Supervised Learning as an Inverse Problem

Kůrková, Věra
2004

Institute of Computer Science
Academy of Sciences of the Czech Republic

Supervised learning as an inverse problem

Věra Kůrková

Technical report No. 906

April 2004

Pod Vodárenskou věží 2, 182 07 Prague 8, phone: (+4202) 6605 1111, fax: (+4202) 86 585 789,
e-mail:ics@cs.cas.cz

Institute of Computer Science
Academy of Sciences of the Czech Republic

Supervised learning as an inverse problem[1]

Věra Kůrková[2]

Technical report No. 906

April 2004

Abstract:

Keywords:
Learning from data, generalization, minimization of empirical error, regularization, kernel methods.

[1]This work was partially supported by GA ČR grant 201/02/0428. The author thanks to P. C. Kainen from Georgetown University and M. Sanguineti from Universita di Genova for fruitful comments and discussions.

[2]Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 2, 182 07 Prague 8, Czech Republic, E-mail: vera@cs.cas.cz

1 Introduction

The goal of supervised learning is to adjust parameters of a neural network so that it approximates with a sufficient accuracy a functional relationship between inputs and outputs. Typically, such a relationship is not known analytically. Instead, a training set is given consisting of a sample of input/output pairs $z = \{(x_i, y_i) \in \mathcal{R}^d \times \mathcal{R}, i = 1, \ldots, m\}$. So the task of learning is to find a function from a hypothesis set formed by functions computable by a given class of neural networks that approximates the sample of *empirical data*. A similar task of finding a function fitting to astronomical data was solved by Gauss and Legendre by the *least square method*, i.e., minimization of the sum of squares of errors [4]. The least square method became popular in statistics and engineering and was also used in many neural network learning algorithms such as backpropagation.

The problem of finding a function from a given parameterized family fitting to empirical data belongs to a wider class of *inverse problems* of determining unknown causes (such as shapes of functions, forces, shapes of distributions) from known consequences (empirical data). Inverse problems are fundamental in various domains of applied science such as medical diagnostics (tomography), seismology and meteorological forecasting. The dependence of consequences on causes is usually modelled by an operator, the simplest type of which is linear. For finite dimensional case, the theory of linear inverse problems is based on *Moore-Penrose pseudoinverse* of a matrix. Pseidoinverse method was generalized to infinite dimensional Hilbert spaces [8], [3] and combined with *regularization* introduced by Tikhonov and Arsenin [16] to develop a theory describing properties of least-squares pseudosolutions, their stability and their relationship to regularized solutions [3]. Modelling of generalization based on Tikhonov's regularization was introduced by Poggio and Girosi [14]. Later Girosi [7] considered regularization in the domain of a special class of Hilbert spaces, called *reproducing kernel Hilbert spaces* (RKHS), the norms on which can play a role of measures of various types of oscillations and thus enable to model a variety of *conceptual data*, which has to be added to the empirical ones to guarantee generalization capability. RKHS, defined by Aronszajn [1], were introduced into interpolation of data by Parzen [13] and Wahba [17]. For a survey of applications of RKHS to learning see, e.g., [5], [15].

In this paper, we reformulate the problem of minimization of an empirical error functional as a linear inverse problem by introducing a suitable operator. We describe properties of this operator (compactness, representation of its adjoint) and apply theory of continuous linear inverse problems in the domain of infinite dimensional Hilbert spaces. We describe relationship between a pseudosolution and regularized solutions for variable regularization parameters and analyze improvements of stability that can be obtained by regularization in terms of condition numbers of Gram matrices and size of data samples.

2 Minimization of empirical error as an inverse problem

Let Ω be a nonempty set, m a positive integer and $z = \{(x_i, y_i) \in \Omega \times \mathcal{R}, i = 1, \ldots, m\}$ be a sample of pairs of data. A standard approach to learning from data used , e.g., in backpropagation is based on minimization of the *empirical error* functional defined as $\mathcal{E}_{z,V}(f) = \frac{1}{m}\sum_{i=1}^{m} V(f(x_i), y_i)$, where $V : \mathcal{R}^2 \to [0, \infty)$ satisfying $V(y, y) = 0$ for all $y \in \mathcal{R}$ is a *loss function* that measures how much is lost when $f(x)$ is computed instead of y. The most common loss function is the *square loss* $V(f(x), y) = (f(x) - y)^2$. To simplify notation, we denote by $\mathcal{E}_z$ the empirical error functional with the square loss function, i.e.,

$$\mathcal{E}_z(f) = \frac{1}{m}\sum_{i=1}^{m}(f(x_i) - y_i)^2.$$

Using a standard terminology from the theory of optimization we denote by (M, Φ) the problem of minimization of a functional Φ over a set M, which is called a *hypothesis set*. Every $f^o \in M$ such that $\Phi(f^o) = \min_{f \in M} \Phi(f)$ is called a *solution* of the problem (M, Φ). We denote by $argmin(M, \Phi) = \{f^o \in M : \Phi(f^o) = \min_{f \in M} \Phi(f)\}$ the set of all solutions of (M, Φ).

