INVERSE PROBLEMS IN LEARNING FROM DATA

Věra Kůrková
Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 2, Prague, Czech Republic

Keywords: Learning from data, Minimization of empirical and expected error functionals, Reproducing kernel Hilbert spaces.

Abstract: It is shown that application of methods from theory of inverse problems to learning from data leads to simple proofs of characterization of minima of empirical and expected error functionals and their regularized versions. The reformulation of learning in terms of inverse problems also enables comparison of regularized and non regularized case showing that regularization achieves stability by merely modifying output weights of global minima. Methods of theory of inverse problems lead to choice of reproducing kernel Hilbert spaces as suitable ambient function spaces.

1 INTRODUCTION

Supervised learning can be formally described as an optimization problem of minimization of error functionals over parameterized sets of input-output functions computable by a given computational model. Various learning algorithms iteratively modify parameters of the model until sufficiently small values of error functionals are achieved and the corresponding input-output function of the model sufficiently well fits to the training data.

However, such algorithms in their best only achieve good fit to the training data. It has been proven that for typical computational units such as sigmoidal perceptrons and Gaussian kernels, sufficiently large networks can exactly interpolate any sample of data (Ito, 1992; Michelli, 1986). Data are often noisy and networks perfectly fitting to randomly chosen training samples may be too much influenced by the noise and may not perform well on data that were not chosen for training. Thus various attempts to modify error functionals to improve so called "generalization capability" of the model has been proposed. In 1990s, Girosi and Poggio (Girosi and Poggio, 1990) introduced into learning theory a method of regularization as a means of improving generalization. They considered modifications of error functionals based on Tikhonov regularization which adds an additional functional, called stabilizer, which penalizes undesired properties of input-output functions such as high-frequency oscillations (Girosi et al., 1995). In practical applications, various simple stabilizers have been successfully used such as seminorms based on derivatives (Bishop, 1995) or sum or square of output weights (Fine, 1999; Kecman, 2001).

Regularization was developed in 1970s as a method of improving stability of solutions of certain problems from physics called *inverse problems*, where *unknown causes* (e.g., shapes of functions, forces or distributions) of *known consequences* (measured data) have to be found. These problems has been studied in applied science, such as acoustics, geophysics and computerized tomography (see, e.g., (Hansen, 1998)). To solve such a problem, one needs to know how unknown causes determine known consequences, which can often be described in terms of an *operator*. In problems originating from physics, dependence of consequences on causes is usually described by integral operators (such as those defining Radon or Laplace transforms (Bertero, 1989; Engl et al., 1999)). As some problems do not always have exact solutions or have solutions which are unstable with respect to noise, various methods of finding approximate solutions and improving its stability has been developed.

Also minimization of empirical and expected error functionals with quadratic loss functions can be formulated as inverse problems. But the operators representing problems of finding unknown input/output functions approximating well training data are quite different from typical operators describing inverse problems from applied science. In this paper, we show that application of methods from theory of inverse problems to learning from data leads to choice

Kůrková V..
INVERSE PROBLEMS IN LEARNING FROM DATA.
DOI: 10.5220/0003079003160321
In *Proceedings of the International Conference on Fuzzy Computation and 2nd International Conference on Neural Computation* (ICNC-2010), pages 316-321
ISBN: 978-989-8425-32-4

od reproducing kernel Hilbert spaces as ambient function spaces. In these spaces, characterization of functions minimizing error functionals follows easily from basic results on pseudosolutions and regularized solutions. These characterizations have been proven earlier using other methods, such as Fréchet derivatives (Wahba, 1990; Cucker and Smale, 2002; Poggio and Smale, 2003) or operators with fractional powers (Cucker and Smale, 2002), but reformulation of minimization of error functionals in terms of inverse problems allows much simpler and transparent proofs. It also provides a unifying framework showing that an optimal regularized solution of the minimization task differs from a non regularized one merely in coefficients of linear combinations of computational units. Thus representation of learning as inverse problems provides a useful tool for theoretical investigation of properties of kernel and radial-basis networks. It characterizes optimal input-output functions of these computational models and enables to estimate effects of regularization.

The paper is organized as follows. Section 2 gives basic concepts and notations on learning from data. In section 3, basic terminology and tools from theory of inverse problems are introduced. In section 4, these tools are applied to description of theoretically optimal input-output functions in learning from data over networks with kernel units.

