Learning with Generalization Capability by Kernel Methods of Bounded Complexity

Kůrková, Věra
2003

Institute of Computer Science
Academy of Sciences of the Czech Republic

Learning with generalization capability by kernel methods of bounded complexity

Věra Kůrková and Marcello Sanguineti

Technical report No. 901

December 2003

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

Learning with generalization capability by kernel methods of bounded complexity [1]

Věra Kůrková[2] and Marcello Sanguineti[3]

Technical report No. 901

December 2003

Abstract:

Learning from data with generalization capability is studied in the framework of minimization of regularized empirical error functionals over nested families of hypothesis sets with increasing model complexity. For Tikhonov's regularization with kernel stabilizers, minimization over restricted hypothesis sets containing for a fixed integer n only linear combinations of all n-tuples of kernel functions is investigated. Upper bounds are derived on the speed of convergence of suboptimal solutions from such sets to the optimal solution achievable without restrictions on model complexity. The bounds are of the form $1/\sqrt{n}$ multiplied by a term that depends on the size of the sample of empirical data, the vector of output data, the Gram matrix of the kernel with respect to the input data, and the regularization parameter.

Keywords:
supervised learning, generalization, model complexity, kernel methods, minimization of regularized empirical errors, upper bounds on rates of approximate optimization

[1] Collaboration between V. K. and M. S. was supported by the Scientific Agreement Italy-Czech Republic, Area MC 6, Project 22. V. K. was partially supported by GA ČR grant 201/02/0428 and M. S. by a CNR - Agenzia 2000 grant.

[2] 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

[3] Department of Communications, Computer, and System Sciences (DIST) University of Genova, Via Opera Pia 13, 16145 Genova, Italy, marcello@dist.unige.it

1 Introduction

In contrast to rule-based methods of classical artificial intelligence, connectionism employs learning based on examples. The goal of supervised learning is to adjust parameters of a connectionistic model so that it approximates with a desired accuracy a functional relationship between inputs and outputs by learning from a set of examples, i.e., a sample $z = \{(x_i, y_i) \in \mathcal{R}^d \times \mathcal{R},\ i = 1, \ldots, m\}$ of m input/output pairs of *empirical data*. In statistical learning theory [3, 41], learning based on empirical data has been modelled as minimization of a functional, called empirical error. For a sample z and a loss function $V : \mathcal{R}^2 \to \mathcal{R}$, the empirical error $\mathcal{E}_{z,V}$ is defined as $\mathcal{E}_{z,V}(f) = \frac{1}{m}\sum_{i=1}^{m} V(f(x_i), y_i)$ for all f from an ambient function space (called a hypothesis space), over which such a minimization is performed. The loss function V measures how much it is lost if to an input x an output $f(x)$ is associated instead of an output y.

Endowing a connectionistic model with a generalization capability requires some *conceptual data*, i.e., some global knowledge of the desired input/output functional relationship such as smoothness or lack of high frequency oscillations. Conceptual data can be formalized either by specifying a subset of the hypothesis space containing only functions with a desired behavior, to which minimization of the empirical error is restricted, or by adding to the empirical error a term penalizing undesired properties, or by combining these two approaches. The first method is an application to learning from data of Ivanov's or Miller's regularization, the second one of Tikhonov's, and the third one of Phillips' [6, pp. 68-78].

Tikhonov's regularization [39, 40], which was introduced into learning theory by Poggio and Girosi [18, 32, 33], leads to minimization over the whole hypothesis space of the regularized empirical error functional, defined as the sum of two functionals $\mathcal{E}_{z,V} + \gamma\Psi$. The first one, the empirical error $\mathcal{E}_{z,V}$, enforces closeness to the sample z of empirical data, while Ψ, called stabilizer expresses requirements on the global behavior of the desired input/output functional relationship. The regularization parameter γ controls the trade-off between fitting to the empirical and the conceptual data.

A large class of hypothesis spaces can be studied in the framework of the theory of Hilbert spaces of a special type, called reproducing kernel Hilbert spaces (RKHSs). Norms on such spaces can play a role of measures of various types of oscillations of input/output mapping. RKHSs were formally defined by Aronszajn [2], but their theory employs work of Schönberg [38] as well as many classical results on kernels and positive definite functions. RKHS were introduced into applications closely related to learning by Parzen [30] and Wahba [43], and they were employed explicitly in learning theory by Vapnik [8] and Girosi [17].

The Representer Theorem [9, p. 42], [16, 18, 32, 34, 36] shows that for Tikhonov's regularization with a stabilizer defined as a strictly increasing function of the norm on a RKHS, the problem of minimization of the regularized empirical error over such a space has a unique solution of the form of a linear combination of the m-tuple of the kernel functions, which are parameterized by the input data $x_1, \ldots, x_m$. For a stabilizer equal to the square of the norm on a RKHS, the vector c of the coefficients of the linear combination is given by the solution of the well-posed linear system of equations $(\gamma m\mathcal{I} + \mathcal{K}[x])c = y$, where $y = (y_1, \ldots, y_m)$ is the output data vector, $\mathcal{I}$ is the $m \times m$ identity matrix, and $\mathcal{K}[x]$ is the Gram matrix of the kernel K with respect to the input data vector $x = (x_1, \ldots, x_m)$.

A paradigmatic example of a kernel is the Gaussian kernel, for which the solution given by the Representer Theorem has the form of an input/output function of a Gaussian radial-basis-function network with m units centered at the input data $x_1, \ldots, x_m$ [16]. The coefficients of the linear combination play the role of output weights of such a network. This interpretation of the Representer Theorem was used in [18, p. 219] to argue that "the regularization principles lead to approximation schemes that are equivalent to networks with one layer of hidden units".

The Representer Theorem was employed to design a learning algorithm (see, e.g., [9, p. 42], [34, pp. 538-539]) that requires to solve the linear system of equations $(\gamma m\mathcal{I} + \mathcal{K}[x])c = y$ (examples of various pattern recognition and binary classification tasks solved using this algorithm are quoted in [34]). An advantage of this algorithm is that it gives the best possible solution of the task of fitting a function to a given sample of empirical data and satisfying a global property, which can be described in terms of a condition on oscillations that can be modelled using a kernel. However, its practical applications are limited by the speed of convergence of iterative algorithms solving the linear system of equations and by the size of the condition number of the matrix $\gamma m\mathcal{I} + \mathcal{K}[x]$. For some methods, the computational complexity of solving such a system grows polynomially with the size m of the sample (e.g., when the Gaussian elimination is used, it grows for m large enough as $m^3/3$ [29, p. 175]). For some data and kernels, keeping the condition number of $\gamma m\mathcal{I} + \mathcal{K}[x]$ small requires a large regularization parameter γ, which causes poor fit to the empirical data.

The learning algorithm based on the Representer Theorem uses a model of complexity equal to the size m of the sample of data and it does not allow any flexibility in choosing the inner parameters of the computational units (as they are set equal to the input data). Typical neural-network learning algorithms differ from this algorithm in two aspects: (1) model complexity determined by the number of network units is either set in advance (typically, it is much smaller than the size of the training set) or it is adjusted dynamically and (2) inner parameters of the units are searched for during learning.

Motivated by the model complexity constrains typical for neural network approaches, in this paper we investigate suboptimal solutions of the problem of minimization of a regularized empirical error over hypothesis sets corresponding to kernel models with limited complexity and flexible choice of parameters. We derive upper bounds on the speed of convergence of sequences of suboptimal solutions achievable by minimization over hypothesis sets formed by linear combinations of at most n kernel functions with arbitrary parameters to the optimal solution given by the Representer Theorem. The upper bounds are of the form $1/\sqrt{n}$ multiplied by a term that depends on the size m of the sample, the l_1- and l_2-norms of the vector $y = (y_1, \ldots, y_m)$ of output data, the minimum and maximum eigenvalues of the Gram matrix $\mathcal{K}[x]$ of the kernel with respect to the input data, and the regularization parameter γ.

We state conditions on the sample, the kernel and the regularization parameter, under which the term multiplying $1/\sqrt{n}$ is “small” and so such suboptimal solutions converge quickly to the optimal one. In such cases, kernel methods with a bounded model complexity give a good approximation of the best possible solution of the learning task. As our estimates are not merely asymptotic, they can be applied to any bound on model complexity smaller than the size of the training set. In particular for the Gaussian kernel, we derive an upper bound of the form $\frac{3(1+\gamma)y_{\max}^2}{n\gamma^2}$, where $y_{\max}$ denotes the maximum of the absolute values of output data.

The paper is organized as follows. Section 2 introduces concepts concerning minimization of functionals and Tikhonov's regularization applied to learning with RKHSs as hypothesis spaces. Section 3 states the Representer Theorem and discusses the condition number of the matrix used in algorithms based on this theorem. Section 4 develops tools for investigating approximate optimization over hypothesis sets corresponding to kernel methods with bounded model complexity, and describes continuity and convexity properties of regularized empirical error functionals with various types of loss functions. Section 5 contains our main results estimating the speed of convergence of sequences of suboptimal solutions with increasing model complexity. Section 6 is a brief discussion. We include an Appendix describing properties of RKSHs and illustrating them on examples of kernels and types of oscillations measured by norms defined by such kernels.

