr	abs val	$\|y - f(\mathbf{x})\|$
regr	ϵ-insensitive	$\|y - f(\mathbf{x})\|_\epsilon$
class	quad	$(y - f(\mathbf{x}))^2$
class	hinge	$\|1 - yf(\mathbf{x})\|_+$
class	logistic	$\ln(1 + e^{-yf(\mathbf{x})})$
class	exponential	$e^{-yf(\mathbf{x})}$

Table 1: Some of the loss function commonly in use. We note that both in regression and classification the loss functions depend on just one variable τ, with $\tau = y - f(x)$ for regression and $\tau = yf(\mathbf{x})$ for classification. Moreover we recall that $|\tau|_\epsilon = \tau$ if $|\tau| > \epsilon$ and $|\tau|_\epsilon = 0$ otherwise.

2.3 Learning Functionals

The *expected risk* of a function f is a functional $I : \mathcal{F} \to [0; +\infty]$ defined as

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

and can be seen as the average error obtained by the function f, where f is a possible solution of the learning problem and the probability measure ρ is unknown. The above integral is well defined for example if $\mathcal{F} = \mathcal{C}(X)$ or $\mathcal{F} = L^2(X)$ since V is continuous and the input and output space are compact.

The expected risk is not computable since the probability measure ρ is fixed but unknown. Given a training set $S = (\mathbf{x}_i, y_i)_{i=1}^{\ell}$, a possible way to estimate $I[f]$ is to evaluate the *empirical risk*

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

Both the expected and empirical risk can be represented in a compact form trough the following expression

$$\mathbb{E}_\rho[V(y, f(\mathbf{x}))],$$

ρ being respectively the unknown distribution describing the relation between $\mathbf{x}$ and y or the empirical measure

$$\rho_S = \frac{1}{\ell}\sum_{i=1}^{\ell}\delta_{(\mathbf{x}_i,y_i)},$$

associated with a training set S. The above definition of expected risk allows us to get some more insight of what we mean by effective, or better generalizing, solution. We recall in fact that the problem of learning is to find, given the training set S, an *estimator* f effectively predicting the label of a new point. This now translates into requiring the expected risk $I[f]$ associated to f to be as small as possible. As we will see shortly, this last statement needs some care since we are working in a probabilistic setting, but before dealing in details with this issue we need to introduce the central notion of hypothesis space.

2.4 Hypothesis space

From the algorithmic point view it is impossible to deal with the entire space $\mathcal{F}$ on which the expected risk $I[f]$ is defined. The search of a solution is hence restricted to an appropriate class of function called *hypothesis space* which is usually indicated with $\mathcal{H}$. Though more general choice can be considered (Cucker and Smale, 2002), Reproducing Kernel Hilbert Spaces (RKHS) proved to be particularly profitable for learning problems.

We briefly recall the basic definitions and facts about RKHS. Let $K : X \times X \to \mathbb{R}$ be a continuous function such that

1. K is symmetric, i.e. $K(\mathbf{x}, \mathbf{s}) = K(\mathbf{s}, \mathbf{x})$
2. K is definite positive, that is given $\mathbf{x}_1, \ldots, \mathbf{x}_l \in X$, let $(K_l)_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$ then K_l is a semidefinite positive matrix (that is, its eigenvalues are non-negative).

The function K is called *Mercer kernel.*

We let $\mathcal{C}(X)$ be the space of continuous functions on X and $\|\cdot\|_\infty$ the uniform norm, that is,

$$\|f\|_\infty = \sup_{\mathbf{x}\in X} |f(\mathbf{x})|.$$

Moreover, given $\mathbf{x} \in X$, we define $K_\mathbf{x}$ be the function on X given by $K_\mathbf{x}(\mathbf{s}) = K(\mathbf{s}, \mathbf{x})$, and

$$C_K = \sup_{\mathbf{x}\in X} \sqrt{K(\mathbf{x}, \mathbf{x})}. \tag{1}$$

The following theorem is essentially due to Aronszajn (1950).

Theorem 2 *Given a Mercer kernel K, there is a unique subspace $\mathcal{H}_K$ of $\mathcal{C}(X)$ such that*

1. *$\mathcal{H}_K$ is a (real) Hilbert space;*

2. $\mathcal{H}_K = \overline{\text{span}}\{K_{\mathbf{x}} \mid \mathbf{x} \in X\}$*;*

3. $\langle K_{\mathbf{x}}, K_{\mathbf{s}} \rangle_{\mathcal{H}_K} = K(\mathbf{s}, \mathbf{x})$*.*

Moreover, one has the following properties

1. for all $\mathbf{x} \in X$*,* $f(\mathbf{x}) = \langle f, K_{\mathbf{x}} \rangle_{\mathcal{H}_K}$*;*

2. the canonical immersion of $\mathcal{H}_K$ *into* $\mathcal{C}(X)$ *is compact and*

$$\|f\|_\infty \leq C_K \, \|f\|_{\mathcal{H}_k}\,, \tag{2}$$

where C_K *is given by Eq.* (2)*.*

The space $\mathcal{H}_K$ is called the *reproducing kernel Hilbert space* with reproducing kernel K.