2 ERROR FUNCTIONALS WITH QUADRATIC LOSS FUNCTIONS

In statistical learning theory, learning from data has been modeled as a search for a function minimizing the expected error functional defined by data described by a probability measure. For X a compact subset of $\mathbb{R}^d$ and Y a bounded subset of $\mathbb{R}$, let ρ be a non degenerate (no nonempty open set has measure zero) probability measure on $Z = X \times Y$. The *expected error functional* (sometimes also called expected risk or theoretical error) determined by ρ is defined for every f in the set $\mathcal{M}(X)$ of all bounded ρ-measurable functions on X as $\mathcal{E}_{\rho,V}(f) = \int_Z V(f(x),y)\,d\rho$, where $V : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ is a *loss function*. The most common loss function is the *quadratic loss* defined as $V(u,v) = (u-v)^2$. We shortly denote by $\mathcal{E}_\rho$ the expected error with the quadratic loss, i.e.,

$$\mathcal{E}_\rho(f) = \int_Z (f(x)-y)^2\,d\rho\,.$$

Learning algorithms use a discretized version of the expected error called the *empirical error*. It is determined by a training sample $z = \{(x_i,y_i) \in \mathbb{R}^d \times \mathbb{R} \,|\, i = 1,\dots,m\}$ of input-output pairs of data and a discrete probability measure $p = \{p(i) \,|\, i = 1,\dots,m\}$ on the set $\{1,\dots,m\}$ (i.e., $\sum_{i=1}^m p(i) = 1$). Empirical error is denoted $\mathcal{E}_{z,p,V}$ and defined as $\mathcal{E}_{z,p,V}(f) = \sum_{i=1}^m p(i)V(f(x_i),y_i)$. Similarly as in the case of expected error, we denote by $\mathcal{E}_{z,p}$ the empirical error with the quadratic loss function, i.e.,

$$\mathcal{E}_{z,p}(f) = \sum_{i=1}^m p(i)\,(f(x_i)-y_i)^2.$$

One of many advantages of the quadratic loss function is that it enables to reformulate minimization of expected and empirical error as minimization of distances from certain “optimal” functions.

It is easy to see and well-known (Cucker and Smale, 2002) that the minimum of $\mathcal{E}_\rho$ over the set $\mathcal{M}(X)$ of all bounded ρ-measurable functions on X is achieved at the *regression function* f_ρ defined for $x \in X$ as $f_\rho(x) = \int_Y y\,d\rho(y|x)$, where $\rho(y|x)$ is the *conditional (w.r.t. x) probability measure* on Y.

Let ρ_X denote the *marginal probability measure* on X defined for every $S \subseteq X$ as $\rho_X(S) = \rho(\pi_X^{-1}(S))$, where $\pi_X : X \times Y \to X$ denotes the projection, and let $\mathcal{L}^2_{\rho_X}(X)$ denote the Lebesgue space of all functions on X satisfying $\int_X f^2 d\rho_X < \infty$ with the $\mathcal{L}^2_{\rho_X}$-norm denoted by $\|.\|_{\mathcal{L}^2}$. It can be easily verified that $f_\rho \in \mathcal{L}^2_{\rho_X}(X)$. So $\min_{f\in\mathcal{M}(X)} \mathcal{E}_\rho(f) = \min_{f\in\mathcal{L}^2_{\rho_X}(X)} \mathcal{E}_\rho(f_\rho) = \mathcal{E}_\rho(f_\rho) = \sigma^2_\rho$. Moreover, for every $f \in \mathcal{L}^2_{\rho_X}(X)$ (Cucker and Smale, 2002, p.5)

$$\mathcal{E}_\rho(f) = \int_X (f(x) - f_\rho(x))^2 d\rho_X + \sigma^2_\rho = \|f - f_{\rho_X}\|^2_{\mathcal{L}^2} + \sigma^2_\rho. \tag{1}$$

So on the function space $\mathcal{L}^2_{\rho_X}(X)$, the expected error functional $\mathcal{E}_\rho$ with the quadratic loss can be represented as the square of the $\mathcal{L}^2_{\rho_X}$-distance from its minimum point f_ρ.