2 Tikhonov's regularization in reproducing kernel Hilbert spaces

By a normed linear space $(X, \|.\|)$ we mean a real normed linear space. $\mathcal{R}$ denotes the set of real numbers.

Let M be a subset of X and $\Phi : X \to \mathcal{R}$ be a functional. Using standard notation (see, e.g., [13]), we denote by

$$(M, \Phi)$$

the problem of minimization of Φ over M. M is called the set of *admissible solutions* or *admissible set.*

By *argmin* $(M,\Phi) = \{g \in M : \Phi(g) = \inf_{g \in M} \Phi(g)\}$ is denoted the set of *argminima* of the problem (M,Φ) and for any $\varepsilon > 0$, $argmin_\varepsilon(M,\Phi) = \{g \in M : \Phi(g) < \inf_{g\in M} \Phi(g) + \varepsilon\}$ is the set of *ε-near argminima* of (M,Φ). An argminimum of (M,Φ) is called a *solution* (or a *minimum point* of the problem (M,Φ). A sequence $\{g_n\}$ of elements of M is called *Φ-minimizing over M* if $\lim_{n\to\infty} \Phi(g_n) = \inf_{g\in M} \Phi(g)$. By the definition of infimum, for any problem (M,Φ) with M non-empty there always exists a minimizing sequence.

Let $z = \{(x_i, y_i),\ i = 1, \ldots, m\}$ be a finite set of input/output pairs of data. A standard approach to learning from empirical data (see, e.g., [41]) is based on minimization of the *empirical error* functional (also called the *empirical risk* functional), defined as

$$\mathcal{E}_V(f) = \mathcal{E}_{z,V}(f) = \frac{1}{m}\sum_{i=1}^{m} V(f(x_i), y_i),$$

where $V : \mathcal{R}\times\mathcal{R} \to [0,\infty)$ satisfying $V(y,y) = 0$ for all $y \in \mathcal{R}$ is called a *loss function.* When z is clear from the context, we write $\mathcal{E}_V$ instead of $\mathcal{E}_{z,V}$.

The most common loss function is the *square loss*, defined as

$$V(f(x), y) = (f(x) - y)^2.$$

In this paper we mostly focus on the empirical error defined using the square loss, for which we write merely $\mathcal{E}$. Other common loss functions are the *absolute value loss* $V(f(x), y) = |f(x)-y|$ and *Vapnik's ε-insensitive loss* $V(f(x), y) = \max(|f(x) - y| - \varepsilon, 0)$.

Tikhonov's regularization replaces the problem

$$(M, \mathcal{E}_V)$$

with the problem

$$(M, \mathcal{E}_V + \gamma\Psi),$$

where Ψ is a functional called *stabilizer* and $\gamma > 0$ is a *regularization parameter* [40].

An important class of stabilizers are squares of norms on reproducing kernel Hilbert spaces (RKHSs). Such stabilizers enable one to penalize high oscillations of various types. For a set Ω and a symmetric positive definite function $K : \Omega\times\Omega \to \mathcal{R}$, called *kernel*, we denote by

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

the RKHS defined by K (see the Appendix). The square $\|.\|_K^2$ is used as a stabilizer instead of the norm $\|.\|_K$ for technical reasons, as the square of the norm on any Hilbert space is a uniformly convex functional (see Proposition 4.1 (iii)), which implies uniqueness of the solution of the regularized problem (see, e.g., [12, p. 10], [9, pp. 27, 42]) and convergence of minimizing sequences to this solution [28]. The role of $\|.\|_K^2$ as a stabilizer is illustrated in the Appendix on two examples of classes of kernels playing the role of high-frequency filters.

Using $\|.\|_K^2$ as a stabilizer, the regularized empirical error functional with a loss function V and a regularization parameter γ has the form

$$\mathcal{E}_{V,\gamma,K}(f) = \frac{1}{m}\sum_{i=1}^{m} V(f(x_i), y_i) + \gamma\, \|f\|_K^2.$$

As in the case of the empirical error, also for the regularized empirical error we use for the square loss a simplified notation

$$\mathcal{E}_{\gamma,K}(f) = \frac{1}{m}\sum_{i=1}^{m} (f(x_i) - y_i)^2 + \gamma\|f\|_K^2$$

instead of $\mathcal{E}_{V,\gamma,K}(f)$.

3 The Representer Theorem

Existence, uniqueness and an explicit formula describing the solution of the problem $(\mathcal{H}_K(\Omega), \mathcal{E}_{\gamma,K})$ of the regularized empirical error with the square loss function over the whole RKHS are given by the *Representer Theorem.* For a kernel K, a positive integer m, and a vector $x \in \Omega^m$, we denote by $\mathcal{K}[x]$ the $m \times m$ matrix defined as

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

which is called the *Gram matrix of the kernel* K *with respect to the vector* $x = (x_1, \dots, x_m)$. By $\mathcal{I}$ is denoted the identity matrix.

Theorem 3.1 (Representer Theorem) *Let* Ω *be a nonempty set,* $K : \Omega \times \Omega \to \mathcal{R}$ *a kernel,* m *a positive integer,* $x = (x_1, \dots, x_m) \in \Omega^m$*,* $y = (y_1, \dots, y_m) \in \mathcal{R}^m$*,* $z = (x, y)$ *and* $\gamma > 0$*. Then there exists a unique solution* g^o *of the problem* $(\mathcal{H}_K(\Omega), \mathcal{E}_{\gamma,K})$ *such that*

$$g^o = \sum_{i=1}^{m} c_i K_{x_i}, \tag{3.1}$$

where $c = (c_1, \dots, c_m)$ *is the unique solution of the well-posed linear system*

$$(\gamma\, m \mathcal{I} + \mathcal{K}[x])c = y. \tag{3.2}$$

The Representer Theorem was originally proven in [21]. An elegant proof using directional derivatives is given in [34, pp. 538-539], while a more sophisticated argument based on the Mercer Theorem (which applies merely to Mercer kernels) is in [9, p. 42]. Inspection of these proofs shows that for any differentiable loss function V, the solution is of the form $g^o = \sum_{i=1}^{m} c_i K_{x_i}$. However, when V is not a polynomial of degree 2, the equation to be solved to compute the coefficients $c_1, \dots, c_m$ is nonlinear [17, p. 1473]. A weaker form of Theorem 3.1 without a formula for computing the coefficients $c_1, \dots, c_m$ even holds for an arbitrary loss function V and a stabilizer of the form $\psi(\|\cdot\|_K)$ with $\psi : [0, +\infty) \to \mathcal{R}$ strictly increasing [36].

The Representer Theorem was exploited to design an algorithm for learning from data. Applications of this algorithm are quoted in [34]. However, feasibility of such applications is limited by the speed of convergence of iterative algorithms solving the linear system of equations (3.2) and by the size of the condition number of the matrix $\gamma m\mathcal{I} + \mathcal{K}[x]$.

Recall that the *condition number* of a nonsingular $m \times m$ matrix A with respect to a norm $\|.\|$ on $\mathcal{R}^m$ is defined as

$$cond(A) = \|A\|\, \|A^{-1}\|,$$

where $\|A\|$ denotes the norm of A as a linear operator A on $(\mathcal{R}^m, \|.\|)$.

Let $\lambda_{\max}(A)$, $\lambda_{\min}(A)$, resp., denote maximal and minimal eigenvalues of the matrix A. To simplify our notation, we write $\lambda_{\max}$ instead of $\lambda_{\max}(\mathcal{K}[x])$ and similarly for $\lambda_{\min}$. As $\mathcal{K}[x]$ is positive semidefinite, all its eigenvalues are nonnegative [29, p. 7].

It is easy to check that for any norm and any nonsingular matrix A, $cond(A) \geq \frac{|\lambda_{\max}(A)|}{|\lambda_{\min}(A)|}$ and for any symmetric nonsingular matrix A, $cond_2(A) = \frac{|\lambda_{\max}(A)|}{|\lambda_{\min}(A)|}$, where $cond_2(A)$ denotes the condition number of A with respect to the l_2-norm.

As λ is an eigenvalue of $\mathcal{K}[x]$ if and only if $\gamma m + \lambda$ is an eigenvalue of $\gamma m\mathcal{I} + \mathcal{K}[x]$, we have

$$cond_2(\gamma m\mathcal{I} + \mathcal{K}[x]) = \frac{\gamma\, m + \lambda_{\max}}{\gamma\, m + \lambda_{\min}} \leq \frac{\lambda_{\max}}{\lambda_{\min}} = cond_2(\mathcal{K}[x]) \tag{3.3}$$

and

$$cond_2(\gamma m\mathcal{I} + \mathcal{K}[x]) = \frac{\gamma\, m + \lambda_{\max}}{\gamma\, m + \lambda_{\min}} \leq 1 + \frac{\lambda_{\max}}{\gamma\, m}. \tag{3.4}$$

So (3.3) shows that when $cond_2(\mathcal{K}[x])$ is sufficiently small, good conditioning of $\gamma m\mathcal{I}+\mathcal{K}[x]$ is guaranteed for any γ. However for large samples, $\mathcal{K}[x]$ might be ill-conditioned. For example, when the data are uniformly distributed on an interval, then the probability that $\mathcal{K}[x]$ is ill-conditioned increases with m (see [10, Theorem 2.2] and [11, Theorem 5.1]).

