Dipartimento di Informatica e
Scienze dell'Informazione

Spectral Methods for Regularization in Learning Theory

Ernesto De Vito, Lorenzo Rosasco, Alessandro Verri

Technical Report DISI-TR-05-18

DISI, Università di Genova
v. Dodecaneso 35, 16146 Genova, Italy
http://www.disi.unige.it/

Spectral Methods for Regularization in Learning Theory

Ernesto De Vito,* Lorenzo Rosasco,† Alessandro Verri‡

January 23, 2006

Abstract

In this paper we show that a large class of regularization methods designed for solving ill-posed inverse problems gives rise to novel learning algorithms. All these algorithms are consistent kernel methods which can be easily implemented. The intuition behind our approach is that, by looking at regularization from a filter function perspective, filtering out undesired components of the target function ensures stability with respect to the random sampling thereby inducing good generalization properties. We present a formal derivation of the methods under study by recalling that learning can be written as the inversion of a linear embedding equation given a stochastic discretization. Consistency as well as finite sample bounds are derived for both regression and classification.

1. Introduction

In the context of learning the term regularization refers to techniques allowing to avoid overfitting. Typically, regularization boils down to a Lagrangian formulation of an appropriate constrained minimization problem - e.g. Tikhonov regularization, ridge regression or regularized least squares. In the context of inverse problems regularization is formally defined and leads to algorithms for determining approximate solutions to ill-posed problems solutions which are *stable* with respect to noise (see for example Tikhonov and Arsenin (1977), Engl et al. (1996), Bertero and Boccacci (1998) and references therein).

In this paper, by restricting the focus on the quadratic loss function and hypothesis spaces which are reproducing kernel Hilbert spaces we follow (De Vito et al., 2005b) and we cast the problem of learning in a functional analytical framework which is ideal to exploit the connection with the theory of inverse problems. We show that a large class of regularization schemes typically used in the context of inverse problems gives rise to consistent kernel methods. We prove finite sample bounds for both regression and classification. We also provide an intuition of the way such algorithms work from a filter function point of view. Since we work with the square loss function, we need to solve a (possibly ill-conditioned) matrix inversion problem. Filtering out the components corresponding to small singular values allows us to stabilize the problem from a numerical point view. In order to understand the filter effect on generalization we have to look at the population case, when the probability underlying the problem is known. In this limit case we have to invert a linear operator and the filter allows us to find a stable solution

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

†DISI, Università di Genova, v. Dodecaneso 35, 16146 Genova, Italy, ROSASCO@DISI.UNIGE.IT

‡DISI, Università di Genova, v. Dodecaneso 35, 16146 Genova, Italy, VERRI@DISI.UNIGE.IT

with respect to perturbations on the problem. The picture is then clear since the sample case can be seen as a perturbation (due to random discretization) of the population case: the true probability measure is replaced by the empirical measure on the sample. Unlike the inverse problem setting, in learning *stability* is meant with respect to perturbations on the problem due to the random sampling (see Rakhlin et al. (2005) and reference therein for different notions of stability).

The remarkable fact of our analysis is that we can treat most of the linear methods for ill-posed inverse problems in a unified framework. We describe a set of simple sufficient conditions allowing an easy proof that algorithms for inverse problems are consistent learning algorithms. As a by-product of this analysis, we find that these algorithms have different properties from both the theoretical and the algorithmic point of view. The price we pay for our generality is that for the two algorithms already studied (see Smale and Zhou (2005), Caponnetto and De Vito (2005) for Tikhonov regularization and Yao et al. (2005) for gradient descent learning) the bounds we find do not match the best available bounds. In a follow-up paper (Bauer et al., 2005) a more technical analysis, based on the same techniques considered here, is given and the best available bounds recovered as special cases.

The idea to exploit regularization algorithms for ill-posed problems in function approximation problem is well known. Indeed, in a deterministic setting (the inputs are fixed and the noise deterministic), interpolation and approximation are standard ill-posed problems (see for example Bertero et al. (1985, 1988) for a review). In the context of statistics the focus was mostly on Tikhonov regularization, also called ridge regression (Hastie et al., 2001) or regularized (penalized) least squares (Wahba, 1990). In this setting the input points are either fixed or sampled and the noise is a random variable. Several results are available (see for example Györfi et al. (1996)) but the probabilistic analysis is usually done in expectation. Some results for general regularization schemes are given in Loubes and Ludena (2004) though for fixed inputs. In machine learning the idea to use regularization goes back to Poggio and Girosi (1992) and the connection between large margin kernel methods such as Support Vector Machines and regularization is well known (see Vapnik (1998), Evgeniou et al. (2000) and reference therein). Again ideas coming from inverse problems regarded mostly the use of Tikhonov regularization and were extended to several error measures other then the quadratic loss function. Concerning this latter loss function a theoretical analysis can be found in Smale and Zhou (2005) and Caponnetto and De Vito (2005). The gradient descent learning algorithm in Yao et al. (2005) can be seen as an instance of Landweber iteration (Engl et al., 1996) and is related to the boosting algorithm, called $L2$ boost in Bühlmann and Yu (2002). For other iterative methods some partial results, which do not take into account the random sampling, are presented in Ong and Canu (2004), where promising experiments on real and simulated data are also presented.

In this paper we build up on the connections between the theory of learning and the theory of inverse problems (De Vito et al., 2005b,a). The interplay between ill-posedeness, stability and generalization is indeed not new to learning (see Poggio and Girosi (1992), Evgeniou et al. (2000), Bousquet and Elisseeff (2002), Mukherjee et al. (2004), Poggio et al. (2004)).

The plan of the paper is the following. In Section 2, after describing the main idea of learning in reproducing kernel Hilbert spaces, we describe the considered class of regularization algorithms from a filter function perspective. In Section 3 we give a more formal and abstract characterization of regularization as well as several examples of algorithms. The main theoretical results are also presented and discussed whereas the proofs can be found in Section 5. In Section

4 we discuss in depth the connection between learning ad inverse problems. Finally, we end with some comments and the main open issues on this subject.

2. Regularization in Reproducing Kernel Hilbert Spaces

We start giving a brief account of learning from examples (see Vapnik (1998), Cucker and Smale (2002b), Evgeniou et al. (2000), Bousquet et al. (2004) and references therein). The focus is on the regression problem and the quadratic loss function though we will recall how some results for classification can be derived. The problem of (supervised) learning can be thought as the problem of finding an unknown input-output relation on the basis of a finite number of input-output instances (the examples). Ideally one would like to find a rule to predict the output once a new input is given, that is to be able to *generalize*. To allow modeling the uncertainty in the learning process the problem is formalized in a probabilistic setting.

The input space X is a closed subset in $\mathbb{R}^d$, the output space is $Y = [-M, M]$ for regression ($Y = \{-1, 1\}$ for classification) and the sample space is simply $Z = X \times Y$. We model the input-output relation endowing Z with a probability measure $\rho(x, y) = \rho(y|x)\rho_X(x)$, where ρ_X is the marginal distribution on X and $\rho(y|x)$ is the conditional distribution of y given x. In this setting what is given is a training set $\mathbf{z} = (\mathbf{x}, \mathbf{y}) = \{(x_1, y_1), \cdots, (x_n, y_n)\}$ drawn i.i.d. according to ρ and the goal is to find an algorithm $\mathbf{z} \to f_{\mathbf{z}}$ such that the function $f_{\mathbf{z}}(x)$ is a good estimate of the output y. The quality of an estimator $f_{\mathbf{z}}$ is assessed by its the expected error

$$\mathcal{E}(f_{\mathbf{z}}) = \int_{X \times Y} (y - f_{\mathbf{z}}(x))^2 d\rho(x, y),$$

which can be interpreted as the average error on all the possible input-output pairs. Clearly we would like to find an estimator with small expected error. The minimizer of the expected error over the space $L^2(X, \rho_X)$ of square integrable functions with respect to ρ_X becomes the regression function