2.5 Learning, Generalization and Consistency

As we mentioned before, to generalize a good estimator f_S should have expected risk as small as possible. However, this statement needs some careful examination: since $I[f_S]$ is a random variable on the sample space Z^ℓ, the most we can require is $I[f_S]$ to be small with high probability. This formally translates into finding a probabilistic bound on $I[f_S]$, that is a function $E = E(\mathcal{H}, \eta, \ell, S)$ such that fixed an hypothesis space $\mathcal{H}$, the number of examples $\ell \in \mathbb{N}$ and the confidence level $\eta \in (0, 1)$

$$\text{Prob}\{S \in Z^\ell \mid I[f^S] > E(\mathcal{H}, \eta, \ell, S)\} \leq \eta. \tag{3}$$

The idea of consistency is formally translated into the following property of the solution, for any $\epsilon > 0$

$$\lim_{\ell \to \infty} \text{Prob}\{I[f_S] - \inf_{f \in \mathcal{H}} I[f] > \epsilon\} = 0.$$

If in the above limit we consider the supreme over all the possible probability measures, we speak of *u*niversal consistency.

3. Some classical results

In this section we briefly outline some classical results of statistical learning theory, for details we refer to (Vapnik, 1988, Evgeniou et al., 2000).

3.1 Empirical Risk Minimization

Fixed an hypothesis space $\mathcal{H}$ and given a certain training set S, probably the simplest and most studied algorithm is Empirical Risk Minimization. According to ERM a solution f_S is found solving the following minimization problem

$$\min_{f \in \mathcal{H}} \{ I_{emp}^{S}[f] \}.$$

The term Empirical Risk Minimization is somewhat misleading since in general such a minimum might not exist. For this reason one should replace ERM with the notion of *almost*-ERM (Mukherjee et al., 2002). For the sake of simplicity we assume that the ERM admits a solution f_S and develop our analysis for the ERM algorithm (we refer to Mukherjee et al. (2002) for a detailed account in the case of almost-ERM). The key problem of learning theory was posed by Vapnik (1988) as the problem of the universal consistency of ERM. It turns out that ERM consistency is ensured if the hypothesis space is sufficiently small so that a solution is close to the best solution both with respect to the continuous and to the empirical measure. To formalize the above statement we need the fundamental notion of *uniform convergence in probability* of a class of functions. Function classes for which uniform convergence in probability holds are called *uniform Glivenko-Cantelli (uGC)* classes of functions.

Definition 3 (uniform Glivenko-Cantelli) *A function class $\mathcal{F}$ is a uGC class if for any $\epsilon > 0$*

$$\lim_{\ell \to \infty} \sup_{\rho} \mathrm{Prob}\{ \sup_{f \in \mathcal{F}} |\mathbb{E}_{\rho_S} f - \mathbb{E}_{\rho} f| > \epsilon \} = 0.$$

Applied to the function class $\ell(\mathbf{z})$, defined by V and $\mathcal{H}$, (see section 2.2) this means that for all the probability distributions we have that

$$\mathrm{Prob}\{ \sup_{f \in \mathcal{F}} |I_{emp}^{S}[f] - I[f]| > \epsilon \} \leq \delta.$$

We are now ready to state the theorem that completely characterizes the consistency of ERM.

Theorem 4 *Given a loss function V and fixed an hypothesis space $\mathcal{H}$, ERM algorithm is consistent if and only if the class of functions $\ell(\mathbf{z})$ is uGC.*

By means of the above theorem the search for conditions for the consistency of ERM now translates into searching conditions for uniform convergence in probability over the hypothesis space. As mentioned before, such conditions can be determined by “measuring” the complexity of the hypothesis space. Before giving the main result in this direction, we need the following definition of complexity

Definition 5 (V_γ-dimension) *Let $A \leq V(y, f(\mathbf{x})) \leq B$, $f \in \mathcal{F}$, with $A, B < 0$. The V_γ-dimension of V in the set $\mathcal{F}$ is defined as the maximum number h of vectors $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_h, y_h)$ that can be separated into two classes in all 2^h possible ways using rules*

$$\textit{class 1 if:} \quad V(y_i, f(\mathbf{x}_i)) \geq s + \gamma$$

$$\textit{class 0 if:} \quad V(y_i, f(\mathbf{x}_i)) \leq s - \gamma$$

for $f \in \mathcal{F}$ and some $s > 0$. If for any number N we can find N points $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$ that can be separated in all the 2^N possible ways, we will say that the V_γ-dimension of V in $\mathcal{F}$ is infinite.

