Journal of Machine Learning Research () Submitted 12/04; Published

Learning from Examples as an Inverse Problem

Ernesto De Vito DEVITO@UNIMO.IT
Dipartimento di Matematica
Università di Modena e Reggio Emilia
Modena, Italy
and INFN, Sezione di Genova,
Genova, Italy

Lorenzo Rosasco ROSASCO@DISI.UNIGE.IT
Andrea Caponnetto CAPONNETTO@DISI.UNIGE.IT
Umberto De Giovannini UMBERTO.DEGIOVANNINI@FASTWEBNET.IT
Francesca Odone ODONE@DISI.UNIGE.IT
DISI
Università di Genova,
Genova, Italy

Editor:

Abstract

Many works related learning from examples to regularization techniques for inverse problems. Nevertheless by now there was no formal evidence neither that learning from examples could be seen as an inverse problem nor that theoretical results in learning theory could be independently derived using tools from regularization theory. In this paper we provide a positive answer to both questions. Indeed, considering the square loss, we translate the learning problem in the language of regularization theory and we show that consistency results and optimal regularization parameter choice can be derived by the discretization of the corresponding inverse problem.

Keywords: Statistical Learning, Inverse Problems, Regularization theory, Consistency.

1. Introduction

The main goal of learning from examples is to infer a decision function from a finite sample of data drawn according to a fixed but unknown probabilistic input-output relation. The desired property of the selected function is to be descriptive also of new data, i.e. it should *generalize*. The fundamental work of Vapnik and further developments (Vapnik, 1998, Alon et al., 1997, Schölkopf and Smola, 2002, Bartlett and Mendelson, 2002) show that the key to obtain a meaningful solution is to control the complexity of the hypothesis space. Interestingly, as pointed out in a number of papers (see Poggio and Girosi (1992), Evgeniou et al. (2000) and references therein), this is in essence the idea underlying regularization techniques for ill-posed problems (Tikhonov and Arsenin, 1977, Engl et al., 1996). Not surprisingly the form of the algorithms proposed in both theories is strikingly similar (Mukherjee et al., 2002). Anyway a careful analysis shows that a rigorous mathematical connection between statistical learning and regularization for inverse problems is not straightforward.

In this paper we show that the problem of learning from examples can be translated into a suitable inverse problem, which for square loss leads to solve a linear inverse problem, and we propose a novel approach to derive consistency results based on regularization theory. Some previous works on this subject consider the special case in which the elements of the input space are fixed and not probabilistically drawn (Mukherjee et al., 2004, Kurkova, 2004). Some weaker results in the same spirit of those presented in this paper can be found in Rudin (2004) where anyway the connections with inverse problems is not discussed. The arguments used to derive our results are close to those used in the study of stochastic inverse problems discussed in Vapnik (1998). Our results are also similar to those presented in Smale and Zhou (2004b) where the connection between learning and sampling theory is investigated. Finally (Ong and Canu, 2004) consider how various regularization techniques apply to the context of learning. In particular several iterative algorithms are considered and convergence with respect to the regularization parameter (semiconvergence) is proved. Finally, in De Vito et al. (2004b) there is a detailed discussion of the problem of estimating the noise for Tikhonov regularization in the context of inverse linear problems. The application to learning theory is also considered with results similar to the ones given here, but based on different probabilistic inequalities.

The paper is organized as follows. After recalling the main concepts and notation of statistical learning (Section 2) and of inverse problems (Section 3), in Section 4 we develop a formal connection between the two theories. In Section 5 the main results are stated and discussed. All the mathematical proofs are contained in the Appendix. Finally in Section 6 we conclude with some remarks and open problems.

2. Learning from examples

We briefly recall some basic concepts of statistical learning theory (for details see Vapnik (1998), Evgeniou et al. (2000), Schölkopf and Smola (2002), Cucker and Smale (2002) and references therein).

In the framework of learning from examples, there are two sets of variables: the input space

X, which we assume to be a compact subset of $\mathbb{R}^n$, and the output space Y, which is a subset of $\mathbb{R}$ contained in $[-M, M]$ for some $M \geq 0$. The relation between the input $x \in X$ and the output $y \in Y$ is described by a probability distribution $\rho(x, y) = \nu(x)\rho(y|x)$ on $X \times Y$. The distribution ρ is known only through a sample $\mathbf{z} = (\mathbf{x}, \mathbf{y}) = ((x_1, y_1), \ldots, (x_\ell, y_\ell))$, called *training set*, drawn independently and identically distributed (i.i.d.) according to ρ. Given the sample $\mathbf{z}$, the aim of learning theory is to find a function $f_{\mathbf{z}} : X \to \mathbb{R}$ such that $f_{\mathbf{z}}(x)$ is a good estimate of the output y when a new input x is given. The function $f_{\mathbf{z}}$ is called *estimator* and the map providing $f_{\mathbf{z}}$, for any training set $\mathbf{z}$, is called *learning algorithm*.

Given a measurable function $f : X \to \mathbb{R}$, the ability of f to describe the distribution ρ is measured by its *expected risk* defined as

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

where $V(f(x), y)$ is the *loss function*, which measures the cost paid by replacing the true label y with the estimate $f(x)$. In this paper we consider the square loss

$$V(f(x), y) = (f(x) - y)^2.$$

With this choice, it is well known that the regression function

$$g(x) = \int_Y y\, d\rho(y|x),$$

is well defined (since Y is compact) and is the minimizer of the expected risk over the space of all the measurable real functions on X. In this sense g can be seen as the ideal estimator of the distribution probability ρ. However, the regression function cannot be reconstructed exactly since only a finite, possibly small, set of examples $\mathbf{z}$ is given.

To overcome this problem, in the framework of the regularized least squares algorithm (Wahba, 1990, Poggio and Girosi, 1992, Cucker and Smale, 2002, Zhang, 2003), an hypothesis space $\mathcal{H}$ of functions is fixed and the estimator ${f_{\mathbf{z}}}^{\lambda}$ is defined as the solution of the regularized least squares problem,

$$\min_{f \in \mathcal{H}} \{ \frac{1}{\ell} \sum_{i=1}^{\ell} (f(x_i) - y_i)^2 + \lambda \Omega(f) \}, \tag{1}$$

where Ω is a penalty term and λ is a positive parameter to be chosen in order to ensure that the discrepancy.

$$I[{f_{\mathbf{z}}}^{\lambda}] - \inf_{f \in \mathcal{H}} I[f]$$

is small with hight probability. Since ρ is unknown, the above difference is studied by means of a probabilistic bound $\mathcal{B}(\lambda, \ell, \eta)$, which is a function depending on the regularization parameter λ, the number ℓ of examples and the confidence level $1 - \eta$, such that

$$\mathbf{P}\left[I[{f_{\mathbf{z}}}^{\lambda}] - \inf_{f \in \mathcal{H}} I[f] \leq \mathcal{B}(\lambda, \ell, \eta) \right] \geq 1 - \eta.$$

We notice that, in general, $\inf_{f \in \mathcal{H}} I[f]$ is larger than $I[g]$ and represents a sort of irreducible error (Hastie et al., 2001) associated with the choice of the space $\mathcal{H}$. We do not require the infimum $\inf_{f \in \mathcal{H}} I[f]$ to be achieved. If the minimum on $\mathcal{H}$ exists, we denote the minimizer by $f_{\mathcal{H}}$.