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

Moreover we recall that for $f \in L^2(X, \rho_X)$ we can write

$$\mathcal{E}(f) = \|f - f_\rho\|_\rho^2 + \mathcal{E}(f_\rho) \tag{1}$$

so that we can restate the problem as that of approximating the regression function in the norm $\|\cdot\|_\rho = \|\cdot\|_{L^2(X,\rho_X)}$. Moreover since $f_{\mathbf{z}}$ is a random variable we need some probabilistic analysis and more precisely we are interested into a worst case analysis through finite sample bounds such that

$$\mathrm{P}\left[\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) > \varepsilon\right] \leq \eta(\varepsilon, n) \qquad \forall \varepsilon > 0, \ n \in \mathbb{N}.$$

where $\eta(\varepsilon, n)$ does not depend on ρ and $\lim_{n \to +\infty} \eta(\varepsilon, n) = 0$.
From the so called "no free lunch" theorem (Devroye et al., 1996) is well-known that we cannot derive this kind of results without furtherly restricting the class of possible problems. A usual way to put restrictions on the possible probability measures is assuming f_ρ belonging to some compact set often characterized in terms of some smoothness or approximation properties (see for example the discussion in DeVore et al. (2004)). In this paper we do this relating the problem to the approximation schemes we consider, that is regularization in reproducing kernel Hilbert

spaces. We devote the rest of this section to illustrate the class of approximation schemes we are going to analyze and discuss a fairly natural way to impose condition on the regression function f_ρ.

2.1 Learning the Regression Function via Regularization: Filter Function Perspective

The algorithms we consider look for an estimator in an hypotheses space $\mathcal{H}$ which is a reproducing kernel Hilbert space (RKHS) on the set X (Aronszajn, 1950). This means that $\mathcal{H}$ is a Hilbert space of functions $f : X \to \mathbb{R}$ such that, for all $x \in X$, there is a function $K_x \in \mathcal{H}$ satisfying the following reproducing property

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

where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is the scalar product in $\mathcal{H}$. The RKHS $\mathcal{H}$ is uniquely characterized by its kernel $K : X \times X \to \mathbb{R}$, $K(t, x) = K_x(t)$, which is symmetric and positive definite. For technical reasons, we assume that the kernel is measurable and bounded

$$\sup_{x \in X} \sqrt{K(x, x)} \leq \kappa, \tag{2}$$

so that $\mathcal{H}$ is a subspace of $L^2(X, \rho_X)$ (however, in general, $\mathcal{H}$ is not closed in $L^2(X, \rho_X)$). Moreover we require $\mathcal{H}$ to be dense in $L^2(X, \rho_X)$ so that

$$\inf_{f \in \mathcal{H}} \mathcal{E}(f) = \inf_{f \in L^2(X, \rho_X)} \mathcal{E}(f) = \mathcal{E}(f_\rho).$$

(however, we do not require that $f_\rho \in \mathcal{H}$). This assumption simplifies the exposition and can be relaxed replacing f_ρ with its projection on the closure of $\mathcal{H}$ in $L^2(X, \rho_X)$.

A classic and yet effective algorithm is regularized least-squares algorithm (RLSA). A family of estimators is found solving the regularized least square problem

$$\min_{f \in \mathcal{H}} \{\mathcal{E}_{\mathbf{z}}(f) + \lambda \|f\|^2_{\mathcal{H}}\} \tag{3}$$

where λ is a positive parameter and

$$\mathcal{E}_{\mathbf{z}}(f) = \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 \tag{4}$$

is the empirical error. The final estimator is defined providing the above scheme with a parameter choice $\lambda_n = \lambda(n, \mathbf{z})$ so that $f_{\mathbf{z}} = f_{\mathbf{z}}^{\lambda_n}$. Understanding the way such an algorithm works allows to develop different regularization schemes. A possible interpretation relates the the penalty $\|f\|^2_{\mathcal{H}}$ to the complexity of the solution. Choosing $\lambda > 0$ we restrict the possible solution in a certain ball in the RKHS and the radius of of the ball is related to complexity measure such as covering numbers (Cucker and Smale, 2002b) or Rademacher complexities on such a spaces (Mendelson, 2003). This way of reasoning looks at the RLSA as an approximate implementation of Structural Risk Minimization Vapnik (1998). To avoid *over-fitting*, i.e. the solution grows in complexity to describe the training set and becomes unable to generalize, we put a constraints on the complexity of the solution. The regularization parameter λ should be chosen in such a way that the empirical error and the complexity of the solution are balanced out.

Another point of view is that of considering the penalty term as a smoothness term which enforces stability of the solution. Here stability has to be thought with respect to the random sampling of the data. This point of view is mostly adopted in the regularization of ill-posed inverse problems where anyway usually only output noise and deterministic sampling is considered. Anyway this point of view is not new to learning theory since the connection between stability and generalization was considered in Bousquet and Elisseeff (2002), Mukherjee et al. (2004), Poggio et al. (2004). As we restrict our analysis to the quadratic loss function we can have some interesting insight. Motivated but recent results on the connection between learning and inverse problems we now try to explain why smoothness is also important for generalization in learning.

Indeed the regularized least-squares algorithm can be seen as implementing a low pass filter on the expansion of the regression function on suitable basis. We recall that the representer theorem Kimeldorf and Wahba (1970) ensures that the solution of problem (3) can be written as

$$f_{\mathbf{z}}^{\lambda} = \sum_{i=1}^{n} \alpha K(x, x_i) \qquad \text{with} \qquad \alpha = (\mathbf{K} + n\lambda I)^{-1}\mathbf{y}, \tag{5}$$

where $\mathbf{K}$ is the kernel matrix $(\mathbf{K})_{ij} = K(x_i, x_j)$. From the explicit form of the coefficients we see that as $\lambda > 0$ we are numerically stabilizing a matrix inversion problem which is possibly ill-conditioned (that is numerically unstable). This is important from the algorithmic point of view, but it is also crucial to ensure the generalization properties of the estimator. For the population version of (3)

$$\min_{f \in \mathcal{H}} \{\mathcal{E}(f) + \lambda \|f\|_{\mathcal{H}}^2\}, \tag{6}$$

the representer theorem (see for example Cucker and Smale (2002a)) gives the explicit form of the solution as

$$f^{\lambda} = (L_K + \lambda I)^{-1} L_K f_{\rho}$$

where L_K is the integral operator of kernel K acting in $L^2(X, \rho_X)$

$$(L_K f)(t) = \int_K K(t, x) f(x) d\rho_X(x).$$

and we considered f^{λ} as a function in $L^2(X, \rho_X)$. Since the kernel is bounded, symmetric and positive definite, L_K is a positive compact operator[1] and the spectral theorem ensures the existence of a basis of eigenfunctions $L_K u_i = \sigma_i u_i$ with $\sigma_i \geq 0$. Then we can rewrite the solution of the above problem as

$$f^{\lambda} = \sum_{i=1}^{\infty} \frac{\sigma_i}{\sigma_i + \lambda} \langle f_{\rho}, u_i \rangle_{\rho} u_i.$$

From the latter expression we see that the effect of regularization is that of a low pass filter which select the components of the regression function corresponding to large eigenvalues. If we slightly perturb ρ, the operator L_K and f_{ρ} change, however the filter ensures that corresponding solution f^{λ} is close to f_{ρ} provided that the perturbation is small and the parameter λ is suitable chosen. The idea is that we can look to the sample case exactly as a perturbation on the

1. This fact is trivial if X is compact, otherwise see Carmeli et al. (2005).

problem due to random sampling. In this case we think of $\mathbf{y}$ and $\mathbf{K}$ as perturbation of f_ρ and L_K respectively. The low pass filter is then a way to ensure stability. This intuition is derived in a more formal way in Section 4 looking at learning in RKHS as an inverse problem.

For regularized least squares algorithm the filter function is $g_\lambda(\sigma) = \frac{1}{\sigma+\lambda}$ but it is natural to extend this approach to other *regularization* g_λ. Each of them defines a corresponding algorithm by means of

$$f_{\mathbf{z}}^{\lambda} = \sum_{i=1}^{n} \alpha_i K(x, x_i) \qquad \text{with} \qquad \alpha = \frac{1}{n} g_\lambda(\frac{\mathbf{K}}{n})\mathbf{y} \tag{7}$$

and again the the final estimator is defined providing the above scheme with a parameter choice $\lambda_n = \lambda(n, \mathbf{z})$ so that $f_{\mathbf{z}} = f_{\mathbf{z}}^{\lambda_n}$. Clearly not all the functions g_λ are admissible and we give a characterization of regularization in the next section.

