Learning as an Inverse Problem in Reproducing Kernel Hilbert Spaces

Kůrková, Věra
2010

Institute of Computer Science
Academy of Sciences of the Czech Republic

Learning as an Inverse Problem in Reproducing Kernel Hilbert Spaces

Věra Kůrková

Technical report No. 1094

November 2010

Pod Vodárenskou věží 2, 182 07 Prague 8, phone: +420 266 051 111, fax: +420 286 585 789,
e-mail:ics@cs.cas.cz

Institute of Computer Science
Academy of Sciences of the Czech Republic

Learning as an Inverse Problem in Reproducing Kernel Hilbert Spaces

Věra Kůrková[1]

Technical report No. 1094

November 2010

Abstract:

Applications of methods from theory of inverse problems to learning from data are studied. It is shown that learning modeled as minimization of error functionals can be reformulated in terms of inverse problems defined by evaluation and inclusion operators. Methods from theory of inverse problems are used to create a theoretical framework for study of behavior of error functionals, to obtain simple proofs of characterizations of their argminima, and to get some insight into an effect of regularization on improving generalization. Moreover, investigation of learning from data in the context of theory of inverse problems shows usefulness of the choice of Hilbert spaces defined by kernels (called reproducing kernel Hilbert spaces) as suitable ambient function spaces.

Keywords:
learning form data, minimization of expected and empirical error functionals, inverse problems, evaluation and inclusion operators, reproducing kernel Hilbert spaces

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

Learning as an Inverse Problem in Reproducing Kernel Hilbert Spaces

Věra Kůrková
Institute of Computer Science, Academy of Sciences of the Czech Republic
Pod Vodárenskou věží 2, 182 07, Prague 8, Czech Republic
vera@cs.cas.cz

Abstract

Applications of methods from theory of inverse problems to learning from data are studied. It is shown that learning modeled as minimization of error functionals can be reformulated in terms of inverse problems defined by evaluation and inclusion operators. Methods from theory of inverse problems are used to create a theoretical framework for study of behavior of error functionals, to obtain simple proofs of characterizations of their argminima, and to get some insight into an effect of regularization on improving generalization. Moreover, investigation of learning from data in the context of theory of inverse problems shows usefulness of the choice of Hilbert spaces defined by kernels (called reproducing kernel Hilbert spaces) as suitable ambient function spaces.

Keywords: learning form data, minimization of expected and empirical error functionals, inverse problems, evaluation and inclusion operators, reproducing kernel Hilbert spaces

1 Introduction

Inverse problems have been encountered in many branches of applied science and methodology for their solutions has been well developed [4, 18, 12, 19]. Recently, applications of concepts and methods from theory of inverse problems have also been extended to learning from data.

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 functions of the model fit well to the training data. But 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 [17] introduced Tikhonov regularization into learning theory as a means of improving generalization. They proposed to add to error functionals stabilizers which penalize undesired properties of input-output functions such as high-frequency oscillations [27]. Girosi, Jones and Poggio [16] considered stabilizers penalizing high frequencies in the Fourier representation of a potential solution. In practical applications, various simple stabilizers (such as norms based on derivatives [5] or ℓ_1 or ℓ_2-norm of output weights) have been used successfully [13, 20].

Later, Girosi [15] showed that stabilizers of this type belong to a wider class formed by the squares of norms on a special type of Hilbert spaces defined by kernels, which are called reproducing kernel Hilbert spaces (RKHS). These spaces were formally defined by Aronszajn [2], but their theory includes many earlier results by Mercer [26] and Schönebrg [30]. In addition to providing a rich variety of stabilizers, kernels can also increase chances for linear separation of more types of data by transforming geometry of input spaces. Aizerman, Braverman and Rozonoer [1] used kernels (under the name potential functions) to solve classification tasks by embedding input spaces into higher dimensional Hilbert spaces. Boser, Guyon and Vapnik [7] and Cortes and Vapnik [8] farther developed this classification method into the concept of the support vector machine, which became a widely used classification algorithm. Cucker and Smale [10] theoretically investigated learning as an optimization of error functionals over RKHSs. They characterized argminima of regularized error functionals over these spaces and used this characterization to design an alternative learning algorithm (see also [28]).

Kůrková [21, 22] and De Vito et al. [32] proposed to represent minimizations of error functionals as inverse problems defined by evaluation and inclusion operators. In this paper, we father develop this representation. Investigation of learning in terms of inverse problems leads to the choice of RKHSs as ambient function spaces because on these spaces the evaluation and inclusion operators are continuous. We show that in addition to continuity, these operators have on RKHSs many other useful properties which allow easy application of methods from theory of inverse problems. Thus we obtain simpler proofs of characterizations of argminima of error functionals than those obtained in [10] which hold under milder conditions on kernels and their domains. We also compare a regularized case with a non regularized one to obtains some insight into the effect of regularization on theoretically optimal solutions. The reformulation of learning in terms of inverse problems shows connections of modern learning theory with many classical problems from physics.

The paper is organized as follows. Section 2 presents basic concepts and notations on learning from data. In section 3, it is shown that minimization of error functionals with the quadratic loss function can be reformulated in terms of inverse problems defined by inclusion and evaluation operators. In section 3, a class of Hilbert spaces, called reproducing kernel Hilbert spaces (RKHSs), is described and its basic properties are briefly recalled. In section 5, properties of inclusion and evaluation operators on these spaces are investigated using methods from functional analysis. In sections 6 and 7, these properties are combined with results from theory of inverse problems to describe theoretically optimal solutions of learning tasks and to compare regularized and non regularized cases.

2 Error functionals with quadratic loss functions

Learning from data has been modeled as an optimization problem of a search for a function computable by a given computational model minimizing certain error functionals defined by the data. In learning theory, the data has been described by probability distributions or samples of so called training data.

For X a measurable subset of $\mathbb{R}^d$ and Y a bounded subset of $\mathbb{R}$, let ρ be a non degenerate (no non empty open set has measure zero) probability measure on $Z = X \times Y$ ($\rho(Z) = 1$). The *expected error functional* (sometimes also called expected risk or theoretical error) $\mathcal{E}_{\rho,V}$ determined by ρ and a *loss function* $V : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ is defined for those f in the set $\mathcal{M}(X)$ of all bounded ρ-measurable functions on X for which the integral

$$\mathcal{E}_{\rho,V}(f) = \int_Z V(f(x), y)\, d\rho$$

is finite. The most common loss function is the *quadratic loss* defined as $V(u,v) = (u-v)^2$. We denote by $\mathcal{E}_\rho$ the *expected error with the quadratic loss*, i.e.,

$$\mathcal{E}_\rho(f) = \int_Z (f(x) - y)^2\, d\rho\,. \tag{1}$$

Various learning algorithms (such as back-propagation or genetic algorithms, see, e.g., [13, 20]) aim to minimize a discretized version of the expected error called the *empirical error*. It is determined by a sample $z = \{(u_i, v_i) \in X \times Y \,|\, i = 1, \ldots, m\}$ of input-output pairs of data. The empirical error is denoted $\mathcal{E}_{z,V}$ and defined as

$$\mathcal{E}_{z,V}(f) = \frac{1}{m} \sum_{i=1}^{m} V(f(u_i), v_i).$$

We denote by $\mathcal{E}_z$ the *empirical error with the quadratic loss function*, i.e.,

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

One of many advantages of the quadratic loss function is that it enables to reformulate minimizations of expected and empirical errors as minimizations of distances from certain "optimal" functions.