The problem of minimization of the empirical error functional can be studied in the framework of theory of inverse problems. Given an operator $A : (X, \|.\|_X) \to (Y, \|.\|_Y)$ between Banach spaces, an *inverse problem* defined by A is to find for $g \in Y$ some $f \in X$ such that $A(f) = g$ [3]. An inverse

problem is called *linear* when A is a linear operator. Elements of X are called *solutions* and elements of Y *data*. When Y is finite dimensional, the inverse problem is called a problem with *discrete data*.

If for every $g \in Y$ there exists a unique solution $f \in X$, then the inverse problem is called *well-posed*. So for a well-posed inverse problem, there exists a unique inverse operator $A^{-1} : Y \to X$. When A is continuous, then by the Banach open map theorem [6, p.141] A^{-1} is continuous, too. Even a continuous dependence of solutions on data does not always guarantee robustness against a noise. As a measure of stability of solutions of an inverse problem is used the *condition number* defined for a well-posed problem given by an operator A as $cond(A) = \|A\| \, \|A^{-1}\|$.

Often, inverse problems are ill-posed or ill-conditioned. When a solution does not exist, one can search for best approximate solution f^o, called a *pseudosolution*, defined by $\|A(f^o) - g\|_Y = \min_{f \in X} \|A(f) - g\|_Y$ and a *normal pseudosolution* f^+, which is a pseudosolution of the minimal norm, i.e., $\|f^+\|_X = \min\{\|f^o\|_X : f^o \in S(g)\}$, where $S(g)$ is the set of all psedosolutions of the inverse problem given by an operator A and data g. When for every $g \in Y$ there exists a normal pseudosolution f^+, then a *pseudoinverse operator* $A^+ : Y \to X$ can be defined as $A^+(g) = f^+$. Similarly as in the case of well-posed problems, the *condition number* of an operator A with a pseudoinverse A^+ is defined as $cond(A) = \|A\| \, \|A^+\|$.

For X and Y finite dimensional, the pseudosolution can be described in terms of Moore-Penrose pseudoinverse of the matrix corresponding to the operator A. The concept of Moore-Penrose pseudoinversion has been extended to the case of linear continuous operators between Hilbert spaces [8]. To take advantage of the theory of generalized inversion in Hilbert spaces, we express as an inverse problem the problem $(X, \mathcal{E}_z)$ of minimization of the empirical error $\mathcal{E}_z$ over a Hilbert space X of functions on some set Ω. Let $z = (x, y)$, where $x = (x_1, \ldots, x_m) \in \Omega^m$ and $y = (y_1, \ldots, y_m) \in \mathcal{R}^m$, be a sample defining the empirical error functional $\mathcal{E}_z$. Consider an operator $L_x : X \to \mathcal{R}^m$ defined as

$$L_x(f) = \left(\frac{f(x_1)}{\sqrt{m}}, \ldots, \frac{f(x_m)}{\sqrt{m}} \right).$$

Then $\mathcal{E}_z$ can be represented as

$$\mathcal{E}_z = \left\| L_x - \frac{y}{\sqrt{m}} \right\|_2^2, \tag{2.1}$$

where $\|.\|_2$ denotes the l_2-norm on $\mathcal{R}^m$. Similarly, $\langle . \rangle_2$ denotes the inner product on $\mathcal{R}^m$, while $\|.\|_X$ and $\langle . \rangle_X$ denote the norm and the inner product, resp., on X.

Thus the problem of minimization of $\mathcal{E}_z$ over X is equivalent to the problem of finding a pseudosolution of the inverse problem given by the operator L_x for the data $\frac{y}{\sqrt{m}}$. As the range of the operator L_x is finite dimensional, this problem belongs to the class of problems with discrete data. When $(X, \|.\|_X)$ is chosen in such a way that L_x is continuous, we can apply the following theorem summarizing properties of the pseudosolution of a continuous linear operator stated in [3, pp. 56-60] and in [8, pp.37-46].