In particular, the learning algorithm is *consistent* if it is possible to choose the regularization parameter, as a function of the available data $\lambda = \lambda(\ell, \mathbf{z})$, in such a way that

$$\lim_{\ell \to +\infty} \mathbf{P}\left[I[{f_{\mathbf{z}}}^{\lambda(\ell,\mathbf{z})}] - \inf_{f \in \mathcal{H}} I[f] \geq \epsilon \right] = 0, \tag{2}$$

for every $\epsilon > 0$. The above convergence in probability is usually called *(weak) consistency* of the algorithm (see Devroye et al. (1996) for a discussion on the different kind of consistencies).

In this paper we assume that the hypothesis space $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) on X with a continuous kernel K. We recall the following facts (Aronszajn, 1950, Schwartz, 1964). The kernel $K : X \times X \to \mathbb{R}$ is a continuous symmetric positive definite function, where *positive definite* means that

$$\sum_{i,j} a_i a_j K(x_i, x_j) \geq 0.$$

for any $x_1, \ldots x_n \in X$ and $a_1, \ldots a_n \in \mathbb{R}$.
The space $\mathcal{H}$ is a real separable Hilbert space whose elements are real continuous functions defined on X. In particular, the functions $K_x = K(\cdot, x)$ belong to $\mathcal{H}$ for all $x \in X$, and

$$\begin{aligned} \mathcal{H} &= \overline{\operatorname{span}\{\, K_x \,|\, x \in X \}} \\ \langle K_x, K_t \rangle_{\mathcal{H}} &= K(x,t) \qquad \forall x, t \in X, \end{aligned}$$

where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is the scalar product in $\mathcal{H}$. Moreover, since the kernel is continuous and X is compact

$$\kappa = \sup_{x \in X} \sqrt{K(x,x)} = \sup_{x \in X} \|K_x\|_{\mathcal{H}} < +\infty, \tag{3}$$

where $\|\cdot\|_{\mathcal{H}}$ is the norm in $\mathcal{H}$. Finally, given $x \in X$, the following *reproducing* property holds

$$f(x) = \langle f, K_x \rangle_{\mathcal{H}} \qquad \forall f \in \mathcal{H}. \tag{4}$$

In particular, in the learning algorithm (1) we choose the penalty term

$$\Omega(f) = {\|f\|_{\mathcal{H}}}^2,$$

so that , by a standard convex analysis argument, the minimizer ${f_{\mathbf{z}}}^{\lambda}$ exists, is unique and can be computed by solving a linear finite dimensional problem, (Wahba, 1990).

With the above choices, we will show that the consistency of the regularized least squares algorithm can be deduced using the theory of linear inverse problems we review in the next section.

3. Ill-Posed Inverse Problems and Regularization

In this section we give a very brief account of the main concepts of linear inverse problems and regularization theory (see Tikhonov and Arsenin (1977), Groetsch (1984), Bertero et al. (1985, 1988), Engl et al. (1996), Tikhonov et al. (1995) and references therein).

Let $\mathcal{H}$ and $\mathcal{K}$ be two Hilbert spaces and $A : \mathcal{H} \to \mathcal{K}$ a linear bounded operator. Consider the equation

$$Af = g \tag{5}$$

where $g \in \mathcal{K}$ is the *exact* datum. Finding the function f satisfying the above equation, given A and g, is the linear inverse problem associated to (5). In general the above problem is ill-posed, that is, the solution either not exists, is not unique or does not depend continuously on the datum g. Existence and uniqueness can be restored introducing the Moore-Penrose generalized solution $f^\dagger$ defined as the minimal norm solution of the least squares problem

$$\min_{f \in \mathcal{H}} \|Af - g\|_{\mathcal{K}}^2 . \tag{6}$$

It can be shown (Tikhonov et al., 1995) that the generalized solution $f^\dagger$ exists if and only if $Pg \in \text{Range}(A)$, where P is the projection on the closure of the range of A. However, the generalized solution $f^\dagger$ does not depend continuously on the datum g, so that finding $f^\dagger$ is again an ill-posed problem. This is a problem since the exact datum g is not known, but only a *noisy* datum $g_\delta \in \mathcal{K}$ is given, where $\|g - g_\delta\|_{\mathcal{K}} \leq \delta$. According to Tikhonov regularization (Tikhonov and Arsenin, 1977) a possible way to find a solution depending continuously on the data is to replace Problem (6) with the following convex problem

$$\min_{f \in \mathcal{H}} \{\|Af - g_\delta\|_{\mathcal{K}}^2 + \lambda \|f\|_{\mathcal{H}}^2\}, \tag{7}$$

where $\lambda > 0$ and the unique minimizer is given by

$$f_\delta^\lambda = (A^*A + \lambda I)^{-1} A^* g_\delta. \tag{8}$$

A crucial issue is the choice of the regularization parameter λ as a function of the noise. A basic requirement is that the *reconstruction error*

$$\left\| f_\delta^\lambda - f^\dagger \right\|_{\mathcal{H}}$$

is small. In particular, λ must be selected, as a function of the noise level δ and the data g_δ, in such a way that the regularized solution $f_\delta^{\lambda(\delta, g_\delta)}$ converges to the generalized solution, that is,

$$\lim_{\delta \to 0} \left\| f_\delta^{\lambda(\delta, g_\delta)} - f^\dagger \right\|_{\mathcal{H}} = 0, \tag{9}$$

for any g such that $f^\dagger$ exists.

Sometimes, another measure of the error, namely the *residual*, is considered according to the following definition

$$\left\| A f_\delta^\lambda - P g \right\|_{\mathcal{K}} = \left\| A f_\delta^\lambda - A f^\dagger \right\|_{\mathcal{K}}, \tag{10}$$

which will be important in our analysis of learning. Comparing (9) and (10), it is clear that while studying the convergence of the residual we do not have to assume that the generalized solution exists.

We conclude this section with some remarks. The above formalism can be easily extended to the case of a noisy operator $A_\delta : \mathcal{H} \to \mathcal{K}$ where

$$\|A - A_\delta\| \leq \delta,$$

and $\|\cdot\|$ is the operator norm (Tikhonov et al., 1995). Moreover, the similarity between regularized least squares and Tikhonov regularization is apparent comparing Problems (1) and (7). However while trying to formalize this analogy several difficulties emerge. First, to treat the problem of learning in the setting of ill-posed inverse problems we have to define a direct problem by means of a suitable operator A between two Hilbert spaces $\mathcal{H}$ and $\mathcal{K}$. Second, the nature of the noise δ in the context of statistical learning is not clear . Finally, we have to clarify the relation between consistency, expressed by (2), and the convergence considered in (9). In the following sections we will show a possible way to tackle these problems.

