Neural Network Learning as an Inverse Problem

VĚRA KŮRKOVÁ, *Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 2, 182 07 Prague 8, Czech Republic.*
Email: vera@cs.cas.cz

Abstract

Capability of generalization in learning of neural networks from examples can be modelled using regularization, which has been developed as a tool for improving stability of solutions of inverse problems. Such problems are typically described by integral operators. It is shown that learning from examples can be reformulated as an inverse problem defined by an evaluation operator. This reformulation leads to an analytical description of an optimal input/output function of a network with kernel units, which can be employed to design a learning algorithm based on a numerical solution of a system of linear equations.

Keywords: learning from data, generalization, empirical error functional, inverse problem, evaluation operator, kernel methods

1 Introduction

Today's technology provides many measuring and recording devices, which supply us with a huge amount of data. Unfortunately, such machine-made data are not comprehensible for our brains. We resemble the king Midas, whose foolish wish to transform everything he touched into gold was fulfilled, which led to his starvation as his body could not be feeded by gold. So we turn again to machines (computers) to help us to transform raw data into patterns that our brains can digest.

As Popper [24] emphasized, no patterns can be derived solely from *empirical data*. Some hypotheses about patterns have to be chosen and among patterns satisfying these hypotheses, a pattern with a good fit to data has to be searched for. History of science presents many examples of this approach, e.g., Kepler's assumption that planets move on the most perfect curves fitting to the data collected by Tycho de Brahe. Kepler found that among perfect curves, ellipses are the ones best fitting to the data (better than circles, which he tried first).

Also Gauss and Legendre searched for curves fitting to astronomical data collected from observations of comets. They developed the *least square method*. In 1805, Legendre wrote "of all the principles that can be proposed, I think that there is none more general, more exact and more easy to apply than that consisting of minimizing the sum of the squares of errors." Many researchers of later generations shared Legendre's enthusiasm and used the least square method in statistical inference, pattern recognition, function approximation, curve or surface fitting, etc.

In all these techniques, minimization of the average of squares of errors is used to find coefficients of a linear combination of some fixed family of functions. Thus the best fitting function is searched for in a *linear hypothesis space*. However, the assumption that empirical

doi:10.1093/jigpal/jzi041

data can be approximated by a linear function is rather restrictive. Especially, it is not suitable for high-dimensional data. It is known from theory of linear approximation that the dimension of a linear space needed for approximation within accuracy ε is $\mathcal{O}\left((\frac{1}{\varepsilon})^d\right)$, where d denotes the number of variables [21]. So complexity of linear models grows exponentially with the data dimension d.

In contrast to the traditional applications of the least square method, in *neurocomputing* this method is applied to *nonlinear hypothesis sets*. The hypothesis sets are formed by linear combinations of parameterized functions, the form of which is given by the type of *computational units*, from which a network is built. Originally the units, called perceptrons, modelled some simplified properties of neurons, but later also other types of units (radial and kernel units) with suitable mathematical properties for function approximation became popular. The *back-propagation algorithm* developed by Werbos in 1970s and reinvented by Rummelhart, Hinton and Williams in 1980s (see [29]) calculates how the gradient of the average square error depends on *coefficients of the linear combination* as well as on *inner parameters* of functions forming their linear combination. The algorithm iteratively modifies all network parameters until a sufficiently well fitting input/output function of the network is found.

Neurocomputing brought to data analysis also a new terminology: searching for parameters of their input/output functions is called *learning*, samples of data *training sets* and a capability to satisfactorily process new data that have not been used for learning is called *generalization*.

Capability of generalization depends on the choice of a hypothesis set of input/output functions, where one searches for a pattern (a functional relationship) fitting to empirical data. So a restriction of the hypothesis set to only physically meaningful functions can improve generalization.

An alternative approach to modelling of generalization was proposed by Poggio and Girosi [22]. They modified the least square method by *Tikhonov's regularization*, which adds to the least square error a term, called *stabilizer*, penalizing undesired input/output functions. Girosi, Jones and Poggio considered stabilizers penalizing high frequencies in the Fourier representation of the hypothetical solutions [11].