For any operator $A : X \to Y$, we denote by $N(A) = \{f \in X : A(f) = 0\}$ its *null space*, by $R(A) = \{g \in Y : (\exists f \in X)(A(f) = g)\}$ its *range*, by $\pi_R : Y \to R(A)$ the projection of Y onto $R(A)$ and if A has an adjoint A^*, by $\pi_N : Y \to N(A^*)$ the projection of Y onto the null space of A^*. For any $g \in Y$, we denote $S(g) = \{f^o \in X : \|A(f^o) - g\|_Y = \min_{f \in X} \|A(f) - g\|_Y\}$.

Theorem 2.1 *Let X, Y be Hilbert spaces, $A : X \to Y$ be a continuous linear operator with a closed range, then:*
(i) A has an adjoint A^*;
(ii) $R(A)$ is closed and $N(A^*) \oplus R(A) = Y$;
(iii) there exists a unique continuous linear operator $A^+ : Y \to X$ such that for every $g \in Y$, $A^+(g) \in S(g)$, $\|A^+(g)\|_X = \min_{f^o \in S(g)} \|f^o\|_X$ and $S(g) = \{A^+(g) + f : f \in N(A)\}$;
(iv) for every $g \in Y$, $AA^+(g) = \pi_R(g)$;
(v) $A^+ = (A^*A)^+A^* = A^*(AA^*)^+$.

3 Minimization of empirical errors over reproducing kernel Hilbert spaces

To apply Theorem 2.1 to learning from data we need to find proper hypothesis spaces (formed by functions defined on some sets Ω), on which the operators L_x are continuous for all $x = (x_1, \ldots, x_m) \in \Omega^m$. $(\mathcal{L}_2(\Omega), \|.\|_{\mathcal{L}_2})$ cannot be used as such a hypothesis space as its elements are not pointwise defined functions. But even the subspace of the space of continuous functions $\mathcal{C}(\Omega)$ containing functions with finite $\mathcal{L}_2$-norms is not suitable as some L_x might not be continuous on this space. For example, for $\Omega = \mathcal{R}^d$, L_0 defined as $L_0(f) = f(0)$ is not bounded and hence it cannot be continuous (L_0 maps the sequence $\left\{n^d e^{-(\frac{\|x\|}{n})^2}\right\}$ of functions with $\mathcal{L}_2$-norms equal to 1 to an unbounded sequence of real numbers). But there exits a large class of Hilbert spaces, on which operators L_x are continuous. Moreover, norms on spaces from this class can play roles of measures of various types of oscillations of input/output mappings.

A reproducing kernel Hilbert space RKHS is a Hilbert space formed by functions defined on a nonempty set Ω such that for every $x \in \Omega$ the evaluation functional $\mathcal{F}_x$, defined for any f in the Hilbert space as $\mathcal{F}_x(f) = f(x)$, is bounded [1], [2], [5]. RKHS can be elegantly characterized in terms of *kernels*, which are *symmetric positive semidefinite functions* $K : \Omega \times \Omega \to \mathcal{R}$, i.e., functions satisfying for all m, all $(w_1, \ldots, w_m) \in \mathcal{R}^m$, and all $(x_1, \ldots, x_m) \in \Omega^m$, $\sum_{i,j=1}^m w_i\, w_j\, K(x_i, x_j) \geq 0$. A kernel is *positive definite* if $\sum_{i,j=1}^m w_i w_j K(x_i, x_j) = 0$ for any distinct $x_1, \ldots, x_m$ implies that for all $i = 1, \ldots, m$, $w_i = 0$ (the terminology is not unified, some authors use the terms positive definite and strictly positive definite instead of positive semidefinite and positive definite, resp.).

To every RKSH one can associate a unique kernel $K : \Omega \times \Omega \to \mathcal{R}$ such that for every f in the RKHS and $x \in \Omega$

$$f(x) = \langle f, K_x \rangle_K, \tag{3.1}$$

where $K_x : \Omega \to \mathcal{R}$ is defined as $K_x(y) = K(x, y)$ for all $y \in \Omega$ ((3.1) is called the *reproducing property*). On the other hand, every kernel $K : \Omega \times \Omega \to \mathcal{R}$ generates a RKHS denoted by $(\mathcal{H}_K(\Omega), \|.\|_K)$, which is defined as the completion of the linear span of the set of functions $\{K_x : x \in \Omega\}$ with the inner product defined by $\langle K_x, K_y \rangle_K = K(x, y)$ (see, e.g., [1], [2, p. 81]).