Also the empirical error functional can be represented in terms of a distance from a certain function. Let $X_z = \{x_1,\dots,x_m\}$ with all x_i distinct and $h_z : X_z \to Y$ be defined as $h_z(x_i) = y_i$. Let $\|.\|_{2,m}$ denote the weighted l^2-norm on $\mathbb{R}^m$ defined as $\|x\|^2_{2,m} = \frac{1}{\sqrt{m}}\sum_{i=1}^m p(i)x_i^2$. Then

$$\mathcal{E}_z(f) = \|f_{|X_z} - h_z\|^2_{2,m}. \tag{2}$$

So minimization of the empirical error $\mathcal{E}_{z,p}$ is a search for a function, the restriction on X_z of which has a smallest l^2_p-distance from the function h_z defined by the sample z.

3 INVERSE PROBLEMS IN LEARNING

The representations (1) and (2) allow us to use in learning theory methods and tools from theory of inverse problems. For a *linear operator* $A : \mathcal{X} \to \mathcal{Y}$ between two Hilbert spaces $(\mathcal{X}, \|.\|_{\mathcal{X}})$, $(\mathcal{Y}, \|.\|_{\mathcal{Y}})$ (in finite-dimensional case, a matrix A) an *inverse problem* (see, e.g., (Bertero, 1989)) determined by A is to find for $g \in \mathcal{Y}$ (called *data*) some $f \in \mathcal{X}$ (called *solution*) such that

$$A(f) = g.$$

If for every $g \in \mathcal{Y}$ there exists a unique solution $f \in \mathcal{X}$, which depends continuously on data, then the inverse problem is called *well-posed*. So for a well-posed inverse problem, there exists a unique inverse operator $A^{-1} : \mathcal{Y} \to \mathcal{X}$. When A is continuous, then by the Banach open map theorem (Friedman, 1982, p.141) A^{-1} is continuous, too. However, a continuous dependence of solutions on data may not always guarantee robustness against a noise. Stability has been measured by behavior of eigenvalues of A and the *condition number* defined for a well-posed problem given by an operator A as $\mathrm{cond}(A) = \|A\| \, \|A^{-1}\|$. Problems with large condition numbers are called *ill-conditioned*.

When for some $g \in \mathcal{Y}$ no solution exists, at least one can search for a *pseudosolution* f^o, for which $A(f^o)$ is a best approximation to g among elements of the range of A, i.e.,

$$\|A(f^o) - g\|_{\mathcal{Y}} = \min_{f \in \mathcal{X}} \|A(f) - g\|_{\mathcal{Y}}.$$

Theory of inverse problems overcomes ill-posedness by using so called normal pseudosolutions instead of solutions and in addition it also overcomes ill-conditioning by using various regularized solutions. Tikhonov's regularization (Tikhonov and Arsenin, 1977) replaces the problem of minimization of the functional $\|A(.) - g\|^2_{\mathcal{Y}}$ with minimization of

$$\|A(.) - g\|^2_{\mathcal{Y}} + \gamma\Psi,$$

where Ψ is a functional called *stabilizer* and the *regularization parameter* γ plays the role of a trade-off between an emphasis on a proximity to data and a penalization of undesired solutions expressed by Ψ. A typical choice of a stabilizer is the square of the norm on $\mathcal{X}$, for which Tikhonov regularization minimizes the functional

$$\|A(.) - g\|^2_{\mathcal{Y}} + \gamma\|.\|^2_{\mathcal{X}}. \qquad (3)$$

Let $(\mathcal{H}, \|.\|_{\mathcal{H}})$ be a Hilbert space, which is a linear subspace of $\mathcal{L}^2_{\rho_X}(X)$ with a possibly different norm than the one obtained by restriction of $\|.\|_{\mathcal{L}^2_{\rho_X}}$ to $\mathcal{H}$, and let $J : (\mathcal{H}, \|.\|_{\mathcal{H}}) \to (\mathcal{L}^2_{\rho_X}(X), \|.\|_{\mathcal{L}^2_{\rho_X}})$ denote the inclusion operator. By the representation (1), we have

$$\mathcal{E}_\rho(f) = \|f - f_\rho\|^2_{\mathcal{L}^2_{\rho_X}} + \sigma^2_\rho = \|J(f) - f_\rho\|^2_{\mathcal{L}^2_{\rho_X}} + \sigma^2_\rho. \qquad (4)$$

So the problem of minimization of $\mathcal{E}_\rho$ over $\mathcal{H}$ is equivalent to the inverse problem defined by the *inclusion operator J* for the data f_ρ.