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 to X, and let $\mathcal{L}^2_{\rho_X}(X)$ denote the space of all functions on X satisfying $\int_X f^2(x) d\rho_X(x) < \infty$ with the norm defined as $\|f\|_{\mathcal{L}^2_{\rho_X}} = \sqrt{\int_X f(x)^2 d\rho_X(x)}$. It is easy to see and well-known [10] that the minimum of $\mathcal{E}_\rho$ over the set $\mathcal{L}^2_{\rho_X}(X)$ is achieved at the *regression function* f_ρ defined for every $x \in X$ as

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

where $\rho(y|x)$ is the *conditional w.r.t. x probability measure* on Y. Setting σ^2_ρ, we get

$$\min_{f \in \mathcal{L}^2_{\rho_X}(X)} \mathcal{E}_\rho(f) = \mathcal{E}_\rho(f_\rho) = \sigma^2_\rho.$$

Moreover, for every $f \in \mathcal{L}^2_{\rho_X}(X)$

$$\mathcal{E}_\rho(f) = \int_X (f(x) - f_\rho(x))^2 d\rho_X + \sigma_\rho^2 = \|f - f_\rho\|^2_{\mathcal{L}^2_{\rho_X}} + \sigma_\rho^2 \tag{3}$$

(see,e.g., [10, p.5]). So on the space $\mathcal{L}^2_{\rho_X}(X)$, minimization of the expected error functional $\mathcal{E}_\rho$ with the quadratic loss is equivalent to minimization of the $\mathcal{L}^2_{\rho_X}$-distance from its minimum point f_ρ.

Also the empirical error functional $\mathcal{E}_z$ can be represented in terms of a distance functional. For a sample $z = ((u_1, v_1), \ldots, (u_m, v_m))$, set $u = (u_1, \ldots, u_m)$, $v = (v_1, \ldots, v_m)$, and let $\|.\|_{2,m}$ denote the weighted ℓ^2-norm on $\mathbb{R}^m$ defined as

$$\|x\|_{2,m} = \sqrt{\frac{1}{m}\sum_{i=1}^{m} x_i^2}.$$

Then for every $f \in \mathcal{L}^2_{\rho_X}(X)$ we have

$$\mathcal{E}_z(f) = \|(f(u_1), \ldots, f(u_m)) - (v_1, \ldots, v_m)\|^2_{2,m}. \tag{4}$$

So minimization of the empirical error $\mathcal{E}_z$ over $\mathcal{L}^2_{\rho_X}(X)$ is equivalent to minimization of the $\|.\|_{2,m}$-distance between the vector of the output data $v = (v_1, \ldots, v_m)$ and a vector $(f(u_1), \ldots, f(u_m))$ obtained by evaluating a function $f \in \mathcal{L}^2_{\rho_X}(X)$ at the input data $u = (u_1, \ldots, u_m)$.

3 Inverse problems in learning

The equivalences (3) and (4) of minimizations of the error functionals $\mathcal{E}_\rho$ and $\mathcal{E}_z$ enable investigation of learning from data in the framework of theory of inverse problems. In this section, we describe operators defining such inverse problems and state properties of solutions of inverse problems which will be used in next sections as tools for characterization of optimal solutions of learning tasks.

Let $(\mathcal{H}, \|.\|_\mathcal{H})$ be a Hilbert space, such that $\mathcal{H}$ is a linear subspace of $\mathcal{L}^2_{\rho_X}(X)$ and the norm $\|.\|_\mathcal{H}$ is any norm induced by an inner product (not necessarily the one inherited from $\|.\|_{\mathcal{L}^2_{\rho_X}}$ by restricting it to $\mathcal{H}$). 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 (3), we have

$$\mathcal{E}_\rho(f) = \|J(f) - f_\rho\|^2_{\mathcal{L}^2_{\rho_X}} + \sigma_\rho^2. \tag{5}$$

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

For any vector $u = (u_1, \ldots, u_m) \in \mathbb{R}^m$ and any space $\mathcal{S}$ of functions on some $X \subseteq \mathbb{R}^d$ such that $u \in X$, let $J_u : \mathcal{S} \to \mathbb{R}^m$ denote an *evaluation operator* defined for all $f \in \mathcal{S}$ as

$$J_u(f) = (f(u_1), \ldots, f(u_m)). \tag{6}$$

The representation (4) implies that

$$\mathcal{E}_z(f) = \|J_u(f) - v\|^2_{2,m}. \tag{7}$$

Thus minimizing the empirical error $\mathcal{E}_z$ with the quadratic loss function over $\mathcal{S}$ is equivalent to solving an *inverse problem given by the evaluation operator J_u for the data v*.

To describe argminima of the error functionals $\mathcal{E}_\rho$ and $\mathcal{E}_z$ and argminima of some of their regularized modifications, we take advantage of the following basic results from theory of inverse problems from [4, pp.68-70] and [18, pp.74-76]. By $R(A)$ is denoted the *range* of an operator $A : (\mathcal{X}, \|.\|_\mathcal{X}) \to (\mathcal{Y}, \|.\|_\mathcal{Y})$, by $\pi_{cl\,R(A)}$ the *projection on the closure* of $R(A)$ in $(\mathcal{Y}, \|.\|_\mathcal{Y})$, and by A^* the adjoint of A (the unique operator $A^* : (\mathcal{Y}, \|.\|_\mathcal{Y}) \to (\mathcal{X}, \|.\|_\mathcal{X})$ such that for all $f \in \mathcal{X}$ and all $g \in \mathcal{Y}$, $\langle A(f), g\rangle_\mathcal{Y} = \langle f, A^*(g)\rangle_\mathcal{X}$).

Theorem 3.1 *Let $A : (\mathcal{X}, \|.\|_\mathcal{X}) \to (\mathcal{Y}, \|.\|_\mathcal{Y})$ be a continuous linear operator between two Hilbert spaces. Then there exists a unique continuous linear pseudoinverse operator A^+ such that*
(i) if $R(A)$ is closed, then $A^+ : \mathcal{Y} \to \mathcal{X}$;
(ii) if $R(A)$ is not closed, then $A^+ : \mathcal{Y}^+ \to \mathcal{X}$, where $\mathcal{Y}^+ = \{g \in \mathcal{Y} \,|\, \pi_{clR(A)}(g) \in R(A)\}$;
(iii) for every g in the domain of A^+, $\|A^+(g)\|_\mathcal{X} = \min_{f^o \in S(g)} \|f^o\|_\mathcal{X}$, where $S(g) = \operatorname{argmin}(\mathcal{X}, \|A(.) - g\|_\mathcal{Y})$, $AA^+(g) = \pi_{clR(A)}(g)$, and

$$A^+ = (A^*A)^+A^* = A^*(AA^*)^+; \tag{8}$$

(iv) 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

$$A^\gamma = (A^*A + \gamma I_\mathcal{X})^{-1}A^* = A^*(AA^* + \gamma I_\mathcal{Y})^{-1} \tag{9}$$

where $I_\mathcal{X}, I_\mathcal{Y}$ denote the identity operators on $\mathcal{X}$ and $\mathcal{Y}$, resp.;
(v) for every g in the domain of A^+, $\lim_{\gamma \to 0} \|A^\gamma(g) - A^+(g)\|_{cX} = 0$.

4 Reproducing kernel Hilbert spaces

The representations (5) and (7) of minimizations of error functionals with the quadratic loss function as inverse problems provide useful tools for description of optimal solutions of learning tasks. However, assumptions of Theorem 3.1 require computational models with input-output functions belonging to Hilbert spaces on which evaluation functionals are continuous and so are also their inclusions to $\mathcal{L}^2_{\rho_X}(X)$.