As by (3.4) $\lim_{\gamma\to\infty} cond_2(\gamma m\mathcal{I}+\mathcal{K}[x]) = 1$, a regularization parameter γ can be always chosen so that $cond_2(\gamma m\mathcal{I}+\mathcal{K}[x])$ is close to 1. But good conditioning of $\gamma m\mathcal{I}+\mathcal{K}[x]$ is not the only requirement on γ, its size must also allow good fit to the empirical data and thus it cannot be too large. Existence of γ guaranteeing a good fit to data as well as good conditioning depends on the speed of convergence of the condition number of $\gamma m\mathcal{I}+\mathcal{K}[x]$ to 1. The smaller $\frac{\lambda_{\max}}{m}$, the faster is this convergence.

When γ guaranteeing both small condition number and a good fit to the empirical data cannot be found, other algorithms for learning from data than the one based on the Representer Theorem have to be applied. A rich variety of learning algorithms have been developed in the field of neurocomputing. Typically, such algorithms operate on networks of a smaller model complexity than the algorithm based on the Representer Theorem. The number of hidden units in such networks is either set in advance or allocated during learning, but typically it is much smaller than the size m of the sample used as a training set. Moreover, the hidden-unit parameters (which are called *centroids* in the case of RBF networks) are not set equal to the input vectors from the data sample but are adjusted during learning. In the next section we derive tools for estimating speed of convergence of suboptimal solutions obtained by neural-network algorithms to the optimal one given by the Representer Theorem.

4 Minimization of regularized empirical errors over hypothesis sets with bounded model complexity

Suboptimal solutions obtainable by neural-network algorithms can be studied in the framework of optimization over nested families of subsets of RKHSs formed by linear combinations of all n-tuples of kernel functions $\{K_x : x \in \Omega\}$. For a subset G of a linear space, let $span_n\, G = \{\sum_{i=1}^n w_i g_i : w_i \in \mathcal{R},\, g_i \in G\}$ denote the set of linear combinations of all n-tuples of elements of G. Then the optimal solution described by the Representer Theorem is an element of $span_m\, G_K$, where $G_K = \{K_x \;:\; x \in \Omega\}$. The set $span_m G_K$ can be interpreted as the set of all input/output functions of a neural network with one hidden layer with m computational units computing functions from G_K. In particular for the Gaussian kernel, the solution has the form of an input/output function of a Gaussian radial-basis function (RBF) network with m hidden units [18].

To compare the optimal solution given by the Representer Theorem with suboptimal ones that can be obtained by minimization of $\mathcal{E}_{\gamma,K}$ over restricted hypothesis sets containing only linear combinations of all n-tuples of elements of the set G_K, we shall employ a version of the Maurey-Jones-Barron Theorem [4, 20, 31] reformulated in [22, 23] in terms of a norm called *G-variation.*

Recall that the *Minkowski functional* of a subset M of a linear space X, denoted by p_M, is defined for every $f \in X$ as $p_M(f) = \inf\{\lambda \in \mathcal{R}_+ : f/\lambda \in M\}$. For M a subset of a normed linear space $(X, \|\cdot\|)$ we denote by $cl\, M$ its *closure* with respect to the topology generated by $\|\cdot\|$, i.e., $cl\, M = \{f \in X : (\forall \varepsilon > 0)\ (\exists g \in M)\, \|f - g\| < \varepsilon)\}$.

G-variation is defined for a subset G of a normed linear space $(X, \|.\|)$ as the Minkowski functional of the closure of the convex hull of the set $G \cup -G$. So denoting G-variation by $\|\cdot\|_G$, for every $f \in X$ we have $\|f\|_G = \{c > 0 \;:\; f/c \in cl\, conv\, (G \cup -G)\}$. For properties of G-variation, see [23, 24, 25, 27].

Maurey-Jones-Barron's theorem reformulated in terms of G-variation [23] gives for a Hilbert space $(X, \|.\|)$, its bounded subset G with $s_G = \sup_{g\in G} \|g\|$ and every $f \in X$ the following upper bound on rate of approximation by $span_n G$.

$$\|f - span_n G\| \leq \sqrt{\frac{(s_G \, \|f\|_G)^2 - \|f\|^2}{n}}. \tag{4.1}$$

Taking advantage of this upper bound, the next theorem estimates rates of convergence of suboptimal solutions of the problems of minimizations of a continuous functional Φ over hypothesis sets of the form $span_n G$ with n increasing. The estimates are formulated in terms of moduli of continuity and convexity of the functional to be minimized.

A functional $\Phi : X \to \mathcal{R}$ is *continuous* at $f \in X$ if for any $\varepsilon > 0$ there exists $\eta > 0$ such that $\|f - g\| < \eta$ implies $|\Phi(f) - \Phi(g)| < \varepsilon$. A *modulus of continuity* of Φ at f is a function $\omega : [0, +\infty) \to [0, +\infty)$ defined as $\omega(a) = \sup\{|\Phi(f) - \Phi(g)| : \|f - g\| \leq a\}$.

Φ is *convex* on a convex set $M \subseteq X$ if for all $h, g \in M$ and all $\lambda \in [0, 1]$, we have $\Phi(\lambda h + (1 - \lambda)g) \leq \lambda\Phi(h) + (1 - \lambda)\Phi(g)$.

Φ is *uniformly convex* on a convex set $M \subseteq X$ if there exists a non-negative function $\delta : \mathcal{R}_+ \to \mathcal{R}_+$, such that $\delta(0) = 0$, for all $t > 0$, $\delta(t) > 0$, and for all $h, g \in M$ and all $\lambda \in [0, 1]$, $\Phi(\lambda h + (1 - \lambda)g) \leq \lambda\Phi(h) + (1 - \lambda)\Phi(g) - \lambda(1 - \lambda)\delta(\|h - g\|)$. Any such function δ is called a *modulus of convexity* of Φ [28] [4].

Before proving the theorem, we state elementary properties of moduli of convexity.

Proposition 4.1 *Let $(X, \|.\|)$ be a normed linear space, $M \subseteq X$ convex and Φ be a uniformly convex functional on M with a modulus of convexity δ. Then the following hold:*
(i) if Ψ is convex on M and $\gamma > 0$, then $\Phi + \gamma\Psi$ is uniformly convex on M with a modulus of convexity $\gamma\,\delta$;
(ii) if $g^o \in argmin(M, \Phi)$, then for every $g \in M$ $\delta(\|g - g^o\|) \leq \Phi(g) - \Phi(g^o)$;
(iii) if $(X, \|.\|)$ is a Hilbert space, then the functional $\|.\|^2 : X \to \mathcal{R}$ is uniformly convex with a modulus of convexity $\delta(t) = t^2$.

Proof. (i) and (ii) follow directly from the definitions.
(ii) By the definition of uniformly convex convex functional, for every $\lambda \in [0, 1]$ we have $\lambda(1 - \lambda)\delta(\|g - g^o\|) \leq \lambda\Phi(g) + (1 - \lambda)\Phi(g^o) - \Phi(\lambda g + (1 - \lambda)g^o)$. As $\Phi(g^o) \leq \Phi(\lambda g + (1 - \lambda)g^o)$, we get $\lambda(1 - \lambda)\delta(\|g - g^o\|) \leq \lambda\Phi(g) + (1 - \lambda)\Phi(g^o) - \Phi(g^o) = \lambda\left(\Phi(g) - \Phi(g^o)\right)$. Hence $(1 - \lambda)\delta(\|g - g^o\|) \leq \Phi(g) - \Phi(g^o)$. Taking the infimum over λ, we obtain $\delta(\|g - g^o\|) \leq \Phi(g) - \Phi(g^o)$.
(iii) It is easy to check that for every $h, g \in X$ and $\lambda \in [0, 1]$, we have $\|\lambda h + (1 - \lambda)g\|^2 \leq \lambda\|h\|^2 + (1 - \lambda)\|g\|^2 - \lambda(1 - \lambda)\|h - g\|^2$. □

Theorem 4.2 *Let $(X, \|.\|)$ be a Hilbert space, G its bounded subset, $s_G = \sup_{g \in G} \|g\|$, $\Phi : X \to (-\infty, +\infty]$ a functional, $g^o \in argmin\,(X, \Phi)$, Φ continuous at g^o with a modulus of continuity α, $\{\varepsilon_n\}$ a sequence of positive reals, $g_n \in argmin_{\varepsilon_n}(span_n\,G, \Phi)$, and let*

$$a = (s_G \|g^o\|_G)^2 - \|g^o\|^2.$$

Then for every integer n the following estimates hold:
(i) $\inf_{g \in span_n\,G} \Phi(g) - \Phi(g^o) \leq \alpha\left(\sqrt{\frac{a}{n}}\right)$;
(ii) if $\|g^o\|_G < \infty$ and $\lim_{n \to \infty} \varepsilon_n = 0$, then $\{g_n\}$ is a Φ-minimizing sequence and $\Phi(g_n) - \Phi(g^o) \leq \alpha\left(\sqrt{\frac{a}{n}}\right) + \varepsilon_n$;
(iii) if Φ is uniformly convex with a modulus of convexity δ, then
$\delta(\|g_n - g^o\|) \leq \alpha\left(\sqrt{\frac{a}{n}}\right) + \varepsilon_n$.