To reformulate minimization of $\mathcal{E}_{z,p}$ as an inverse problem, define for the input sample $x = (x_1, \dots, x_m)$ an *evaluation operator* $L_x : (\mathcal{H}, \|.\|_{\mathcal{H}}) \to (\mathbb{R}^m, \|.\|_{2,m})$ as

$$L_x(f) = (f(x_1), \dots, f(x_m)).$$

It is easy to check that for every $f : X \to \mathbb{R}$,

$$\mathcal{E}_{z,m}(f) = \sum_{i=1}^{m} p(i)(f(x_i) - y_i)^2 = \|L_x(f) - y\|^2_{2,m}. \qquad (5)$$

4 PSEUDOSOLUTIONS AND REGULARIZED SOLUTIONS

Originally, properties of pseudoinverse and regularized inverse operators were described for operators between finite dimensional spaces, where such operators can be represented by matrices (Moore, 1920; Penrose, 1955). In 1970s, the theory of pseudoinversion was extended to the infinite-dimensional case – it was shown that similar properties as the ones of Moore-Penrose pseudoinverses of matrices also hold for pseudoinverses of *continuous linear operators* between Hilbert spaces (Groetch, 1977). The reason is that continuous operators have *adjoint operators* $A^* : \mathcal{Y} \to \mathcal{X}$ satisfying the equation $\langle A(f), g\rangle_{\mathcal{Y}} = \langle f, A^*(g)\rangle_{\mathcal{X}}$. These adjoints play an important role in a characterization of pseudosolutions and regularized solutions. In the next section, we will see that that for proper function spaces, adjoints of evaluation and inclusion operators used in representations (4) and (5) can be easily described.

First we recall some basic results from theory of inverse problems from (Bertero, 1989, pp.68-70) and (Groetch, 1977, pp.74-76). For every *continuous linear operator* $A : (\mathcal{X}, \|.\|_{\mathcal{X}} \to (\mathcal{Y}, \|.\|_{\mathcal{Y}})$ between two Hilbert spaces there exists a unique continuous linear pseudoinverse operator $A^+ : \mathcal{Y} \to \mathcal{X}$ (when the range $R(A)$ is closed, otherwise A^+ is defined only for those $g \in \mathcal{Y}$, for which $\pi_{clR(A)}(g) \in R(A)$). The pseudoinverse A^+ satisfies for every $g \in \mathcal{Y}$, $\|A^+(g)\|_{\mathcal{X}} = \min_{f^o \in S(g)} \|f^o\|_{\mathcal{X}}$, where $S(g) = \mathrm{argmin}(\mathcal{X}, \|A(.) - g\|_{\mathcal{Y}})$, for every $g \in \mathcal{Y}$, $AA^+(g) = \pi_{clR}(g)$, and

$$A^+ = (A^*A)^+A^* = A^*(AA^*)^+. \qquad (6)$$

Moreover, for every $\gamma > 0$, there exists a unique operator

$$A^{\gamma} : \mathcal{Y} \to \mathcal{X}$$

such that for every $g \in \mathcal{Y}$, $\{A^{\gamma}(g)\} = \operatorname{argmin}(\mathcal{X}, \|A(.) - g\|^2_{\mathcal{Y}} + \gamma\|.\|^2_{\mathcal{X}})$ and

$$A^{\gamma} = (A^*A + \gamma I_{\mathcal{X}})^{-1}A^* = A^*(AA^* + \gamma I_{\mathcal{Y}})^{-1} \qquad (7)$$

where $I_{\mathcal{X}}, I_{\mathcal{Y}}$ denote the identity operators. For every $g \in \mathcal{Y}$, for which $A^+(g)$ exists, $\lim_{\gamma \to 0} A^{\gamma}(g) = A^+(g)$.