Spaces $(\mathcal{L}^2_\mu(X), \|.\|_{\mathcal{L}^2_\mu})$ cannot be used as such ambient function spaces because evaluation functionals on them are not continuous. Indeed, one can easily construct many sequences of functions in $\mathcal{L}^2_\mu(X)$ which all have the same values of their $\mathcal{L}^2_\mu$-norms but their evaluations at some points diverge (for example, some sequences of

functions converging to the Dirac delta function). On the other hand on the space $\mathcal{C}(X)$ of bounded continuous functions with the supremum norm $\|.\|_{\sup}$, all evaluation functionals are continuous, but $(\mathcal{C}(X), \|.\|_{\sup})$ is not a Hilbert space. Thus we need some function spaces combining good properties of two types of spaces: a Hilbert space structure as in the case of $(\mathcal{L}^2_\mu(X), \|.\|_{\mathcal{L}^2_\mu})$ and continuity of all evaluation functionals as in the case of $(\mathcal{C}(X), \|.\|_{\sup})$.

Fortunately, the space $\mathcal{L}^2_{\rho_X}(X)$ contains many subspaces with suitable inner products on which all evaluation functionals are continuous. Moreover, some of such spaces contain input-output functions of widely used computational models. These spaces are called *reproducing kernel Hilbert spaces* (RKHSs). They were defined by Aronszajn [2] as *Hilbert spaces of pointwise defined functions on which all evaluation functionals are continuous*. They are called RKHSs because each such space is uniquely determined by a *symmetric positive semidefinite kernel*. Recall that a function $K : X \times X \to \mathbb{R}$ is called *positive semidefinite* if for any positive integer m, any $x_1, \ldots, x_m \in X$ and any $a_1, \ldots, a_m \in \mathbb{R}$

$$\sum_{i=1}^{m} \sum_{j=1}^{m} a_i a_j K(x_i, x_j) \geq 0.$$

RKHSs became popular in soft-computing due to the use of kernels in support vector machines [31, 9, 29] but they have been studied in mathematics since 1950 and their theory includes many earlier results by Schönberg [30] and Mercer [26]. Since 1990s, RKHS have been used as useful ambient function spaces in data analysis [33]. For their theory see, e.g., [2, 3, 10]. Here we just recall that a RKHS determined by $K : X \times X \to \mathbb{R}$, denoted

$$\mathcal{H}_K(X),$$

is formed by all linear combinations of functions of the form $K_x : X \to \mathbb{R}$, $x \in X$, defined as

$$K_x(y) = K(x, y)$$

together with limits of Cauchy sequences in the norm $\|.\|_K$ of these linear combinations. The functions K_x are called *representers*. The norm $\|.\|_K$ is induced by the inner product $\langle ., . \rangle_K$, which is defined on representers as

$$\langle K_x, K_y \rangle_K = K(x, y).$$

The most important property of reproducing kernel Hilbert spaces is so called *reproducing property* guaranteeing that for all $f \in \mathcal{H}_K(X)$ and all $x \in X$

$$\langle f, K_x \rangle_K = f(x). \tag{10}$$

So the representers play a similar role as the Dirac delta in the distribution theory [34], but in contrast to the Dirac distribution, representers are real-valued functions.

A paradigmatic example of a positive semidefinite kernel is the *Gaussian kernel* $K(x, y) = e^{-\|x-y\|^2}$. A reproducing kernel Hilbert space defined by the Gaussian kernel contains all linear combinations of translations of the Gaussian function. Such linear combinations can be computed as input-output functions of an important computational

model called *network with Gaussian radial units with varying centroids and fixed width* (for a survey on properties and applications of such networks see, e.g., [20]).

A simplest type of positive semidefinite functions are *product kernels* which have the form

$$K(x,y) = k(x)\,k(y)$$

where $k : X \to \mathbb{R}$ is a one-variable function. Another large class of kernels are *convolution kernels*. These kernels have the form

$$K(x,y) = k(x-y) \tag{11}$$

where $k : \mathbb{R}^d \to \mathbb{R}$ is an even function. By the Bochner theorem [6], when the Fourier transform $\tilde{k}$ is positive then K defined in (11) is positive semidefinite. It was shown in [25] (see also [15]) that for such convolution kernels with $k \in \mathcal{L}^2(\mathbb{R}^d) \cap \mathcal{L}^1(\mathbb{R}^d)$, the value of $\|f\|_K^2$ at any $f \in \mathcal{H}_K(\mathbb{R}^d)$ can be expressed as

$$\|f\|_K^2 = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \frac{\tilde{f}(\omega)^2}{\tilde{k}(\omega)}\, d\omega. \tag{12}$$

Note that the set of all symmetric positive semidefinite functions is quite large as it is closed under various operations such as finite linear combinations with positive coefficients, pointwise limits or tensor products [3]. So norms on RKHSs offer a rich class of stabilizers suitable for Tikhonov's regularization of inverse problems modeling minimization of expected and empirical error functionals.

5 Properties of inclusion and evaluation operators on RKHSs

To apply results from theory of inverse problems to learning from data we first need to derive some properties of evaluation and inclusion operators on RKHSs.

For $X \subseteq \mathbb{R}^d$, a kernel $K : X \times X \to \mathbb{R}$, a σ-finite measure μ on X, define an integral operator $L_{K,\mu} = L_K$ on the subspace of $\mathcal{L}^2_\mu(X)$ formed by those g for which for every $x \in X$ the integral

$$L_K(g)(x) := \int_X g(y)\,K(x,y)\,d\mu(y) \tag{13}$$

is finite.

The following proposition gives a condition on a kernel K which implies that the reproducing kernel Hilbert space $\mathcal{H}_K(X)$ induced by the kernel K is a subspace of $\mathcal{L}^2_\mu(X)$ and the *inclusion operator* $J_K : (\mathcal{H}_K(X), \|.\|_K) \to (\mathcal{L}^2_\mu(X), \|.\|_{\mathcal{L}^2_\mu})$ is continuous.

Recall that every bounded linear operator $T : (\mathcal{X}, \|\cdot\|_\mathcal{X}) \to (\mathcal{Y}, \|\cdot\|_\mathcal{Y})$ between two Hilbert spaces has an *adjoint* operator $T^* : (\mathcal{Y}, \|\cdot\|_\mathcal{Y}) \to (\mathcal{X}, \|\cdot\|_\mathcal{X})$ [14]. An operator T is called a *Hilbert-Schmidt* operator if for any orthonormal basis $\{e_j \mid j \in I\}$ of $(\mathcal{X}, \|.\|_\mathcal{X})$, $\sum_{j\in I} \|T(e_j)\|^2_\mathcal{Y} < \infty$.

Proposition 5.1 *Let $X \subseteq \mathbb{R}^d$ be measurable, μ be a σ-finite measure on X, $K : X \times X \to \mathbb{R}$ be a symmetric positive semidefinite kernel such that $\int_X K(x,x)\,d\mu(x) < \infty$. Then*

(i) $\mathcal{H}_K(X) \subseteq \mathcal{L}^2_\mu(X)$ and the inclusion operator $J_K : (\mathcal{H}_K(X), \|.\|_K) \to (\mathcal{L}^2_\mu(X), \|.\|_{\mathcal{L}^2_\mu})$ is continuous;
(ii) $L_K = J_K^* : (\mathcal{L}^2_\mu(X), \|.\|_{\mathcal{L}^2_\mu}) \to (\mathcal{H}_K(X), \|.\|_K)$ and so L_K is continuous;
(iii) J_K is a Hilbert-Schmidt operator and both J_K and L_K are compact.