Proof. (i) For every n and every $\varepsilon > 0$, choose an ε-near best approximation f_n^ε of g^o in $span_n G$. So $\|g^o - f_n^\varepsilon\| < \|g^o - span_n\,G\| + \varepsilon$. As $f_n^\varepsilon \in span_n G$, we have $\inf_{g \in span_n\,G} \Phi(g) - \Phi(g^o) \leq \Phi(f_n^\varepsilon) - \Phi(g^o)$. Estimating the right-hand side of this inequality in terms of the modulus of continuity α of Φ at g^o, we obtain $\inf_{g \in span_n\,G} \Phi(g) - \Phi(g^o) \leq \alpha\big(\|f_n^\varepsilon - g^o\|\big) \leq \alpha\big(\|g^o - span_n G\| + \varepsilon\big)$. By (4.1) we get

[4] The terminology is not unified: some authors use the term strictly uniformly convex instead of uniformly convex, while they reserve the term uniformly convex for the case when $\delta : [0, +\infty) \to [0, +\infty)$ merely satisfies $\delta(0) = 0$ and for some $t_0 > 0$, $\delta(t_0) > 0$ (see, e.g., [42] and [12, p. 10]).

$$\inf_{g \in span_n G} \Phi(g) - \Phi(g^o) \le \alpha \left(\sqrt{\frac{a}{n}} + \varepsilon \right). \tag{4.2}$$

Infimizing (4.2) over ε we obtain (i).

(ii) By the definition of ε_n-argminimum, we have
$\Phi(g_n) - \Phi(g^o) \le \inf_{g \in span_n G} \Phi(g) - \Phi(g^o) + \varepsilon_n$. So by the item (i) we get

$$\Phi(g_n) - \Phi(g^o) \le \alpha \left(\sqrt{\frac{a}{n}} \right) + \varepsilon_n. \tag{4.3}$$

If $\lim_{n\to\infty} \varepsilon_n = 0$ and $\|g^o\|_G$ is finite, then the right-hand side of (4.3) converges to zero and so $\{g_n\}$ is Φ-minimizing.

(iii) By the item (i), the definition of ε_n-argmin, and Proposition 4.1 (iii), we have $\delta(\|g_n - g^o\|) \le \Phi(g_n) - \Phi(g^o) < \inf_{g \in span_n G} \Phi(g) - \Phi(g^o) + \varepsilon_n \le \alpha \left(\sqrt{\frac{a}{n}} \right) + \varepsilon_n$. □

Theorem 4.2 can be derived as a corollary of a more general theorem from [26], which has consequences also for other types of regularization including the Ivanov's one. However, the direct argument stated here in the proof of Theorem 4.2 is much simpler than the proof of the more general result from [26].

To employ Theorem 4.2 for deriving rates of approximate minimization of regularized empirical error functionals with kernel stabilizers, we need to estimate moduli of continuity and convexity of these functionals. The next proposition describes convexity and continuity properties of regularized empirical error functionals with various loss functions.

Proposition 4.3 *Let Ω be a nonempty set, $K : \Omega \times \Omega$ a kernel, $\gamma > 0$, m a positive integer, $x = (x_1, \ldots, x_m) \subseteq \Omega^m$, $y = (y_1, \ldots, y_m) \in \mathcal{R}^m$, $z = (x, y)$, $y_{\min} = \min\{|y_i| : i = 1, \ldots, m\}$, and $V : \Omega \times \mathcal{R} \to \mathcal{R}$ a loss function. Then the following hold:*
(i) if for every $i = 1, \ldots, m$ the functions $V(\cdot, y_i) : \mathcal{R} \to \mathcal{R}$ are convex, then $\mathcal{E}_{V,\gamma,K}$ is uniformly convex on $\mathcal{H}_K(\Omega)$ with a modulus of convexity $\delta(t) = \gamma t^2$;
(ii) if V is either the square or the absolute value loss function, then at every $f \in \mathcal{H}_K(\Omega)$ the functional $\mathcal{E}_{V,\gamma,K}$ is continuous with a modulus of continuity bounded from above by the quadratic function $\beta(t) = b_2 t^2 + b_1 t$, where for the square loss $b_2 = s_K^2 + \gamma$ and $b_1 = 2\big(\|f\|_K\,(s_K^2 + \gamma) + y_{\min}\, s_K^2\big)$, while for the absolute value loss $b_2 = \gamma$ and $b_1 = s_K + 2\gamma\|f\|_K$;
(iii) if V is the square loss function, then there exists a unique argminimum g^o of the problem $(\mathcal{H}_K(\Omega), \mathcal{E}_{V,\gamma,K})$ and for every $f \in \mathcal{H}_K(\Omega)$

$$\|f - g^o\|_K^2 \le \frac{\mathcal{E}_{V,\gamma,K}(f) - \mathcal{E}_{V,\gamma,K}(g^o)}{\gamma}.$$

Proof. (i) It is easy to show that for such loss functions the empirical error functional $\mathcal{E}_V = 1/m \sum_{i=1}^m V(f(x_i), y_i)$ is convex and so the statement follows from Proposition 4.1 (i) and (iii).

(ii) For the square loss, by the inequality (7.1) we obtain $|\mathcal{E}_{V,\gamma,K}(f) - \mathcal{E}_{V,\gamma,K}(g)| = \left| \frac{1}{m} \sum_{i=1}^m \big((f(x_i) - y_i)^2 - (g(x_i) - y_i)^2\big) + \gamma \left(\|f\|_K^2 - \|g\|_K^2\right) \right| \le \left| \frac{1}{m} \sum_{i=1}^m \big(f(x_i) - g(x_i) \big) \big(f(x_i) + g(x_i) - 2y_i \big) \right| + \gamma \big| \|f\|_K - \|g\|_K \big| (\|f\|_K + \|g\|_K) \le \sup_{x\in\Omega} |f(x) - g(x)| \big| \sup_{x\in\Omega} |f + g| - 2y_{\min} \big| + \gamma \|f - g\|_K (\|f\|_K + \|g\|_K) \le t\, s_K \big| s_K \|f + g\|_K - 2y_{\min} \big| + t\,\gamma\, (\|f\|_K + \|g\|_K)$.

Let $t > 0$ and f, g be such that $\|f - g\|_K \le t$. Then $|\mathcal{E}_{V,\gamma,K}(f) - \mathcal{E}_{V,\gamma,K}(g)| \le t\, s_K \big|2\|f\|_K\, s_K + t\, s_K - 2y_{\min}\big| + \gamma t\, (2\|f\|_K + t) \le t^2\, (s_K^2 + \gamma) + 2t\big(\|f\|_K\, s_K^2 + y_{\min}\, s_K + \gamma\|f\|_K\big)$. Thus, $\|f - g\|_K < t$ implies $|\mathcal{E}_{V,\gamma,K}(f) - \mathcal{E}_{V,\gamma,K}(g)| \le \beta(t) = b_2 t^2 + b_1 t$, where $b_2 = s_K^2 + \gamma$ and $b_1 = 2\big(\|f\|_K\, (s_K + \gamma) + y_{\min}\, s_K^2\big)$.

Similarly, for the absolute value loss we have $|\mathcal{E}_{V,\gamma,K}(f) - \mathcal{E}_{V,\gamma,K}(g)| = \left| \frac{1}{m} \sum_{i=1}^m \big|f(x_i) - g(x_i)\big| + \gamma \left(\|f\|_K^2 - \|g\|_K^2\right) \right| \le \sup_{x\in\Omega} |f(x) - g(x)| + \gamma \big| \|f\|_K - \|g\|_K \big| (\|f\|_K + \|g\|_K) \le s_K\, \|f - g\|_K + \gamma\, \|f - g\|_K\, (\|f\|_K + \|g\|_K) \le s_K\, t + t\,\gamma\, (\|f\|_K + \|g\|_K) \le s_K\, t + t\,\gamma\, (t +$

$2\,\|f\|_K$). Thus, $\|f-g\|_K < t$ implies $|\mathcal{E}_{V,\gamma,K}(f) - \mathcal{E}_{V,\gamma,K}(g)| \leq \beta(t) = b_2 t^2 + b_1 t$, where $b_2 = \gamma$ and $b_1 = s_K + 2\gamma\|f\|_K$.