By the Cauchy-Schwartz inequality, for every $f \in \mathcal{H}_K(\Omega)$ and $x \in \Omega$ we have $|f(x)| = |\langle f, K_x \rangle_K| \leq \|f\|_K \sqrt{K(x, x)} \leq \|f\|_K s_K$, where $s_K = \sup_{x \in \Omega} \sqrt{K(x, x)}$. Thus for every kernel K, we have

$$\sup_{x \in \Omega} |f(x)| \leq s_K \|f\|_K. \tag{3.2}$$

For a kernel $K : \Omega \times \Omega \to \mathcal{R}$, a positive integer m and a vector $x = (x_1, \ldots, x_m)$, by $\mathcal{K}[x]$ is denoted the $m \times m$ matrix defined as $\mathcal{K}[x]_{i,j} = K(x_i, x_j)$, which is called the *Gram matrix of the kernel K with respect to the vector x*.

A paradigmatic example of a kernel is the Gaussian kernel $G_\rho(u, v) = e^{-\rho \|u - v\|^2}$ on $\mathcal{R}^d \times \mathcal{R}^d$. For this kernel, the space $\mathcal{H}_{G_\rho}(\mathcal{R}^d)$ contains all functions computable by radial-basis function networks with a fixed width equal to ρ.

The following theorem describes properties of inverse problems defined by operators L_x on RKHSs.

Proposition 3.1 *Let $K : \Omega \times \Omega \to \mathcal{R}$ be a kernel, m be a positive integer, and $z = (x, y)$, where $x = (x_1, \ldots, x_m) \in \Omega^m, y = (y_1, \ldots, y_m) \in \mathcal{R}^m$, then:*
(i) $L_x : \mathcal{H}_K(\Omega) \to \mathcal{R}^m$ is a Lipschitz continuous compact linear operator with a closed range;
(ii) the adjoint operator $L_x{}^* : \mathcal{R}^m \to \mathcal{H}_K(\Omega)$ is compact and satisfies for every $u \in \mathcal{R}^m$, $L_x{}^*(u) = \frac{1}{\sqrt{m}} \sum_{i=1}^m u_i K_{x_i}$;
(iii) $R(L_x)$ is closed and $N(L_x^*) \oplus R(L_x) = \mathcal{R}^m$ and when K is positive definite, then $N(L_x{}^*) = \{0\}$ and $R(L_x) = \mathcal{R}^m$;
(iv) $L_x L_x^* : \mathcal{R}^m \to \mathcal{R}^m$ can be represented by the matrix $\frac{1}{m}\mathcal{K}[x]$;
(v) there exists a continuous linear pseudoinverse operator $L_x^+ : \mathcal{R}^m \to \mathcal{H}_K(\Omega)$ such that for every $u \in \mathcal{R}^m$, $L_x L_x^+(u) = \pi_R(u)$ and when K is positive definite, then $L_x L_x^+(u) = u$;
(vi) $L_x^+ = (L_x^* L_x)^+ L_x^* = L_x^*(L_x L_x^*)^+$.

Proof. (i) Linearity follows directly from the definition of L_x. By the reproducing property (3.1) and the relashionship (3.2) between the supremum norm and the norm $\|.\|_K$, for every $f \in \mathcal{H}_K(\Omega)$, $\|L_x(f)\|_2^2 = \frac{1}{m}\sum_{i=1}^m f(x_i)^2 \leq \|f\|_K^2 s_K^2$. Thus $\|L_x(f)\|_2 \leq s_K\|f\|_K$ and so L_x is Lipschitz continuous. As $\mathcal{R}^m$ is finite-dimensional, L_x has closed range and is compact [6, p.188].

(ii) By (i) L_x is compact and as the adjoint of a compact operator is compact [6, p.187], also L_x^* is compact. The representation of L_x^* follows from the reproducing property (3.1), which implies that $\langle L_x(f), u\rangle_2 = \frac{1}{\sqrt{m}}\sum_{i=1}^m u_i f(x_i) = \frac{1}{\sqrt{m}}\langle f, \sum_{i=1}^m u_i K_{x_i}\rangle_K = \langle f, {L_x}^*(u)\rangle_K$.

(iii) The first statement follows from Theorem 2.1 (ii). When K is positive definite, then $\{K_{x_1}, \ldots, K_{x_m}\}$ are linearly independent and thus by (ii) $N(L_x^*) = \{0\}$. Hence $R(L_x) = \mathcal{R}^m$.

(iv) For every $u \in \mathcal{R}^m$, $L_x L_x^*(u) = \frac{1}{m}\sum_{i=1}^m u_i K(x_i, x_j)$. So $L_x L_x^*(u) = \frac{1}{m}\mathcal{K}[x]u$.

