Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Complexity 21 (2005) 350–367
Journal of COMPLEXITY
www.elsevier.com/locate/jco

Learning with generalization capability by kernel methods of bounded complexity

Věra Kůrková[a,1,2], Marcello Sanguineti[b,*,1,3]

[a] *Institute of Computer Science, Academy of Sciences of the Czech Republic Pod Vodárenskou věží 2, 182 07, Prague 8, Czech Republic*

[b] *Department of Communications, Computer, and System Sciences (DIST) University of Genoa, Via Opera Pia 13, 16145 Genova, Italy*

Received 23 November 2003; accepted 15 November 2004
Available online 12 February 2005

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 rate 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

* Corresponding author.
E-mail addresses: vera@cs.cas.cz (V. Kůrková), marcello@dist.unige.it (M. Sanguineti).

[1] Collaboration between V. K. and M. S. was supported by the 2004–2006 Scientific Agreement among University of Genoa, National Research Council of Italy, and Academy of Sciences of the Czech Republic, project "Learning from Data by Neural Networks and Kernel Methods: an Approach Based on Approximate Optimization".

[2] Partially supported by Project 1ET100300419 "Intelligent Models, Algorithms, Methods, and Tools for Semantic Web Realization" of the program "Information Society" of the National Research Program of the Czech Republic.

[3] Partially supported by a PRIN grant from the Italian Ministry of University and Research (Project "New Techniques for the Identification and Adaptive Control of Industrial Systems").

0885-064X/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.jco.2004.11.002

1. Introduction

A key property of systems performing intelligent computing, such as feature extraction, pattern recognition, semantic web realization, and classification, is learning ability. The goal of supervised learning is to adjust the parameters of a computational model so that it approximates to a desired accuracy a functional relationship between inputs and outputs by learning from a set of examples, i.e., a sample $\mathbf{z} = \{(x_i, y_i) \in \Omega \times \Re,\ i = 1, \ldots, m\}$ of m input/output pairs of *empirical data*. It is desirable that a model trained on a sample of empirical data also has a *generalization capability*, i.e., it is able to satisfactorily process new data, which were not used for learning. To endow a model with this capability, one needs some global knowledge of the desired input/output functional relationship, such as smoothness or lack of high-frequency oscillations.

In statistical learning theory [9,45], learning from empirical data is modelled as minimization of a functional, called *empirical error*. For a *sample* $\mathbf{z}$ of data and a *loss function* $V : \Re^2 \to [0, +\infty)$, the empirical error $\mathcal{E}_{\mathbf{z},V}$ is defined as $\mathcal{E}_{\mathbf{z},V}(f) = \frac{1}{m}\sum_{i=1}^{m} V(f(x_i), y_i)$, where f belongs to a function space, called *hypothesis space*, over which such a minimization is performed.

Mathematical modeling of generalization requires some *prior information* on the behavior of potential solutions. Such information is already expressed by the choice of a hypothesis space, over which the empirical error is minimized. It can be further specified by restricting minimization of the empirical error to a subset of the hypothesis space (containing only functions with some desired behavior). Alternatively, one can add to the empirical error a term penalizing undesired properties, or combine these two approaches. The first method is an application to learning of Ivanov's regularization, the second one of Tikhonov's, and the third one of Miller's [6, pp. 68–78].

Tikhonov's regularization [43,44], which was introduced into learning theory by Poggio and Girosi [20,35,36], leads to minimization over the whole hypothesis space of the *regularized empirical error functional*, defined as the sum of two functionals $\mathcal{E}_{\mathbf{z},V} + \gamma\Psi$. The first one, the empirical error $\mathcal{E}_{\mathbf{z},V}$, enforces closeness to the sample $\mathbf{z}$ of empirical data, whereas 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 empirical data and penalizing undesired behavior.

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 often play the role of measures of various types of oscillations of input/output mappings. RKHSs were formally defined by Aronszajn [2], but their theory employs work by Schönberg [41], as well as many classical results on kernels and positive definite functions. RKHSs were introduced into applications closely related to learning by Parzen [33] and Wahba [47], and into learning theory by Cortes and Vapnik [8] and Girosi [19].