4. Learning as an Inverse Problem

We now show how the problem of learning can be rephrased in a framework close to the one presented in the previous section.
We let $L^2(X, \nu)$ be the Hilbert space of square integrable functions on X with respect to the marginal measure ν and we define the operator $A : \mathcal{H} \to L^2(X, \nu)$ as

$$(Af)(x) = \langle f, K_x \rangle_{\mathcal{H}},$$

where K is the reproducing kernel of $\mathcal{H}$. The fact that K is bounded, see (3), ensures that A is a bounded linear operator. Two comments are in order. First, from (4) we see that the action of A on an element f is simply

$$(Af)(x) = f(x) \quad \forall x \in x,\ f \in \mathcal{H},$$

that is, A is the canonical inclusion of $\mathcal{H}$ into $L^2(X, \nu)$. However it is important to note that A changes the norm since $\|f\|_{\mathcal{H}}$ is different to $\|f\|_{L^2(X,\nu)}$. Second, to avoid pathologies connected with subsets of zero measure, we assume that ν is not degenerate[1]. This condition and the fact that K is continuous ensure that A is injective (see Appendix for the proof).

It is known that, considering the quadratic loss function, the expected risk can be written as

$$\begin{aligned} I[f] &= \int_X (f(x) - g(x))^2 \, d\nu(x) + \int_{X \times Y} (y - g(x))^2 \, d\rho(x, y) \\ &= \|f - g\|^2_{L^2(X,\nu)} + I[g], \end{aligned}$$

where g is the regression function (Cucker and Smale, 2002) and f is any function in $L^2(X, \nu)$. If f belongs to the hypothesis space $\mathcal{H}$, the definition of the operator A allows to write

$$I[f] = \|Af - g\|^2_{L^2(X,\nu)} + I[g]. \tag{11}$$

Moreover, if P is the projection on the closure of the range of A, that is, the closure of $\mathcal{H}$ into $L^2(X, \nu)$, then the definition of projection gives

$$\inf_{f \in \mathcal{H}} \|Af - g\|^2_{L^2(X,\nu)} = \|g - Pg\|^2_{L^2(X,\nu)} .$$

Given $f \in \mathcal{H}$, clearly $PAf = Af$, so that

$$I[f] - \inf_{f \in \mathcal{H}} I[f] = \|Af - g\|^2_{L^2(X,\nu)} - \|g - Pg\|^2_{L^2(X,\nu)} = \|Af - Pg\|^2_{L^2(X,\nu)} , \tag{12}$$

which is the square of the residual of f.

Now, comparing (11) and (6), it is clear that the expected risk admits a minimizer $f_{\mathcal{H}}$ on the hypothesis space $\mathcal{H}$ if and only if $f_{\mathcal{H}}$ is precisely the generalized solution $f^\dagger$ of the linear inverse problem

$$Af = g. \tag{13}$$

The fact that $f_{\mathcal{H}}$ is the minimal norm solution of the least squares problem is ensured by the fact that A is injective.

Let now $\mathbf{z} = (\mathbf{x}, \mathbf{y}) = ((x_1, y_1), \ldots, (x_\ell, y_\ell))$ be the training set. The above arguments can be repeated replacing the set X with the finite set $\{x_1, \ldots, x_\ell\}$. We now get a discretized version of A by defining the *sampling operator* (Smale and Zhou, 2004a)

$$A_{\mathbf{x}} : \mathcal{H} \to \mathbf{E}^\ell \qquad (A_{\mathbf{x}} f)_i = \langle f, K_{x_i} \rangle_{\mathcal{H}} = f(x_i),$$

where $\mathbf{E}^\ell = \mathbb{R}^\ell$ is the finite dimensional euclidean space endowed with the scalar product

$$\left\langle \mathbf{w}, \mathbf{w}' \right\rangle_{\mathbf{E}^\ell} = \frac{1}{\ell} \sum_{i=1}^{\ell} w_i w'_i.$$

1. This means that all the open non-void subsets of X have strictly positive measure.

It is straightforward to check that

$$\frac{1}{\ell}\sum_{i=1}^{\ell}(f(x_i)-y_i)^2 = \|A_{\mathbf{x}}f-\mathbf{y}\|^2_{\mathbf{E}^\ell},$$

so that the estimator ${f_{\mathbf{z}}}^\lambda$ given by the regularized least squares algorithm, see Problem (1), is the Tikhonov regularized solution of the discrete problem

$$A_{\mathbf{x}}f = \mathbf{y}. \tag{14}$$

At this point it is useful to remark the following two facts. First, in learning from examples rather than finding an approximation to the solution of the noisy (discrete) Problem (14), we want to find a stable approximation to the solution of the exact (continuous) Problem (13) (compare with Kurkova (2004)). Second, in statistical learning theory, the key quantity is the residual of the solution, which is a weaker measure than the reconstruction error, usually studied in the inverse problem setting. In particular, consistency requires a weaker kind of convergence than the one usually studied in the context of inverse problems . Moreover we observe that in the context of learning the existence of the minimizer $f_{\mathcal{H}}$, that is, of the generalized solution, is no longer needed to define good asymptotic behavior.
After this preliminary considerations in the next section we further develop our analysis stating the main results of this paper.

5. Regularization, Stochastic Noise and Consistency

Table 1 compares the classical framework of inverse problems (see Section 3) with the formulation of learning proposed above. We note some differences. First, the noisy data space $\mathbf{E}^\ell$ is different from the exact data space $L^2(X,\nu)$ so that A and $A_{\mathbf{x}}$ belong to different spaces, as well as g and $\mathbf{y}$. A measure of the difference between $A_{\mathbf{x}}$ and A, and between g and $\mathbf{y}$ is then required. Second, both $A_{\mathbf{x}}$ and $\mathbf{y}$ are random variables and we need to relate the noise δ to the number ℓ of examples in the training set $\mathbf{z}$. Given the above premise our derivation of consistency results is developed in two steps: we first study the residual of the solution by means of a measure of the noise due to discretization, then we show a possible way to give a probabilistic evaluation of the noise previously introduced.