(v) By Theorem 2.1 (iii) and (iv), there exists a pseudoinverse operator L_x^+ satisfying for all $u \in \mathcal{R}^m$, $L_x L_x^+(u) = \pi_R(u)$. When K is positive definite, then by (iii) $R(L_x) = \mathcal{R}^m$ and so $L_x L_x^*(u) = u$.

(vi) follows from Theorem 2.1 (v). □

The next theorem states properties of the solutions of the problem $(\mathcal{H}_K(\Omega), \mathcal{E}_z)$.

Theorem 3.2 *Let $K : \Omega \times \Omega \to \mathcal{R}$ be a kernel, m be a positive integer and $z = (x, y)$, where $x = (x_1, \ldots, x_m) \in \Omega^m$, $x_1, \ldots, x_m$ are distinct and $y = (y_1, \ldots, y_m) \in \mathcal{R}^m$, then:*
(i) $L_x^+(\frac{y}{\sqrt{m}}) \in argmin(\mathcal{H}_K(\Omega), \mathcal{E}_z)$, for every $f^o \in argmin(\mathcal{H}_K(\Omega), \mathcal{E}_z)$, $\|L_x^+(\frac{y}{\sqrt{m}})\|_K \leq \|f^o\|_K$ and $argmin(\mathcal{H}_K(\Omega), \mathcal{E}_z) = L_x^+(\frac{y}{\sqrt{m}}) + N(L_x)$;
(ii) for every $f^o \in argmin(\mathcal{H}_K(\Omega), \mathcal{E}_z)$, $L_x(f^o) = \pi_R(\frac{y}{\sqrt{m}})$ and when K is positive definite, $L_x(f^o) = \frac{y}{\sqrt{m}}$;
(iv) $\min_{f\in\mathcal{H}_K(\Omega)} = \frac{1}{m}\|\pi_R(y) - y\|_2^2$ and when K is positive definite, then $\min_{f\in\mathcal{H}_K(\Omega)} \mathcal{E}_z(f) = 0$;
(v) $L_x^+(\frac{y}{\sqrt{m}}) = \sum_{i=1}^m c_i K_{x_i}$, where $c = (c_1, \ldots, c_m) = \mathcal{K}[x]^+ y$;
(vi) for every $f^o \in argmin(\mathcal{H}_K(\Omega), \mathcal{E}_z)$, $\sum_{i=1}^m f^o(x_i)K_{x_i} = \sum_{i=1}^m y_i K_{x_i}$ and when K is positive definite, then f^o interpolates the data z, i.e., $f^o(x_i) = y_i$ for all $i = 1, \ldots, m$.

Proof. (i) and (ii) follows from Theorem 2.1 (iii) and Proposition 3.1 (iii) and (v).

(iii) By the representation (2.1) and by (ii), $\min_{f\in\mathcal{H}_K(\Omega)} \mathcal{E}_z(f) = \min_{f\in\mathcal{H}_K(\Omega)} \|L_x(f) - \frac{y}{\sqrt{m}}\|_2^2 = \|L_x L_x^+(\frac{y}{\sqrt{m}}) - \frac{y}{\sqrt{m}}\|_2^2 = \frac{1}{m}\|\pi_R(y) - y\|_2^2$. As by Proposition 3.1 (iii) for K positive definite, $R(L_x) = \mathcal{R}^m$, we have $\min_{f\in\mathcal{H}_K(\Omega)} \mathcal{E}_z(f) = 0$.

(iv) By Proposition 3.1 (vi) and (iv), $L_x^+ = L_x^*(L_x L_x^*)^+$, where $(L_x L_x^*)^+$ can be represented by the matrix $m\mathcal{K}[x]^+$. Thus for every $u \in \mathcal{R}^m$, $L_x^+(u) = \frac{1}{\sqrt{m}}\sum_{i=1}^m a_i K_{x_i}$, where $a = m\mathcal{K}[x]^+ u$. In particular, $L_x^+(\frac{y}{\sqrt{m}}) = \sum_{i=1}^m c_i K_{x_i}$, where $c = \mathcal{K}[x]^+ y$.