Here we note that the filter function point of view suggests a natural way to describe regularity of the regression function. Indeed since $f_\rho \in L^2(X, \rho_X)$ we can consider the expansion on the eigensystem of L_K to write

$$f_\rho = \sum_{i=1}^{\infty} \langle f_\rho, u_i \rangle_\rho \, u_i$$

and clearly

$$\sum_{i=1}^{\infty} \langle f_\rho, u_i \rangle_\rho^2 < \infty, \tag{8}$$

that is, the Fourier coefficients of f_ρ with respect to the basis have to go sufficiently fast to zero. A natural way to enforce some more regularity on f_ρ is assuming something more on how fast the Fourier coefficients go to zero. The easier way to do this is to replace (8) with

$$\sum_{i=1}^{\infty} \frac{\langle f_\rho, u_i \rangle_\rho^2}{\sigma_i^r} < \infty$$

where $\{\sigma_i\}$ are the eigenvalues of L_K and $r > 0$. In other words we assume that

$$f_\rho \in \Omega_{r,R} = \{f \in L^2(X, \rho_X) : f = L_K^r v, \|v\|_\rho \leq R\}. \tag{9}$$

Such a condition was first used in the context of learning in Cucker and Smale (2002b) but as noted in De Vito et al. (2005b,a) is a slightly generalization of the classical regularity condition in ill-posed inverse problems, namely Hölder source condition (Engl et al., 1996). For $r = 1/2$ it amounts to assume that the regression function can be seen as as function in the RKHS. In general it depends on the marginal measure ρ_X. The bigger is the smoothness parameter r the easier it is to approximate f_ρ. Intuitively the faster the Fourier coefficients go to zero less information has to be recovered and the fewer examples are needed.

In the following section first, we study under which conditions on $g_\lambda(\sigma)$ we can define sensible learning algorithms and discuss several examples. Then we state and discuss finite sample bounds as well as consistency for such a class of algorithms.

3. Regularization Algorithms for Learning

We now present the class of regularization algorithms we are going to study. Regularization is essentially defined according to what is usual done for ill-posed inverse problems. The main difference is that we require an extra condition, namely a Lipschitz condition, which enables us to show that the obtained learning algorithms are stable.

Definition 1 (Regularization) *We say that a family $g_\lambda : [0, \kappa^2] \to \mathbb{R}$, $0 < \lambda \leq \kappa^2$, is regularization if the following conditions hold*

1. *There exists a constant D such that*

$$\sup_{0<\sigma\leq\kappa^2} |\sigma g_\lambda(\sigma)| \leq D \tag{10}$$

2. *There exists a constant B such that*

$$\sup_{0<\sigma\leq\kappa^2} |g_\lambda(\sigma)| \leq \frac{B}{\lambda} \tag{11}$$

3. *There is a constant $\overline{\nu} > 0$, namely he qualification of the regularization g_λ such that*

$$\sup_{0<\sigma\leq\kappa^2} |1 - g_\lambda(\sigma)\sigma|\sigma^\nu \leq \gamma_\nu \lambda^\nu, \qquad \forall\ 0 < \nu \leq \overline{\nu} \tag{12}$$

   where the constant $\gamma_\nu > 0$ does not depend on λ.

4. *The following Lipschitz condition holds*

$$|(g_\lambda(\sigma) - g_\lambda(\sigma'))| \leq \frac{L}{\lambda^\mu}|\sigma - \sigma'| \tag{13}$$

   where L is a constant independent to λ and μ a positive coefficient.

Let us briefly discuss such conditions. The first three conditions are standard in theory of inverse problems (Engl et al., 1996) whereas the last one is added to deal with the learning setting. The first two conditions are of technical nature, however the constants B and D will enter in the form of the bounds. Basically they ensure that the obtained algorithm can be seen as family of linear continuous maps, parameterized by the regularization parameter λ. The third condition ensures that the solution of the population problem

$$f^\lambda = g_\lambda(L_K) L_K f_\rho$$

converges to f_ρ when λ goes to zero. In other words this ensures that the bias (approximation error) goes to zero as λ goes to zero. Moreover it is also sufficient to derive the corresponding convergence rate if f_ρ satisfies some a priori condition like (9). The meaning of the qualification will be apparent from Theorem 9. Here we just mention the fact that methods with finite qualification cannot fully exploit the possible regularity of the solution and the results no longer improve beyond a certain regularity level.
The fourth condition is quite natural since it ensures stability with respect to perturbations of

the operator L_K and in practice we can only have approximation of L_K based on the training set. Indeed Theorem 8.1 in Birman and Solomyak (2003) ensures that Condition 13 implies

$$\|g_\lambda(B_1) - g_\lambda(B_2)\| \leq \frac{L}{\lambda^\mu} \|B_1 - B_2\|$$

where B_1, B_2 belongs to the Banach space of normal operators endowed with the uniform norm and have spectrum in $[0, \kappa^2]$. The exponent μ will essentially determine the rate of convergence of each algorithm.

3.1 Some Examples of Regularization Algorithms and Semiiterative Regularization

In this Section we describe several algorithms satisfying the above definition. For details on the derivation of the various conditions we refer to Engl et al. (1996) whereas the Lipschitz constant can be directly evaluated as the maximum of the first derivative of g_λ.

Tikhonov Regularization

We start our discussion reviewing Tikhonov regularization. In this case the regularization is $g_\lambda(\sigma) = \frac{1}{\sigma+\lambda}$ so that (10) and (11) hold with $B = D = 1$. Condition (12) is verified with $\gamma_\nu = 1$ for $0 < \nu \leq 1$ and hence the qualification equals to 1. A straightforward computation shows that (13) holds with $L = 1$ and $\mu = 2$. The algorithm amount to a matrix inversion problem as can be seen from (5).

Landweber Iteration

Landweber iteration is characterized by

$$g_t(\sigma) = \tau \sum_{i=0}^{t-1} (1 - \tau\sigma)^i$$

where we identify $\lambda = t^{-1}, t \in \mathbb{N}$ and take $\tau = 1/\kappa^2$. In this case we have $B = D = 1$ and the qualification is infinite since (12) holds with $\gamma_\nu = 1$ if $0 < \nu \leq 1$ and $\gamma_\nu = \nu^\nu$ otherwise. A simple computation shows that $L = 1$ and $\nu = 2$. As shown in Yao et al. (2005) this method corresponds to empirical risk minimization via gradient descent and τ determines the step-size. Early stopping of the iterative procedure allows to avoid over-fitting so that the iteration number plays the role of the regularization parameter. In Yao et al. (2005) the fixed step-size $\tau = 1/\kappa^2$ was shown to be the best choice among the variable step-size $\tau = \frac{1}{\kappa^2(t+1)^\theta}$, with $\theta \in [0, 1)$. This suggests that τ does not play any role for regularization. From the algorithmic point of view we can rewrite the algorithm as the following iterative map

$$\alpha_i = \alpha_{i-1} + \frac{\tau}{n}(\mathbf{y} - \mathbf{K}\alpha_{i-1}), \qquad i = 1, \ldots, t$$

setting $\alpha_0 = 0$.