The *Representer Theorem* [10, p. 42], [18,20,26,35,37,39] states that for Tikhonov's regularization with a stabilizer defined as a strictly increasing function of the norm on an RKHS, the problem of minimization of the regularized empirical error over the whole 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 vector $\mathbf{x} = (x_1, \ldots, x_m)$. In particular, for a stabilizer equal to the square of the norm on an RKHS, the vector $\mathbf{c}$ of the coefficients

of the linear combination is obtained as the solution of the well-posed linear system of equations $(\gamma m\mathcal{I} + \mathcal{K}[\mathbf{x}])\mathbf{c} = \mathbf{y}$, where $\mathcal{I}$ is the $m \times m$ identity matrix, $\mathcal{K}[\mathbf{x}]$ is the Gram matrix of the kernel K with respect to $\mathbf{x}$, and $\mathbf{y} = (y_1, \ldots, y_m)$ is the output data vector [3].

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 $x_1, \ldots, x_m$ [18]. The coefficients of the linear combination play the role of output weights of such a network. On the basis of this interpretation of the Representer Theorem, in [20, p. 219] it was argued that "the regularization principles lead to approximation schemes that are equivalent to networks with one layer of hidden units."

The Representer Theorem was used to design a learning algorithm (see, e.g., [10, p. 42] and [37, pp. 538–539]) that requires one to solve the linear system of equations $(\gamma m\mathcal{I} + \mathcal{K}[\mathbf{x}])\mathbf{c} = \mathbf{y}$. 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 describable in terms of a condition on smoothness that can be modelled in terms of a kernel.

However, practical applications of this algorithm are limited by the rate of convergence of iterative methods solving the system of equations and by the size of the condition number of the matrix $\gamma m\mathcal{I} + \mathcal{K}[\mathbf{x}]$. For some methods, the computational requirements for solving such a system grow polynomially with the size m of the sample (e.g., for the Gaussian elimination and m large enough, they grow as $m^3/3$ [32, p. 175]). For some data and kernels, keeping the condition number of $\gamma m\mathcal{I} + \mathcal{K}[\mathbf{x}]$ small requires a large value of the regularization parameter γ, which may cause poor fit to the empirical data.

The learning algorithm based on the Representer Theorem uses a computational model of complexity determined by the size m of the sample of data, and does not allow any flexibility in choosing the inner parameters of the computational units (as they are set equal to the input data).

In this paper, we investigate suboptimal solutions of the problems of minimization of regularized empirical error functionals over hypothesis sets corresponding to kernel models with limited complexity and flexible choice of parameters. We derive upper bounds on the rates of convergence of sequences of suboptimal solutions achievable by minimization over hypothesis sets formed by linear combinations of at most n kernel functions (either with arbitrary parameters or with parameters drawn from the data set) 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_2-norm of the vector $\mathbf{y} = (y_1, \ldots, y_m)$ of output data, the minimal and the maximal eigenvalues of the Gram matrix $\mathcal{K}[\mathbf{x}]$ 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 suboptimal solutions converge "quickly" to the optimal one. Under such conditions, kernel methods with bounded model complexity provide good approximations to 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. For the Gaussian kernel we derive an upper bound of the form $\frac{b}{\sqrt{n}}$, where $b = \frac{y_{\max}^2}{\gamma}\left(3\frac{1+\gamma}{\gamma} + 2\right)$ and $y_{\max}^2$ is 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 from data with RKHSs as hypothesis spaces. Section 3 states the Representer Theorem and explores the condition numbers of the matrices used in algorithms based on this theorem. Section 4 develops tools for investigating approximate optimization over hypothesis sets with bounded model complexity. Section 5 describes continuity and convexity properties of regularized empirical error functionals with various types of loss functions and estimates rates of convergence of sequences of suboptimal solutions to the problem of learning by kernel methods with increasing model complexity. Section 6 illustrates the estimates on RKHSs defined by convolution kernels. Section 7 is a brief discussion.

2. Tikhonov's regularization of the learning problem in reproducing kernel Hilbert spaces

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

Let M be a subset of X and $\Phi : X \to \Re$ be a functional. Using standard notation [15], we denote by

$$(M, \Phi)$$

the problem of minimizing Φ over M; M is called *hypothesis set*.

By $\operatorname{argmin}(M, \Phi) = \{g \in M : \Phi(g) = \inf_{g\in M} \Phi(g)\}$ is denoted the set of *minimum points* of the problem (M, Φ) and for any $\varepsilon > 0$, $\operatorname{argmin}_\varepsilon (M, \Phi) = \{g \in M : \Phi(g) < \inf_{g\in M} \Phi(g) + \varepsilon\}$ is the set of *ε-near minimum points* of (M, Φ). A minimum point of (M, Φ) is called a *solution* of the problem (M, Φ). A sequence $\{g_n\}$ of elements of M is called *Φ-minimizing over M* if $\lim_{n\to\infty} \Phi(g_n) = \inf_{g\in M} \Phi(g)$.