Proof. (i) By the reproducing property (10) and the Cauchy-Schwartz inequality, for every $f \in \mathcal{H}_K(X)$ we have

$$\|J_K(f)\|^2_{\mathcal{L}^2_\mu} = \int_X f(x)^2 d\mu(x) = \int \langle f, K_x\rangle^2_K \, d\mu(x) \leq \|f\|^2_K \int_X K(x,x) d\mu(x). \qquad (14)$$

When $\int_X K(x,x)\, d\mu(x) < \infty$, (14) implies $f = J_K(f) \in \mathcal{L}^2_\mu(X)$ and continuity of J_K.

(ii) Every continuous linear operator between two Hilbert spaces has a continuous adjoint operator (see,e.g., [14]) and so J_K has an adjoint J_K^*. It follows from the reproducing property (10) and the definition of an adjoint operator that for all $x \in X$ and all $g \in \mathcal{L}^2_\mu(X)$,

$$J_K^*(g)(x) = \langle J_K^*(g), K_x\rangle_K = \langle g, J_K(K_x)\rangle_{\mathcal{L}^2_\mu} = \int_X g(y)\, K(x,y) d\mu(y) = L_K(g)(x)$$

and so $L_K = J_K^* : (\mathcal{L}^2_\mu(X), \|.\|_{\mathcal{L}^2}) \to (\mathcal{H}_K(X), \|.\|_K)$.

(iii) By the reproducing property (10) and the assumption $\int_X K(x,x) d\mu(x) < \infty$, for every orthonormal basis $\{e_i \,|\, i \in I\}$ of $\mathcal{H}_K(X)$ we have

$$\sum_{i\in I} \|J_K(e_i)\|^2_{\mathcal{L}^2_\mu} = \sum_{i\in I} \int_X e_i(x)^2 \, d\mu(x) = \int_X \sum_{i\in I} \langle e_i, K_x\rangle_K \, e_i(x) d\mu(x) = \int_X K(x,x) d\mu(x) < \infty.$$

Thus J_K is a Hilbert-Schmidt operator which implies that its adjoint is a Hilbert-Schmidt operator too and that both these operators are compact [14, p.187] □

Note that for any convolution kernel $K(x,y) = k(x-y)$, the assumption of Proposition 5.1

$$\int_X K(x,x) d\mu(x) = \int_X k(0) d\mu(x) = k(0)\, \mu(X) < \infty$$

holds if and only if $\mu(X)$ is finite.

In learning theory, it is assumed that the measure ρ is a probabilistic measure and hence $\rho_X(X) = 1$. Thus for any bounded kernel K, we have $\int_X K(x,x) d\rho_X(x) < \infty$. So by Proposition 5.1(ii), for every bounded symmetric positive semidefinite kernel K, the integral operator $L_K : (\mathcal{L}^2_{\rho_X}, \|.\|_{\mathcal{L}^2_{\rho_X}}) \to (\mathcal{H}_K(X), \|.\|_K)$ is a compact operator.

To derive some useful properties of RKHSs, we apply spectral theory to the operator

$$T_K := J_K L_K : (\mathcal{L}^2_\mu(X), \|.\|_{\mathcal{L}^2_\mu}) \to (\mathcal{L}^2_\mu(X), \|.\|_{\mathcal{L}^2_\mu})$$

obtained by composing L_K with J_K. The next theorem summarizes some properties of the operators J_K and T_K.

Theorem 5.2 *Let $X \subseteq \mathbb{R}^d$ be measurable, μ be a σ-finite measure on X, $K : X \times X \to \mathbb{R}$ be a symmetric positive semidefinite kernel such that $\int_X K(x,x)d\mu(x) < \infty$. Then*
(i) $T_K : (\mathcal{L}^2_\mu(X), \|.\|_{\mathcal{L}^2_\mu}) \to (\mathcal{L}^2_\mu(X), \|.\|_{\mathcal{L}^2_\mu})$ is a compact, self-adjoint, and positive operator;
(ii) there exists at most countable orthonormal family $\{\psi_j \,|\, j \in I\}$ in $(\mathcal{L}^2_\mu(X), \|.\|_{\mathcal{L}^2_\mu})$ formed by eigenfunction of T_K with the corresponding family of non negative eigenvalues $\{\lambda_j \,|\, j \in I\}$ ordered non increasingly, which in the case of I infinite converges to zero, such that for every $f \in \mathcal{L}^2_\mu(X)$,

$$T_K(f) = \sum_{j \in I} \lambda_j \langle f, \psi_j \rangle_{\mathcal{L}^2_\mu} \psi_j; \tag{15}$$

and

$$K(x,y) = \sum_{j \in I} \lambda_j \psi_j(x) \psi_j(y); \tag{16}$$

(iii) $\{\sqrt{\lambda_j}\psi_j \,|\, j \in I\}$ is an orthonormal basis of $(\mathcal{H}_K(X), \|.\|_K)$ and $\sum_{j \in I} \lambda_j < \infty$;
(iv) $\mathrm{cl}_{\mathcal{L}^2_\mu} R(J_K) = \mathrm{cl}_{\mathcal{L}^2_\mu} J_K(\mathcal{H}_K(X)) = \{f \in \mathcal{L}^2_\mu(X) \,|\, f = \sum_{j \in I} \langle f, \psi_j \rangle_{\mathcal{L}^2_\mu} \psi_j\}$;
(v) $R(J_K)$ is closed if and only if I is finite.

Proof. (i) By Proposition 5.1 (iii), J_K and L_K are compact and thus also T_K is compact. As $T_K = J_K L_K = J_K J_K^*$, it is self-adjoint. We have $\langle T_K(f), f \rangle_{\mathcal{L}^2_\mu} = \langle J_K L_K(f), f \rangle_{\mathcal{L}^2_\mu} = \langle L_K(f), L_K(f) \rangle_K \geq 0$ and so T_K is positive.

(ii) The representation (15) of T_K follows from the Spectral theorem [11, p. 683], which holds for all compact self-adjoint operators.

(iii) For all $j \in I$, $\lambda_j \psi_j = T_K(\psi_j) = J_K L_K(\psi_j)$ and thus $\psi_j \in \mathcal{H}_K(X)$ and $L_K(\psi_j) = \lambda_j \psi_j$. So we have $1 = \langle J_K(\psi_j), \psi_j \rangle_{\mathcal{L}^2_\mu} = \langle \psi_j, L_K(\psi_j) \rangle_K = \langle \psi_j, \lambda_j \psi_j \rangle_K = \lambda_j \|\psi_j\|^2_K$. As J_K and L_K are adjoints we have for all $i, j \in I$ such that $i \neq j$, $\langle \lambda_i \psi_i, \psi_j \rangle_K = \langle L_K(\psi_i), \psi_j \rangle_K = \langle \psi_i, J_K(\psi_j) \rangle_{\mathcal{L}^2_\mu} = \langle \psi_i, \psi_j \rangle_{\mathcal{L}^2_\mu} = 0$. Thus $\{\sqrt{\lambda_j}\psi_j \,|\, j \in I\}$ is an orthonormal family in $\mathcal{H}_K(X)$. By (ii) and the reproducing property (10), all representers K_x can be expressed as $K_x = \sum_{j \in I} \lambda_j \psi_j(x) \psi_j = \sum_{j \in I} \lambda_j \langle K_x, \psi_j \rangle_K \psi_j = \sum_{j \in I} \langle K_x, \sqrt{\lambda_j}\psi_j \rangle_K \sqrt{\lambda_j}\psi_j$ and so $\{\sqrt{\lambda_j}\psi_j \,|\, j \in I\}$ is a basis of $\mathcal{H}_K(X)$. By Proposition 5.1 (iii), J_K is a Hilbert-Schmidt operator. Thus $\sum_{j \in I} \lambda_j = \sum_{j \in I} \|J_K(\sqrt{\lambda_j}\psi_j)\|^2_{\mathcal{L}^2_\mu} < \infty$.