Formulas (6) and (7) combined with representations (4) and (5) provide useful tools for description of optimal input-output functions in learning from data. However, there is a restriction on computational models requiring that input-output functions belong to suitable function spaces with an inner product where evaluation functionals are continuous. The space $\mathcal{L}^2_{\rho_X}(X)$ is a Hilbert space (it has an inner product) but it is easy to see that evaluation functionals on this space are not continuous (it contains sequences of functions with the same norm with diverging evaluations at zero). This space is too large, but it contains suitable subspaces formed by functions with limited oscillations. These spaces contain input-output functions computable by networks with kernel and radial units, in particular Gaussian radial-basis networks with any fixed width. They have became popular due to the use of kernels in support vector machines, but they were studied in mathematics since 1950 (Aronszajn, 1950). Since 1990s, they were considered as useful ambient function spaces in data analysis (Wahba, 1990). These spaces are called *Reproducing Kernel Hilbert Spaces* (RKHS) because each such space is uniquely determined by a symmetric positive semidefinite kernel $K : X \times X \to \mathbb{R}$. For their theory and applications see, e.g., (Schölkopf and Smola, 2002). Here we just recall that a RKHS determined by K, denoted $\mathcal{H}_K(X)$, contains all linear combinations of functions of the form $K(.,v) : X \to \mathbb{R}$, $v \in X$, defined as $K(.,v)(u) = K(u,v)$ (these functions are called *representers* and they are generators of the linear space $\mathcal{H}_K(X)$, see, e.g., (Wahba, 1990; Cucker and Smale, 2002)). The RKHS $\mathcal{H}_K(X)$ is endowed with an inner product defined on generators as $\langle K_u, K_v \rangle_K = K(u,v)$, which induces the norm $\|K_u\|^2_K = K(u,u)$.

A paradigmatic example of a kernel is the *Gaussian kernel* $K(u,v) = e^{-\|u-v\|^2}$. A RKHS defined by the Gaussian kernel contains all linear combinations of translations of Gaussians, so it contains input-output functions of radial-basis networks with Gaussian radial functions with a fixed width.

The role of kernel norms as stabilizers in Tikhonov's regularization (3) can be intuitively well understood in the case of *convolution kernels*, i.e., kernels $K(x,y) = k(x-y)$ defined as translations of a function $k : \mathbb{R}^d \to \mathbb{R}$, for which the Fourier transform $\tilde{k}$ is positive. For such kernels, the value of the stabilizer $\|.\|^2_K$ at any $f \in \mathcal{H}_K(\mathbb{R}^d)$ can be expressed as

$$\|f\|^2_K = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \frac{\tilde{f}(\omega)^2}{\tilde{k}(\omega)} d\omega.$$

So when $\lim_{\|\omega\| \to \infty} 1/\tilde{k}(\omega) = \infty$, the stabilizer $\|.\|^2_K$ plays a role of a high-frequency filter. Examples of convolution kernels with positive Fourier transforms are the Gaussian and Bessel kernel. Note that for any convolution kernel K with $k(0) = 1$, all functions $f = \sum_{i=1}^m w_i K_{u_i}$, which are computable by one-hidden layer networks with units computing translations of k, satisfy $\|f\|_K \leq \sum_{i=1}^m |w_i| \, \|K_{u_i}\|_K = \sum_{i=1}^m |w_i| k(0) = \sum_{i=1}^m |w_i|$. So the output-weight regularization widely used for its simplicity in practical applications guarantees a decrease of the stabilizer $\|.\|^2_K$ and thus penalizes solutions with high-frequency oscillations.

5 MINIMA OF ERROR FUNCTIONALS

Reformulations (4) and (5) of minimizations of error functionals as inverse problems together with equations (6) and (7) allow us to describe properties of theoretically optimal solutions of minimization of error functionals with the quadratic loss.

For a kernel K, we denote by $L_K : \mathcal{L}^2_{\rho_X}(X) \to \mathcal{L}^2_{\rho_X}(X)$ the integral operator defined as

$$L_K(f)(y) = \int_X f(x) K(y,x) d\rho_X(x).$$

The next theorem describes minimima of the expected error $\mathcal{E}_\rho$ and of its Tikhonov regularization $\mathcal{E}_{\rho,\gamma,K} = \mathcal{E}_\rho + \gamma\|.\|^2_K$. It shows that for every $\gamma > 0$ there exists a unique function f^{γ} minimizing $\mathcal{E}_{\rho,\gamma,K}$ and that this function is the image of the regression function f_ρ under an integral operator L_{K_γ} with a modified kernel K_γ, defined as $K_\gamma(x,y) = \sum_{i=1}^{\infty} \frac{\lambda_i}{\lambda_i + \gamma} \phi_i(x)\phi_i(y)$, where λ_i and ϕ_i are the eigenvalues of the operator L_K.