Semiiterative Regularization and the ν-method

An interesting class of algorithms are the so called semiiterative regularization or accelerated Landweber iteration. This class of methods can be seen as a generalization of Landweber iteration where the regularization is now

$$g_t(\sigma) = p_t(\sigma)$$

with p_t polynomial of degree $t-1$. In this case we can identify $\lambda = t^{-2}, t \in \mathbb{N}$ and we assume $\kappa = 1$ for simplicity. We have $B = 2$ and $D = 1$ and a directly application of Markov inequality for polynomial of degree t shows $L = 4$ and $\mu = 4$. The qualification of this class of method is usually finite. An example which turns out to be particularly interesting is the so called $\nu - method$. We refer to Engl et al. (1996) for a derivation of this method. In the $\nu - method$ the qualification is ν (fixed) with $\gamma_\nu = c$ for some positive constant c. The algorithm amounts to solving, for $\alpha_0 = 0$, the following map

$$\alpha_i = \alpha_{i-1} + u_i(\alpha_{i-1} - \alpha_{i-2}) + \frac{\omega_i}{n}(\mathbf{y} - \mathbf{K}\alpha_{i-1}), \qquad i = 1, \dots, t$$

where

$$\begin{aligned} u_i &= \frac{(i-1)(2i-3)(2i+2\nu-1)}{(i+2\nu-1)(2i+4\nu-1)(2i+2\nu-3)} \\ \omega_i &= 4\frac{(2i+2\nu-1)(i+\nu-1)}{(i+2\nu-1)(2i+4\nu-1)} \qquad t > 1. \end{aligned}$$

The interest of this method lies in the fact that since the regularization parameter here is $\lambda = t^{-2}$ we just need the square root of the number of iterations needed by Landweber iteration. In inverse problems this method proved to be extremely fast and is often used as valid alternative to conjugate gradient (see Engl et al. (1996), Chapter 6 for details).

Iterated Tikhonov

As we have seen while discussing Tikhonov regularization such method has finite qualification and this reflects in the impossibility to exploit the regularity of the solution beyond a certain regularity level. To overcome this problems the following regularization can be considered

$$g_{\lambda,t}(\sigma) = \frac{(\sigma+\lambda)^t - \sigma^t}{\lambda(\sigma+\lambda)^t}$$

In this case we have $D = 1$ and $B = t$ and the qualification of the method is now t with $\gamma_\nu = 1$. A direct computation shows that $L = t(2\kappa)^{t-1}$ and $\mu = 2t$. The algorithm is described by the following iterative map

$$(\mathbf{K} + n\lambda I)\alpha_i = \mathbf{y} + n\lambda\alpha_{i-1} \qquad i = 1, \dots, t$$

choosing $\alpha_0 = 0$. It is easy to see that for $t = 1$ we simply recover the standard Tikhonov regularization but as we let $t > 0$ we improve the qualification of the method. Moreover we note that by fixing λ we can think of the above algorithms as an iterative regularization with t regularization parameter.

3.2 Finite Sample Bounds for Regression and Classification

In this section we fix a regularization scheme g_λ as in Definition 1 and we define the family of algorithms

$$f_{\mathbf{z}}^\lambda = \sum_{i=1}^n \alpha_i K(x, x_i) \qquad \text{with} \qquad \alpha = \frac{1}{n} g_\lambda(\frac{\mathbf{K}}{n})\mathbf{y} \tag{14}$$

parametrized by $0 < \lambda \leq \min\{1, \kappa\}$. Recalling that $\kappa^2 = \sup_{x \in X} K(x, x)$ and $M = \sup |y|$, the following result holds.

Theorem 2 (Finite Sample Bounds) *Suppose that $f_\rho \in \Omega_{r,R}$ and r is smaller or equal than the qualification of g_λ. If we let $\beta = \max\{1, 2\mu\}$ and choose*

$$\lambda_n = n^{-\frac{1}{2r+\beta}} \tag{15}$$

then for $0 < \eta \leq 1$ the following inequality holds with probability at least $1 - \eta$

$$\mathcal{E}(f_{\mathbf{z}}^{\lambda_n}) - \mathcal{E}(f_\rho) \leq \log \frac{4}{\eta}(2C_1^2 + 2\gamma_r^2 R^2) n^{-\frac{2r}{2r+\beta}} \tag{16}$$

where $C_1 = 4\sqrt{2}\kappa M \left(\sqrt{DB} + \kappa^{\frac{5}{2}} L\right)$.

We postpone the proof to Section 5 and add some remarks and corollaries.

For essentially all the methods discussed in Section 3 we have $\mu = 2$, so that our analysis give a bound of order $n^{-\frac{2r}{2r+4}}$. For example if we just know that $f_\rho \in \mathcal{H}$ then $r = 1/2$ and we have a bound of order $n^{-1/5}$, clearly if r and the qualification of the method are sufficiently big the rate can be close to $1/n$. For some regularization algorithms better results than the those presented here are available. For example for Tikhonov regularization bounds of order $n^{-\frac{2r}{2r+1}}$ where proved in Smale and Zhou (2005) and improved in Caponnetto and De Vito (2005) if more information on the structure of the kernel is available. Anyway since this method has finite qualification the results does not improve if $r > 1$. For Landweber iteration bounds of order $n^{-\frac{2r}{2r+3}}$, $r > 0$, where proved in Yao et al. (2005). These results require ad hoc proofs for each algorithm. Here we trade-off generality with the quality of the rates. Our main goal is not finding the best achievable bounds but giving a set of sufficient conditions which allows to derive finite sample bounds for a broad class of algorithms with a relatively simple proof. Up-to our knowledge iterated Tikhonov regularization as well as the class of semiiterative methods are not used in learning. We also note that the above result shows a data independent choice of the regularization parameter. As usual such a choice requires the knowledge of the regularity of the solution so that a data dependent choice would be preferable. In practice selection of the regularization parameter minimizing some validation or cross validation error can be considered.

Consistency for the class of considered algorithms easily follows as a corollary.

Corollary 3 (Consistency) *Under the same assumptions of Theorem 2 let $\mathcal{M}(\Omega_{r,R})$ the set of all Borel probability measure on Z such that $f_\rho \in \Omega_{r,R}$. then*

$$\lim_{\tau \to \infty} \limsup_{n \to \infty} \sup_{\rho \in \mathcal{M}(\Omega_{r,R})} \mathrm{P}\left[\mathcal{E}(f_{\mathbf{z}}^{\lambda_n}) - \mathcal{E}(f_\rho) > \tau n^{-\frac{2r}{2r+\beta}}\right] = 0.$$

The above results have a direct application if we consider classification, that is $Y = \{-1, 1\}$ (Bousquet et al., 2004). In this case we consider $\mathrm{sign} f_{\mathbf{z}}^{\lambda}$ as our decision rule and the error measures is usually the misclassification risk defined as

$$R(f) = \mathrm{P}\left[(x, y) \in X \times Y \ : \ \mathrm{sign} f(x) \neq y\right],$$

whose minimizer is the *Bayes rule* $\mathrm{sign} f_\rho$ (Devroye et al., 1996). A straightforward result can be obtained recalling that the following relation between the risk and the expected error (with respect to the square loss)

$$R(f) - R(f_\rho) \leq \sqrt{\mathcal{E}(f) - \mathcal{E}(f_\rho)}.$$

see (Bartlett et al., 2003). Anyway such a result can be improved if some more information on the problem is available. To this aim it is interesting to consider Tsybakov noise condition

$$\mathrm{P}\left[\, x \in X \;:\; |f_\rho(x)| \leq L\right] \leq B_q L^q, \quad \forall L \in [0,1], \tag{17}$$

where $q \in [0,\infty]$ (Tsybakov, 2004). The meaning of such a condition is better understood noting that $f_\rho(x) = 2\rho(1|x) - 1$ so that if q goes to ∞ the problem is separable (realizable setting). In this case the following comparison result is available

$$R(f) - R(f_\rho) \leq 4c_\alpha \left(\mathcal{E}(f) - \mathcal{E}(f_\rho)\right)^{\frac{1}{2-\alpha}} \tag{18}$$

with $\alpha = \frac{q}{q+1}$ and $c_\alpha = B_q + 1$, see Bartlett et al. (2003) or Yao et al. (2005). The following corollary is straightforward.

Corollary 4 (Bayes Consistency) *Under the same assumptions of Theorem 2 assume that Tsybakov noise condition holds. If we choose λ_n according to (15) and use sign$f_{\mathbf{z}}^{\lambda_n}$ as our decision rule then the following bound holds with probability at least $1-\eta$*

$$R(f_{\mathbf{z}}^{\lambda_n}) - R(f_\rho) \leq C(\mathcal{H},\eta,\rho) n^{-\frac{2r}{(2r+\beta)(2-\alpha)}}$$

where $C(\mathcal{H},\eta,\rho) = 4c_\alpha(\log\frac{4}{\eta}(2C_1^2 + 2\gamma_r^2 R^2))^{\frac{1}{2-\alpha}}$ with c_α as in (18) and C_1 given in Theorem 2.