(iv) By (ii) and (iii), $\{\psi_j \,|\, j \in I\}$ is an orthonormal basis of $(\mathrm{cl}_{\mathcal{L}^2_\mu} R(J_K), \|.\|_{\mathcal{L}^2_\mu})$. Thus for every $f \in \mathcal{L}^2_\mu(X)$, $f \in \mathrm{cl}_{\mathcal{L}^2_\mu} R(J_K)$ if and only if $f = \sum_{j \in I} \langle f, \psi_j \rangle_{\mathcal{L}^2_\mu} \psi_j$.

(v) By (iii), for every $f \in R(J_K) = J_K(\mathcal{H}_K(X))$, $\sum_{j \in I} \lambda_j \langle f, \psi_j \rangle^2_K < \infty$. As J_K and L_K are adjoints, we have $\sum_{j \in I} \frac{1}{\lambda_j} \langle f, \lambda_j \psi_j \rangle^2_K = \sum_{j \in I} \frac{1}{\lambda_j} \langle f, L_K(\psi_j) \rangle^2_K = \sum_{j \in I} \frac{1}{\lambda_j} \langle f, \psi_j \rangle^2_{\mathcal{L}^2_\mu} < \infty$. Thus by (iv), if $\mathrm{cl}_{\mathcal{L}^2_\mu} R(J_K) = R(J_K)$ then for all $f \in \mathcal{L}^2_\mu(X)$, $\sum_{j \in I} \langle f, \psi_j \rangle^2_{\mathcal{L}^2_\mu} < \infty$ implies $\sum_{j \in I} \frac{1}{\lambda_j} \langle f, \psi_j \rangle^2_{\mathcal{L}^2_\mu} < \infty$. This property holds if and only if I is finite. □

Recall that kernels with the representation $K(x,y) = \sum_{j \in I} \lambda_j \psi_j(x) \psi_j(y)$ with I finite are called *degenerate*.

Corollary 5.3 *Let $X \subseteq \mathbb{R}^d$ be measurable, μ be a σ-finite measure on X, $K : X \times X \to \mathbb{R}$ be a symmetric positive semidefinite kernel such that $\int_X K(x,x)d\mu(x) < \infty$. If K is degenerate, then the domain of the pseudoinverse operator J_K^+ of the inclusion operator $J_K : (\mathcal{H}_K(X), \|.\|_K) \to (\mathcal{L}^2_\mu(X), \|.\|_{\mathcal{L}^2_\mu})$ is the whole space $L^2_\mu(X)$. If K is non degenerate, then the domain of J_K^+ is the subspace $\{f \in \mathcal{L}^2_\mu(X) \,|\, \sum_{j\in I} \frac{1}{\lambda_j}\langle f, \psi_j\rangle^2_{\mathcal{L}^2_\mu} < \infty\}$.*

Proof. If K is degenerate, then by Theorem 5.2 (v), $R(J_K)$ is closed and thus by Theorem 3.1 (i), the domain of J_K^+ is the whole space $\mathcal{L}^2_\mu(X)$. If K is non degenerate, then by Theorem 5.2 (v), $R(J_K)$ is not closed and so by Theorem 3.1 (ii), the domain of J_K^+ is the subspace $\{f \in \mathcal{L}^2_\mu(X) \,|\, \pi_{\mathrm{cl}R(J_K)}(f) \in J_K(\mathcal{H}_K(X))\}$. By Theorem 5.2 (iv), it is equal to the subspace $\{f \in \mathcal{L}^2_\mu(X) \,|\, \sum_{j\in I} \frac{1}{\lambda_j}\langle f, \psi_j\rangle^2_{\mathcal{L}^2_\mu} < \infty\}$. □

The next proposition describes properties of evaluation operators on RKHSs.

Proposition 5.4 *Let $X \subseteq \mathbb{R}^d$, $K : X \times X \to \mathbb{R}$ be a symmetric positive semidefite kernel. Then for every positive integer m and every $u \in X^m$*
(i) $J_u : (\mathcal{H}_K(X), \|.\|_K) \to (\mathbb{R}^m, \|.\|_{2,m})$ is continuous;
(ii) $R(J_u)$ is closed in $(\mathbb{R}^m, \|.\|_{2,m})$;
(iii) J_u is compact;
(iv) the adjoint $J_u^* : (\mathbb{R}^m, \|.\|_{2,m}) \to (\mathcal{H}_K(X), \|.\|_K)$ satisfies for all $x \in X$ and all $w \in \mathbb{R}^m$,

$$J_u^*(w)(x) = \frac{1}{m}\sum_{i=1}^m w_i K(x, u_i).$$

Proof. (i) Continuity of J_u follows from the definition of a RKHS.

(ii) Every linear subspace of a finite dimensional space is closed and so is $R(J_u)$.

(iii) As every continuous operator with a finite range is compact [14, p. 188], so is J_u.

(iv) By the reproducing property (10) and the definition of an adjoint operator, for every $x \in X$ and every $w \in \mathbb{R}^m$ we have $J_u^*(w)(x) = \langle J_u^*(w), K_x\rangle_K = \langle w, J_u(K_x)\rangle_{2,m} = \frac{1}{m}\sum_{i=1}^m w_i K(x, u_i)$. □

6 Minimization of expected error

In this section we apply results from theory of inverse problems to inclusion and evaluation operators on reproducing kernel Hilbert spaces.

First, we show that the expected error $\mathcal{E}_\rho$ achieves its minimum over a RKHS $\mathcal{H}_K(X)$ if and only if the projection $\bar{f}_\rho$ of the regression function f_ρ on the $\mathcal{L}^2_{\rho_X}$-closure of $\mathcal{H}_K(X)$ is contained in $\mathcal{H}_K(X)$.

Theorem 6.1 *Let $X \subseteq \mathbb{R}^d$ be measurable, $Y \subset \mathbb{R}$ bounded, ρ be a non degenerate probability measure on $X \times Y$, $K : X \times X \to \mathbb{R}$ be a symmetric positive semidefinite*

kernel such that $\int_X K(x,x)d\rho_X(x) < \infty$, $\{\lambda_j \,|\, j \in I\}$ *and* $\{\psi_j \,|\, j \in I\}$ *be eigenvalues and eigenfunctions, resp., of the operator* $T_K : (\mathcal{L}^2_{\rho_X}(X), \|.\|_{\mathcal{L}^2_{\rho_X}}) \to (\mathcal{L}^2_{\rho_X}(X), \|.\|_{\mathcal{L}^2_{\rho_X}})$, *and* $\bar{f}_\rho = \sum_{j\in I} \langle f_\rho, \psi_j \rangle_{\mathcal{L}^2_{\rho_X}} \psi_j$. *Then*
(i) in the case of K is degenerate, $\bar{f}_\rho \in \mathcal{H}_K(X)$ *and* $\min_{f\in\mathcal{H}_K(X)} \mathcal{E}_\rho(f) = \mathcal{E}_\rho(\bar{f}_\rho) = \|\bar{f}_\rho - f_\rho\|^2_{\mathcal{L}^2_{\rho_X}} + \sigma^2_\rho$;
(ii) in the case of K is non degenerate, $\mathcal{E}_\rho$ *achieves minimum over* $\mathcal{H}_K(X)$ *if and only if* $\bar{f}_\rho \in \mathcal{H}_K(X)$ *which is equivalent to* $\sum_{j\in I} \frac{1}{\lambda_j} \langle f_\rho, \psi_j \rangle^2_{\mathcal{L}^2_{\rho_X}} < \infty$. *If* $\bar{f}_\rho \in \mathcal{H}_K(X)$, *then* $\min_{f\in\mathcal{H}_K(X)} \mathcal{E}_\rho(f) = \mathcal{E}_\rho(\bar{f}_\rho) = \|\bar{f}_\rho - f_\rho\|^2_{\mathcal{L}^2_{\rho_X}} + \sigma^2_\rho$.