Theorem 1. *Let $X \subset \mathbb{R}^d$ be compact, $Y \subset \mathbb{R}$ be bounded, $K : X \times X \to \mathbb{R}$ be a continuous symmetric positive definite kernel, ρ be a non degenerate probability measure on $X \times Y$. Then*
(i) if K is degenerate, then the inclusion operator $J_K : (\mathcal{H}_K(X), \|.\|_K) \to (\mathcal{L}^2_{\rho_X}(X), \|.\|_{\mathcal{L}^2})$ has a pseudoinverse operator $J^+_K : (\mathcal{L}^2_{\rho_X}(X), \|.\|_{\mathcal{L}^2}) \to (\mathcal{H}_K(X), \|.\|_K)$ such that for every $g \in \mathcal{L}^2_{\rho_X}(X)$, $J^+_K(g) = \pi_K(g)$, where π_K is the projection on $\mathcal{H}_K(X)$,

while if K is non degenerate, then the pseudoinverse operator J_K^+ is only defined on $R(J_K) = \mathcal{H}_K(X)$ and for all $g \in \mathcal{H}_K(X)$, $J_K^+(g) = g$.
(ii) for every $\gamma > 0$, there exists a unique function $f^\gamma \in \mathcal{H}_K(X)$ minimizing $\mathcal{E}_{\rho,\gamma,K}$, $f^\gamma = L_{K_\gamma}(f_{\rho_X})$, $\lim_{\gamma\to 0} \|f^\gamma - f_\rho\|_{\mathcal{L}^2} = 0$, and $\lim_{\gamma\to 0} \mathcal{E}_\rho(f^\gamma) = \mathcal{E}_\rho(f_\rho)$.

The essential part of the proof of this theorem is showing that the adjoint J_K^* of the inclusion

$$J_K : (\mathcal{H}_K(X), \|.\|_K) \to (\mathcal{L}^2_{\rho_X}(X), \|.\|_{\mathcal{L}^2_{\rho_X}})$$

is the integral operator L_K. This allows application of the equations (6) and (7). The formula

$$f_\gamma = L_{K_\gamma}(f_\rho) \tag{8}$$

describing the regularized solution can be called the Representer Theorem in analogy to a similar result describing the minimum point of the regularized empirical error functional. This theorem was derived in (Cucker and Smale, 2002, p.42), where it was formulated as

$$f_\gamma = (\mathcal{I} + \gamma L_K^{-1}) f_\rho. \tag{9}$$

However, the formulation (9) might be misleading as the inverse L_K^{-1} to L_K is defined only on a subspace of $\mathcal{L}^2_{\rho_X}(X)$. Note that it cannot be defined on any complete subspace of $\mathcal{L}^2_{\rho_X}(X)$ because in such a case by Banach open map theorem it should be bounded, but for a non degenerate kernel K, the eigenvalues $\frac{1}{\lambda_i}$ of the inverse L_K^{-1} diverge. Here we derived the description of the regularized solution (8) easily as a straightforward consequence of well-known properties of Tikhonov's regularization, while in (Cucker and Smale, 2002, pp.27-28) the formula (9) was derived using results on operators with fractional powers.

By Theorem 1, when the regression function f_{ρ_X} is not in the function space $\mathcal{H}_K(X)$ defined by the kernel K, then $\mathcal{E}_\rho$ does not achieve a minimum on $\mathcal{H}_K(X)$. However, for every regularization parameter $\gamma > 0$, the regularized functional $\mathcal{E}_{\rho,\gamma,K}$ achieves a unique minimum equal to $L_{K_\gamma}(f_\rho)$. Theorem 1 also shows how regularization modifies coefficients $\{w_i\}$ in the representation of the regression $f_{\rho_X} = \sum_{i=1}^\infty w_i \phi$ in terms of the basis $\{\phi_i\}$ formed by eigenvalues of the operator L_K. Regularization replaces these coefficients with coefficients $\{\frac{w_i \lambda_i}{\lambda_i + \gamma}\}$. The function $\alpha(i) = \frac{w_i \lambda_i}{\lambda_i + \gamma}$ is decreasing to 0, so the higher frequency coefficients are more reduced. With the regularization parameter γ decreasing to zero, the coefficients $\frac{w_i \lambda_i}{\lambda_i + \gamma}$ converge to w_i and so the regularized solutions f_γ converge in $\mathcal{L}^2_{\rho_X}$-norm to the regression function f_ρ.