5.1 Bounding the Residual of Tikhonov Solution

In this section we study the dependence of the minimizer of Tikhonov functional on the operator A and the data g. We indicate with $\mathcal{L}(\mathcal{H})$ and $\mathcal{L}(\mathcal{H},\mathcal{K})$ the Banach space of bounded linear operators from $\mathcal{H}$ into $\mathcal{H}$ and from $\mathcal{H}$ into $\mathcal{K}$ respectively. We denote with $\|\cdot\|_{\mathcal{L}(\mathcal{H})}$ the uniform norm in $\mathcal{L}(\mathcal{H})$ and, if $A \in \mathcal{L}(\mathcal{H},\mathcal{K})$, we recall that A^* is the adjoint

Inverse problem	Learning theory
input space $\mathcal{H}$	hypothesis space RKHS $\mathcal{H}$
data space $\mathcal{K}$	target space $L^2(X,\nu)$
norm in $\mathcal{K}$ $\|f\|_{\mathcal{K}}$	expected risk $I[f]$
exact operator A	inclusion of $\mathcal{H}$ into $L^2(X,\nu)$
exact datum g	regression function $g(x)=\int_Y y\,d\rho(y\|x)$
generalized solution $f^\dagger$	ideal solution $f_{\mathcal{H}}$
reconstruction error $\left\|f-f^\dagger\right\|_{\mathcal{H}}$	residual $\|Af-Af_{\mathcal{H}}\|^2_{L^2(X,\nu)}=I[f]-I[f_{\mathcal{H}}]$
noisy data space $\mathcal{K}$	$\mathbf{E}^\ell$
noisy data $g_\delta\in\mathcal{K}$	$\mathbf{y}\in\mathbf{E}^\ell$
noisy operator $A_\delta:\mathcal{H}\to\mathcal{K}$	sampling operator $A_{\mathbf{x}}:\mathcal{H}\to\mathbf{E}^\ell$
Tikhonov regularization	Regularized least squares algorithm

Table 1: The above table summarizes the relation between the theory of inverse problem and the theory of learning from examples.

operator. The Tikhonov solutions of Problems (13) and (14) can be written as

$$\begin{aligned} f^\lambda &= (A^*A+\lambda I)^{-1}A^*g, \\ f_{\mathbf{z}}^\lambda &= (A_{\mathbf{x}}^*A_{\mathbf{x}}+\lambda I)^{-1}A_{\mathbf{x}}^*\mathbf{y}. \end{aligned}$$

The above equations show that $f_{\mathbf{z}}^\lambda$ and f^λ depend only on $A_{\mathbf{x}}^*A_{\mathbf{x}}$ and A^*A, which are operators from $\mathcal{H}$ into $\mathcal{H}$, and on $A_{\mathbf{x}}^*\mathbf{y}$ and A^*g, which are elements of $\mathcal{H}$. This observation suggests that noise levels could be evaluated controlling $\|A_{\mathbf{x}}^*A_{\mathbf{x}}-A^*A\|_{\mathcal{L}(\mathcal{H})}$ and $\|A_{\mathbf{x}}^*\mathbf{y}-A^*g\|_{\mathcal{H}}$.

To this purpose, for every $\delta=(\delta_1,\delta_2)\in\mathbb{R}_+^2$, we define the collection of training sets

$$\mathcal{U}_\delta := \{\mathbf{z}\in(X\times Y)^\ell \mid \|A_{\mathbf{x}}^*\mathbf{y}-A^*g\|_{\mathcal{H}}\le\delta_1,\ \|A_{\mathbf{x}}^*A_{\mathbf{x}}-A^*A\|_{\mathcal{L}(\mathcal{H})}\le\delta_2\}.$$

Recalling that P is the projection on the closure of the range of A and $Y\subset[-M,M]$, we are ready to state the following theorem.

Theorem 1 *Given $\lambda>0$, the following inequality holds*

$$\left|\left\|Af_{\mathbf{z}}^\lambda-Pg\right\|_{L^2(X,\nu)}-\left\|Af^\lambda-Pg\right\|_{L^2(X,\nu)}\right|\le\frac{\delta_1}{2\sqrt{\lambda}}+\frac{M\delta_2}{4\lambda}$$

for any training set $\mathbf{z}\in\mathcal{U}_\delta$.

We postpone the proof to Appendix A and briefly comment on the above result. The first term in the l.h.s. of the inequality is exactly the residual of the regularized solution whereas the second term represents the approximation error, which does not depend on the sample. Our bound quantifies the difference between the residual of the regularized solutions of the exact and noisy problems in terms of the noise level $\delta = (\delta_1, \delta_2)$. As mentioned before this is exactly the kind of result needed to derive consistency. Our result bounds the residual both from above and below and is obtained introducing the collection $\mathcal{U}_\delta$ of training sets compatible with a certain noise level δ. It is left to quantify the noise level corresponding to a training set of cardinality ℓ. This will be achieved in a probabilistic setting in the next section, where we also discuss a standard result on the approximation error.

5.2 Stochastic Evaluation of the Noise and Approximation Term

In this section we give a probabilistic evaluation of the noise levels δ_1 and δ_2 and we analyze the behavior of the term $\left\|Af^\lambda - Pg\right\|_{L^2(X,\nu)}$. In the context of inverse problems a noise estimate is a part of the available data whereas in learning problems we need a probabilistic analysis.

Theorem 2 *Let* $0 < \eta < 1$. *Then*

$$\mathbf{P}\left[\|A^*g - A_\mathbf{x}{}^*\mathbf{y}\|_\mathcal{H} \leq \delta_1(\ell,\eta),\ \|A^*A - A_\mathbf{x}{}^*A_\mathbf{x}\|_{\mathcal{L}(\mathcal{H})} \leq \delta_2(\ell,\eta)\right] \geq 1-\eta$$

where $\kappa = \sup_{x\in X}\sqrt{K(x,x)}$,

$$\delta_1(\ell,\eta) = \frac{M\kappa}{2}\, g\left(\frac{8}{\ell}\log\frac{4}{\eta}\right) \qquad\qquad \delta_2(\ell,\eta) = \frac{\kappa^2}{2}\, g\left(\frac{8}{\ell}\log\frac{4}{\eta}\right)$$

with $g(t) = \frac{1}{2}(t + \sqrt{t^2+4t}) = \sqrt{t} + o(\sqrt{t})$.

We refer again to Appendix A for the complete proof and add few comments. The one proposed is just one of the possible probabilistic tools that can be used to study the above random variables. For example union bounds and Hoeffding's inequality can be used introducing a suitable notion of covering numbers on $X \times Y$.

An interesting aspect in our approach is that the collection of training sets compatible with a certain noise level δ does not depend on the regularization parameter λ. This last fact allows us to consider indifferently data independent parameter choices $\lambda = \lambda(\ell)$ as well as data dependent choices $\lambda = \lambda(\ell, \mathbf{z})$. Since through data dependent parameter choices the regularization parameter becomes a function of the given sample $\lambda(\ell, \mathbf{z})$, in general some further analysis is needed to ensure that the bounds hold uniformly w.r.t. λ.

We now consider the term $\left\|Af^\lambda - Pg\right\|_{L^2(X,\nu)}$ which does not depend on the training set $\mathbf{z}$ and plays the role of an approximation error (Smale and Zhou, 2003, Niyogi and Girosi, 1999). The following is a classical result in the context of inverse problems (see for example Engl et al. (1996)).