Proof. By the representation (3), any argminimum of $\mathcal{E}_\rho$ over $\mathcal{H}_K(X)$ is a pseudosolution of an inverse problem defined by the operator J_K for the data f_ρ and $\mathcal{E}_\rho(f) = \|\bar{f}_\rho - f_\rho\|^2_{\mathcal{L}^2_{\rho_X}} + \sigma^2_\rho$. Then the statement follows from Corollary 5.3. □

As $\rho_X(X) = 1$, for every bounded kernel K, in particular for every convolution kernel, we have $\int_X K(x,x)d\rho_X(x) < \infty$. Thus we can apply Theorem 6.1 to minimization of an expected error functional $\mathcal{E}_\rho$ over any RKHS $\mathcal{H}_K(X)$ where K is a bounded or convolution kernel.

By Theorem 6.1, the expected error $\mathcal{E}_\rho$ achieves its minimum over $\mathcal{H}_K(X)$ only when the projection $\bar{f}_\rho$ of the regression function f_ρ to the $\mathcal{L}^2_{\rho_X}$-closure of $\mathcal{H}_K(X)$ belongs to $\mathcal{H}_K(X)$. The next theorem shows that when a stabilizer $\gamma\|.\|^2_K$ is added to the expected error $\mathcal{E}_\rho$, then the modified functional always achieves a minimum at a unique function in $\mathcal{H}_K(X)$. Let

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

denote the *Tikhonov regularization of the expected error* $\mathcal{E}_\rho$ *with the stabilizer* $\|.\|^2_K$ *and the parameter* γ.

To describe argminima of $\mathcal{E}_{\rho,\gamma,K}$ over $\mathcal{H}_K(X)$, we introduce a modified kernel $K_\gamma : X \times X \to \mathbb{R}$, which is defined for all $x, y \in X$ as

$$K_\gamma(x,y) = \sum_{j\in I}^{\infty} \frac{\lambda_j}{\lambda_j + \gamma} \psi_j(x)\psi_j(y), \tag{17}$$

where λ_j and ψ_j are the eigenvalues and eigenfunctions, resp., of the operator T_K described in Theorem 5.2 (ii). Let

$$L_{K_\gamma} : (\mathcal{L}^2_\mu(X), \|.\|_{\mathcal{L}^2_\mu}) \to (\mathcal{H}_{K_\gamma}(X), \|.\|_{K_\gamma})$$

denote the integral operator with the kernel K_γ. The next proposition states some properties of K_γ and L_{K_γ}.

Proposition 6.2 *Let* $X \subseteq \mathbb{R}^d$ *and* $K : X \times X$ *be is a symmetric positive semidefinite kernel. Then for all* $\gamma > 0$, $K_\gamma : X \times X \to \mathbb{R}$ *is a symmetric positive semidefinite kernel. If X is measurable and* μ *is a* σ*-finite measure on X, then* L_{K_γ} *maps* $\mathcal{L}^2_\mu(X)$ *to* $\mathcal{H}_K(X)$, *i.e.,* $R(L_{K_\gamma}) \subseteq \mathcal{H}_K(X)$.

Proof. By Theorem 5.2 (ii), all λ_j are non negative and so we have

$$K_\gamma(x,y) = \sum_{j\in I} \frac{\lambda_j}{\lambda_j+\gamma}\psi_j(x)\psi_j(y) \leq \frac{1}{\gamma}\sum_{j\in I}\lambda_j\psi_j(x)\psi_j(y) = \frac{1}{\gamma}K(x,y)$$

for all $x,y \in X$. Thus the sum of the series (17) is finite. Moreover, for all $j \in I$, the functions $\frac{\lambda_j}{\lambda_j+\gamma}\psi_j(x)\psi_j(y)$ are positive semidefinite because they are product kernels. It is easy to check that a pointwise limit of a sequence of positive semidefinite functions is positive semidefinite. Thus K_γ is a symmetric positive semidefinite kernel.

For all $f \in \mathcal{L}^2_\mu(X)$, $L_{K_\gamma}(f) = \sum_{j\in I}\frac{\lambda_j}{\lambda_j+\gamma}\langle f,\psi_j\rangle_{\mathcal{L}^2_\mu}\psi_j$. By Theorem 5.2 (ii), $L_{K_\gamma}(f) \in \mathcal{H}_K(X)$ if and only if $\sum_{j\in I}\frac{1}{\lambda_j}\left(\frac{\lambda_j}{\lambda_j+\gamma}\langle f,\psi_j\rangle_{\mathcal{L}^2_\mu}\right)^2 < \infty$. We have $\sum_{j\in I}\frac{1}{\lambda_j}\left(\frac{\lambda_j}{\lambda_j+\gamma}\langle f,\psi_j\rangle_{\mathcal{L}^2_\mu}\right)^2 \leq \sum_{j\in I}\frac{\lambda_j}{\gamma^2}\langle f,\psi_j\rangle^2_{\mathcal{L}^2_\mu} \leq \frac{\lambda_1}{\gamma^2}\sum_{j\in I}\langle f,\psi_j\rangle^2_{\mathcal{L}^2_\mu} = \frac{\lambda_1}{\gamma^2}\|f\|^2_{\mathcal{L}^2_\mu} < \infty$ and so $R(L_{K_\gamma}) \subseteq \mathcal{H}_K(X)$. □

Theorem 6.3 *Let $X \subseteq \mathbb{R}^d$ be measurable, $Y \subset \mathbb{R}$ be bounded, $K : X \times X \to \mathbb{R}$ be a continuous symmetric positive semidefinite kernel, ρ be a non degenerate probability measure on $X \times Y$. Then for every $\gamma > 0$, there exists a unique function $f^\gamma \in \mathcal{H}_K(X)$ minimizing $\mathcal{E}_{\rho,\gamma,K}$ such that*
(i) $f^\gamma = L_{K_\gamma}(f_\rho)$;
(ii) $\mathcal{E}_\rho(f^\gamma) - \mathcal{E}_\rho(f_\rho) = \|f^\gamma - f_\rho\|^2_{\mathcal{L}^2_{\rho_X}}$;
(iii) $\lim_{\gamma\to 0}\|f^\gamma - f_\rho\|_{\mathcal{L}^2_{\rho_X}} = 0$.

Proof. (i) By the representation (3), the argminimum of $\mathcal{E}_{\rho,\gamma,K}$ over $\mathcal{H}_K(X)$ is equal to the regularized solution of the inverse problem given by the operator J_K with the stabilizer $\|.\|^2_K$ and the parameter γ for the data f_ρ. By Proposition 5.1 (i) J_K is continuous and so by Theorem 3.1, $f^\gamma = J^\gamma_K(f_\rho)$ where $J^\gamma_K = J^*_K(J_KJ^*_K + \gamma I_{\mathcal{L}^2_{\rho_X}})^{-1}$.

By Proposition 5.1 (ii), $J^*_K = L_K$ and so we have $J^\gamma_K = L_K(T_K + \gamma I_{\mathcal{L}^2_{\rho_X}})^{-1}$. By Theorem 5.2(ii), J^γ_K has eigenvalues $\frac{\lambda_j}{\lambda_j+\gamma}$ and hence J^γ_K is equal to the operator L_{K_γ}. By Proposition 6.2, $R(L_{K_\gamma}) \subseteq \mathcal{H}_K(X)$ and so $L_{K_\gamma}(f_\rho) \in \mathcal{H}_K(X)$.