(v) By Proposition 3.1 (ii), $\frac{1}{m}\sum_{i=1}^m f^o(x_i)K_{x_i} = L_x^* L_x(f^o)$. On the other hand, by (ii) and Proposition 3.1 (ii), $L_x^* L_x(f^o) = L_x^* L_x L_x^+(\frac{y}{\sqrt{m}}) = L_x^*(\frac{y}{\sqrt{m}}) = \frac{1}{m}\sum_{i=1}^m y_i K_{x_i}$. Hence $\sum_{i=1}^m f^o(x_i)K_{x_i} = \sum_{i=1}^m y_i K_{x_i}$. When K is positive definite, then $\{K_{x_1}, \ldots, K_{x_m}\}$ are linearly independent and thus for all $i = 1, \ldots, m$, $f^o(x_i) = y_i$. □

So for every kernel K and every sample of empirical data z, there exists a solution of the problem of minimization of the empirical error functional $\mathcal{E}_z$ over the space $\mathcal{H}_K(\Omega)$. The set of such solutions $argmin(\mathcal{H}_K(\Omega), \mathcal{E}_z)$ is a closed convex set of the form $\sum_{i=1}^m c_i K_{x_i} + N(L_x)$, where $c = \mathcal{K}[x]^+ y$ and $N(L_x)$ is the null space of the operator L_x. Minimum of $\mathcal{E}_z$ over $\mathcal{H}_K(\Omega)$ is equal to $\frac{1}{m}\|\pi_R(y) - y\|_2^2$, where π_R is the projection of $\mathcal{R}^m$ onto $R(L_x)$. For K positive definite, the solution interpolates the data and minimum of $\mathcal{E}_z$ over $\mathcal{H}_K(\Omega)$ is equal to zero.

Stability of the solution $\sum_{i=1}^m c_i K_{x_i}$ with respect to a small perturbation of the vector of output data y depends on the condition number of the matrix $\mathcal{K}[x]$. The solution is robust against noise only when the condition number is close to 1. For Gaussian kernels G_ρ, upper bounds on such condition numbers growing with the dimension d of the input data and the product ρq^2, where q is the separation radius of the input data x (which is defined as $q = \frac{1}{2}\min\{\|x_i - x_j\|_2 : i, j = 1, \ldots, m, i \neq j\}$), but independent on the size m of the data sample, were derived in [12].

4 Learning with generalization as a regularized inverse problem

The function $\sum_{i=1}^{m} c_i K_{x_i}$ with $c = \mathcal{K}[x]^+ y$ guarantees the best fit to the sample of data z that can be achieved using functions from the space $\mathcal{H}_K(\Omega)$. By choosing as a hypothesis space a RKHS, we impose a condition on oscillations of potential solutions. The type of such a condition can be illustrated on convolution kernels $K : \mathcal{R}^d \times \mathcal{R}^d \to \mathcal{R}$ satisfying $K(u,v) = k(u-v)$ for some $k : \mathcal{R} \to \mathcal{R}$ with positive Fourier transform $\tilde{k}$. For such kernels $\|f\|_K^2 = \frac{1}{(2\pi)^{d/2}} \int_{\mathcal{R}^d} \frac{\tilde{f}(\omega)^2}{\tilde{k}(\omega)} d\omega$ [7].

The restriction on potential solutions can be further strengthened by penalizing the size of the norm $\|.\|_K$ of the solution. This approach to constraining solutions of ill-posed inverse problems has been developed in 1960th by several authors. It is called *Tikhonov's regularization* due to Tikhonov's unifying formulation [16]. Tikhonov's regularization replaces the problem of minimization of the functional $\|A(.) - g\|_Y^2$ over X with minimization of $\|A(.) - g\|_Y^2 + \gamma \|.\|_X^2$, where the *regularization parameter* γ plays the role of a trade-off between fitting to empirical and conceptual data. The following theorem summarizes properties of solutions of regularized inverse problems stated in [3, pp.68-70] and [8, pp.74-76]. By I is denoted the identity operator $I : \mathcal{R}^d \to \mathcal{R}^d$ and by $\mathcal{I}$ the corresponding $m \times m$ matrix.

Theorem 4.1 *Let X, Y be Hilbert spaces, $A : X \to Y$ be a continuous linear operator with a closed range, then:*
(i) for every $\gamma > 0$, there exists a unique operator $A^\gamma : Y \to X$ such that for every $g \in Y$, $\{A^\gamma(g)\} = argmin(X, \|A(.) - g\|_Y^2 + \gamma \|.\|_X^2)$;
(ii) for every $\gamma > 0$, $A^\gamma = (A^*A + \gamma I)^{-1} A^* = A^*(AA^* + \gamma I)^{-1}$;
(iii) for every $g \in Y$, $e_g : (0,\infty) \to (0,\infty)$ defined as $e_g(\gamma) = \|AA^\gamma(g) - g\|_Y$ is strictly increasing, $\lim_{\gamma \to 0} e_g(\gamma) = \|\pi_N(g)\|_Y$ and $\lim_{\gamma \to \infty} e_g(\gamma) = \|g\|_Y$;
(iv) for every $g \in Y$, $E_g : (0,\infty) \to (0,\infty)$ defined as $E_g(\gamma) = \|A^\gamma(g)\|_X$ is strictly decreasing, $\lim_{\gamma \to 0} E_g(\gamma) = \|A^+(g)\|_X$ and $\lim_{\gamma \to \infty} E_g(\gamma) = 0$.