Proposition 3 *Let f^λ the Tikhonov regularized solution of the problem $Af = g$, then the following convergence holds*

$$\lim_{\lambda \to 0^+} \left\| Af^\lambda - Pg \right\|_{L^2(X,\nu)} = 0.$$

We report the proof in Appendix A for completeness. The above proposition ensures that, independently of the probability measure ρ, the approximation term goes to zero as $\lambda \to 0$. Unfortunately it is well known, both in learning theory (see for example Devroye et al. (1996), Vapnik (1998), Smale and Zhou (2003), Steinwart (2004)) and inverse problems theory (Groetsch, 1984), that such a convergence can be arbitrarily slow and convergence rates can be obtained only under some assumptions either on the regression function g or on the probability measure ρ (Smale and Zhou, 2003).

We are now in the position to derive the consistency result that we present in the following section.

5.3 Consistency and Regularization Parameter Choice

Combining Theorems 1 and 2 with Proposition 3, we easily derive the following result (see Appendix A for the proof).

Theorem 4 *Given $0 < \eta < 1$, then*

$$\begin{aligned} I[f_{\mathbf{z}}{}^\lambda] - \inf_{f \in \mathcal{H}} I[f] &\leq \left[\left(\frac{M\kappa}{2\sqrt{\lambda}} + \frac{M\kappa^2}{4\lambda} \right) g\left(\frac{8}{\ell} \log \frac{4}{\eta} \right) + \left\| Af^\lambda - Pg \right\|_{L^2(X,\nu)} \right]^2 \qquad (15) \\ &= \left[M\kappa^2 \sqrt{\frac{\log \frac{4}{\eta}}{2\lambda^2 \ell}} + \left\| Af^\lambda - Pg \right\|_{L^2(X,\nu)} + o\left(\sqrt{\frac{1}{\lambda^2 \ell} \log \frac{4}{\eta}} \right) \right]^2 \end{aligned}$$

with probability greater that $1 - \eta$. Moreover, if $\lambda = O(l^{-b})$ with $0 < b < \frac{1}{2}$, then

$$\lim_{\ell \to +\infty} \mathbf{P} \left[I[f_{\mathbf{z}}{}^{\lambda(\ell,\mathbf{z})}] - \inf_{f \in \mathcal{H}} I[f] \geq \epsilon \right] = 0.$$

for every $\epsilon > 0$.

As mentioned before, the second term in the r.h.s. of the above inequality is an approximation error and vanishes as λ goes to zero. The first term in the r.h.s. of Inequality (15) plays the role of sample error. It is interesting to note that since $\delta = \delta(\ell)$ we have an equivalence between the limit $\ell \to \infty$, usually studied in learning theory, and the limit $\delta \to 0$, usually considered for inverse problems. Our result presents the formal connection between the consistency approach considered in learning theory, and the regularization-stability convergence property used in ill-posed inverse problems. Although it is known that connections already exist, as far as we know, this is the first full connection between the two areas, for the specific case of square loss.

We now briefly compare our result with previous work on the consistency of the regularized least squares algorithm. Recently, several works studied the consistency property and the related convergence rate of learning algorithms inspired by Tikhonov regularization. For the classification setting, a general discussion considering a large class of loss functions can be found in Steinwart (2004), whereas some refined results for specific loss functions can be found in Chen et al. (2004) and Scovel and Steinwart (2003). For regression problems in Bousquet and Elisseeff (2002) a large class of loss functions is considered and a bound of the form

$$I[{f_{\mathbf{z}}}^{\lambda}] - I_{\mathbf{z}}[{f_{\mathbf{z}}}^{\lambda}] \leq O\left(\frac{1}{\sqrt{\ell}\lambda}\right)$$

is proved, where $I_{\mathbf{z}}[{f_{\mathbf{z}}}^{\lambda}]$ is the empirical error [2]. Such a bound allows to prove consistency using the error decomposition in Steinwart (2004). The square loss was considered in Zhang (2003) where, using leave-one out techniques, the following bound in expectation was proved

$$E_{\mathbf{z}}(I[{f_{\mathbf{z}}}^{\lambda}]) \leq O\left(\frac{1}{\ell\lambda}\right).$$

In De Vito et al. (2004a) a bound of the form

$$I[{f_{\mathbf{z}}}^{\lambda}] - \inf_{f\in\mathcal{H}} I[f] \leq \left(S(\lambda,\ell) + \left\|Af^{\lambda} - Pg\right\|_{L^2(X,\nu)}\right)^2$$

is derived using similar techniques to those presented in this paper. In that case $S(\lambda,\ell) \leq O\left(\frac{1}{\sqrt{\ell}\lambda^{\frac{3}{2}}}\right)$ and we see that Theorem 4 gives $S(\lambda,\ell) \leq O\left(\frac{1}{\sqrt{\ell}\lambda}\right)$. Finally our results were recently improved in Smale and Zhou (2004b), where, using again techniques similar to those presented here, a bound of the form $S(\lambda,\ell) \leq O\left(\frac{1}{\sqrt{\ell}\lambda}\right) + O\left(\frac{1}{\ell\lambda^{\frac{3}{2}}}\right)$ is obtained. It is worth noting that in general working on the square root of the error leads to better overall results.

6. Conclusions

In this paper we analyse the connection between the theory of statistical learning and the theory of ill-posed problems. More precisely, we show that, considering the quadratic loss function, the problem of finding the best solution $f_{\mathcal{H}}$ for a given hypothesis space $\mathcal{H}$ is a linear inverse problem and that the regularized least squares algorithm is the Tikhonov regularization of the discretized version of the above inverse problem. As a consequence, the consistency of the algorithm is traced back to the well known convergence property of the

2. We recall that the empirical error is defined as $I_{\mathbf{z}}[f] = \frac{1}{\ell}\sum_{i=1}^{\ell} V(f(x_i), y_i)$.

Tikhonov regularization. A probabilistic estimate of the noise is given based on a elegant concentration inequality in Hilbert spaces due to Yurinsky (1995).

An open problem is to extend the above results to arbitrary loss functions, like the hinge loss (Vapnik, 1998). For other choices of loss functions the problem of finding the best solution gives rise to a non linear ill-posed problem and the theory for this kind of problems is much less developed than the corresponding theory for linear problems. Moreover, since, in general, the expected risk $I[f]$ for arbitrary loss function does not define a metric, it is not so clear the relation between the expected risk and the residual. Some result in this direction can be found in Steinwart (2004) for classification setting. Further problems are the choice of the regularization parameter, for example by means of the generalized Morozov principle (Engl et al., 1996) and the extension of our analysis to a wider class of regularization algorithms.

Acknowledgments

We would like to thank M.Bertero, C. De Mol, M. Piana, T. Poggio, S. Smale, G. Talenti and A. Verri for useful discussions and suggestions. This research has been partially funded by the INFM Project MAIA, the FIRB Project ASTAA and the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778.

Appendix A. Proofs of the Main Results

First, we collect some useful properties of the operators A and $A_{\mathbf{x}}$.