(ii) By the representation (3), $\mathcal{E}_\rho(f^\gamma) - \mathcal{E}_\rho(f_\rho) = \|f^\gamma - f_\rho\|^2_{\mathcal{L}^2_{\rho_X}}$.

(iii) For every $\gamma > 0$, we have $\|f_\rho - L_{K_\gamma}(f_\rho)\|_{\mathcal{L}^2_{\rho_X}} = \|\sum_{j\in I}(1-\frac{\lambda_j}{\lambda_j+\gamma})\langle f_\rho,\psi_j\rangle_{\mathcal{L}^2_{\rho_X}}\psi_j\|_{\mathcal{L}^2_{\rho_X}}$ $= \|\sum_{j\in I}\frac{\gamma}{\lambda_j+\gamma}\langle f_\rho,\psi_j\rangle_{\mathcal{L}^2_{\rho_X}}\psi_j\|_{\mathcal{L}^2_{\rho_X}} \leq \frac{\gamma}{\lambda_1+\gamma}\|\sum_{j\in I}\langle f_\rho,\psi_j\rangle_{\mathcal{L}^2_{\rho_X}}\psi_j\|_{\mathcal{L}^2_{\rho_X}} = \frac{\gamma}{\lambda_1+\gamma}\|\bar{f}_\rho\|_{\mathcal{L}^2_{\rho_X}}$. Thus $\lim_{\gamma\to 0}\|f^\gamma - f_\rho\|_{\mathcal{L}^2_{\rho_X}} = \lim_{\gamma\to 0}\|L_{K_\gamma}(f_\rho) - f_\rho\|_{\mathcal{L}^2_{\rho_X}} = 0$. □

By Theorem 6.3, for every $\gamma > 0$ there exists a unique function f^γ minimizing the regularized expected error $\mathcal{E}_{\rho,\gamma,K}$ over $\mathcal{H}_K(X)$. This function is the image of the regression function f_ρ under the integral operator L_{K_γ} which maps $\mathcal{L}^2_{\rho_X}(X)$ to $\mathcal{H}_K(X)$. The regularization modifies coefficients $w_j = \langle f_\rho,\psi_j\rangle_{\mathcal{L}^2_{\rho_X}}$ in the representation $\bar{f}_\rho = \sum_{j\in I} w_i\psi_j$ of the projection $\bar{f}_\rho$ of the regression function f_ρ on the $\mathcal{L}^2_{\rho_X}$-closure of

$\mathcal{H}_K(X)$. Regularization replaces these coefficients with coefficients $\frac{w_j\lambda_j}{\lambda_j+\gamma}$. For a fixed regularization parameter $\gamma > 0$, the function $\alpha_\gamma(j) = \frac{\lambda_j}{\lambda_j+\gamma}$ is decreasing monotonically to 0, so higher frequency coefficients are more reduced. For each $j \in I$, $\lim_{\gamma\to 0} \frac{w_j\lambda_j}{\lambda_j+\gamma} = w_j$ and so with the regularization parameter γ decreasing to zero, the coefficients $\frac{w_j\lambda_j}{\lambda_j+\gamma}$ converge to w_j.

The role of kernel norms as stabilizers in Tikhonov's regularization 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 $k \in \mathcal{L}^2(\mathbb{R}^d) \cap \mathcal{L}^1(\mathbb{R}^d)$ for which the Fourier transform $\tilde{k}$ is positive. It was shown in [25] (see also [15]) that for such kernels, the value of the stabilizer $\|.\|_K^2$ at any $f \in \mathcal{H}_K(\mathbb{R}^d)$ can be expressed as

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

So when $\lim_{\|\omega\|\to\infty} 1/\tilde{k}(\omega) = \infty$, the stabilizer $\|.\|_K^2$ plays a role of a high-frequency filter. Examples of convolution kernels with positive Fourier transforms are the Gaussian and the Bessel kernel (the kernel induced by β_r with $\hat{\beta}_r(s) = (1+\|s\|^2)^{-r/2}$).

The characterization of the regularized solution f^γ given in Theorem 6.3 (i) was derived earlier in [10, pp.27-28] using properties of operators with fractional powers for the case of X compact and K continuous. However, in [10], it was formulated as

$$f^\gamma = (I + \gamma T_K^{-1}) f_\rho. \tag{18}$$

The formulation (18) might be misleading as for a non degenerate kernel K the inverse L_K^{-1} is defined only on a subspace of $\mathcal{L}^2_{\rho_X}(X)$ which cannot be complete. Indeed if it were complete, then by the Banach open map theorem [14], L_K^{-1} would be bounded. But for a non degenerate kernel, the eigenvalues $\frac{1}{\lambda_j}$ of the inverse operator L_K^{-1} diverge. Our formulation

$$f^\gamma = L_{K_\gamma}(f_\rho)$$

given in Theorem 6.3 in terms of the integral operator L_{K_γ} with a modified kernel K_γ is rigorous and includes also non compact and non continuous cases. Moreover, as our proof takes advantage of results from theory of inverse problems, it is quite short and simple.

7 Minimization of empirical error

The next theorem describes minima od empirical error functional $\mathcal{E}_z$ and its regularized modification. We denote by

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

the *Tikhonov regularization of the empirical error* $\mathcal{E}_z$ *with the stabilizer* $\|.\|_K^2$ *and the regularization parameter* γ. For a kernel $K : X \times X \to \mathbb{R}$ and a vector $u \in X^m$, we denote by $\mathcal{K}[u]$ the *Gram matrix of the kernel K with respect to the vector u*, i.e., the matrix

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

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

Theorem 7.1 *Let $X \subseteq \mathbb{R}^d$, $K : X \times X \to \mathbb{R}$ be a symmetric positive semidefinite kernel, m be a positive integer, $z = (u,v)$ with $u = (u_1,\ldots,u_m) \in X^m$, $v = (v_1,\ldots,v_m) \in \mathbb{R}^m$, then*
(i) there exists an argminimum f^+ of $\mathcal{E}_z$ over $\mathcal{H}_K(X)$, which satisfies

$$f^+ = J_u^+(v) = \sum_{i=1}^{m} c_i K_{u_i}, \tag{19}$$

where

$$c = (c_1,\ldots,c_m) = \mathcal{K}[u]^+ v,$$

and for all $f^o \in \operatorname{argmin}(\mathcal{H}_K(X), \mathcal{E}_z)$, $\|f^+\|_K \leq \|f^o\|_K$;
(ii) for all $\gamma > 0$, there exists a unique argminimum f^γ of $\mathcal{E}_{z,\gamma,K}$ over $\mathcal{H}_K(X)$, which satisfies

$$f^\gamma = J_u^\gamma(v) = \sum_{i=1}^{m} c_i^\gamma K_{u_i}, \tag{20}$$

where

$$c^\gamma = (c_1^\gamma,\ldots,c_m^\gamma) = \left(\mathcal{K}_m[u] + \gamma I_m\right)^{-1} v\,;$$

(iii) $\lim_{\gamma\to} \|f^\gamma - f^+\|_K = 0$.