Let Ω be a set and $\mathbf{z} = \{(x_i, y_i) \in \Omega \times \Re,\ i = 1, \ldots, m\}$ an m-tuple of input/output pairs of data, called a *sample*. A standard approach to learning from empirical data [9,45] is based on minimization of the *empirical error* functional (also called the *empirical risk* functional), defined for any f in the hypothesis set as

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

where $V : \Re^2 \to [0, \infty)$ satisfying for all $y \in \Re$, $V(y, y) = 0$ is called a *loss function*. When the sample $\mathbf{z}$ is clear from the context, we write merely $\mathcal{E}_V$ instead of $\mathcal{E}_{\mathbf{z},V}$.

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 merely write $\mathcal{E}$. So we let

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

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* [43,44].

An important class of stabilizers are squares of norms on reproducing kernel Hilbert spaces (RKHSs) (see, e.g., [5,10,40]). Such stabilizers often enable one to penalize high oscillations of various types. For a set Ω and a symmetric positive semidefinite function $K : \Omega \times \Omega \to \Re$, called *kernel*, we denote by $(\mathcal{H}_K(\Omega), \|.\|_K)$ the RKHS defined by K. The squared norm $\|.\|_K^2$ is used as a stabilizer instead of $\|.\|_K$ for technical reasons, as the square of the norm on any Hilbert space is a uniformly convex functional (see Proposition 4.1 (iii)); this implies uniqueness of the solution of the regularized problem (see, e.g., [14, p. 10; 10, pp. 27, 42]) and convergence of minimizing sequences to this solution [31].

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, when the square loss is employed in the regularized empirical error we use the simplified notation

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

Thus we denote by

$$(\mathcal{H}_K(\Omega), \mathcal{E}_{\gamma,K})$$

the problem of minimizing over the RKHS $\mathcal{H}_K(\Omega)$ the regularized empirical error with the square loss and the stabilizer $\|.\|_K^2$.

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 minimizing over the whole RKHS the regularized empirical error with

the square loss and the stabilizer $\|.\|_K^2$ are given by the *Representer Theorem*. For a kernel K, a positive integer m, and a vector $\mathbf{x} = (x_1, \ldots, x_m) \in \Omega^m$ of input data, we denote by $\mathcal{K}[\mathbf{x}]$ the $m \times m$ matrix defined as

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

which is called the *Gram matrix of the kernel K with respect to the vector* $\mathbf{x}$. We denote by $\mathcal{I}$ the $m \times m$ identity matrix.

Theorem 3.1 (*Representer Theorem*). *Let Ω be a nonempty set, $K : \Omega \times \Omega \to \Re$ a kernel, m a positive integer, $\mathbf{x} = (x_1, \ldots, x_m) \in \Omega^m$, $\mathbf{y} = (y_1, \ldots, y_m) \in \Re^m$, and $\gamma > 0$. Then the problem $(\mathcal{H}_K(\Omega), \mathcal{E}_{\gamma,K})$ has the unique solution*

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

where $\mathbf{c} = (c_1, \ldots, c_m)$ is the unique solution of the well-posed linear system

$$(\gamma\, m\, \mathcal{I} + \mathcal{K}[\mathbf{x}])\mathbf{c} = \mathbf{y}. \tag{2}$$

The Representer Theorem was originally proven in [23]. An elegant proof using functional derivatives was given in [37, pp. 538–539], while for Mercer kernels a more sophisticated argument based on the Mercer Theorem was provided in [10, p. 42]. In [26] it was derived from the theory of inverse problems. Inspection of 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 $c_1, \ldots, c_m$ is nonlinear [19, p. 1473]. A weaker form of the Representer Theorem, without a formula for computing the coefficients $c_1, \ldots, c_m$, even holds for an arbitrary loss function V and a stabilizer of the form $\psi(\|\cdot\|_K)$, where $\psi : [0, +\infty) \to \Re$ is a strictly increasing function [39].

The Representer Theorem was exploited to design algorithms for learning from data (see, e.g., [10, p. 42, 37, pp. 538–539]). However, its applications are limited by the rates of convergence of iterative methods solving the linear system of equations (2) and by the size of the condition number of the matrix $\gamma m\mathcal{I} + \mathcal{K}[\mathbf{x}]$.

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

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

where $\|\mathcal{A}\|$ denotes the norm of $\mathcal{A}$ as a linear operator on $(\Re^m, \|.\|)$. We denote by $\lambda_{\max}(\mathcal{A})$ and $\lambda_{\min}(\mathcal{A})$ the maximal and minimal eigenvalues of the matrix $\mathcal{A}$, respectively.