4. Regularization Operators, an Inverse Problems Perspective

In this section we clarify the role of the regularization looking at learning algorithm as an inverse problem as showed in De Vito et al. (2005b). For backgrounds and details on inverse problems we refer to (Tikhonov and Arsenin, 1977, Engl et al., 1996, Bertero and Boccacci, 1998).

In the framework of learning, if an hypothesis space $\mathcal{H}$ is given, the *ideal* estimator is the solution of the minimization problem

$$\inf_{f\in\mathcal{H}} \mathcal{E}(f) = \inf_{f\in\mathcal{H}} \|I_K f - f_\rho\|_\rho^2 + \mathcal{E}(f_\rho). \tag{19}$$

The above equality is a consequence of (1) and we have stressed the fact that f is an element of $\mathcal{H}$, but its relevant norm is the norm in $L^2(X,\rho_X)$, writing explicitly the inclusion operator $I_K : \mathcal{H} \to L^2(X,\rho_X)$. We notice that the action of I_K is trivial since it maps f into itself, but the norm changes from $\|\cdot\|_{\mathcal{H}}$ to $\|\cdot\|_\rho$.
It follows that (19) is equivalent to the least square problem associated to the linear inverse problem

$$I_K f = f_\rho. \tag{20}$$

In a similar way, given a training set $\mathbf{z} = (\mathbf{x},\mathbf{y})$, we have that

$$\min_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n (f(x_i) - y_i)^2 = \min_{f\in\mathcal{H}} \|S_{\mathbf{x}} f - \mathbf{y}\|_n^2, \tag{21}$$

where $\|\cdot\|_n$ is $1/n$ times the euclidean norm in $\mathbb{R}^n$ and $S_{\mathbf{x}} : \mathcal{H} \to \mathbb{R}^n$ is the sampling operator

$$(S_{\mathbf{x}} f)_i = f(x_i).$$

Again we can see that empirical risk minimization is the least square problem associated to the linear inverse problem

$$S_{\mathbf{x}} f = \mathbf{y} \tag{22}$$

(here we recover the problem of approximating a function from finite data, that is finding f such that $f(x_i) = y_i$ with $i = 1, .., n$).

A simple calculation shows that the least square solutions of (19) and (21) are solutions of the following linear equations

$$I_K^* I_K f = I_K^* f_\rho \tag{23}$$

and

$$S_{\mathbf{x}}^* S_{\mathbf{x}} f = S_{\mathbf{x}}^* \mathbf{y}. \tag{24}$$

Notice that in the above formulation $I_K^* I_K$ and $S_{\mathbf{x}}^* S_{\mathbf{x}}$ are operators from $\mathcal{H}$ to $\mathcal{H}$, whereas $I_K^* f_\rho$ and $S_{\mathbf{x}}^* \mathbf{y}$ are elements of $\mathcal{H}$. Moreover, if the number n of data goes to infinity, as a consequence of the law of large numbers, $S_{\mathbf{x}}^* S_{\mathbf{x}}$ and $S_{\mathbf{x}} \mathbf{y}$ converge to $I_K^* I_K$ and $I_K^* f\rho$, respectively (see Lemma 5 below). However, since $I_K^* I_K$ is a compact operator, in general the (Moore-Penrose) inverse of $I_K^* I_K$ is not continuous and, hence, the solution of (24) does not converge to the solution of (23), which is simply f_ρ in the present framework (under the assumption that $\mathcal{H}$ is dense in $L^2(X, \rho_X)$).

The key idea of inverse problems is to regularize (23) by considering a family of regularized solutions

$$g_\lambda(I_K^* I_K) I_K^* f_\rho \tag{25}$$

depending of a positive parameter λ in such a way that

1. $g_\lambda(\sigma)$ is bounded for σ in $[0, \kappa^2]$, so the spectral theorem ensures that $g_\lambda(I_K^* I_K)$ is bounded, too;

2. $g_\lambda(\sigma)$ approximates the function $\frac{1}{\sigma}$ as λ goes to 0, that is, $g_\lambda(I_K^* I_K)$ is a family of operators approximating the inverse of $I_K^* I_K$ when λ goes to 0. This allows recovering the exact solution f_ρ in the limit.

Moreover in learning we also require $g_\lambda(\sigma)$ to be a Lipschitz function of σ, so that the discretized solution

$$g_\lambda(S_{\mathbf{x}}^* S_{\mathbf{x}}) S_{\mathbf{x}}^* \mathbf{y}$$

converges to $g_\lambda(I_K^* I_K) I_K^* f_\rho$ for n going to infinity and given λ. Within this setting the final step of the regularization procedure is the choice of the regularization parameter $\lambda = \lambda_n$ as a function of n so that $g_{\lambda_n}(S_{\mathbf{x}}^* S_{\mathbf{x}}) S_{\mathbf{x}}^* \mathbf{y}$ converges to f_ρ.

We end the section with a remark about the notion of convergence we are interested into. Usually in the framework of inverse problems the convergence is considered with respect to the norm in $\mathcal{H}$ (reconstruction error), so that it is necessary to require the existence of at least a solution of (23), namely the Moore-Penrose solution. In learning theory we are interested into convergence in $L^2(X, \rho_X)$-norm (residual), hence we do not require the existence of the Moore-Penrose solution, which in our context is equivalent to the assumption that $f_\rho \in \mathcal{H}$. Moreover, since both $S_{\mathbf{x}}$ and $\mathbf{y}$ are random variables, the convergence has to be understood in probability or in expectation.

5. Error Estimates and Proof of the Main Result

In this section we prove the main results of the paper stated in Section 3. The idea is to show that error of the estimator, for a fixed value of the regularization parameter, can be suitably decomposed in a probabilistic term, sample error, and a deterministic term, approximation error. If explicit bounds on the two terms are available we can find the value of the regularization parameter which solve the bias-variance trade-off, that is the value of λ balancing out the sample and approximation errors. Most of this section is devoted to prove such bounds. Before actually proving such results it is convenient to define some operators on the RKHS $\mathcal{H}$.

5.1 Sampling and Covariance Operators

We recall that the main intuition behind the considered class of algorithm is that they ensure stability with respect to the random sampling. In particular we regarded the sample case, that is $\mathbf{y}$ and $\mathbf{K}$, as a perturbation of the population case, that is of f_ρ and L_K. To give a formal proof to the above intuition we would like to give a quantitative measure of the discrepancy between the sample and population case. Rather than comparing $\mathbf{K}$ and L_K it is useful to define the following operators. For details we refer to Carmeli et al. (2005).
We let $I_K : \mathcal{H} \to L^2(X, \rho_X)$ be the inclusion operator, which is continuous by (2), $I_K^* : L^2(X, \rho_X) \to \mathcal{H}$ the adjoint operator and $T := I_K^* I_K : \mathcal{H} \to \mathcal{H}$ the covariance operator. It can be proved that $L_K = I_K I_K^*$ and

$$T \;=\; \int_X \langle \cdot, K_x \rangle_{\mathcal{H}} \, K_x d\rho_X(x).$$

Since the kernel is bounded and positive definite, both L_K and T are trace class positive operator and there is a sequence of vectors $(e_i)_{i\geq 1}$ in $\mathcal{H}$ and a sequence of numbers $(\sigma_i)_{i\geq 1}$ (possibly finite) such that

$$Tf = \sum_{i=1} \sigma_i \langle f, e_i \rangle_{\mathcal{H}} \, e_i \qquad \langle e_i, e_j \rangle_{\mathcal{H}} = \delta_{ij} \qquad \sum_i \sigma_i \leq \kappa^2 \qquad \sigma_{i+1} \geq \sigma_i > 0$$

for all $f \in \mathcal{H}$ and, letting $u_i = \frac{1}{\sqrt{\sigma_i}} e_i \in L^2(X, \rho_X)$

$$L_K f = \sum_{i=1}^{n} \sigma_i \langle f, u_i \rangle_\rho \, u_i \qquad \langle u_i, u_j \rangle_\rho = \delta_{ij}.$$

In particular, $\|L_K\|_{\mathcal{L}(L^2(X,\rho_X))} = \|T\|_{\mathcal{L}(\mathcal{H})} \leq \sum_i \sigma_i \leq \kappa^2$.