We are now ready to state the following theorem

Theorem 6 *Finiteness of the V_γ dimension is a necessary and sufficient condition for uniform convergence in probability over the hypothesis space $\mathcal{H}$.*

3.2 Structural Risk Minimization Principle

We start noting that, although ERM itself does not generally provide an effective solution to the learning problem, the above results suggest how it can be modified to get better performances: a good algorithm, beside minimizing the empirical error on the training set, should be able to control the complexity of the solution. This last observation is at the basis of the so called Structural Risk Minimization Principle (Vapnik, 1988). According to SRM a modified algorithm can be obtained exploiting ERM on a structure of nested spaces of increasing complexity. The solution is the one for which the best trade-off between empirical error and complexity is attained.

4. Preliminary results: Ivanov Regularizarion

A classical way to build a structure of spaces of increasing complexity is to consider balls of increasing radius in a RKHS $\mathcal{H}$,

$$\mathcal{H}_R := \{f \in \mathcal{H} \mid \|f\|_h h^2 < R^2\}.$$

The obtained algorithm corresponds in RT to Ivanov Regularization (Tikhonov and Arsenin, 1977). The first part of our work is dedicated to the study of consistency and generalization properties of such an algorithm. As a byproduct of our analysis we are able to investigate the impact of choosing different loss functions from the theoretical viewpoint. Moreover we derive a general result on the minimizer of the expected risk for convex loss functions 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 (see Tab. 2.2). Furthermore, if the hypothesis space is sufficiently rich, the bounds obtained for the hinge loss are not loosened by the thresholding stage. For details about the results presented in this section we refer to Rosasco et al. (2003a).

4.1 Estimation error bounds for Ivanov Regularization

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

$$I[f_D] - I[f_0] = (I[f_D] - I[f_R]) + (I[f_R] - I[f_0]) \tag{4}$$

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 (4) 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 (2002). 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* $\epsilon > 0$,

$$\text{Prob}\{D \in Z^\ell | \sup_{f \in \mathcal{H}_R} |I[f] - I_{emp}[f]| \leq \epsilon\} \geq 1 - 2N(\frac{\epsilon}{4L_M})e^{-\frac{\ell\epsilon^2}{8B^2}} \quad . \tag{5}$$

The above Lemma can be compared to the classic result in Vapnik (1988) (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 in Zhou (2002), Pontil (2003).

We are now in a position to generalize Theorem C in Cucker and Smale (2002) and obtain the probabilistic bound by observing that for a fixed η the confidence term in Eq. (5) 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 $\lim_{\ell \to \infty} \epsilon(\eta, \ell, R) = 0$.

4.2 Statistical properties of loss functions

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

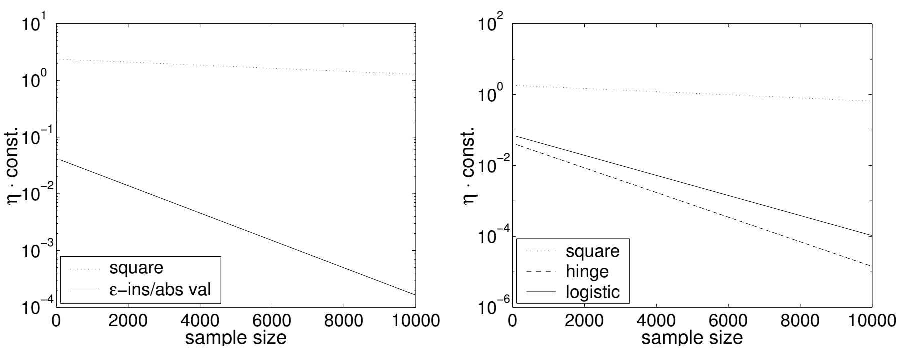

Figure 1: Semilogarithmic plots of the convergence rates of various loss functions for regression (left) and classification (right). In both cases, $R = 1.5$, $\epsilon = 0.2$ and the dimensionality of the input-space used to estimate the covering number is 10. For regression we set $\delta = 1.5$.

4.3 Comparing convergence rates

Using Eq. (5) 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.,

$$2N\left(\frac{\epsilon}{4}\right)\exp\left(-\frac{\ell\epsilon^2}{8(R+\delta)^2}\right) \tag{6}$$

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

$$2N\left(\frac{\epsilon}{4(2R+\delta)}\right)\exp\left(-\frac{\ell\epsilon^2}{8(R(2R+\delta)+\delta^2)^2}\right). \tag{7}$$

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 (6) and (7). In Figure (1) on the left 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. The behavior of the convergence rates for these three loss functions is depicted in Figure (1) on the right (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.