Girosi [10] showed that stabilizers of this type belong to a wider class formed by the squares of norms on a special type of Hilbert spaces called *reproducing kernel Hilbert spaces* (RKHS). Such norms can play a role of *measures of various types of oscillations* and thus enable to model a variety of prior knowledge (*conceptual data*), which has to be added to the empirical ones to guarantee a generalization capability. RKHS were formally defined by Aronszajn [2], but their theory includes many classical results on positive definite functions, matrices and integral operators with kernels. In data analysis, kernels were first used by Parzen [19] and Wahba [28], who applied them to data smoothing by splines.

Aizerman, Braverman and Rozonoer [1] used kernels (under the name potential functions) to solve classification tasks by transforming geometry of input spaces by embedding them into higher dimensional inner product spaces [1]. Boser, Guyon and Vapnik [5] and Cortes and Vapnik [6] farther developed this method of classification into the concept of the *support vector machine* (a one-hidden-layer network with kernel units in the hidden layer and one threshold output unit).

A variety of kernel methods and algorithms are of current interest, and their potential uses are wide ranging; see, e.g., the recent applications-oriented monographs by Schöllkopf and Smola [25] and by Cristianini and Shawe-Taylor [7], a theoretical article by Cucker and

Smale's [8] or a brief survey by Poggio and Smale [23].

So use of kernels in machine learning has two reasons: (1) norms on RKHSs can play roles of undesirable attributes of input/output functions, the penalization of which improves generalization, and (2) kernels define embeddings of input spaces into feature spaces with geometries that allow to separate linearly data to be classified. In this paper, we add to these two reasons a third one. We show that the Aronszajn's definition [2] of RKHSs (as Hilbert spaces with all evaluation operators continuous) determines precisely a class of Hilbert spaces, where solutions of the least square problems of finding input/output functions fitting to empirical data (minimization of *empirical error functionals*) can be obtained as *Moore-Penrose pseudosolutions of linear inverse problems*.

Tasks of finding *unknown causes* (such as shapes of functions, forces or distributions) from *known consequences* (measured data) have been studied in applied science (such as acoustics, geophysics and computerized tomography see, e.g., [13]) under the name *inverse problems*. To solve such a problem, one needs to know how unknown causes determine known consequences, which can often be described in terms of *operators*. In problems originating from physics, such operators are typically integral (such as those defining Radon or Laplace transforms [3], [9]).

In this paper, we reformulate minimization of the least square error as an inverse problem defined by an *evaluation operator*. A crucial property for an application of tools from theory of inverse problems is *continuity of an operator*. As the evaluation operator is defined in terms of the evaluation functionals at the sample of input data, the class of Hilbert spaces, where the theory can be applied, corresponds exactly to the class of reproducing kernel Hilbert spaces. Thus reformulation of a learning problem as an inverse problem justifies the use of RKHSs as suitable hypothesis spaces.

We show that this reformulation leads to analytical description of the optimal input/output function of a network with kernel units, which can be employed to design a learning algorithm based on a numerical solution of a system of linear equations.

The paper is organized as follows. In Section 2, learning from data is reformulated as an inverse problem. Section 3 describes properties of reproducing kernel Hilbert spaces. In Sections 4 and 5, main results on functions minimizing an empirical error functional and its regularizations are presented.

2 Learning from empirical data as an inverse problem

A standard approach to learning from data used, e.g., in neurocomputing in the back-propagation algorithm, is based on minimization of the least square error. For a *sample of input/output pairs of data (training set)* $z = \{(u_i, v_i) \in \Omega \times \mathcal{R},\, i = 1, \ldots, m\}$, where $\Omega \subset \mathcal{R}^d$, a functional $\mathcal{E}_z$ called *empirical error* is defined for a function $f : \Omega \to \mathcal{R}$ as

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

We show that minimization of this functional can be reformulated as an inverse problem.

For a linear operator $A : X \to Y$ between two Hilbert spaces X, Y (in finite-dimensional case, a matrix A) an *inverse problem* determined by A is to find for $g \in Y$ some $f \in X$ such that

$$A(f) = g,$$

where g is called a *data* and f a *solution* [3].