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

To simplify the notation, we write $\lambda_{\max}$ instead of $\lambda_{\max}(\mathcal{K}[\mathbf{x}])$ and similarly for $\lambda_{\min}$. As $\mathcal{K}[\mathbf{x}]$ is positive semidefinite, all its eigenvalues are nonnegative [32, p. 7]. As λ is an eigenvalue of $\mathcal{K}[\mathbf{x}]$ if and only if $\gamma m + \lambda$ is an eigenvalue of $\gamma m\mathcal{I} + \mathcal{K}[\mathbf{x}]$, we have

$$cond_2(\gamma m\mathcal{I} + \mathcal{K}[\mathbf{x}]) = \frac{\gamma m + \lambda_{\max}}{\gamma m + \lambda_{\min}} \leqslant \frac{\lambda_{\max}}{\lambda_{\min}} = cond_2(\mathcal{K}[\mathbf{x}]) \tag{3}$$

and

$$cond_2(\gamma m\mathcal{I} + \mathcal{K}[\mathbf{x}]) \leqslant 1 + \frac{\lambda_{\max}}{\gamma m}. \tag{4}$$

Eq. (3) shows that when $cond_2(\mathcal{K}[\mathbf{x}])$ is sufficiently small, good conditioning of $\gamma m\mathcal{I} + \mathcal{K}[\mathbf{x}]$ is guaranteed for any value of γ. However, for large values of m the matrix $\mathcal{K}[\mathbf{x}]$ might be ill-conditioned. For example, when the data are uniformly distributed over an interval, then the probability that $\mathcal{K}[\mathbf{x}]$ is ill-conditioned increases with m (see [12, Theorem 2.2], [13, Theorem 5.1]). On the other hand, Eq. (4) shows that $\lim_{\gamma\to\infty} cond_2(\gamma m\mathcal{I}+\mathcal{K}[\mathbf{x}]) = 1$ and thus the regularization parameter γ can always be chosen such that $cond_2(\gamma m\mathcal{I}+\mathcal{K}[\mathbf{x}])$ is close to 1. But good conditioning of $\gamma m\mathcal{I} + \mathcal{K}[\mathbf{x}]$ is not the only requirement for γ, as its value must also allow a good fit to the empirical data and thus it cannot be too large. Existence of a value of γ guaranteeing a good fit to data as well as good conditioning depends on the rate of convergence of the condition number of $\gamma m\mathcal{I} + \mathcal{K}[\mathbf{x}]$ to 1. The smaller $\frac{\lambda_{\max}}{m}$, the faster such convergence. The problem of choosing γ in order to minimize the expected error was investigated in [11].

When a value of γ guaranteeing both a small condition number and a good fit to the empirical data cannot be found, algorithms for learning from data that differ from the one based on the Representer Theorem have to be applied. A variety of learning algorithms have been developed in the field of neurocomputing. Typically, such algorithms operate on networks of lower model complexity than the algorithm based on the Representer Theorem. The number of computational units in such networks is either set in advance or adjusted during learning, but, typically, it is much smaller than the size m of the sample used as a training set. Moreover, the values of the computational units' 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 searched for during learning.

4. Minimization of functionals over hypothesis sets with bounded model complexity

In this section, we derive tools for estimating rates of convergence of suboptimal solutions over computational models with n units (the case of interest is $n < m$) to the optimal solution given by the Representer Theorem. Such suboptimal solutions can be studied in terms of optimization over nested families of subsets of RKHSs formed by linear combinations of all n-tuples of kernel functions chosen from the sets $\{K_x : x \in \Omega\}$ or $\{K_{x_1}, \ldots, K_{x_m}\}$.

For a subset G of a linear space, let $span_n\, G = \left\{\sum_{i=1}^n w_i g_i : w_i \in \Re,\ g_i \in G\right\}$ denote the set of linear combinations of all n-tuples of elements of G. The optimal solution to the problem $(\mathcal{H}_K(\Omega), \mathcal{E}_{\gamma,K})$ described by the Representer Theorem is an element of

$span_m\, G_{K,\mathbf{x}} \subseteq span_m\, G_K$, where $G_{K,\mathbf{x}} = \{K_{x_1}, \ldots, K_{x_m}\}$ and $G_K = \{K_x : x \in \Omega\}$. The set $span_m G_K$ can be interpreted as the set of all input/output functions of a computational model with one hidden layer of 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 computational units [20].