(iii) The existence of a unique argminimum g^o follows from the Representer Theorem. By Proposition 4.1(i), (iii), and (iv) for every $f \in \mathcal{H}_K(\Omega)$, we have $\gamma\|f - g^o\|_K^2 \leq |\mathcal{E}_{V,\gamma,K}(f) - \mathcal{E}_{V,\gamma,K}(g^o)|$. □

The assumptions of Proposition 4.3 (i) are satisfied by both the square and the absolute value loss. So these two loss functions determine uniformly convex functionals $\mathcal{E}_{V,\gamma,K}$ with quadratic moduli of convexity. Their moduli of continuity at any $f \in \mathcal{H}_K(\Omega)$ are also bounded from above by a quadratic function, which has the form $\beta(t) = b_2 t^2 + b_1 t$, where in both cases b_2 depends on γ and, for the square loss, on s_K, while b_1 depends on γ, s_K, $\|f\|_K$ and, for the square loss, on $y_{\min}$. The larger the regularization parameter γ, the larger the coefficients of the quadratic function bounding the moduli of continuity. Generally, the modulus of continuity of $\mathcal{E}_{V,\gamma,K}$ depends on the moduli of continuity of the functions $V(\cdot, y_i)$, $i = 1, \ldots, m$.

5 Suboptimal solutions over kernel models with bounded complexity

In this section, we derive estimates of rates of convergence of suboptimal solutions of the problems $(span_n G_K, \mathcal{E}_{\gamma,K})$ to the optimal solution g^o of the problem $(\mathcal{H}_K(\Omega), \mathcal{E}_{\gamma,K})$ given by the Representer Theorem. In contrast to the optimal solution g^o, which is a linear combinations of the representers $K_{x_1}, \ldots, K_{x_m}$ determined by the sample $x_1, \ldots, x_m$ of input data, suboptimal solutions are formed by linear combinations of *arbitrary n-tuples* of elements of $G_K = \{K_x : x \in \Omega\}$. In practical applications, a proper n-tuple together with coefficients of the linear combination are adjusted during learning by some neural-network algorithm (see, e.g., [1, 7, 19]).

Without loss of generality we can assume that $y_{\min} = \min\{|y_i| : i = 1, \ldots, m\} = 0$, as by shifting the sample as well as the solution we can always reduce the problem to this case. Note that although the next theorem holds for any integer n, it is useful only for $n < m$ since by the Representer Theorem, the minimum over $span_m G_K$ is equal to the minimum over the whole space.

Theorem 5.1 *Let Ω be a nonempty set, $K : \Omega \times \Omega \to \mathcal{R}$ a kernel, $s_K = \sup_{x\in\Omega} \sqrt{K(x,x)}$, m a positive integer, $x = (x_1, \ldots, x_m) \in \Omega^m$, $y = (y_1, \ldots, y_m) \in \mathcal{R}^m$, $z = (x, y)$, $y_{min} = \min\{|y_i| : i = 1, \ldots, m\} = 0$, $g^o = \sum_{i=1}^m c_i\, K_{x_i}$ the unique argminimum of $(\mathcal{H}_K(\Omega), \mathcal{E}_{\gamma,K})$, $\{\varepsilon_n\}$ a sequence of positive reals such that $\lim_{n\to\infty} \varepsilon_n = 0$, and $\{g_n\}$ a sequence of ε_n-argminima of $(span_n\, G_K, \mathcal{E}_K)$. Let $u = (s_K^2 + \gamma)a$ and $v = 2(s_K^2 + \gamma)\|g^o\|_K \sqrt{a}$, where $a = (s_K\, \|g^o\|_{G_K})^2 - \|g^o\|_K^2$. Then for every integer n the following estimates hold:*
(i) $\inf_{g\in span_n\, G_K} \mathcal{E}_{\gamma,K}(g) - \mathcal{E}_{\gamma,K}(g^o) \leq \frac{u}{n} + \frac{v}{\sqrt{n}}$;
(ii) $\mathcal{E}_{\gamma,K}(g_n) - \mathcal{E}_K(g^o) \leq \frac{u}{n} + \frac{v}{\sqrt{n}} + \varepsilon_n$;
(iii) $\|g_n - g^o\|_K^2 \leq \frac{1}{\gamma}\left(\frac{u}{n} + \frac{v}{\sqrt{n}} + \varepsilon_n\right)$;
(iv) $\sup_{x\in\Omega} |g_n(x) - g^o(x)|^2 \leq \frac{s_K^2}{\gamma}\left(\frac{u}{n} + \frac{v}{\sqrt{n}} + \varepsilon_n\right)$.

Proof. (i) Combining Theorem 4.2 (i) with Proposition 4.3 (ii), we get $\inf_{g\in span_n\, G_K} \mathcal{E}_{\gamma,K}(g) - \mathcal{E}_{\gamma,K}(g^o) \leq \beta\left(\sqrt{\frac{a}{n}}\right)$, where $\beta(t) = (s_K^2 + \gamma)\,(t^2 + 2\,\|g^o\|_K\, t)$, which gives the upper bound $(s_K^2 + \gamma)\left(\frac{a}{n} + 2\|g^o\|_K \sqrt{\frac{a}{n}}\right)$.

Similarly, the item (ii) follows from Theorem 4.2 (ii) and Proposition 4.3 (ii), the item (iii) from (ii) and Proposition 4.3 (iii), and the item (iv) from (iii) and the inequality (7.1). □

So when u and v are not too large, it is possible to choose n small enough so that networks with n hidden units are implementable and a suboptimal solution over sets of functions computable by such networks is a good approximation of the optimal solution given by the Representer Theorem.

The only terms in the above formulas defining u and v, which cannot be derived directly from the data sample z, the kernel K and the regularization parameter γ, are the values of the two norms of the optimal solution g^o: its G_K-variation and its K-norm. The next proposition estimates these two values in terms of the size m of the sample, the regularization parameter γ, the l_2-norm of the output vector y, and the maximum and the minimum eigenvalues, $\lambda_{\max}$ and $\lambda_{\min}$, of the Gram matrix $\mathcal{K}[x]$ of the kernel K with respect to the input data vector x. By $\|\cdot\|_1$ and $\|\cdot\|_2$ are denoted the l_1 and l_2-norm, resp., on $\mathcal{R}^m$.

Proposition 5.2 *Let Ω be a nonempty set, $K : \Omega \times \Omega \to \mathcal{R}$ a kernel, $s_K = \sup_{x\in\Omega} \sqrt{K(x,x)}$, $\gamma > 0$, m a positive integer, $x = (x_1, \ldots, x_m) \in \Omega^m$, $y = (y_1, \ldots, y_m) \in \mathcal{R}^m$, $g^o = \sum_{i=1}^m c_i K_{x_i}$ the unique solution of the problem $(\mathcal{H}_K(\Omega), \mathcal{E}_{\gamma,K})$. Then the following estimates hold:*
(i) $\|g^o\|_{G_K} \leq \frac{\sqrt{m}\|y\|_2}{\gamma m+\lambda_{\min}}$;
(ii) $\|g^o\|_K \leq \frac{\sqrt{\lambda_{\max}}\|y\|_2}{\gamma m+\lambda_{\min}}$;
(iii) $s_K^2 \, \|g^o\|_{G_K}^2 - \|g^o\|_K^2 \leq \frac{(s_K^2\, m-\lambda_{\min})\,\|y\|_2^2}{(\gamma m+\lambda_{\min})^2}$.

Proof. (i) It follows from the Representer Theorem, the definition of G_K-variation, and the Cauchy-Schwartz inequality that

$$\|g^o\|_{G_K} \leq \sum_{i=1}^m |c_i| = \|c\|_1 \leq \sqrt{m}\,\|c\|_2, \tag{5.1}$$

where $c = (\gamma\, m \mathcal{I} + \mathcal{K}[x])^{-1} y$. By the definition of the norm of an operator, $\|c\|_2 \leq \|(\gamma\, m\mathcal{I} + \mathcal{K}[x])^{-1}\|_2\, \|y\|_2$. As $(\gamma m\mathcal{I} + \mathcal{K}[x])^{-1}$ is symmetric and positive definite, its l_2-norm is equal to its maximal eigenvalue, which is $\frac{1}{\gamma m+\lambda_{\min}}$. So we have

$$\|c\|_2 \leq \frac{\|y\|_2}{\gamma m + \lambda_{\min}} \tag{5.2}$$

and thus $\|g^o\|_{G_K} \leq \frac{\sqrt{m}\|y\|_2}{\gamma m+\lambda_{\min}}$.

(ii) By the Representer Theorem, $\|g^o\|_K^2 = \left\langle \sum_{i=1}^m c_i\, K_{x_i}, \sum_{j=1}^m c_j K_{x_j} \right\rangle_K = \sum_{i,j=1}^m c_i\, c_j K(x_i, x_j) = c^T\, \mathcal{K}[x]c$, where c^T denotes the transpose of the vector c. Hence [29, p. 21]

$$\lambda_{\min}\|c\|_2^2 \leq \|g^o\|_K^2 \leq \lambda_{\max}\|c\|_2^2\,. \tag{5.3}$$

Thus by (5.2), $\|g^o\|_K \leq \frac{\sqrt{\lambda_{\max}}\|y\|_2}{\gamma\, m+\lambda_{\min}}$.