The next theorem describes minima od empirical error functional and its regularized modification. We use the notation

$$\mathcal{E}_{z,p,\gamma,K} = \mathcal{E}_{z,p} + \gamma \|.\|_K^2.$$

By $\mathcal{K}[x]$ is denoted the *Gram matrix of the kernel K with respect to the vector x*, i.e., the matrix

$$\mathcal{K}[x]_{i,j} = K(x_i, x_j),$$

by $\mathcal{K}_m[x]$ the matrix $\frac{1}{m}\mathcal{K}[x]$, and by $\mathcal{I}$ the identity $m \times m$ matrix.

Theorem 2. *Let $K : X \times X \to \mathbb{R}$ be a symmetric positive semidefinite kernel, m be a positive integer, $z = (x,y)$ with $x = (x_1, \dots, x_m) \in X^m$, $y = (y_1, \dots, y_m) \in \mathbb{R}^m$, with $x_1, \dots, x_m$ distinct, and p be a discrete probability measure on $\{1, \dots, m\}$, then*
(i) $f^+ = L_x^+(y)$ is a minimum point of $\mathcal{E}_{z,p}$ at $\mathcal{H}_K(X)$, $\|f^+\|_K \le \|f^o\|_K$ for all $f^o \in \operatorname{argmin}(\mathcal{H}_K(X), \mathcal{E}_{z,p})$, and $f^+ = L_x^+(y) = \sum_{i=1}^m c_i K_{x_i}$, where $c = (c_1, \dots, c_m) = \mathcal{K}[x]^+ y$,
(ii) for all $\gamma > 0$, there exists a unique f^γ minimizing $\mathcal{E}_{z,p,\gamma,K}$ over $\mathcal{H}_K(X)$, $f^\gamma = \sum_{i=1}^m c_i^\gamma K_{x_i}$, where $c^\gamma = (\mathcal{K}_m[x] + \gamma \mathcal{I})^{-1} y$.

The essential part of the proof of this theorem is description of the adjoint $L_x^* : (\mathbb{R}^m, \|.\|_{2,m}) \to (\mathcal{H}_K(X), \|.\|_K)$ as $L_x^*(u) = \frac{1}{m}\sum_{i=1}^m u_i K_{x_i}$. This leads to representation of $L_x L_x^* : \mathbb{R}^m \to \mathbb{R}^m$ by the matrix $\mathcal{K}_m[x]$, which together with equations (6) and (7) gives the description of functions minimizing the empirical error and its regularization.

Theorem 2 shows that for every kernel K, every sample of empirical data z and discrete probability measure p, there exists a function f^+ minimizing the empirical error functional $\mathcal{E}_{z,p}$ over the whole RKHS defined by K. This function is formed by a linear combination of the representers $K_{x_1}, \dots, K_{x_m}$ of input data x_i, i.e.,it has the form

$$f^+ = \sum_{i=1}^m c_i K_{x_i}. \tag{10}$$

Thus f^+ can be interpreted as an *input-output function of a neural network with one hidden layer of kernel units and a single linear output unit*. The coefficients $c = (c_1, \dots, c_m)$ of the linear combination (corresponding to network output weights) satisfy $c = \mathcal{K}[x]^+ y$, so the output weights can be obtained by solving the system of linear equations. However, as the operator L_x has finite dimensional range, it is compact and thus its pseudoinverse is unbounded. So the optimal solution of minimization od empirical error is unstable. The regularized solution

$$f^\gamma = \sum_{i=1}^m c_i^\gamma K_{x_i}$$

is also a linear combination of functions $K_{x_1},\dots,K_{x_m}$. But the coefficients of these two linear combinations are different: in the regularized case $c^\gamma = (\mathcal{K}[x] + \gamma \mathcal{I})^{-1} y$, while in the non-regularized one $c = \mathcal{K}[x]^+ y$.