Let now $\mathbf{x} = (x_i)_{i=1}^n$ with $x_i \in X$, we define the sampling operator $S_\mathbf{x} : \mathcal{H} \to \mathbb{R}^n$ as

$$(S_\mathbf{x} f)_i = f(x_i) = \langle f, K_{x_i} \rangle_{\mathcal{H}} \qquad i = 1, \ldots, n,$$

where the norm $\|\cdot\|_n$ in $\mathbb{R}^n$ is $1/n$ times the euclidean norm, and the empirical covariance operator $T_\mathbf{x} : \mathcal{H} \to \mathcal{H}$ as $T_\mathbf{x} := S_\mathbf{x}^* S_\mathbf{x}$. It can be proved that

$$T_\mathbf{x} := \frac{1}{n} \sum_{i=1}^{n} \langle \cdot, K_{x_i} \rangle_{\mathcal{H}} \, K_{x_i}.$$

and $S_{\mathbf{x}}S_{\mathbf{x}}^* = 1/n\mathbf{K}$. Clearly $T_{\mathbf{x}}$ is a positive operator with finite rank (hence it is a trace class operator) and $\|T_{\mathbf{x}}\|_{\mathcal{L}(\mathcal{H})} \leq \kappa^2$.

The above operators allow to write f^λ and $f_{\mathbf{z}}^\lambda$ in a suitable form, that is,

$$f^\lambda = g_\lambda(T)I_K^* f_\rho \qquad f_{\mathbf{z}}^\lambda = g_\lambda(T_{\mathbf{x}})S_{\mathbf{x}}^*\mathbf{y}. \tag{26}$$

where both f^λ and $f_{\mathbf{z}}^\lambda$ are regarded as elements of $\mathcal{H}$.
Now we can look at $T_{\mathbf{x}}$ and $S_{\mathbf{x}}^*\mathbf{y}$ as approximation of T and $I_K^* f_\rho$ respectively. The advantage is that we are now dealing with operators acting on $\mathcal{H}$ and functions in $\mathcal{H}$ which can be more easily compared.

To prove the main error estimates in next Section we recall some facts. Due to the assumption that $\mathcal{H}$ is dense, the best model $f_{\mathcal{H}}^\dagger$ exists if and only if f_ρ is an element of $\mathcal{H}$, so that $I_K f_{\mathcal{H}} = f_\rho$. Moreover it is easy to see that we can relate the norm in $\mathcal{H}$ and $L^2(X, \rho_X)$ by means of the operator T. For $f \in \mathcal{H}$ we can write explicitly the embedding operator I_K to get

$$\|I_K f\|_\rho = \left\|\sqrt{T} f\right\|_{\mathcal{H}}. \tag{27}$$

This fact can be easily proved recalling that the inclusion operator is continuous and hence admits a polar decomposition $I_K = U\sqrt{T}$, where U is a partial isometry (Rudin, 1991).

Finally for sake of completeness we show how (7) and (26) are related. To this aim we recall that by polar decomposition the following equalities hold $S_{\mathbf{x}} = \sqrt{1/n\mathbf{K}}U_{\mathbf{x}}^*$, $S_{\mathbf{x}}^* = U_{\mathbf{x}}\sqrt{1/n\mathbf{K}}$ and clearly $T_{\mathbf{x}} = U_{\mathbf{x}} 1/n\mathbf{K} U_{\mathbf{x}}^*$. Then we can write

$$f_{\mathbf{z}}^\lambda = g_\lambda(T_{\mathbf{x}})S_{\mathbf{x}}^* y = U_{\mathbf{x}} g_\lambda(\frac{1}{n}\mathbf{K})\sqrt{\frac{1}{n}\mathbf{K}}y \tag{28}$$

where we used $U_{\mathbf{x}}^* U_{\mathbf{x}}$ is the identity on the range of $\mathbb{K}$. From the above formula we immediately see that $f_{\mathbf{z}}^\lambda$ is an element of the range of $U_{\mathbf{x}}$, which is the linear span of the vectors K_{x_i}. Hence $f_{\mathbf{z}}^\lambda = \sum_{i=1}^n \alpha_i K_{x_i}$ and, if we apply the sampling operator on both sides of (28), we get

$$S_{\mathbf{x}} f_{\mathbf{z}}^\lambda = S_{\mathbf{x}} \sum_{i=1}^n \alpha_i K_{x_i} = \mathbf{K}\alpha$$

where α denotes the vector of the coefficients and

$$S_{\mathbf{x}} U_{\mathbf{x}} g_\lambda(\frac{1}{n}\mathbf{K})\sqrt{\frac{1}{n}\mathbf{K}}y = \sqrt{\frac{1}{n}\mathbf{K}}g_\lambda(\frac{1}{n}\mathbf{K})\sqrt{\frac{1}{n}\mathbf{K}}y.$$

Then the following equality holds

$$\mathbf{K}\alpha = \frac{1}{n}\mathbf{K} g_\lambda(\frac{1}{n}\mathbf{K})y.$$

5.2 Approximation and Sample Error

We can now derive the error estimates which are the key to the prove of Theorem 2. The bias-variance problem follows considering, for fixed λ, the following error decomposition

$$\sqrt{\mathcal{E}(f_{\mathbf{z}}^\lambda) - \mathcal{E}(f_\rho)} \leq \left\|f_{\mathbf{z}}^\lambda - f^\lambda\right\|_\rho + \left\|f^\lambda - f_\rho\right\|_\rho \tag{29}$$

where we used (1) and triangle inequality. In this case one term $\left\|f_{\mathbf{z}}^{\lambda} - f^{\lambda}\right\|_{\rho}$ accounts for the presence of a perturbation (sample or estimation error) whereas the other term $\left\|f^{\lambda} - f_{\rho}\right\|_{\rho}$ accounts for the fact that, though considering the unperturbed problem, we are limiting the approximation property of our algorithm by fixing λ (approximation error).

If the best in the model $f_{\mathcal{H}}^{\dagger}$ exists besides the expected error we can also consider the error measured with respect to the norm in the RKHS $\mathcal{H}$. This can be interesting since convergence in $\mathcal{H}$-norm implies point-wise convergence and moreover by choosing different kernels we might get convergence in different norms (for example Sobolev norms). In this case the decomposition is simply

$$\left\|f_{\mathbf{z}}^{\lambda} - f_{\mathcal{H}}^{\dagger}\right\|_{\mathcal{H}} \leq \left\|f_{\mathbf{z}}^{\lambda} - f^{\lambda}\right\|_{\mathcal{H}} + \left\|f^{\lambda} - f_{\mathcal{H}}^{\dagger}\right\|_{\mathcal{H}}.$$

We first consider the estimation error. Our approach is divided into two steps. Recalling (26) we prove analytically that that the difference $f^{\lambda} - f_{\mathbf{z}}^{\lambda}$ can be expressed in terms of the perturbation measures $T - T_{\mathbf{x}}$ and $I_K^* f_{\rho} - S_{\mathbf{x}}^* \mathbf{y}$. Then we need to to give probabilistic estimates for such perturbation measures. For the latter we make use of the following result from De Vito et al. (2005a) based on concentration of Hilbert space valued random variables (Pinelis and Sakhanenko, 1985).