(iii) By (5.1) and (5.3),

$$s_K^2\|g^o\|_{G_K}^2 - \|g^o\|_K^2 \leq s_K^2\sqrt{m}\|c\|_2^2 - \lambda_{\min}\|c\|_2^2 \leq \left(s_K^2 m - \lambda_{\min}\right)\|c\|_2^2 \leq \frac{(s_K^2\, m - \lambda_{\min})\,\|y\|_2^2}{(\gamma m + \lambda_{\min})^2}.$$

□

As both $\lambda_{\min}$ and $\lambda_{\max}$ are nonnegative, we can farther simplify the upper bounds from Proposition 5.2:
(i) $\|g^o\|_{G_K} \leq \frac{\|y\|_2}{\gamma\sqrt{m}}$,
(ii) $\|g^o\|_K \leq \frac{\sqrt{\lambda_{\max}}\|y\|_2}{\gamma m}$,
(iii) $s_K^2\, \|g^o\|_{G_K}^2 - \|g^o\|_K^2 \leq \frac{s_K^2\|y\|_2^2}{\gamma^2 m}$.

Combining Proposition 5.2 with Theorem 5.1, we derive upper bounds on rates of convergence of approximate solutions of the problems $(span_n\, G_K, \mathcal{E}_{\gamma,K})$ to the solution of the problem $(\mathcal{H}_K(\Omega), \mathcal{E}_{\gamma,K})$ in terms of s_K, m, γ, $\|y\|_2$, $\lambda_{\min}$ and $\lambda_{\max}$.

Corollary 5.3 *Let Ω be a nonempty set, $K : \Omega \times \Omega \to \mathcal{R}$ a kernel, $s_K = \sup_{x\in\Omega} \sqrt{K(x,x)}$, $\gamma > 0$, m a positive integer, $x = (x_1, \dots, x_m) \in \Omega^m$, $y = (y_1, \dots, y_m) \in \mathcal{R}^m$, $\min\{|y_i| : i = 1, \dots, m\} = 0$, $g^o = \sum_{i=1}^m c_i K_{x_i}$ the unique solution of $(\mathcal{H}_K(\Omega), \mathcal{E}_K)$, $\{\varepsilon_n\}$ a sequence of positive reals, $\{g_n\}$ a sequence of ε_n-argminima of $(span_n\, G_K, \mathcal{E}_{\gamma,K})$. Let*

$$\bar{u} = \left(s_K^2 + \gamma\right) \frac{(s_K^2\, m - \lambda_{\min})\, \|y\|_2^2}{(\gamma m + \lambda_{\min})^2} \leq \left(s_K^2 + \gamma\right) \frac{s_K^2\, \|y\|_2^2}{\gamma^2 m} \qquad \textit{and}$$

$$\bar{v} = 2\left(s_K^2 + \gamma\right) \frac{\sqrt{\lambda_{\max}}\|y\|_2}{(\gamma\, m + \lambda_{\min})^2} \sqrt{(s_K^2\, m - \lambda_{\min})\, \|y\|_2^2} \leq 2\left(s_K^2 + \gamma\right) \frac{\sqrt{\lambda_{\max}} s_K \|y\|_2^2}{\gamma^2 m^2}.$$

Then for every positive integer n the following estimates hold:
(i) $\inf_{g\in span_n\, G_K} \mathcal{E}_{\gamma,K}(g) - \mathcal{E}_{\gamma,K}(g^o) \leq \frac{\bar{u}}{n} + \frac{\bar{v}}{\sqrt{n}}$*;*
(ii) $\mathcal{E}_{\gamma,K}(g_n) - \mathcal{E}_K(g^o) \leq \frac{\bar{u}}{n} + \frac{\bar{v}}{\sqrt{n}} + \varepsilon_n$*;*
(iii) $\|g_n - g^o\|_K^2 \leq \frac{1}{\gamma}\left(\frac{\bar{u}}{n} + \frac{\bar{v}}{\sqrt{n}} + \varepsilon_n\right)$*;*
(iv) $\sup_{x\in\Omega} |g_n(x) - g^o(x)|^2 \leq \frac{s_K^2}{\gamma}\left(\frac{\bar{u}}{n} + \frac{\bar{v}}{\sqrt{n}} + \varepsilon_n\right)$*.*

Thus to obtain a good approximation of the optimal solution given by the Representer Theorem by a suboptimal solution computable by a neural network, both $\frac{\hat{u}}{n}$ and $\frac{\hat{v}}{\sqrt{n}}$ have to be sufficiently small for some n, for which networks with n hidden units computing functions from G_K are implementable.

The next corollary illustrates behavior of $\frac{\hat{u}}{n}$ and $\frac{\hat{v}}{\sqrt{n}}$ in the case of convolution kernels $K(u,v) = \psi(\|u - v\|)$ with $\psi : \mathcal{R} \to [0,1]$ monotonically decreasing and satisfying $\psi(0) = 1$ (so it applies to the Gaussian kernel). The corollary estimates rates of convergence of suboptimal solutions for input/output pairs of data $(x_1, y_1), \dots, (x_m, y_m)$, for which the inputs are sufficiently separated so that there exists some $a \in [0,1]$ such that for all distinct $i, j \in \{1, \dots, m\}$, $\psi(\|x_i - x_j\|) \leq a$.

Corollary 5.4 *Let $K : \mathcal{R}^d \times \mathcal{R}^d \to \mathcal{R}$ be a kernel such that $K(u,v) = \psi(\|u - v\|)$ with $\psi : \mathcal{R} \to [0,1]$ monotonically decreasing, satisfying $\psi(0) = 1$, and such that for all distinct $i, j \in \{1, \dots, m\}$, $\psi(\|x_i - x_j\|) \leq a$. Let $\gamma > 0$, m be a positive integer, $x = (x_1, \dots, x_m) \in \mathcal{R}^m$, $y = (y_1, \dots, y_m) \in \mathcal{R}^m$, $y_{\min} = \min\{|y_i| : i = 1, \dots, m\} = 0$, $g^o = \sum_{i=1}^m c_i K_{x_i}$ the unique solution of $(\mathcal{H}_K(\mathcal{R}^d), \mathcal{E}_K)$, $\{\varepsilon_n\}$ a sequence of positive reals, and $\{g_n\}$ a sequence of ε_n-argminima of $(span_n\, G_K, \mathcal{E}_{\gamma,K})$. Let*

$$\hat{u} = (1 + \gamma) \frac{\|y\|_2^2}{\gamma^2 m} \qquad \textit{and}$$

$$\hat{v} = 2\,(1 + \gamma) \frac{\sqrt{1 + (m-1)a}\|y\|_2^2}{\gamma^2 m^2}.$$

Then for every positive integer n the following estimates hold:
(i) $\inf_{g\in span_n\, G_K} \mathcal{E}_{\gamma,K}(g) - \mathcal{E}_{\gamma,K}(g^o) \leq \frac{\hat{u}}{n} + \frac{\hat{v}}{\sqrt{n}}$*;*
(ii) $\mathcal{E}_{\gamma,K}(g_n) - \mathcal{E}_K(g^o) \leq \frac{\hat{u}}{n} + \frac{\hat{v}}{\sqrt{n}} + \varepsilon_n$*;*
(iii) $\|g_n - g^o\|_K^2 \leq \frac{1}{\gamma}\left(\frac{\hat{u}}{n} + \frac{\hat{v}}{\sqrt{n}} + \varepsilon_n\right)$*;*
(iv) $\sup_{x\in\Omega} |g_n(x) - g^o(x)|^2 \leq \frac{1}{\gamma}\left(\frac{\hat{u}}{n} + \frac{\hat{v}}{\sqrt{n}} + \varepsilon_n\right)$*.*

Proof. The estimates follow from Corollary 5.3 combined with the following upper bounds on $\bar{u}$ and $\bar{v}$:
As $s_K = 1$ and $\lambda_{\max} \leq \|\mathcal{K}[x]\|_1 = \max_{j=1,\dots,m} \sum_{i=1}^m |K[x]_{i,j}|$ [29, pp. 6, 21-23], we have $\lambda_{\max} \leq 1 + (m-1)a$ and so we get

$$\bar{u} = (1 + \gamma) \frac{\|y\|_2^2}{\gamma^2 m} = \hat{u}$$

$$\bar{v} \leq 2\,(1+\gamma)\,\frac{\sqrt{1+(m-1)a}\|y\|_2^2}{\gamma^2 m^2} = \hat{v}.$$

□

Estimating from above formulas from Corollary 5.4 in terms of the maximum of the absolute values of output data, we get the following corollary.