When for some $g \in Y$ no solution exists, one can at least 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.,

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

Using standard terminology from optimization theory, for a functional $\Phi : X \to \mathcal{R}$ and $M \subseteq X$ we denote by

$$(M, \Phi)$$

the *problem of minimizitation of* Φ *over* M; the set M is called a *hypothesis set*. An element $f^o \in M$ such that

$$\Phi(f^o) = \min_{f \in M} \Phi(f)$$

is called a *solution* of the problem (M, Φ) and

$$argmin(M, \Phi) = \{f^o \in M : \Phi(f^o) = \min_{f \in M} \Phi(f)\}$$

denotes the set of all solutions of (M, Φ). So a *pseudosolution* is a solution of the problem

$$(X, \|A(.) - g\|_Y).$$

The set $argmin(X, \|A(.) - g\|_Y)$ is convex and, hence, if it is nonempty, there exists a unique pseudosolution of minimal norm, called the *normal pseudosolution* [12]). This normal pseudosolution, denoted by f^+, satisfies

$$\|f^+\|_X = \min\{\|f^o\|_X : f^o \in argmin(X, \|A(.) - g\|_Y)\}.$$

When for every $g \in Y$, the set $argmin(X, \|A(.) - g\|_Y)$ is non-empty (and thus there exists a normal pseudosolution f^+), then a *pseudoinverse operator*

$$A^+ : Y \to X$$

can be defined by setting

$$A^+(g) = f^+.$$

In 1920, Moore [18] described properties of the pseudoinverse of a matrix, which were rediscovered in 1955 by Penrose [20]. In 1970s, the theory of Moore-Penrose pseudoinversion has been extended to the infinite-dimensional case – it was shown that similar properties as the ones of Moore-Penrose pseudoinverses of matrices also are possessed by *continuous linear operators with closed ranges* [12].

We can employ theory of inverse problems to investigation of minimization of an empirical error functional. To reformulate minimization of such a functional as an inverse problem, for an input data vector $u = (u_1, \ldots, u_m) \in \Omega^m$ and a Hilbert space X of functions on Ω, define an *evaluation operator*

$$L_u : X \to \mathcal{R}^m$$

by

$$L_u(f) = \left(\frac{f(u_1)}{\sqrt{m}}, \ldots, \frac{f(u_m)}{\sqrt{m}} \right).$$

It is easy to check that for every f on Ω,

$$\mathcal{E}_z(f) = \frac{1}{m}\sum_{i=1}^{m}(f(u_i) - v_i)^2 = \|L_u(f) - \frac{v}{\sqrt{m}}\|_2^2. \tag{2.1}$$

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

$$\mathcal{E}_z = \left\| L_u - \frac{v}{\sqrt{m}} \right\|_2^2,$$

where $\|.\|_2$ denotes the l_2-norm on $\mathcal{R}^m$.

The representation (2.1) allows one to express the problem of minimization of the empirical error functional $\mathcal{E}_z$ as the problem of finding a pseudosolution $L_u^+(\frac{v}{\sqrt{m}})$ of the inverse problem given by the operator L_u for the data $\frac{v}{\sqrt{m}}$. As the range of the operator L_u is finite dimensional, it is closed in $\mathcal{R}^m$. Thus to employ the extension of Moore-Penrose pseudoinversion to the domain of infinite dimensional spaces, it remains to find proper hypothesis spaces, on which operators L_u are *continuous* for all input data vectors u.

3 Reproducing kernel Hilbert spaces

Neither the space $\mathcal{C}(\Omega)$ of all continuous functions on $\Omega \subset \mathcal{R}^d$ nor the Lebesgue space $\mathcal{L}_2(\Omega)$ of square integrable functions are suitable as the hypothesis spaces: the first one is not a Hilbert space and the second one is not formed by pointwise defined functions. Moreover, L_u is not continuous on any subspace of $\mathcal{L}_2(\mathcal{R}^d)$ containing the sequence of functions $\left\{n^d e^{-(\frac{\|x\|}{n})^2}\right\}$: all elements of this sequence have $\mathcal{L}_2$-norms equal to 1, but the evaluation functional at zero maps this sequence to an unbounded sequence of real numbers and thus it is not continuous.