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 or $G_{K,\mathbf{x}}$), we shall employ a version of the Maurey–Jones–Barron Theorem [3,22,34], reformulated in [24] in terms of a norm called *G-variation*.

We 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 $p_M(f) = \inf\{\lambda \in \Re_+ : f/\lambda \in M\}$. If M is 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 norm, denoted by $\|.\|_G$, 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 for every $f \in X$ we have

$$\|f\|_G = \inf\{c > 0 : f/c \in cl\ conv\,(G \cup -G)\}.$$

For properties of G-variation, see [24,25,27,28,30].

Maurey–Jones–Barron's Theorem stated in terms of G-variation [24,25] 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 the rate of approximation of f by $span_n G$: $\|f - span_n G\| \leqslant \sqrt{\frac{(s_G\|f\|_G)^2 - \|f\|^2}{n}}$.

Taking advantage of this upper bound, we shall estimate rates of convergence of suboptimal solutions over $span_n G$ to the optimal solution of the problem (X, Φ) of minimization of a continuous functional Φ over a normed linear space X.

A functional $\Phi : X \to \Re$ 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 $\omega(a) = \sup\{|\Phi(f) - \Phi(g)| : \|f - g\| \leqslant a\}$.

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 $\Phi(\lambda h + (1 - \lambda)g) \leqslant \lambda\Phi(h) + (1 - \lambda)\Phi(g)$ and it is *uniformly convex* if there exists a nonnegative function $\delta : \Re_+ \to \Re_+$ such that $\delta(0) = 0$, $\delta(t) > 0$ for all $t > 0$, and for all $h, g \in M$ and all $\lambda \in [0, 1]$, $\Phi(\lambda h + (1-\lambda)g) \leqslant \lambda\Phi(h) + (1-\lambda)\Phi(g) - \lambda(1-\lambda)\delta(\|h-g\|)$. Any such function δ is called a *modulus of convexity* of Φ [31][4].

Next proposition states some elementary properties of uniformly convex functionals and moduli of convexity.

[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 where $\delta : \Re_+ \to \Re_+$ merely satisfies $\delta(0) = 0$ and $\delta(t_0) > 0$ for some $t_0 > 0$ (see, e.g., [46,14, p. 10]).

Proposition 4.1. *Let $(X, \|.\|)$ be a normed linear space, $M \subseteq X$ convex, and Φ 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 $\Psi + \gamma\Phi$ is uniformly convex on M with a modulus of convexity $\gamma\,\delta$*;
(ii) *if $g^o \in argmin(M, \Phi)$, then for every $g \in M$, $\delta(\|g - g^o\|) \leqslant \Phi(g) - \Phi(g^o)$*;
(iii) *if $(X, \|.\|)$ is a Hilbert space, then the functional $\|.\|^2 : X \to \Re$ is uniformly convex with a modulus of convexity $\delta(t) = t^2$.*

Proof. (i) follows directly from the definitions.

(ii) By the definition of uniformly convex functional, for every $\lambda \in [0, 1]$ we have $\lambda(1-\lambda)\delta(\|g-g^o\|) \leqslant \lambda\Phi(g)+(1-\lambda)\Phi(g^o)-\Phi(\lambda g+(1-\lambda)g^o)$. As $\Phi(g^o) \leqslant \Phi(\lambda g+(1-\lambda)g^o)$, we get $\lambda(1-\lambda)\delta(\|g-g^o\|) \leqslant \lambda\Phi(g)+(1-\lambda)\Phi(g^o)-\Phi(g^o) = \lambda\,(\Phi(g) - \Phi(g^o))$. Hence $(1-\lambda)\delta(\|g-g^o\|) \leqslant \Phi(g)-\Phi(g^o)$ for every $\lambda \in [0, 1]$. So we obtain $\delta(\|g-g^o\|) \leqslant \Phi(g)-\Phi(g^o)$.