So even if the original inverse problem is ill-posed (it does not have a unique solution), for every $\gamma > 0$, the regularized problem has a unique solution. This is due to uniform convexity of the functional $\|.\|_Y^2$ (see, e.g., [11]). With γ going to zero, the solutions of regularized problem converge to the pseudosolution with the minimal norm $A^+(g)$. The next theorem describes properties of regularized solutions, their relationship to the pseudosolution and improvement of stability achievable using regularization for the problem of minimization of $\mathcal{E}_z$ over a RKHS.

Theorem 4.2 *Let $K : \Omega \times \Omega$ be a kernel, m be a positive integer, $z = (x,y)$, where $x = (x_1, \ldots, x_m) \in \Omega^m$, $x_1, \ldots, x_m$ are distinct, $y = (y_1, \ldots, y_m) \in \mathcal{R}^m$ and $\gamma > 0$, then:*
(i) there exists a unique solution f^γ of the problem $(\mathcal{H}_K(\Omega), \mathcal{E}_z + \gamma \|.\|_K^2)$;
(ii) $f^\gamma = \sum_{i=1}^{m} c_i K_{x_i}$, where $c = (\mathcal{K}[x] + \gamma m \mathcal{I})^{-1} y$;
(iii) $e : (0,\infty) \to [0,\infty)$ defined as $e(\gamma) = \mathcal{E}_z(f^\gamma)$ is strictly increasing, $\lim_{\gamma \to \infty} e(\gamma) = \frac{1}{\sqrt{m}} \|y\|_2$ and $\lim_{\gamma \to 0} e(\gamma) = \|\pi_R(y) - y\|_2$, which is equal to 0 for K positive definite;
(iv) $E : (0,\infty) \to [0,\infty)$ defined as $E(\gamma) = \|f^\gamma\|_K$ is strictly decreasing, $\lim_{\gamma \to 0} E(\gamma) = \| \sum_{i=1}^{m} a_i K_{x_i} \|_K$, where $a = \mathcal{K}[x]^+ y$, and $\lim_{\gamma \to \infty} E(\gamma) = 0$;
(v) when K is positive definite, then $cond(\mathcal{K}[x] + \gamma m \mathcal{I}) = 1 + \frac{(cond(\mathcal{K}[x]) - 1)\lambda_{\min}}{\lambda_{\min} + \gamma m}$, where $\lambda_{\min}$ is the minimal eigenvalue of $\mathcal{K}[x]$.

Proof. (i) follows from Theorem 4.1(i).

(ii) By Theorem 4.1(ii), $f^\gamma = L_x^\gamma(\frac{y}{\sqrt{m}}) = L_x^*(L_x L_x^* + \gamma I)^{-1}(\frac{y}{\sqrt{m}})$. So by Proposition 3.1(iv), $f^\gamma = \sum_{i=1}^{m} c_i K_{x_i}$, where $c\sqrt{m} = (\frac{1}{m}(\mathcal{K}[x] + \gamma m \mathcal{I})^{-1} \frac{y}{\sqrt{m}} = m(\mathcal{K}[x] + \gamma m \mathcal{I})^{-1} \frac{y}{\sqrt{m}}$. Thus $c = (\mathcal{K}[x] + \gamma m \mathcal{I})^{-1} y$.

(iii) and (iv) follow from Theorem 4.1(iii) and (iv).

(v) For every nonsingular $m \times m$ matrix A, the condition number $cond(A)$ with respect to the l_2-norm on $\mathcal{R}^m$ is equal to $\frac{|\lambda_{\max}(A)|}{|\lambda_{\min}(A)|}$, where $\lambda_{\max}(A), \lambda_{\min}(A)$ denote the maximal and minimal eigenvalues, resp. of A, and for every positive definite matrix all eigenvalues are positive. So denoting

$\lambda_{\max}$, $\lambda_{\min}$ the maximal and minimal eigenvalues of $\mathcal{K}[x]$, we get $cond(\mathcal{K}[x] + \gamma m\mathcal{I}) = \frac{\lambda_{\max}+\gamma m}{\lambda_{\min}+\gamma m} = 1 + \frac{\lambda_{\max}-\lambda_{\min}}{\lambda_{\min}+\gamma m} = 1 + \frac{(cond(\mathcal{K}[x])-1)\lambda_{\min}}{\lambda_{\min}+\gamma m}$. □

Theorem 4.2 (ii) shows that the Representer Theorem [15], [5], [17] on learning from data in RKHS is a special case of a more general result from theory of regularization of inverse problems in Hilbert spaces. Note that some direct proofs of the Representer Theorem such as the one in [15] use the same argument based on annihilation of all directional derivatives as the proof of Theorem 4.1(ii) [8, pp.74-75], [3, pp.68-69].