Fortunately, there exists a large class of Hilbert spaces, on which *all evaluation functionals are continuous* and moreover, norms on such spaces can play roles of measures of various types of oscillations of input/output mappings.

Spaces from this class are called *reproducing kernel Hilbert spaces (RKHSs)*. A RKHS is a Hilbert space formed by functions 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 [2]. RKHS can be 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 $i = 1, \ldots, m$, $w_i = 0$. In such a case $\{K_x : x \in \Omega\}$ is a linearly independent set.

The RKHS defined by a kernel K on $\Omega \times \Omega$ is denoted by

$$\mathcal{H}_K(\Omega).$$

It is formed by linear combinations of functions from $\{K_x : x \in \Omega\}$ and limits of all Cauchy sequences with respect to the norm $\|.\|_K$ on the RKHS induced by the inner product defined on generators by $\langle K_x, K_y \rangle_K = K(x, y)$.

For a positive integer m and a vector $u = (u_1, \ldots, u_m)$, by $\mathcal{K}[x]$ is denoted the $m \times m$ matrix defined as

$$\mathcal{K}[u]_{i,j} = K(u_i, u_j),$$

which is called the *Gram matrix of the kernel K with respect to the vector u.*

A paradigmatic example of a kernel is the *Gaussian kernel* $G_\rho(x, y) = e^{-\rho\|x-y\|^2}$ 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.

4 Optimal solution of the learning task

Using tools from the theory of inverse problems, we can describe the unique function in the reproducing kernel Hilbert space given by a kernel K minimizing an empirical error functional, which has the smallest K-norm. The next theorem states that this function is the image of the vector of normalized output data $\frac{v}{\sqrt{m}}$ under the pseudoinverse operator L_u^+ (for the proof see [14]).

THEOREM 4.1
Let $K : \Omega \times \Omega \to \mathcal{R}$ be a kernel, m be a positive integer and $z = (u, v)$, where $u = (u_1, \ldots, u_m) \in \Omega^m$, $u_1, \ldots, u_m$ are distinct and $v = (v_1, \ldots, v_m) \in \mathcal{R}^m$, then:
(i) $L_u^+(\frac{v}{\sqrt{m}}) \in argmin(\mathcal{H}_K(\Omega), \mathcal{E}_z)$, and for every $f^o \in argmin(\mathcal{H}_K(\Omega), \mathcal{E}_z)$,
$\|L_u^+(\frac{v}{\sqrt{m}})\|_K \leq \|f^o\|_K$;
(ii) $L_u^+(\frac{v}{\sqrt{m}}) = \sum_{i=1}^m c_i K_{u_i}$, where $c = (c_1, \ldots, c_m) = \mathcal{K}[u]^+ v$.

So for every kernel K and every sample of empirical data z, there exists a function f^+ minimizing the empirical error functional $\mathcal{E}_z$ over the whole RKHS defined by K. This function is formed by a linear combination functions $K_{u_1}, \ldots, K_{u_m}$:

$$f^+ = L_u^+ \left(\frac{v}{\sqrt{m}} \right) = \sum_{i=1}^m c_i K_{u_i}.$$

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 of the linear combination $c = (c_1, \ldots, c_m)$ (corresponding to so called output weights) can be computed applying the pseudoinverse of the Gram matrix of the kernel K with respect to the input data vector u on the output data vector v:

$$c = \mathcal{K}[u]^+ v.$$

Capability of generalization of such a network can be studied in terms of stability of f^+ against a noise. Recall that the *condition number* of a symmetric positive definite matrix is the ratio between the largest and the smallest eigenvalue. So when this ratio is for the matrix $\mathcal{K}[u]$ small, the solution f^+ is robust against a noise. The condition number measures the worst magnification of data errors, while a more realistic description of ill-conditioning can be derived from inspection of all the eigenvalues of $\mathcal{K}[u]$. For K positive definite, the row vectors $K_{u_1}, \ldots, K_{u_m}$ of the matrix $\mathcal{K}[u]$ are linearly independent. But when the distances

between the data $u_1, \ldots, u_m$ are small, the row vectors might be nearly parallel and the small eigenvalues of $\mathcal{K}[u]$ might cluster near zero. In such a case, small changes of v, can cause large changes of f^+.