Proof. (i) By the representation (7), argminimum of $\mathcal{E}_z$ over $\mathcal{H}_K(X)$ is a pseudosolution of an inverse problem given by the operator J_u for the data v. By Proposition 5.4 (i) and (ii), J_u is continuous and has a closed range, thus we can apply Theorem 3.1(i) to obtain $J_u^+ = J_u^*(J_u J_u^*)^+$. Proposition 5.4(iii) implies that $J_u J_u^* : \mathbb{R}^m \to \mathbb{R}^m$ can be expressed by the matrix $\mathcal{K}_m[u]$. So $f^+ = J_u^+(v) = \sum_{i=1}^m c_i K_{u_i}$, where $c = \mathcal{K}_m[u]^+ v$.

(ii) By Theorem 3.1 (ii), $f^\gamma = J_u^\gamma(v) = J_u^*(J_u J_u^* + \gamma I_m)^{-1} v$, where I_m denotes the identity operator on $\mathbb{R}^m$. Thus $f^\gamma = \sum_{i=1}^m c_i^\gamma K_{u_i}$, where $c^\gamma = (\mathcal{K}_m[u] + \gamma I_m)^{-1} v$.

(iii) By Theorem 3.1 (v), $\lim_{\gamma\to} \|f^\gamma - f^+\|_K = \lim_{\gamma\to} \|J_u^\gamma(v) - J_u^+(v)\|_K = 0$ □

Theorem 7.1 shows that for every symmetric positive semidefinite 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 space $\mathcal{H}_K(X)$. This function is formed by a linear combination of the representers $K_{u_1},\ldots,K_{u_m}$ of the input data $u_1,\ldots,u_m$. Such pseudosolution f^+ can be interpreted as an *input-output function of a network with one hidden layer with kernel units and a single linear output unit*. The coefficients $c = (c_1,\ldots,c_m)$ of the linear combination (corresponding to the output weights of the network) satisfy $c = \mathcal{K}_m[u]^+ v$, so the output weights can be obtained by solving the system of linear equations.

However, as the operator J_u has finite dimensional range, it is compact and thus its pseudoinverse J_u^+ is unbounded. So the optimal solution of minimization od the empirical error $\mathcal{E}_z$ is unstable with respect to a change of output data v. Stability can be improved by replacing the pseudosolution $f^+ = J_u^+(v)$ with the regularized solution $f^\gamma = J_u^\gamma(v)$, which is a linear combination of the same functions $K_{u_1},\ldots,K_{u_m}$. But

the coefficients of these two linear combinations are different: in the regularized case $c^\gamma = (\mathcal{K}_m[u] + \gamma I)^{-1} v$, while in the non-regularized one $c = \mathcal{K}_m[u]^+ v$.

Note that for any convolution kernel $K(x,y) = k(x-y)$ with $k(0) = 1$, all functions of the form $f = \sum_{i=1}^m w_i K_{u_i}$, which are computable by one-hidden layer networks with kernel units computing translations of k, satisfy

$$\|f\|_K \leq \sum_{i=1}^m |w_i| \, \|K_{u_i}\|_K = \sum_{i=1}^m |w_i| \, K(u_i,u_i) = \sum_{i=1}^m |w_i| \, k(0) = \sum_{i=1}^m |w_i|.$$

So

$$\|f\|_K^2 \leq \left(\sum_{i=1}^m |w_i| \right)^2 . \tag{21}$$

In practical learning tasks, an output-weight regularization (which penalizes input-output functions with large ℓ_1 or ℓ_2-norms of output-weight vectors) has been widely used for its simplicity [13]. The inequality (21) shows that an output-weight regularization also penalizes solutions with large $\|.\|_K$-norms.

For X compact and K continuous, Theorem 7.1(ii) was derived by several authors using Fréchet derivatives (see, e.g., [33] [10], [28]). It became well-known under the name "Representer Theorem". Our proof of Theorem 7.1 shows that one can obtain the characterization (20) easily as a straightforward consequence of theory of inverse problems. Moreover, Theorem 7.1 characterizes argminima of empirical error also for non continuous kernels or kernels defined on non compact domains such as $\mathbb{R}^d$. Theorem 7.1 also provides a comparison of a regularized case with a non regularized one. It shows that regularization merely modifies coefficients of the linear combination of functions composing the argminimum. In the non regularized case, the coefficients are obtained from the vector v of output data by applying to it the Moore-Penrose pseudoinverse of the Gram matrix $\mathcal{K}_m[u]$, while in the regularized case, the coefficients are obtained by applying to v the inverse of the modified matrix $\mathcal{K}_m[u] + \gamma I_m$. So the regularization merely changes amplitudes, but it preserves the finite set of basis functions from which the solution is composed.

In learning from large sets of data, typically networks with much smaller number n of computational units than the size m of the training sample are used. Various learning algorithms minimize error functionals over sets of functions formed by linear combinations of n computational units, where $n \ll m$. In [23, 24], we derived some estimates of speed of convergence of minima of error functionals over networks with increasing number of units. Theory of inverse problems provides tools for comparison of these minima with the global ones over the whole RKHSs.

Acknowledgements

This work was partially supported by the Institutional Research Plan AV0Z10300504 and projects of Ministry of Education of the Czech Republic INTELLI OC10047 and Center of Applied Cybernetics 1M684077004 (1M0567).

References

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

[2] N. Aronszajn. Theory of reproducing kernels. *Transactions of AMS*, 68:337–404, 1950.

[3] S. Berg, J. P. R. Christensen, and P. Ressel. *Harmonic Analysis on Semigroups*. Springer-Verlag, New York, 1984.

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

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

[6] S. Bochner. *Lectures on Fourier Integrals*. Princeton University Press, Princeton, 1959.

[7] B. E. Boser, I. M. Guyon, and V. Vapnik. A training algorithms for optimal margin classifiers. In D. Haussler, editor, *Proceedings of the 5th Annual ACM Workshop on Computational Learlung Theory*, pages 144–152, Pittsburg, 1992. ACM Press.

[8] C. Cortes and V. Vapnik. Support vector networks. *Machine Learning*, 20:273–297, 1995.

[9] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, Cambridge, 2000.

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

[11] R. E. Edwards. *Functional Analysis - Theory and Applications*. Dover, New York, 1995.

[12] E. W. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*. Kluwer, Dordrecht, 1999.

[13] T. L. Fine. *Feedforward Neural Network Methodology*. Springer-Verlag, Berlin, Heidelberg, 1999.

[14] A. Friedman. *Modern Analysis*. Dover, New York, 1982.

[15] F. Girosi. An equivalence between sparse approximation and support vector machines. *Neural Computation*, 10:1455–1480, 1998.

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

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

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

[19] P. C. Hansen. *Rank-Deficient and Discrete Ill-Posed Problems*. SIAM, Philadelphia, 1998.

[20] V. Kecman. *Learning and Soft Computing*. MIT Press, Cambridge, 2001.

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

[22] V. Kůrková. Neural network learning as an inverse problem. *Logic Journal of IGPL*, 13:551–559, 2005.

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

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

[25] S. Loustau. Aggregation of SVM classifiers using Sobolev spaces. *Journal of Machine Learning Research*, 9:1559–1582, 2008.

[26] J. Mercer. Functions of positive and negative type and their connection with theory of integral equations. *Philosophical Transactions of Royal Society London*, 209:415–446, 1909.

[27] T. Poggio and F. Girosi. Networks for approximation and learning. *Notices of AMS*, 78:1481–1497, 1990.

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

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

[30] I. J. Schönberg. Metric spaces and completely monotone functions. *Mathematische Analen*, 39:811–841, 1938.

[31] V. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, New york, 1998.

[32] 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, 2005.

[33] G. Wahba. *Splines Models for Observational Data*. SIAM, Philadelphia, 1990.

[34] A.H. Zemanian. *Distribution Theory and Transform Analysis*. Dover, New York, 1965.