Moreover, Theorem 4.2(v) shows how much ill-conditioning of the problem of minimization of $\mathcal{E}_z$ over a RKHS can be improved by regularization. As $\lim_{\gamma m\to\infty}(1+\frac{cond(\mathcal{K}[x]-1)\lambda_{\min}}{\lambda_{\min}+\gamma m}) = 1$, for sufficiently large γm, the condition number of the matrix $\mathcal{K}[x] + \gamma m\mathcal{I}$ is close to 1. The size of γ is limited by requirements of fitting to the sample of empirical data z, while the size m of the sample can be enlarged. Thus for a sufficiently large m, regularization improves stability of the solution.

5 Discussion

Using theory of generalized inversion in Hilbert spaces, we have described solutions of the learning task modelled as the least square problem in the domain of reproducing kernel Hilbert spaces. Such spaces can be used to model radial-basis networks with various types of radial function with fixed width. Practical applications of formulas for computing pseudosolution and regularized solutions given in Theorems 3.2(v) and 4.2(ii) are limited by computational efficiency of iterative methods for solving systems of linear equations $c = \mathcal{K}[x]^+ y$ and $c = (\mathcal{K}[x]+\gamma m\mathcal{I})^{-1} y$ and by the condition numbers of the matrices $\mathcal{K}[x]$ and $\mathcal{K}[x]+\gamma m\mathcal{I}$. We have shown how regularization improves properties of solutions of the learning task: it guarantees uniqueness and might improve stability when the size of the sample of data, their separation radius and the kernel defining the hypothesis space are properly chosen.

The requirement of continuity of the operator L_x do not allow to extend our results to the space of continuous functions on $\mathcal{R}^d$ with finite $\mathcal{L}_2$-norms. Another limiting factor is strong dependence of theory of pseudoinversion on Hilbert space setting. Thus most of our results apply only to empirical error with the square loss function, while, e.g., in the case of absolute value loss, only much weaker results holding for inverse problems with range in $\mathcal{R}^m$ with l_1-norm can be used.

Bibliography

[1] Aronszajn N. (1950). *Theory of reproducing kernels.* Transactions of AMS **68**, 33 – 404.

[2] Berg C., Christensen J. P. R., Ressel P. (1984). *Harmonic Analysis on Semigroups.* Springer-Verlag, New York.

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

[4] Bjorck A. (1996). *Numerical methods for least squares problem.* SIAM.

[5] Cucker F., Smale S. (2001). *On the mathematical foundations of learning.* Bulletin of AMS **39**, 1 – 49.

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

[7] Girosi F. (1998). *An equivalence between sparse approximation and support vector machines.* Neural Computation **10**, 1455 – 1480 (AI Memo No 1606, MIT).

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

[9] Kůrková V. (2003). *High-dimensional approximation by neural networks.* Chapter 4 in *Advances in Learning Theory: Methods, Models and Applications* (J. Stuykens et al., Ed.), 69 – 88. IOS Press, Amsterdam.

[10] Kůrková V., Sanguineti, M. (2004).*Error estimates for approximate optimization by teh extended Ritz method.* SIAM Journal on Optimization (to appear).

[11] Kůrková V., Sanguineti M. (2003). *Learning with generalization capability by kernel methods with bounded complexity.* Research Report ICS-2003-901, submitted to Journal of Complexity.

[12] Narcowich F. J., Sivakumar N., Ward J. D. (1994). *On condition numbers associated with radial-function interpolation.* Journal of Mathematical Analysis and Applications **186**, 457 – 485.

[13] Parzen E. (1966). *An approach to time series analysis.* Annals of Math. Statistics **32**, 951 – 989.

[14] Poggio T., Girosi F. (1990). *Networks for approximation and learning.* Proceedings IEEE **78**, 1481 – 1497.

[15] Poggio T., Smale S. (2003). *The mathematics of learning: dealing with data.* Notices of the AMS **50**, 536 – 544.

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

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