5 Generalization modelled as regularization

The function $f^+ = \sum_{i=1}^m c_i K_{u_i}$ with $c = \mathcal{K}[u]^+ v$ provides the best fit to the sample of data z that can be obtained using functions from the reproducing kernel space defined by the kernel K. By choosing this RKHS as a hypothesis space, one imposes a condition on oscillations of potential solutions. The type of such a condition can be illustrated by convolution kernels $K : \mathcal{R}^d \times \mathcal{R}^d \to \mathcal{R}$ satisfying $K(x, y) = k(x - y)$ for some $k : \mathcal{R} \to \mathcal{R}$ with positive Fourier transform $\tilde{k}$. For such kernels, K-norms can be expressed as high-frequency filters

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

[10]. So to keep this norm small, $\tilde{f}(\omega)$ should decrease rather quickly with $\|s\|$ increasing.

The restriction on high frequency oscillations can be strengthened by *Tikhonov's regularization* with $\|.\|_K^2$ as a *stabilizer*, i.e., by replacing minimization of $\mathcal{E}_z$ with minimization of

$$\mathcal{E}_z + \gamma \|.\|_K^2,$$

where $\gamma > 0$ is called a *regularization parameter*.

The next theorem describes properties of regularized solutions f^γ, their relationship to the pseudosolution f^+ and improvement of stability achievable using Tikhonov's regularization (for the proof see [14]).

THEOREM 5.1
Let $K : \Omega \times \Omega$ be a kernel, m be a positive integer, $z = (u, v)$, where $u = (u_1, \ldots, u_m) \in \Omega^m$, $u_1, \ldots, u_m$ are distinct, $v = (v_1, \ldots, v_m) \in \mathcal{R}^m$ 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) $f^\gamma = \sum_{i=1}^m c_i K_{u_i}$, where $c = (\mathcal{K}[u] + \gamma m \mathcal{I})^{-1} v$;
(iii) when K is positive definite, then $cond(\mathcal{K}[u] + \gamma m \mathcal{I}) = 1 + \frac{(cond(\mathcal{K}[u]) - 1)\lambda_{\min}}{\lambda_{\min} + \gamma m}$, where $\lambda_{\min}$ is the minimal eigenvalue of $\mathcal{K}[u]$.

Similarly as in the case of the function f^+ minimizing the empirical error, the function f^γ minimizing the regularized empirical error is a linear combination of the functions $K_{u_1}, \ldots, K_{u_m}$ defined by the input data $u_1, \ldots, u_m$:

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

But the coefficients of the linear combination are different, their vector $c^\gamma = (c_1^\gamma, \ldots, c_m^\gamma)$ is the image of the output data vector v under the inverse operator $(\mathcal{K}[u] + \gamma m \mathcal{I})^{-1}$:

$$c^\gamma = (\mathcal{K}[u] + \gamma m \mathcal{I})^{-1} v.$$

This formula has been derived by several authors [28], [8], [23] using Fréchet derivatives and called the Representer Theorem. Here, we obtained it directly from properties of regularized solutions of inverse problems.

So increase of "smoothness" of the regularized solution f^γ is achieved by merely changing the coefficients of the linear combination: while in the non regularized case, the coefficients are obtained from the output data vector v using the Moore-Penrose pseudoinverse of the Gram matrix $\mathcal{K}[u]$, in the regularized one, they are obtained using the inverse of a modified matrix $\mathcal{K}[u] + \gamma m\mathcal{I}$. So the regularization merely changes amplitudes, but it preserves the finite set of basis functions from which the solution is composed. These basis functions can be, for example, Gaussians with centers given by the input data.