Corollary 5.5 *Let $K : \mathcal{R}^d \times \mathcal{R}^d \to \mathcal{R}$ be a kernel such that $K(u,v) = \psi(\|u-v\|)$ with $\psi : \mathcal{R} \to [0,1]$ monotonically decreasing, satisfying $\psi(0) = 1$, and such that for all distinct $i,j \in \{1,\dots,m\}$, $\psi(\|x_i - x_j\|) \leq a$. Let $\gamma > 0$, m be a positive integer, $x = (x_1,\dots,x_m) \in \mathcal{R}^m$, $y = (y_1,\dots,y_m) \in \mathcal{R}^m$, $y_{\min} = \min\{|y_i| : i = 1,\dots,m\} = 0$, $y_{\max} = \max\{|y_i| : i = 1,\dots,m\}$, $g^o = \sum_{i=1}^m c_i K_{x_i}$ the unique solution of $(\mathcal{H}_K(\mathcal{R}^d), \mathcal{E}_K)$, $\{\varepsilon_n; n = 1,\dots,m\}$ positive real numbers, $\{g_n : n = 1,\dots,m\}$ ε_n-argminima of $(span_n\, G_K, \mathcal{E}_{\gamma,K})$, and let*

$$c = \frac{3(1+\gamma)y_{\max}^2}{\gamma^2}.$$

Then for every positive integer $n \leq m$ the following estimates hold:
(i) $\inf_{g \in span_n\, G_K} \mathcal{E}_{\gamma,K}(g) - \mathcal{E}_{\gamma,K}(g^o) \leq \frac{c}{n}$*;*
(ii) $\mathcal{E}_{\gamma,K}(g_n) - \mathcal{E}_K(g^o) \leq \frac{c}{n} + \varepsilon_n$*;*
(iii) $\|g_n - g^o\|_K^2 \leq \frac{1}{\gamma}\left(\frac{c}{n} + \varepsilon\right)$*;*
(iv) $\sup_{x \in \Omega} |g_n(x) - g^o(x)|^2 \leq \frac{1}{\gamma}\left(\frac{c}{n} + \varepsilon\right)$.

Proof. As $\|y\|_2^2 \leq m\, y_{\max}^2$, by Corollary 5.4, we get
$\frac{\hat{u}}{n} + \frac{\hat{v}}{\sqrt{n}} \leq \frac{(1+\gamma)y_{\max}^2}{\gamma^2}\left(\frac{1}{n} + \frac{2\sqrt{1+(m-1)a}}{m\sqrt{n}}\right)$, which for $a \in [0,1]$ and $n \leq m$ is bounded from above by $\frac{(1+\gamma)y_{\max}^2}{\gamma^2}\left(\frac{1}{n} + \frac{2}{\sqrt{mn}}\right) \leq \frac{3(1+\gamma)y_{\max}^2}{\gamma^2\, n}$. □

So when γ is not too small and $y_{\max}$ is not too large, Corollary 5.5 guarantees good approximation of the optimal solution by the suboptimal ones.

In particular for the Gaussian kernel, the minimum of the regularized empirical error functional over the set of functions computable by Gaussian radial-basis function networks with n hidden units approximates the global minimum over the whole RKHS within $\frac{c}{n}$, where $c = \frac{3(1+\gamma)y_{\max}^2}{\gamma^2}$. For example, for $\gamma = 0.5$, we have $c = 18y_{\max}^2$ as $\frac{1+\gamma}{\gamma^2} = 6$.

6 Discussion

We have compared two approaches to learning from data with generalization capability, both modelling learning as a minimization of an empirical error functional regularized by the square of a norm on a RKHS, but differing in the hypothesis set over which minimization takes place. The first approach, which is based on the Representer Theorem, considers minimization over the whole RKHS, while the second one only over its subsets formed by functions computable by neural networks with n hidden units computing functions defined by the kernel.

We have derived upper bounds on error of approximation of the optimal solution by the suboptimal ones obtainable using such networks with n increasing. We have shown that when absolute values of output data are not too large and the regularization parameter is not too small, then suboptimal solutions approximate the optimal one within $\frac{c}{n}$ with c moderate. In such cases, neural network algorithms operating on networks with n hidden units can approximate the optimal solution quite well. As the upper bounds from corollaries 5.4 and 5.5 do not depend on the number of variables d, approximation of the optimal solution by neural networks does not exhibit the curse of dimensionality (which is a frequent cause of problems in the case of linear approximators).

So when the solution of the system of linear equations described in the Representer Theorem is either not computationally feasible or when it is ill-conditioned, neural networks represent a useful and quite accurate alternative to the learning algorithms built on the Representer Theorem.

Minimization over sets of neural-network parameters is a nonlinear programming problem, which can be solved, as discussed in [32, p. 1489], by iterative methods such as gradient descent [7, pp. 103-106, 173-174] (possibly with additive stochastic terms to avoid local minima, due to nonconvexity of $\mathcal{E}_K$ as a function of the parameters), genetic algorithms [19], and simulated annealing [1].

Acknowledgements

The authors thank Paul C. Kainen and A. Vogt from Georgetown University, Washington, D.C., USA, for fruitful comments and discussions.

7 Appendix

A Reproducing Kernel Hilbert Space (RKHS) is a Hilbert space $(X, \langle , \rangle)$ formed by functions defined on a nonempty set Ω such that for every $u \in \Omega$ the evaluation functional $\mathcal{F}_u$, defined for any $f \in X$ as $\mathcal{F}_u(f) = f(u)$, is bounded [2, 5, 9].

RKHSs 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 positive integers m, all $(w_1, \dots, w_m) \in \mathcal{R}^m$, and all $(u_1, \dots, u_m) \in \Omega^m$,

$$\sum_{i,j=1}^{m} w_i \, w_j \, K(u_i, u_j) \geq 0.$$

By the Riesz Representation Theorem [15, p. 200], for every $u \in \Omega$ there exists a unique element $K_u \in X$, called the *representer* of u, such that $\mathcal{F}_u(f) = \langle f, K_u \rangle$ for all $f \in X$ (this property is called the *reproducing property*). It is easy to check that the function $K : \Omega \times \Omega$ defined for all $u, v \in \Omega$ as $K(u, v) = \langle u, v \rangle$ is a kernel.

On the other hand, every kernel $K : \Omega \times \Omega \to \mathcal{R}$ generates a RKHS, which is denoted by $\mathcal{H}_K(\Omega)$ with the norm $\| \cdot \|_K$ and the inner product $\langle \cdot, \cdot \rangle_K$. $\mathcal{H}_K(\Omega)$ is defined as the completion of the linear span of the set $\{K_u : u \in \Omega\}$ with the inner product $\langle K_u, K_v \rangle_K = K(u, v)$ (see, e.g., [2] and [5, p. 81]).

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

$$\sup_{u \in \Omega} |f(u)| \leq s_K \|f\|_K \, . \tag{7.1}$$

A paradigmatic example of a kernel is the *Gaussian kernel* $K : \mathcal{R}^d \times \mathcal{R}^d \to \mathcal{R}$, defined as $K(u, v) = \exp(-\|u - v\|^2)$. Other examples of kernels are $K(u, v) = \exp(-\|u - v\|)$, $K(u, v) = \langle u, v \rangle^p$ (*homogeneous polynomial* of degree p), where $\langle \cdot, \cdot \rangle$ is any inner product on $\mathcal{R}^d$, $K(u, v) = (1 + \langle u, v \rangle)^p$ (*inhomogeneous polynomial* of degree p), and $K(u, v) = (a^2 + \|u - v\|^2)^{-\alpha}$ with $\alpha > 0$ [9, p. 38].

The role of $\|.\|_K^2$ as a stabilizer can be illustrated on two examples of classes of kernels. The first one is formed by *Mercer kernels*, i.e., continuous kernels defined on compact $\Omega \subset \mathcal{R}^d$. For a Mercer kernel K, $\|.\|_K^2$ can be expressed using eigenvectors and eigenvalues of the compact linear operator $L_K : \mathcal{L}_2(\Omega) \to \mathcal{C}(\Omega)$ defined for every $f \in \mathcal{L}_2(\Omega)$ as $L_K(f)(x) = \int_\Omega K(x, u) \, f(u) \, du$, where $\mathcal{L}_2(\Omega)$ and $\mathcal{C}(\Omega)$ denote the spaces of square integrable and of continuous functions on

Ω, resp. By the Mercer Theorem (see, e.g., [9, p.36])

$$\|f\|_K^2 = \sum_{i=1}^{\infty} \frac{c_i^2}{\lambda_i}\,,$$

where the λ_i's are the eigenvalues of L_K and the c_i's are the coefficients of the representation $f = \sum_{i=1}^{\infty} c_i \phi_i$, where $\{\phi_i\}$ is the orthonormal basis of $\mathcal{H}_K(\Omega)$ formed by the eigenvectors of L_K.

Note that the sequence $\{\lambda_i\}$ is either finite or it converges to zero (for K smooth, the convergence to zero is rather fast [14, p. 1119]). Thus the stabilizer $\|.\|_K^2$ penalizes functions, for which the sequence of coefficients $\{c_i\}$ does not converge to zero sufficiently quickly. So the stabilizer $\|.\|_K^2$ plays the role of a high-frequency filter.