Lemma 5 *Let $\kappa = \sup_{x \in X} \|K_x\|_{\mathcal{H}}$, $M = \sup_{y \in Y} |y|$. For $n \in \mathbb{N}$ and $0 < \eta \leq 1$ the following inequalities hold with probability at least $1 - \eta$*

$$\begin{aligned} \left\|I_K^* f_{\rho} - S_{\mathbf{x}}^* \mathbf{y}\right\|_{\mathcal{H}} &\leq \delta_1(n, \eta), \quad \delta_1(n, \eta) = \frac{2\sqrt{2}\kappa M}{\sqrt{n}} \sqrt{\log \frac{4}{\eta}} \\ \left\|T - T_{\mathbf{x}}\right\|_{\mathcal{L}(\mathcal{H})} &\leq \delta_2(n, \eta), \quad \delta_2(n, \eta) = \frac{2\sqrt{2}\kappa^2}{\sqrt{n}} \sqrt{\log \frac{4}{\eta}}. \end{aligned} \tag{30}$$

We are now ready to derive our estimates for the sample error. The following result is a natural generalization of Theorem 1 in De Vito et al. (2005b) (see also Theorems 4.2 in Engl et al. (1996)).

Theorem 6 (Estimation Error) *Let g_{λ} as in Definition 1 and $f_{\mathbf{z}}^{\lambda}$, f^{λ} as defined in (26), with $0 < \lambda \leq 1$. Moreover recall $\kappa^2 = \sup_{x \in X} K(x, x)$ and $M = \sup |y|$. Then for $n \in \mathbb{N}$ and $0 < \eta \leq 1$ the following inequality holds with probability at least $1 - \eta$*

$$\left\|f_{\mathbf{z}}^{\lambda} - f^{\lambda}\right\|_{\rho} \leq C_1 \frac{1}{\lambda^{\theta} \sqrt{n}} \sqrt{\log \frac{4}{\eta}} \tag{31}$$

where $C_1 = 4\sqrt{2}\kappa M \left(\sqrt{DB} + \kappa^{\frac{5}{2}} L\right)$ and $\theta = \max\{1/2, \mu\}$.

Moreover with probability at least $1 - \eta$

$$\left\|f_{\mathbf{z}}^{\lambda} - f^{\lambda}\right\|_{\mathcal{H}} \leq C_2 \frac{1}{\lambda^{\gamma} \sqrt{n}} \log \frac{4}{\eta} \tag{32}$$

where $C_2 = 4\sqrt{2}\kappa M \left(B + \kappa^2 L\right)$ and $\gamma = \max\{1, \mu\}$.

Proof The prove of the two bounds is essentially the same. We consider the following decomposition

$$\begin{aligned} f_{\mathbf{z}}^{\lambda} - f^{\lambda} &= g_{\lambda}(T_{\mathbf{x}})S_{\mathbf{x}}^{*}\mathbf{y} - g_{\lambda}(T)I_{K}^{*}f_{\rho} \\ &= (g_{\lambda}(T_{\mathbf{x}}) - g_{\lambda}(T))S_{\mathbf{x}}^{*}\mathbf{y} + g_{\lambda}(T)(S_{\mathbf{x}}^{*}\mathbf{y} - I_{K}^{*}f_{\rho}). \end{aligned} \tag{33}$$

The bound in the $\mathcal{H}$-norm follows from triangle inequality, in fact from Conditions (11) and (13) and spectral theorem (Lang, 1993) we get

$$\left\| f_{\mathbf{z}}^{\lambda} - f^{\lambda} \right\|_{\mathcal{H}} \leq \frac{\kappa M L}{\lambda^{\mu}} \left\| T - T_{\mathbf{x}} \right\| + \frac{B}{\lambda} \left\| S_{\mathbf{x}}^{*}\mathbf{y} - I_{K}^{*}f_{\rho} \right\|_{\mathcal{H}} \tag{34}$$

where we used $\|S_{\mathbf{x}}^{*}\mathbf{y}\|_{\mathcal{H}} = \|1/n \sum_{i=1}^{n} k_{x_i} y_i\|_{\mathcal{H}} \leq \kappa M$.

For the bound on the expected error we recall that using (27) we can write

$$\left\| f_{\mathbf{z}}^{\lambda} - f^{\lambda} \right\|_{\rho} = \left\| \sqrt{T}(f_{\mathbf{z}}^{\lambda} - f^{\lambda}) \right\|_{\mathcal{H}}$$

where we omit writing explicitly I_K. Moreover we have that

$$\left\| \sqrt{T} g_{\lambda}(T) \right\|_{\mathcal{L}(\mathcal{H})} \leq \sqrt{\frac{BD}{\lambda}}$$

in fact Conditions (10), (11) and spectral theorem ensure that $\forall f \in \mathcal{H}$

$$\begin{aligned} \left\| \sqrt{T} g_{\lambda}(T) f \right\|_{\mathcal{H}}^{2} = & \\ &= \left\langle \sqrt{T} g_{\lambda}(T) f, \sqrt{T} g_{\lambda}(T) f \right\rangle \\ &= \langle g_{\lambda}(T) f, T g_{\lambda}(T) f \rangle_{\mathcal{H}} \\ &\leq \left\| g_{\lambda}(T) f \right\|_{\mathcal{H}} \left\| T g_{\lambda}(T) f \right\|_{\mathcal{H}} \leq \frac{B}{\lambda} D \left\| f \right\|_{\mathcal{H}}^{2}. \end{aligned}$$

The following estimate for the sample error follows

$$\left\| f_{\mathbf{z}}^{\lambda} - f^{\lambda} \right\|_{\rho} \leq \frac{\kappa^{2} M L}{\lambda^{\mu}} \left\| T - T_{\mathbf{x}} \right\| + \frac{\sqrt{DB}}{\sqrt{\lambda}} \left\| S_{\mathbf{x}}^{*}\mathbf{y} - I_{K}^{*}f_{\rho} \right\|_{\mathcal{H}} \tag{35}$$

where we used $\sqrt{T} \leq \kappa$. To finish the proof we simply have to plug the probabilistic estimates of Lemma 5 into (34) and (35). ■

Remark 7 *The condition $\lambda < 1$ is considered only to simplify the results and can be replaced by $\lambda < a$ for some positive constant a that would eventually appear in the bound.*

Remark 8 *Inspecting the proof of the above theorem we see that the set of "good" training sets such that the above bound holds does not depend on λ so that the bound still holds if we take $\lambda = \lambda(\mathbf{z})$. This might be helpful while looking for a data-dependent parameter choice.*

Next theorem consider the approximation error. It can be proved by means of minor modification from standard results in inverse problem. In fact its proof can be directly derived from Theorem 4.3 in Engl et al. (1996).

Theorem 9 (Approximation Error) *Let g_λ as in Definition 1, f^λ as defined in (26). If $f_\rho \in \Omega_{r,R}$ and r is smaller then the qualification of g_λ then*

$$\left\| f^\lambda - f_\rho \right\|_\rho \leq \gamma_r R \lambda^r. \tag{36}$$

If $r > 1/2$, then $f^\dagger_{\mathcal{H}}$ exists and

$$\left\| f^\lambda - f^\dagger_{\mathcal{H}} \right\|_{\mathcal{H}} \leq \gamma_c R \lambda^c \tag{37}$$

where $c = r - 1/2$.

Proof We recall that

$$\mathcal{E}(f^\lambda) - \mathcal{E}(f_\rho) = \left\| f_\rho - I_K f^\lambda \right\|^2_\rho,$$

where we wrote explicitly the embedding operator I_K since f^λ belongs to $\mathcal{H}$, and we also recall the following useful inequality

$$g_\lambda(I_K^* I_K) I_K^* = I_K^* g_\lambda(I_K I_K^*).$$

Since $f_\rho \in \Omega_{r,R}$ (and $L_K = I_K I_K^*$) we can write

$$\|f_\rho - I_K f_\lambda\|_\rho = \|f_\rho - I_K g_\lambda(I_K^* I_K) I_K^* f_\rho\| = \|(I - L_K g_\lambda(L_K)) L_K^r u\|. \tag{38}$$

Then Condition(12) ensures that the inequality

$$\left\| f^\lambda - f_\rho \right\|_\rho \leq \gamma_r \lambda^r$$

holds true if r is smaller or equal then the qualification of g_λ.