Theorem 5.1 shows that continuous growth of the regularization parameter γ leads from "under smoothing" to "over smoothing" and estimates how ill-conditioning can be improved by the Tikhonov's regularization. As

$$\lim_{\gamma m\to\infty}(1 + \frac{cond(\mathcal{K}[u] - 1)\lambda_{\min}}{\lambda_{\min} + \gamma m}) = 1,$$

the larger the product γm, the better improvement of stability.

However, the size of the regularization parameter γ is limited by the requirement of fitting f^γ to the sample of empirical data z, so γ cannot be too large. But m can be increased by choosing a larger sample of empirical data. However for large m, computational efficiency of iterative methods for solving systems of linear equations $c = \mathcal{K}[u]^+ v$ and $c = (\mathcal{K}[u] + \gamma m\mathcal{I})^{-1} v$ might limit practical applications of learning algorithms based on the Representer Theorem (Theorem 5.1 (ii)); see [23] for references to such applications.

In typical neural-network algorithms, networks with the number of hidden units n much smaller than the size m of the training set are used. In [16] and [17], estimates of speed of convergence of infima of empirical error over sets of functions computable by such networks to the minimum described in Theorem 5.1 were derived. For reasonable data sets, such convergence is rather fast.

Acknowledgements

This work was partially supported by the project 1ET100300517 of the program "Information Society" of the National Research Program of the Czech Republic.

References

[1] Aizerman M. A., Braverman E. M., Rozonoer L. I. (1964). *Theoretical foundations of potential function method in pattern recognition learning.* Automation and Remote Control **28**, 821 – 837.

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

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

[4] Björck A. (1996). *Numerical Methods for Least Squares Problem.* SIAM.

[5] Boser B., Guyon I., Vapnik V. N. (1992). A training algorithm for optimal margin clasifiers. In *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory* (Ed. D. Haussler) (pp. 144 - 152). ACM Press.

[6] Cortes C., Vapnik V. N. (1995). *Support-vector networks.* Machine Learning **20**, 273 – 297.

[7] Cristianini N., Shawe-Taylor J. (2000). *An Introduction to Support Vector Machines.* Cambridge: Cambridge University Press.

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

[9] Engl H. W., Hanke M., Neubauer A. (2000). *Regularization of Inverse Problems.* Dordrecht: Kluwer.

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

[11] Girosi F., Jones M., Poggio T. (1995). *Regularization theory and neural network architectures.* Neural Computaion **7**, 219 – 269.

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

[13] Hansen P. C. (1998). *Rank-Deficient and Discrete Ill-Posed Problems.* Philadelphia: SIAM.

[14] Kůrková V. (2004). *Supervised learning as an inverse problem.* Research Report ICS-2004-960, Institute of Computer Science, Prague.

[15] Kůrková V. (2004). *Learning from data as an inverse problem.* In *COMPSTAT 2004 - Proceedings on Computational Statistics* (J. Antoch, Ed.), 1377 – 1384. Heidelberg: Physica-Verlag/Springer.

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

[17] Kůrková V., Sanguineti M. (2005). *Learning with generalization capability by kernel methods with bounded complexity.* Journal of Complexity (to appear).

[18] Moore E. H. (1920). *Abstract.* Bull. Amer. Math. Soc. **26**, 394 – 395.

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

[20] Penrose R. (1955). *A generalized inverse for matrices.* Proc. Cambridge Philos. Soc. **5**1, 406 – 413.

[21] Pinkus, A. (1985). *n-width in Approximation Theory.* Berlin: Springer-Verlag.

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

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

[24] Popper K. (1968). *The Logic of Scientific Discovery.* New York: Harper Torch Book.

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

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

[27] Vapnik V. N. (1995). *The Nature of Statistical Learning Theory.* New York: Springer-Verlag.

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

[29] Werbos, P. J. (1995). *Backpropagation: Basics and New Developments.* In The Handbook of Brain Theory and Neural Networks (Ed. Arbib M.) (pp. 134 – 139). Cambridge: MIT Press.

Received 5 January 2005.