Proposition 5 *The operator A is a Hilbert-Schmidt operator from $\mathcal{H}$ into $L^2(X,\nu)$ and*

$$A^*\phi = \int_X \phi(x) K_x \, d\nu(x), \tag{16}$$

$$A^*A = \int_X \langle \cdot, K_x\rangle_{\mathcal{H}} K_x \, d\nu(x), \tag{17}$$

where $\phi \in L^2(X,\nu)$, the first integral converges in norm and the second one in trace norm.

Proof The proof is standard and we report it for completeness.
Since the elements $f \in \mathcal{H}$ are continuous functions defined on a compact set and ν is a probability measure, then $f \in L^2(X,\nu)$, so that A is a linear operator from $\mathcal{H}$ to $L^2(X,\nu)$. Moreover the Cauchy-Schwartz inequality gives

$$|(Af)(x)| = |\langle f, K_x\rangle_{\mathcal{H}}| \le \kappa \, \|f\|_{\mathcal{H}},$$

so that $\|Af\|_{L^2(X,\nu)} \le \kappa \|f\|_{\mathcal{H}}$ and A is bounded.
We now show that A is injective. Let $f \in \mathcal{H}$ and $W = \{x \in X \mid f(x) \neq 0\}$. Assume

$Af = 0$, then W is a open set, since f is continuous, and W has null measure, since $(Af)(x) = f(x) = 0$ for ν-almost all $x \in X$. The assumption that ν is not degenerate ensures W be the empty set and, hence, $f(x) = 0$ for all $x \in X$, that is, $f = 0$.
We now prove (16). We first recall the map

$$X \ni x \mapsto K_x \in \mathcal{H}$$

is continuous since $\|K_t - K_x\|_{\mathcal{H}}^2 = K(t,t) + K(x,x) - 2K(x,t)$ for all $x, t \in X$, and K is a continuous function. Hence, given $\phi \in L^2(X,\nu)$, the map $x \mapsto \phi K_x$ is measurable from X to $\mathcal{H}$. Moreover, for all $x \in X$,

$$\|\phi(x)K_x\|_{\mathcal{H}} = |\phi(x)|\sqrt{K(x,x)} \leq |\phi(x)|\kappa.$$

Since ν is finite, ϕ is in $L^1(X,\nu)$ and, hence, ϕK_x is integrable, as a vector valued map. Finally, for all $f \in \mathcal{H}$,

$$\int_X \phi(x) \langle K_x, f\rangle_{\mathcal{H}} \, d\nu(x) = \langle \phi, Af\rangle_{L^2(X,\nu)} = \langle A^*\phi, f\rangle_{\mathcal{H}},$$

so, by uniqueness of the integral, Equation (16) holds.
Equations (17) is a consequence of Equation (16) and the fact that the integral commutes with the scalar product.
We now prove that A is a Hilbert-Schmidt operator. Let $(e_n)_{n\in\mathbb{N}}$ be a Hilbert basis of $\mathcal{H}$. Since A^*A is a positive operator and $|\langle K_x, e_n\rangle_{\mathcal{H}}|^2$ is a positive function, by monotone convergence theorem, we have that

$$\begin{aligned}
\operatorname{Tr}(A^*A) &= \sum_n \int_X |\langle e_n, K_x\rangle_{\mathcal{H}}|^2 \, d\nu(x) \\
&= \int_X \sum_n |\langle e_n, K_x\rangle_{\mathcal{H}}|^2 \, d\nu(x) \\
&= \int_X \langle K_x, K_x\rangle_{\mathcal{H}} \, d\nu(x) \\
&= \int_X K(x,x) \, d\nu(x) < \kappa^2
\end{aligned}$$

and the thesis follows. ■

Proposition 6 *The sampling operator* $A_{\mathbf{x}} : \mathcal{H} \to \mathbf{E}^\ell$ *is a Hilbert-Schmidt operator and*

$$A_{\mathbf{x}}^*\mathbf{y} = \frac{1}{\ell}\sum_{i=1}^{\ell} y_i K_{x_i} \tag{18}$$

$$A_{\mathbf{x}}^* A_{\mathbf{x}} = \frac{1}{\ell}\sum_{i=1}^{\ell} \langle \cdot, K_{x_i}\rangle_{\mathcal{H}} K_{x_i}. \tag{19}$$

Proof The content of the proposition is a restatement of Proposition 5 and the fact that the integrals reduce to sums. ■

We now give the proofs of our main results. For sake of completeness we preliminary report a standard proof on the convergence of the approximation error.

Proof [of Proposition 3] Consider the polar decomposition $A = U|A|$ of A (see, for example, Lang (1993)), where $|A|^2 = A^*A$ is a positive operator on $\mathcal{H}$ and U is a partial isometry such that the projector P on the range of A is $P = UU^*$. Let $dE(t)$ be the spectral measure of $|A|$. Recalling that

$$f^\lambda = (A^*A + \lambda)^{-1}A^*g = (|A|^2 + \lambda)^{-1}|A|U^*g$$

the spectral theorem gives

$$\begin{aligned}
\left\|Af^\lambda - Pg\right\|_{\mathcal{K}}^2 &= \left\|U|A|(|A|^2+\lambda)^{-1}|A|U^*g - UU^*g\right\|_{\mathcal{K}}^2 = \\
&= \left\|\left(|A|^2\left(|A|^2+\lambda\right)^{-1} - 1\right)U^*g\right\|_{\mathcal{H}}^2 = \\
&= \int_0^{\||A|\|}\left(\frac{t^2}{t^2+\lambda} - 1\right)^2 d\langle E(t)U^*g, U^*g\rangle_{\mathcal{H}}.
\end{aligned}$$

Let $r_\lambda(t) = \frac{t^2}{t^2+\lambda} - 1 = -\frac{\lambda}{t^2+\lambda}$, then

$$|r_\lambda(t)| \le 1 \quad \text{and} \quad \lim_{\lambda\to 0^+} r_\lambda(t) = 0 \quad \forall t > 0,$$

so that the dominated convergence theorem gives that

$$\lim_{\lambda\to 0^+}\left\|Af^\lambda - Pg\right\|_{\mathcal{K}}^2 = 0.$$

■

We now prove the bound on the residual for the Tikhonov regularization.