The characterization of regularized solution minimization of empirical error in reproducing kernel Hilbert spaces was derived in (Wahba, 1990) using Fréchet derivatives (see also (Cucker and Smale, 2002), (Poggio and Smale, 2003)). Our proof based on characterization of the adjoint L_x^* of the evaluation operator L_x is much simpler and it also include the non regularized case and thus it shows the effect of regularization.

Theorem 2 shows that increase of "smoothness" of the regularized solution f^γ is achieved by merely changing the coefficients of the linear combination. In the non regularized case, the coefficients are obtained from the output data vector y using the Moore-Penrose pseudoinverse of the Gram matrix $\mathcal{K}[x]$, while in the regularized one, they are obtained using the inverse of the modified matrix $\mathcal{K}[x] + \gamma \mathcal{I}$. So the regularization merely changes amplitudes, but it preserves the finite set of basis functions from which the solution is composed.

In many practical applications, there are used networks with much smaller number n of units than the size of the training sample of data m. However, characterization of theoretically optimal solutions achievable over networks with large numbers of units (equal to the sizes m of training data) can be useful in investigation on dependence of quality of approximation of such optimal solutions by suboptimal ones obtainable over smaller models (Vito et al., 2005; Kůrková and Sanguineti, 2005a; Kůrková and Sanguineti, 2005b).

As mentioned above, for convolution kernels we have $\|f^\gamma\|_K \leq \sum_{i=1}^m |c_i^\gamma|$. Instead of calculating $\|.\|_K^2$ norm, it is easier to use as a stabilizer the ℓ_1-norm of an output weight vector. For linear combinations of functions of the form K_x, this also leads to minimization of $\|.\|_K^2$ norm.

ACKNOWLEDGEMENTS

This work was partially supported by MŠMT grant COST Intelli OC10047 and the Institutional Research Plan AV0Z10300504.

REFERENCES

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of AMS*, 68:337–404.

Bertero, M. (1989). Linear inverse and ill-posed problems. *Advances in Electronics and Electron Physics*, 75:1–120.

Bishop, C. (1995). Training with noise is equivalent to Tikhonov regularization. *Neural Computation*, 7(1):108–116.

Cucker, F. and Smale, S. (2002). On the mathematical foundations of learning. *Bulletin of AMS*, 39:1–49.

Engl, E. W., Hanke, M., and Neubauer, A. (1999). *Regularization of Inverse Problems*. Kluwer, Dordrecht.

Fine, T. L. (1999). *Feedforward Neural Network Methodology*. Springer-Verlag, Berlin, Heidelberg.

Friedman, A. (1982). *Modern Analysis*. Dover, New York.

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

Girosi, F. and Poggio, T. (1990). Regularization algorithms for learning that are equivalent to multilayer networks. *Science*, 247(4945):978–982.

Groetch, C. W. (1977). *Generalized Inverses of Linear Operators*. Dekker, New York.

Hansen, P. C. (1998). *Rank-Deficient and Discrete Ill-Posed Problems*. SIAM, Philadelphia.

Ito, Y. (1992). Finite mapping by neural networks and truth functions. *Mathematical Scientist*, 17:69–77.

Kecman, V. (2001). *Learning and Soft Computing*. MIT Press, Cambridge.

Kůrková, V. and Sanguineti, M. (2005a). Error estimates for approximate optimization by the extended Ritz method. *SIAM Journal on Optimization*, 15:461–487.

Kůrková, V. and Sanguineti, M. (2005b). Learning with generalization capability by kernel methods with bounded complexity. *Journal of Complexity*, 13:551–559.

Michelli, C. A. (1986). Interpolation of scattered data: Distance matrices and conditionally positive definite functions. *Constructive Approximation*, 2:11–22.

Moore, E. H. (1920). Abstract. *Bulletin of AMS*, 26:394–395.

Penrose, R. (1955). A generalized inverse for matrices. *Proceedings of Cambridge Philosophical Society*, 51:406–413.

Poggio, T. and Smale, S. (2003). The mathematics of learning: dealing with data. *Notices of AMS*, 50:537–544.

Schölkopf, B. and Smola, A. J. (2002). *Learning with Kernels – Support Vector Machines, Regularization, Optimization and Beyond*. MIT Press, Cambridge.

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

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

Wahba, G. (1990). *Splines Models for Observational Data*. SIAM, Philadelphia.