(iii) For every $h, g \in X$ and every $\lambda \in [0, 1]$, we have $\|\lambda h + (1-\lambda)g\|^2 \leqslant \lambda\|h\|^2 + (1-\lambda)\|g\|^2 - \lambda(1-\lambda)\|h-g\|^2$ and thus $\delta(t) = t^2$ is a modulus of convexity of $\|.\|^2$. □

Next theorem gives upper bounds on rates of convergence of suboptimal solutions over $span_n G$ to the optimal solution of the problem (X, Φ) of minimization of a continuous functional Φ over a Hilbert space X. The estimates are formulated in terms of moduli of continuity and convexity of the functional to be minimized. □

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, $g^o \in \operatorname{argmin}(X, \Phi)$, Φ continuous at g^o with a modulus of continuity α, $\{\varepsilon_n\}$ a sequence of positive real numbers, $g_n \in \operatorname{argmin}_{\varepsilon_n}(span_n\, G, \Phi)$, and $a = (s_G\|g^o\|_G)^2 - \|g^o\|^2$. Then, for every positive integer n the following estimates hold*:
(i) $\inf_{g\in span_n\, G} \Phi(g) - \Phi(g^o) \leqslant \alpha\left(\sqrt{\frac{a}{n}}\right)$;
(ii) *if $\|g^o\|_G < \infty$ and $\lim_{n\to\infty} \varepsilon_n = 0$, then $\{g_n\}$ is a Φ-minimizing sequence and* $\Phi(g_n) - \Phi(g^o) \leqslant \alpha\left(\sqrt{\frac{a}{n}}\right) + \varepsilon_n$'
(iii) *if Φ is uniformly convex with a modulus of convexity δ, then* $\delta(\|g_n - g^o\|) \leqslant \alpha\left(\sqrt{\frac{a}{n}}\right) + \varepsilon_n$.

Proof. (i) For every positive integer n and every $\varepsilon > 0$, choose an ε-near best approximation f_n^ε of g^o in $span_n G$. So $\|g^o - f_n^\varepsilon\| < \|g^o - \operatorname{span}_n G\| + \varepsilon$. As $f_n^\varepsilon \in span_n G$, we have $\inf_{g\in span_n\, G} \Phi(g) - \Phi(g^o) \leqslant \Phi(f_n^\varepsilon) - \Phi(g^o)$. Estimating the right-hand side of this inequality in terms of the modulus of continuity α of Φ at g^o, we obtain $\inf_{g\in span_n\, G} \Phi(g) - \Phi(g^o) \leqslant \alpha(\|f_n^\varepsilon - g^o\|) \leqslant \alpha(\|g^o - span_n G\| + \varepsilon)$. By the upper bound from Maurey–Jones–Barron's Theorem reformulated in terms of G-variation we get

$$\inf_{g\in span_n\, G} \Phi(g) - \Phi(g^o) \leqslant \alpha\left(\sqrt{\frac{a}{n}} + \varepsilon\right). \tag{5}$$

Infimizing (5) over ε we obtain (i).

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

$$\Phi(g_n) - \Phi(g^o) \leqslant \alpha\left(\sqrt{\frac{a}{n}}\right) + \varepsilon_n. \tag{6}$$

If $\|g^o\|_G$ is finite and $\lim_{n\to\infty} \varepsilon_n = 0$, then the right-hand side of (6) converges to zero and so $\{g_n\}$ is Φ-minimizing.

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

Theorem 4.2 can be also obtained as a corollary of [29, Theorem 4.2], which applies to other types of regularization, too, such as Ivanov's one. However, the direct argument used here is much simpler than the proof of [29, Theorem 4.2].

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 given by the Representer Theorem for the problem $(\mathcal{H}_K(\Omega), \mathcal{E}_{\gamma,K})$. In contrast to the optimal solution g^o, which is a linear combination of the representers $K_{x_1}, \ldots, K_{x_m}$ determined by the sample $\mathbf{x} = (x_1, \ldots, x_m)$ 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 applications, a proper *n*-tuple together with coefficients of the linear combination can be adjusted by a suitable nonlinear programming algorithm (see, e.g., [1,7,21]).

To employ Theorem 4.2 to estimate rates of approximate minimization of regularized empirical error functionals with kernel stabilizers, we need upper bounds on the 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 5.1. *Let Ω be a nonempty set, $K : \Omega \times \Omega$ a kernel, $s_K = \sup_{x\in\Omega} \sqrt{K(x,x)}$, $\gamma > 0$, m a positive integer, $\mathbf{x} = (x_1, \ldots, x_m) \in \Omega^m$, $\mathbf{y} = (y_1, \ldots, y_m) \in \Re^m$, $y_{\max} = \max\{|y_i| : i = 1, \ldots, m\}$, and $V : \Re^2 \to \Re$ a loss function. Then the following hold*:
(i) *if for every $i = 1, \ldots, m$ the functions $V(\cdot, y_i) : \Re \to \Re$ 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 $b_1 = 2\left(\|f\|_K (s_K^2 + \gamma) + y_{\max} s_K\right)$, 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 minimum point g^o of the problem $(\mathcal{H}_K(\Omega), \mathcal{E}_{V,\gamma,K})$ and for every $f \in \mathcal{H}_K(\Omega)$*