Finally (37) can be proved recalling that each bounded operator admits a polar decomposition $A = U|A|$, where U is a partial isometry and $|A|$ is the positive square root of A^*A (Rudin, 1991). If we let $I_K^* = U(I_K I_K^*)^{\frac{1}{2}}$ be the polar decomposition of I_K^*, then for $r > 1/2$

$$f_\rho = (I_K I_K^*)^r \phi = (I_K I_K^*)^{1/2} (I_K I_K^*)^c \phi = (I_K I_K^*)^{1/2} U^* U (I_K I_K^*)^c U^* U \phi = I_K (T)^c U \phi,$$

where $c = r - 1/2$. It follows that $P f_\rho \in Im(I_K)$, so that $f^\dagger_{\mathcal{H}}$ exists and since $P f_\rho = I_K f^\dagger_{\mathcal{H}}$ clearly $f_{\mathcal{H}} = (T)^c U \phi$. Now we can mimic the proof of the first bound and using $T f^\dagger_{\mathcal{H}} = I_K^* f_\rho$ we can write

$$f^\dagger_{\mathcal{H}} - f^\lambda = (I - g_\lambda(I_K^* I_K) T) f^\dagger_{\mathcal{H}} = (I - g_\lambda(I_K^* I_K) T)(T)^c U \phi$$

If we take the norm of the above expression Condition (12) in Def.(1) and spectral theorem ensures that

$$\left\| f^\dagger_{\mathcal{H}} - f^\lambda \right\|_{\mathcal{H}} \leq \gamma_c \lambda^c R$$

where we used the fact that $\|U\phi\|_{\mathcal{H}} = \|\phi\|_\rho$ since U is a partial isometry. ■

The proof of Theorem 2 and the corollaries are straightforward.
Proof [Finite sample bounds] We simply plug the above estimates into the following inequality

$$\mathcal{E}(f_{\mathbf{z}}^\lambda) - \mathcal{E}(f_\rho) \leq 2 \left\| f_{\mathbf{z}}^\lambda - f^\lambda \right\|_\rho^2 + 2 \left\| f^\lambda - f_\rho \right\|_\rho^2.$$

The proof follows taking the value of λ balancing out the two terms that is the value such that

$$\lambda^{2r} = \frac{1}{\lambda^\beta n}$$

where $\beta = \max\{1, 2\mu\}$. ■

Remark 10 *Clearly we can easily get a similar results for the estimates in $\mathcal{H}$. Interestingly it turns out that the parameter choice does not change.*

Finally we can prove consistency.
Proof [Consistency] We let $\tau = (2C_1^2 + 2\gamma_r^2) \log \frac{4}{\eta}$ and solve with respect to η to get

$$\eta_\tau = 4e^{-\frac{\tau}{2C_1^2 + 2\gamma_r^2}}.$$

Then we know from Theorem 2 that

$$\mathrm{P}\left[\mathcal{E}(f_{\mathbf{z}}^{\lambda_n}) - \mathcal{E}(f_\rho) > \tau n^{\frac{2r}{2r+\beta}} \right] \leq \eta_\tau$$

and clearly

$$\limsup_{n\to\infty} \sup_{\rho \in \mathcal{M}(\Omega_{r,R})} \mathrm{P}\left[\mathcal{E}(f_{\mathbf{z}}^{\lambda_n}) - \mathcal{E}(f_\rho) > \tau n^{\frac{2r}{2r+\beta}} \right] \leq \eta_\tau.$$

The theorem is proved since $\eta_\tau \to 0$ as $\tau \to \infty$. ■

6. Conclusions

In this paper we build upon the mathematical relation between inverse problems and learning theory. It is well known that Tikhonov regularization can be profitably used for learning and enjoys good theoretical properties. In our analysis we show that a large number of algorithms well known to the inverse problems community can be casted in the learning framework. All these algorithms are kernel methods easy to implement and their theoretical properties can be derived by adapting standard results of regularization theory. Our analysis confirms the deep connection between learning and inverse problems.

Current work concentrates on assessing strengths and weaknesses of these new learning algorithms in real applications. From a more theoretical viewpoint we aim to improve the probabilistic bounds (Bauer et al., 2005). Finally, we are studying the extension of the presented analysis to the case of other regularization principles like sparsity enhancing regularization and regularization with differential operators.

Acknowledgments

We would like to thank A. Caponnetto, F. Bauer, S. Pereverzev and Y. Yao for useful discussions and suggestions. This research has been partially funded by the FIRB Project ASTAA and the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778.

References

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

P. L. Bartlett, M. J. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. Technical Report Technical Report 638, Department of Statistics, U.C. Berkeley, 2003.

F. Bauer, S. V. Pereverzev, and L. Rosasco. On regularization algorithms in leanring theory. Technical Report DISI-TR-05-18, DISI Universitá di Genova, december 2005. retrievable at http://www.disi.unige.it/person/RosascoL/.

M. Bertero and P. Boccacci. *Introduction to Inverse Problems in Imaging*. IOP Publishing, Bristol, 1998. A Wiley-Interscience Publication.

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.

M. S. Birman and M. Solomyak. Double operators integrals in hilbert spaces. *Integr. Equ. Oper. Theory*, pages 131–168, 2003.

O. Bousquet, S. Boucheron, and G. Lugosi. *Introduction to Statistical Learning Theory*, volume Lectures Notes in Artificial Intelligence 3176, pages 169, 207. Springer, Heidelberg, Germany, 2004. A Wiley-Interscience Publication.

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

P. Bühlmann and B. Yu. Boosting with the l_2-loss: Regression and classification. *Journal of American Statistical Association*, 98:324–340, 2002.

A. Caponnetto and E. De Vito. Optimal rates for regularized least-squares algorithm. *submitted*, 2005.

C. Carmeli, E. De Vito, and A. Toigo. Reproducing kernel hilbert spaces and mercer theorem. *eprint arXiv: math/0504071*, 2005. available at http://arxiv.org.

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

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

E. De Vito, L. Rosasco, and A. Caponnetto. Discretization error analysis for tikhonov regularization. *to appear in Analysis and Applications*, 2005a.

E. De Vito, L. Rosasco, A. Caponnetto, U. De Giovannini, and F. Odone. Learning from examples as an inverse problem. *Journal of Machine Learning Research*, 6:883–904, May 2005b.

R. DeVore, G. Kerkyacharian, D. Picard, and V. Temlyakov. On mathematical methods of learning. Technical Report 2004:10, Industrial Mathematics Institute, Dept. of Mathematics University of South Carolina, 2004. retrievable at http://www.math/sc/edu/ imip/04papers/0410.ps.

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.

M. Györfi, L.and Kohler, A. Krzyzak, and H. Walk. *A Distribution-free Theory of Non-parametric Regression*. Springer Series in Statistics, New York, 1996, 1996.

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

G. Kimeldorf and G. Wahba. A correspondence between Bayesian estimation of stochastic processes and smoothing by splines. *Ann. Math. Stat.*, 41:495–502, 1970.

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

J.M. Loubes and C. Ludena. Model selection for non linear inverse problems. *submitted to Probability Theory and Related Fields*, 2004.

S. Mendelson. Estimating the performance of kernel classes. *Journal of Machine Learning Research*, 4:759–771, 2003.

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.

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

I. F. Pinelis and A. I. Sakhanenko. Remarks on inequalities for probabilities of large deviations. *Theory Probab. Appl.*, 30(1):143–148, 1985. ISSN 0040-361X.

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.

T. Poggio, R. Rifkin, S. Mukherjee, and P. Niyogi. General conditions for predictivity in le arning theory. *Nature*, 428:419–422, 2004.

A. Rakhlin, S. Mukherjee, and T. Poggio. Stability results in learning theory. *Analysis and Applications*, 3:397–419, 2005.

W. Rudin. *Functional Analysis*. International Series in Pure and Applied Mathematics. Mc Graw Hill, Princeton, 1991.

S. Smale and D.X. Zhou. Learning theory estimates via integral operators and their approximations. *submitted*, 2005. retrievable at http://www.tti-c.org/smale.html.

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

A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, 32:135–166, 2004.

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.

Y. Yao, L. Rosasco, and A. Caponnetto. On early stopping in gradient descent learning. *submitted*, 2005. retrievable at http://mathberkeley.edu/ yao/publications/earlystop.pdf.