4.4 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

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

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. (2002))

$$I[f_0] = I[f_b]. \tag{8}$$

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

$$I[sign(f_D)] - I[f_b], \tag{9}$$

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

$$I[sign(f_D)] \leq I[f_D]. \tag{10}$$

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. (8) and (10) and the theorem, we obtain that for $0 < \eta < 1$ and $R > 0$ with probability at least $1 - \eta$

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

[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 $V'_-(0) = V'_+(0)$.

We stress that the simple derivation of the above bound follows naturally from the special property of the hinge loss expressed in Eq. (8). For other loss functions similar results can be derived through a more complex analysis (Lugosi and Vayatis, 2003, Zhang, 2001).

5. Future Works

The results presented here are the first step towards the study and theoretical comprehension of learning algorithms inspired by regularization techniques. In this first part of our research we focused on Ivanov regularization which amounts to solve a minimization problem of the form:

$$\min\{I_{emp}^{S}\}$$

subject to

$$\|f\|_{\mathcal{H}}^{2} < R^{2}.$$

A class of algorithms, called Regularization Networks, considers the following lagrangian formulation of the above minimization problem

$$\min_{f\in\mathcal{H}}\{I_{emp}^{S}[f] + \lambda \|f\|_{\mathcal{H}}^{2} < R^{2}\}$$

that can be seen as a particular instance of Tikhonov regularization. Regularization Networks (Evgeniou et al., 2000) and in particular Support Vector Machines (Cristianini and Shawe Taylor, 2000) proved to be extremely effective in many different application field but despite of the very good experimental results the theoretical properties of such algorithms are far from being completely investigated. This last fact is mostly due to the fact that for a long time RN have been simply considered as a particular instance of SRM while it is possible to see that, though clearly inspired from SRM, RN cannot be mathematically derived from it. In particular it can be shown that in general it is not trivial to export results obtained in the Ivanov regularization setting to the context of Tikhonov regularization. Future work will be essentially divided in three parts. In the first part we mean to focus on Tikhonov regularization techniques in the context of learning, dealing in particular with the following issues: a) the study of existence, uniqueness and explicit form of the minimizers (some preliminary results can be found in (Rosasco et al., 2003b)); b) the problem of consistency, generalization, and model selection (choice of regularization parameters). Then in the second part we are going to focus on the square loss and try to recast the problem of learning as an inverse problem. We conjecture that some of the notions in the context of RT of inverse problem have to be modified in order to deal with the probabilistic setting of learning problems. It seems that special care has to be taken of the role of random sampling of the data and that a suitable definition of noise has to be given (in particular for classification problems). In the last part of our work we plan to test the outcome of the theoretical investigations on both toy and real problems.

5.1 Future Work Organization

1. Ivanov and Tikhonov regularization in learning from examples.
   (a) Consistency and generalization of Ivanov regularization.
   (b) Solution of Tikhonov regularization.
   (c) Consistency and generalization of Tikhonov regularization.
2. Learning from examples as an inverse problem.
   (a) Study of the role of random sampling.
   (b) *A*d hoc definition of noise, stability and regularizing algorithms.
3. Experiments on toy and real problems.

References

N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence and learnability. *Journal of he ACM*, 44(4):615–631, 1997.

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

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

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

E. De Vito, L. Rosasco, and A. and Caponnetto. Some properties of regularized kernel methods. *submitted to JMLR*, 2003.

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

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

Y Lin, G. Wahba, H. Zhang, and Y Lee. Statistical properties and adaptive tuning of support vector machines. *Machine Learning*, 48:115–136, 2002.

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

S. Mukherjee, T. Niyogi, P.and Poggio, and R. Rifkin. Statistical learning: Loo stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *CBCL Paper 223, Massachusetts Institute of Technology, Cambridge, MA*, 2002.

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

L. Rosasco, E. De Vito, A. Caponnetto, M. Piana, and A. Verri. Are loss functions all the same. *to appear in Neural Computation*, 2003a.

L. Rosasco, E. De Vito, A. Caponnetto, M. Piana, and A. Verri. Minimization of tiklhonov functional: the continuos setting. Technical Report 3232, DISI, 2003b.

A.N. Tikhonov and V.Y. Arsenin. *Solutions of Ill Posed Problems*. W. H. Winston, Washington, D.C., 1977.

V. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1988.

T. Zhang. Convergence of large margin separable linear classification. In T.G. Len, T.K. Dietterich and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 357–363. MIT Press, 2001.

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