The second class of kernels, on which we illustrate the role of $\|.\|_K^2$ as a stabilizer, contains *convolution kernels*, i.e., kernels $K(x,y) = k(x-y)$, for which the Fourier transform $\tilde{k}$ is positive. For such kernels, the stabilizer can be represented as

$$\|f\|_K^2 = \frac{1}{(2\,\pi)^{d/2}} \int_{R^d} \frac{\tilde{f}(\omega)^2}{\tilde{k}(\omega)}\, d\omega \tag{7.2}$$

(see [17], [37, p. 97]). So the function $\frac{1}{\tilde{k}}$ plays an analogous role as the sequence $\{\frac{1}{\lambda_i}\}$ in the case of a Mercer kernel.

For example, the Gaussian kernel is a convolution kernel with positive Fourier transform (its Fourier transform is $\tilde{k}(\omega) = \exp(-\|\omega\|^2/2)$).

Another example of a convolution kernel with positive Fourier transform is $K(u,v) = k(u-v) = \exp(-a\,\|u-v\|)$, where $k(t) = \exp(-a\,\|t\|)$, $\tilde{k}(\omega) = 2^{d/2}\,a\,\pi^{-1/2}\Gamma(d/2+1)\,(a^2 + \|\omega\|^2)^{-(d+1)/2}$ [37, p. 107], and Γ denotes the gamma function defined for a complex number s with $\mathrm{Re}(s) > 0$, as $\Gamma(s) = \int_0^\infty \exp(-r)\,r^{s-1}\,d\,r$ (for all non-negative integers n, $\Gamma(n+1) = n!$). In this case, the rate of decay of $\tilde{k}(\omega)$ is of the order of $\|\omega\|^{-(d+1)}$.

For $d = 1$ and $a = 1$, one gets as a special case a kernel $K : \mathcal{R} \times \mathcal{R} \to \mathcal{R}$ defined as $K(u,v) = k(u-v) = \exp(-|u-v|)$. Since $\Gamma(1) = 1$, $\Gamma(1/2) = \sqrt{\pi}$, and $\Gamma(s+1) = s\,\Gamma(s)$, $\tilde{k}(\omega) = \left(\sqrt{2\pi}(1+\omega^2)\right)^{-1}$. Thus $\|f\|_K^2 = 1/2\pi \int_{\mathcal{R}} \tilde{f}(\omega)^2 \left(\sqrt{2\pi}(1+\omega^2)\right) d\omega = 1/\sqrt{2\pi} \int_{\mathcal{R}} \tilde{f}(\omega)^2\, d\omega + 1/\sqrt{2\pi} \int_{\mathcal{R}} \omega^2\, \tilde{f}(\omega)^2\, d\omega$. As $\tilde{f'} = \omega\,\tilde{f}(\omega)$ and $\int_{\mathcal{R}} f(t)^2\, dt = 1/2\pi \int_{\mathcal{R}} \tilde{f}(\omega)^2\, d\omega$, by Parseval's formula [35, p. 172], $\|f\|_K^2 = \sqrt{2\pi}\left(\|f\|_{\mathcal{L}_2}^2 + \|f'\|_{\mathcal{L}_2}^2\right)$. So as noticed in [17], in this case the norm on the RKHS is equal to the Sobolev norm $\|.\|_{1,2}$.

For more information on kernels and their role in learning theory see, e.g., [37].

Bibliography

[1] E. Aarts and J. Korst. *Simulated Annealing and Boltzmann Machines: A Stochastic Approach to Combinatorial Optimization and Neural Computing.* John Wiley & Sons, 1989.

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

[3] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods.* Cambridge University Press, 2000.

[4] A. R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. on Information Theory*, 39(3):930-945, 1993.

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

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

[7] D. P. Bertsekas. *Nonlinear Programming.* Athena Scientific, Belmont, MA, 1999.

[8] C. Cortes and V. Vapnik, Support vector networks. *Machine Learning*, 20:1-25, 1995.

[9] F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of AMS*, 39:1-49, 2001.

[10] J. A. Cuesta-Albertos and M. Wschebor. Some remarks on the condition number of a real random square matrix. *J. of Complexity*, 19:548-554, 2003.

[11] J. Demmel. The geometry of ill-conditioning. *J. of Complexity*, 3:201-229, 1987.

[12] A. L. Dontchev. *Perturbations, Approximations and Sensitivity Analysis of Optimal Control Systems.* Lecture Notes in Control and Information Sciences, vol. 52. Springer-Verlag, Berlin Heidelberg, 1983.

[13] A. L. Dontchev and T. Zolezzi. *Well-Posed Optimization Problems.* Lecture Notes in Math., vol. 1543. Springer-Verlag, Berlin Heidelberg, 1993.

[14] N. Dunford, J. T. Schwartz: *Linear Operators. Part II: Spectral Theory.* Interscience Publishers, 1963.

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

[16] F. Girosi. Regularization theory, Radial Basis Functions and networks. In *From Statistics to Neural Networks. Theory and Pattern Recognition Applications*, V. Cherkassky, J. H. Friedman, and H. Wechsler, Eds. Springer-Verlag, Subseries F, Computer and Systems Sciences, 1994.

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

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

[19] D. E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning.* Addison-Wesley, 1989.

[20] L. K. Jones. A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training. *Ann. of Statistics*, 20:608-613, 1992.

[21] G. S. Kimeldorf and G. Wahba. A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Ann. Math. Statist.*, 41:495-502, 1970.

[22] V. Kůrková. Dimension-independent rates of approximation by neural networks. In *Computer-Intensive Methods in Control and Signal Processing: The Curse of Dimensionality* (K. Warwick, M. Kárný, Eds.), pp. 261-270. Birkhäuser, Boston, 1997.

[23] V. Kůrková. High-dimensional approximation by neural networks. In *Advances in Learning Theory: Methods, Models and Applications* (J. Stuykens et al., Eds.), pp. 69-88. Amsterdam: IOS Press, 2003.

[24] V. Kůrková and M. Sanguineti. Bounds on rates of variable-basis and neural-network approximation. *IEEE Trans. on Information Theory*, 47:2659-2665, 2001.

[25] V. Kůrková and M. Sanguineti. Comparison of worst case errors in linear and neural network approximation. *IEEE Trans. on Information Theory*, 48:264-275, 2002.

[26] V. Kůrková and M. Sanguineti. Error estimates for approximate optimization by the extended Ritz method. Submitted to *SIAM J. on Optimization.*

[27] V. Kůrková, P. Savický, and K. Hlaváčková. Representations and rates of approximation of real–valued Boolean functions by neural networks. *Neural Networks*, 11:651-659, 1998.

[28] E. S. Levitin and B. T. Polyak. Convergence of minimizing sequences in conditional extremum problems. *Dokl. Akad. Nauk SSSR*, 168(5):764-767, 1966.

[29] J. M. Ortega. *Numerical Analysis: A Second Course.* SIAM, Philadelphia, 1990.

[30] E. Parzen. An approach to time series analysis. *Ann. Math. Statist.*, 32:951-989, 1961.

[31] Pisier, G.: Remarques sur un résultat non publié de B. Maurey. *Séminaire d'Analyse Fonctionnelle* 1980-81, Exposé no. V, pp. V.1-V.12, École Polytechnique, Centre de Mathématiques, Palaiseau, France.

[32] T. Poggio and F. Girosi, Networks for approximation and learning. *Proc. IEEE* 78(9):1481-1497, 1990.

[33] T. Poggio and F. Girosi, Regularization algorithms for learning that are equivalent to multilayer networks. *Science* 247:978-982, 1990.

[34] T. Poggio and S. Smale. The mathematics of learning: dealing with data. *Notices of the AMS* 50(5):536-544, 2003.

[35] W. Rudin. *Functional Analysis.* McGraw-Hill, New York, N.Y., 1973.

[36] B. Schölkopf, R. Herbrich, A. J. Smola, and R. C. Williamson. A generalized Representer Theorem. *Proc. COLT'01, Lecture Notes in Artificial Intelligence*, pp. 416-424. Springer, 2001.

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

[38] I. J. Schönberg. Metric spaces and completely monotone functions. *Ann. of Math.*, 39:811-841, 1938.

[39] A. N. Tikhonov. Solutions of incorrectly formulated problems and the regularization method. *Soviet Math. Dokl.*, 4:1035-1038, 1963.

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

[41] V. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, 1998.

[42] A. A. Vladimirov, Yu. E. Nesterov, and Yu. N. Chekanov. On uniformly convex functionals. *Vestnik Moskovskogo Universiteta. Seriya 15 - Vychislitel'naya Matematika i Kibernetika*, 3:12-23, 1979. (English translation: *Moscow University Computational Mathematics and Cybernetics*, pp. 10-21, 1979).

[43] G. Wahba, *Splines Models for Observational Data*. Series in Applied Mathematics, vol. 59. SIAM, Philadelphia, PA, 1990.