Proof [of Theorem 1] The idea of the proof is to note that, by triangular inequality, we can write

$$\left|\left\|Af_{\mathbf{z}}^\lambda - Pg\right\|_{L^2(X,\nu)} - \left\|Af^\lambda - Pg\right\|_{L^2(X,\nu)}\right| \le \left\|Af_{\mathbf{z}}^\lambda - Af^\lambda\right\|_{L^2(X,\nu)} \tag{20}$$

so that we can focus on the difference between the discrete and continuous solutions. By a simple algebraic computation we have that

$$\begin{aligned}
f_{\mathbf{z}}^\lambda &- f^\lambda = (A_{\mathbf{x}}^*A_{\mathbf{x}} + \lambda I)^{-1}A_{\mathbf{x}}^*\mathbf{y} - (A^*A+\lambda I)^{-1}A^*g = \\
&= [(A_{\mathbf{x}}^*A_{\mathbf{x}} + \lambda I)^{-1} - (A^*A+\lambda I)^{-1}]A_{\mathbf{x}}^*\mathbf{y} + (A^*A+\lambda I)^{-1}(A_{\mathbf{x}}^*\mathbf{y} - A^*g) = \qquad (21)\\
&= (A^*A+\lambda I)^{-1}(A^*A - A_{\mathbf{x}}^*A_{\mathbf{x}})(A_{\mathbf{x}}^*A_{\mathbf{x}} + \lambda I)^{-1}A_{\mathbf{x}}^*\mathbf{y} + (A^*A+\lambda I)^{-1}(A_{\mathbf{x}}^*\mathbf{y} - A^*g).
\end{aligned}$$

and we see that the relevant quantities for the definition of the noise appear.
We claim that

$$\left\|A(A^*A+\lambda I)^{-1}\right\|_{\mathcal{L}(\mathcal{H})} = \frac{1}{2\sqrt{\lambda}} \tag{22}$$

$$\left\|(A_{\mathbf{x}}^*A_{\mathbf{x}}+\lambda I)^{-1}A_{\mathbf{x}}^*\right\|_{\mathcal{L}(\mathcal{H})} = \frac{1}{2\sqrt{\lambda}}. \tag{23}$$

Indeed, let $A = U|A|$ be the polar decomposition of A. The spectral theorem implies that

$$\begin{aligned}\|A(A^*A+\lambda I)^{-1}\|_{\mathcal{L}(\mathcal{H})} &= \|U|A|(|A|^2+\lambda I)^{-1}\|_{\mathcal{L}(\mathcal{H})} = \||A|(|A|^2+\lambda I)^{-1}\|_{\mathcal{L}(\mathcal{H})} \\ &= \sup_{t\in[0,\||A|\|]} \frac{t}{t^2+\lambda}.\end{aligned}$$

A direct computation of the derivative shows that the maximum of $\frac{t}{t^2+\lambda}$ is $\frac{1}{2\sqrt{\lambda}}$ and (22) is proved. Formula (23) follows replacing A with $A_{\mathbf{x}}$.

Last step is to plug Equation (21) into (20) and use Cauchy-Schwartz inequality. Since $\|\mathbf{y}\|_{\mathbf{E}^\ell} \leq M$, (22) and (23) give

$$\left|\|Af_{\mathbf{z}}^\lambda - Pg\|_{L^2} - \|Af^\lambda - Pg\|_{L^2}\right| \leq \frac{M}{4\lambda}\|A^*A - A_{\mathbf{x}}^*A_{\mathbf{x}}\|_{\mathcal{L}(\mathcal{H})} + \frac{1}{2\sqrt{\lambda}}\left\|A_{\mathbf{x}}^*\mathbf{y} - A^*g\right\|_{\mathcal{H}}.$$

so that the theorem is proved. ■

To prove our estimate of the noise we need the following probabilistic inequality due to Yurinsky (1995).

Lemma 7 *Let Z be a probability space and ξ be a random variable on X taking value in a real separable Hilbert space $\mathcal{H}$. Assume that the expectation value $v^* = \mathbb{E}[\xi]$ exists and there are two positive constants H and σ such that*

$$\begin{aligned}\|\xi(z) - v^*\|_{\mathcal{H}} &\leq H \quad \text{a.}s \\ \mathbb{E}[\|\xi - v^*\|_{\mathcal{H}}^2] &\leq \sigma^2.\end{aligned}$$

If z_i are drawn i.i.d. from Z, then, with probability greater than $1-\eta$,

$$\left\|\frac{1}{\ell}\sum_{i=1}^{\ell}\xi(z_i) - v^*\right\| \leq \frac{\sigma^2}{H}\, g\!\left(\frac{2H^2}{\ell\sigma^2}\log\frac{2}{\eta}\right) = \delta(\ell,\eta) \tag{24}$$

where $g(t) = \frac{1}{2}(t + \sqrt{t^2+4t})$. In particular

$$\delta(\ell,\eta) = \sigma\sqrt{\frac{2}{\ell}\log\frac{2}{\eta}} + o\left(\sqrt{\frac{1}{\ell}\log\frac{2}{\eta}}\right)$$

Proof It is just a testament to Th. 3.3.4 of Yurinsky (1995), see also Steinwart (2003). Consider the set of independent random variables with zero mean $\xi_i = \xi(z_i) - v^*$ defined on the probability space Z^ℓ. Since, ξ_i are identically distributed, for all $m \geq 2$ it holds

$$\sum_{i=1}^{\ell} \mathbb{E}[\|\xi_i\|_{\mathcal{H}}^m] \leq \frac{1}{2} m! B^2 H^{m-2},$$

with the choice $B^2 = \ell\sigma^2$. So Th. 3.3.4 of Yurinsky (1995) can be applied and it ensures

$$\mathbf{P}\left[\frac{1}{\ell}\left\|\sum_{i=1}^{\ell}(\xi(z_i) - v^*)\right\| \geq \frac{xB}{\ell}\right] \leq 2\exp\left(-\frac{x^2}{2(1+xHB^{-1})}\right).$$

for all $x \geq 0$. Letting $\delta = \frac{xB}{\ell}$, we get the equation

$$\frac{1}{2}(\frac{\ell\delta}{B})^2 \frac{1}{1+\ell\delta H B^{-2}} = \frac{\ell\delta^2\sigma^{-2}}{2(1+\delta H\sigma^{-2})} = \log\frac{2}{\eta},$$

since $B^2 = \ell\sigma^2$. Defining $t = \delta H \sigma^{-2}$

$$\frac{\ell\sigma^2}{2H^2}\frac{t^2}{1+t} = \log\frac{2}{\eta}.$$

The thesis follows, observing that g is the inverse of $\frac{t^2}{1+t}$ and that $g(t) = \sqrt{t} + o(\sqrt{t})$. ■

We notice that, if ξ is bounded by L almost surely, then v^* exists and we can choose $H = 2L$ and $\sigma = L$ so that

$$\delta(\ell,\eta) = \frac{L}{2}\, g\left(\frac{8}{\ell}\log\frac{2}{\eta}\right). \tag{25}$$

In Smale and Y. (2004) a better estimate is given, replacing the function $\frac{t^2}{1+t}$ with $t\log(1+t)$, anyway the asymptotic rate is the same.

The proof of Theorem 2 is now straightforward.

Proof [Theorem 2] The proof is a simple consequence of estimate (25) applied to the random variables

$$\begin{aligned} \xi_1(x,y) &= yK_x \\ \xi_2(x,y) &= \langle \cdot, K_x\rangle_{\mathcal{H}} K_x = K_x \otimes K_x \end{aligned}$$

where

1. ξ_1 takes value in $\mathcal{H}$, $L_1 = \kappa M$ and $v_1^* = A^* g$, see (16), (18);

2. ξ_2 takes vales in the Hilbert space of Hilbert-Schmidt operators, which can be identified with $\mathcal{H} \otimes \mathcal{H}$, $L_2 = \kappa^2$ and $v_2^* = T$, see (17), (19).

Replacing η with $\eta/2$, (25) gives

$$\begin{aligned} \|A^*g - A_{\mathbf{x}}{}^*\mathbf{y}\|_{\mathcal{H}} \leq \delta_1(\ell,\eta) &= \frac{M\kappa}{2}\, g\left(\frac{8}{\ell}\log\frac{4}{\eta}\right) \\ \|A^*A - A_{\mathbf{x}}{}^*A_{\mathbf{x}}\|_{\mathcal{L}(\mathcal{H})} \leq \delta_2(\ell,\eta) &= \frac{\kappa^2}{2}\, g\left(\frac{8}{\ell}\log\frac{4}{\eta}\right), \end{aligned}$$

respectively, so that the thesis follows. ■

Finally we combine all the above results to prove the consistency of the regularized least squares algorithm.

Proof [Theorem 4] Theorem 1 gives

$$\|Af_{\mathbf{z}}^{\lambda} - Pg\|_{L^2(X,\nu)} \leq \left(\frac{1}{2\sqrt{\lambda}}\delta_1 + \frac{M}{4\lambda}\delta_2\right) + \|Af^{\lambda} - Pg\|_{L^2(X,\nu)}.$$

Equation (12) and the estimates for the noise levels δ_1 and δ_2 given by Theorem 2 ensure that

$$\sqrt{I[f_{\mathbf{z}}{}^{\lambda}] - \inf_{f\in\mathcal{H}} I[f]} \leq \left(\frac{M\kappa}{2\sqrt{\lambda}} + \frac{M\kappa^2}{4\lambda}\right) g\left(\frac{8}{\ell}\log\frac{4}{\eta}\right) + \left\|Af^{\lambda} - Pg\right\|_{L^2(X,\nu)}$$

and (15) simply follows taking the square of the above inequality. Let now $\lambda = 0(\ell^{-b})$ with $0 < b < \frac{1}{2}$, the consistency of the regularised least squares algorithm is proven by inverting the relation between ϵ and η. ■

References

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

N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68:337–404, 1950.

P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities. *Journal of Machine Learning Research*, 3:463–482, 2002.

M. Bertero, C. De Mol, and E. R. Pike. Linear inverse problems with discrete data. I. General formulation and singular system analysis. *Inverse Problems*, 1(4):301–330, 1985.

M. Bertero, C. De Mol, and E. R. Pike. Linear inverse problems with discrete data. II. Stability and regularisation. *Inverse Problems*, 4(3):573–594, 1988.

O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

D. Chen, Q. Wu, Y. Ying, and D. Zhou. Support vector machine soft margin classifiers: Error analysis. *Journal of Machine Learning research*, 5:1143–1175, 2004.

F. Cucker and S. Smale. On the mathematical foundations of learning. *Bull. Amer. Math. Soc. (N.S.)*, 39(1):1–49 (electronic), 2002.

E. De Vito, A. Caponnetto, and L. Rosasco. Model selection for regularized least squares algorithm in learning theory. *to be published in Foundations of Computational Mathematics*, 2004a.

E. De Vito, A. Caponnetto, and L. Rosasco. Discretization Error Analysis for Tikhonov Regularization in Learning Theory. *preprint*, 2004b.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Number 31 in Applications of mathematics. Springer, New York, 1996.

H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of inverse problems*, volume 375 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1996.

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

C. W. Groetsch. *The theory of Tikhonov regularization for Fredholm equations of the first kind*, volume 105 of *Research Notes in Mathematics*. Pitman (Advanced Publishing Program), Boston, MA, 1984.

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

V. Kurkova. Learning from data as an inverse problem. In J. Antoch, editor, *COMPSTAT2004*, pages 1377–1384. Springer-Verlag, 2004.

S. Lang. *Real and Functional Analysis*. Springer, New York, 1993.

P. Mathé and S. V. Pereverzev. Moduli of continuity for operator valued functions. *Numer. Funct. Anal. Optim.*, 23(5-6):623–631, 2002.

S. Mukherjee, P. Niyogi, T. Poggio, and R. Rifkin. Statistical learning: Stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. Technical Report CBCL Paper 223, Massachusetts Institute of Technology, January revision 2004.

S. Mukherjee, R. Rifkin, and T. Poggio. Regression and classification with regularization. *Lectures Notes in Statistics: Nonlinear Estimation and Classification, Proceedings from MSRI Workshop*, 171:107–124, 2002.

P. Niyogi and F. Girosi. Generalization bounds for function approximation from scattered noisy data. *Adv. Comput. Math.*, 10:51–80, 1999.

C.S. Ong and S. Canu. Regularization by early stopping. Technical report, Computer Sciences Laboratory, RSISE, ANU, 2004.

T. Poggio and F. Girosi. A theory of networks for approximation and learning. In C. Lau, editor, *Foundation of Neural Networks*, pages 91–106. IEEE Press, Piscataway, N.J., 1992.

C. Rudin. A different type of convergence for statistical learning algorithms. Technical report, Program in Applied and Computational Mathematics Princeton University, 2004.

B. Schölkopf and A.J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002. URL `http://www.learning-with-kernels.org`.

L. Schwartz. Sous-espaces hilbertiens d'espaces vectoriels topologiques et noyaux associés (noyaux reproduisants). *J. Analyse Math.*, 13:115–256, 1964.

C. Scovel and I. Steinwart. Fast rates support vector machines. *submitted to Annals of Statistics*, 2004.

S. Smale and Yao Y. On-line learning algorithms. Technical report, Toyota Technological Institute, Chicago, 2004.

S. Smale and D. Zhou. Estimating the approximation error in learning theory. *Analysis and Applications*, 1(1):1–25, 2003.

S. Smale and D. Zhou. Shannon sampling and function reconstruction from point values. *Bull. Amer. Math. Soc. (N.S.)*, 41(3):279–305 (electronic), 2004a.

S. Smale and D. Zhou. Shannon sampling II : Connections to learning theory. *preprint*, 2004b.

I. Steinwart. Sparseness of support vector machines. *Journal of Machine Learning Research*, 4:1071–1105, 2003.

I. Steinwart. Consistency of support vector machines and other regularized kernel machines. *accepted on IEEE Transaction on Information Theory*, 2004.

A. N. Tikhonov, A. V. Goncharsky, V. V. Stepanov, and A. G. Yagola. *Numerical methods for the solution of ill-posed problems*, volume 328 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1995. Translated from the 1990 Russian original by R. A. M. Hoksbergen and revised by the authors.

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

V. N. Vapnik. *Statistical learning theory*. Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons Inc., New York, 1998. A Wiley-Interscience Publication.

G. Wahba. *Spline models for observational data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1990.

V. Yurinsky. *Sums and Gaussian vectors*, volume 1617 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1995.

T. Zhang. Leave-one-out bounds for kernel methods. *Neural Computation*, 13:1397–1437, 2003.