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

Proof. (i) It is easy to check 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) Using the Cauchy–Schwartz inequality and reproducing property one can show that for every kernel K, $\sup_{u\in\Omega} |f(u)| \leqslant s_K \|f\|_K$, where $s_K = \sup_{u\in\Omega} \sqrt{K(u,u)}$. Thus for the square loss, $|\mathcal{E}_{V,\gamma,K}(f) - \mathcal{E}_{V,\gamma,K}(g)| = \left|\frac{1}{m} \sum_{i=1}^m \left((f(x_i) - y_i)^2 - (g(x_i) - y_i)^2\right) + \gamma \left(\|f\|_K^2 - \|g\|_K^2\right)\right| \leqslant \left|\frac{1}{m} \sum_{i=1}^m (f(x_i) - g(x_i))(f(x_i) + g(x_i) - 2y_i)\right| + \gamma \left|\|f\|_K - \|g\|_K\right| (\|f\|_K + \|g\|_K) \leqslant \sup_{x\in\Omega} |f(x) - g(x)| \left(\sup_{x\in\Omega} |f(x) + g(x)| + 2y_{\max}\right) + \gamma \|f - g\|_K (\|f\|_K + \|g\|_K)$.

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

Similarly, for the absolute value loss we have $|\mathcal{E}_{V,\gamma,K}(f) - \mathcal{E}_{V,\gamma,K}(g)| = \left|\frac{1}{m} \sum_{i=1}^m (|f(x_i) - y_i| - |g(x_i) - y_i|) + \gamma \left(\|f\|_K^2 - \|g\|_K^2\right)\right| \leqslant \sup_{x\in\Omega} |f(x) - g(x)| + \gamma \left|\|f\|_K - \|g\|_K\right| (\|f\|_K + \|g\|_K) \leqslant s_K \|f - g\|_K + \gamma \|f - g\|_K (\|f\|_K + \|g\|_K)$. If $\|f - g\|_K \leqslant t$, then $|\mathcal{E}_{V,\gamma,K}(f) - \mathcal{E}_{V,\gamma,K}(g)| \leqslant s_K\, t + t\,\gamma\,(\|f\|_K \|g\|_K) \leqslant s_K\, t + t\,\gamma\,(t + 2\|f\|_K)$. Hence $|\mathcal{E}_{V,\gamma,K}(f) - \mathcal{E}_{V,\gamma,K}(g)| \leqslant \beta(t) = b_2 t^2 + b_1 t$, where $b_2 = \gamma$ and $b_1 = s_K + 2\gamma\|f\|_K$.

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

The assumptions of Proposition 5.1(i) are satisfied by both the square loss 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 bounded from above by the quadratic function $\beta(t) = b_2 t^2 + b_1 t$, where for both losses b_2 depends on γ and for the square loss, also on s_K, while b_1 depends on γ, s_K, $\|f\|_K$ and for the square loss, also on $y_{\max}$. 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 $V(\cdot, y_i)$, $i = 1, \ldots, m$.

Although the next theorem holds for any positive integer n, it is useful only for $n < m$ since by the Representer Theorem, the minimum point of $\mathcal{E}_{\gamma,K}$ over $span_m G_K$ is equal to the minimum point over the whole space $\mathcal{H}_K(\Omega)$.

Theorem 5.2. *Let Ω be a nonempty set, $K : \Omega \times \Omega \to \Re$ a kernel, $s_K = \sup_{x\in\Omega} \sqrt{K(x,x)}$, m a positive integer, $\mathbf{x} = (x_1, \ldots, x_m) \in \Omega^m$, $\mathbf{y} = (y_1, \ldots, y_m) \in \Re^m$, $y_{\max} = \max\{|y_i| : i = 1, \ldots, m\}$, $g^o = \sum_{i=1}^m c_i K_{x_i}$ the unique solution of $(\mathcal{H}_K(\Omega), \mathcal{E}_{\gamma,K})$, $\{\varepsilon_n\}$ a sequence of positive real numbers such that $\lim_{n\to\infty} \varepsilon_n = 0$, and $\{g_n\}$ a sequence of*

ε_n-near minimum points of $(span_n G_K, \mathcal{E}_K)$. *Let* $a = (s_K \|g^o\|_{G_K})^2 - \|g^o\|_K^2$, $u = (s_K^2 + \gamma)a$, *and* $v = 2\left((s_K^2+\gamma)\|g^o\|_K + y_{\max}s_K\right)\sqrt{a}$. *Then, for every positive integer n the following estimates hold*:

(i) $\inf_{g\in span_n G_K} \mathcal{E}_{\gamma,K}(g) - \mathcal{E}_{\gamma,K}(g^o) \leqslant \frac{u}{n} + \frac{v}{\sqrt{n}}$;

(ii) $\mathcal{E}_{\gamma,K}(g_n) - \mathcal{E}_K(g^o) \leqslant \frac{u}{n} + \frac{v}{\sqrt{n}} + \varepsilon_n$;

(iii) $\|g_n - g^o\|_K^2 \leqslant \frac{1}{\gamma}\left(\frac{u}{n} + \frac{v}{\sqrt{n}} + \varepsilon_n\right)$;

(iv) $\sup_{x\in\Omega} |g_n(x) - g^o(x)|^2 \leqslant \frac{s_K^2}{\gamma}\left(\frac{u}{n} + \frac{v}{\sqrt{n}} + \varepsilon_n\right)$.

Proof. (i) Combining Theorem 4.2 (i) with Proposition 5.1 (ii), we get $\inf_{g\in span_n G_K} \mathcal{E}_{\gamma,K}(g) - \mathcal{E}_{\gamma,K}(g^o) \leqslant \beta\left(\sqrt{\frac{a}{n}}\right)$, where $\beta(t) = (s_K^2+\gamma)\,t^2 + 2\left((s_K^2+\gamma)\|g^o\|_K + y_{\max}s_K\right)t$, which gives for $\inf_{g\in span_n G_K} \mathcal{E}_{\gamma,K}(g) - \mathcal{E}_{\gamma,K}(g^o)$ the upper bound $(s_K^2+\gamma)\frac{a}{n} + 2((s_K^2 + \gamma)\|g^o\|_K + y_{\max}s_K)\sqrt{\frac{a}{n}} = \frac{u}{n} + \frac{v}{\sqrt{n}}$.

Similarly, item (ii) follows from Theorem 4.2(ii) and Proposition 5.1(ii), item (iii) follows from (ii) and Proposition 5.1(iii), and item (iv) from (iii) and the inequality $\sup_{u\in\Omega} |f(u)| \leqslant s_K\|f\|_K$, which is obtained using the Cauchy-Schwartz inequality and the reproducing property. □

Thus when u and v are not too large, it is possible to choose n small enough so that a computational model with n units is implementable and a suboptimal solution over such a model approximates well the optimal solution given by the Representer Theorem.

Only two terms in the above formulas defining u and v cannot be derived directly from the data sample $\mathbf{z}$, the kernel K and the regularization parameter γ: the values of the two norms of the optimal solution g^o, i.e., its G_K-variation and its norm $\|.\|_K$. 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 $\mathbf{y}$, and the maximal and minimal eigenvalues, $\lambda_{\max}$ and $\lambda_{\min}$, of the Gram matrix $\mathcal{K}[\mathbf{x}]$ of the kernel K with respect to the input data vector $\mathbf{x}$. The l_1- and l_2-norm on $\Re^m$ are denoted by $\|\cdot\|_1$ and $\|\cdot\|_2$, respectively.

The estimates in the rest of the paper (Proposition 5.3, Theorem 5.4, and Corollaries 6.1 and 6.2) involve an upper bound on $\|g^o\|_{G_K}$, which is also an upper bound on $\|g^o\|_{G_{K,\mathbf{x}}}$. Thus, all these estimates can be applied also to approximate solutions over hypothesis sets formed by functions from $\text{span}_n G_{K,\mathbf{x}}$. Such solutions are obtained when n representers are chosen from the set $G_{K,\mathbf{x}}$, as, e.g., in [42], where approximation techniques were proposed that reduce the Gram matrix $\mathcal{K}[\mathbf{x}]$ to a sparse matrix of lower rank.

Proposition 5.3. *Let* Ω *be a nonempty set,* $K : \Omega\times\Omega \to \Re$ *a kernel,* $s_K = \sup_{x\in\Omega}\sqrt{K(x,x)}$, $\gamma > 0$, *m a positive integer,* $\mathbf{x} = (x_1,\ldots,x_m) \in \Omega^m$, $\mathbf{y} = (y_1,\ldots,y_m) \in \Re^m$, $g^o = \sum_{i=1}^m c_i\, K_{x_i}$ *the unique solution of* $(\mathcal{H}_K(\Omega), \mathcal{E}_{\gamma,K})$. *Then the following estimates hold*:

(i) $\|g^o\|_{G_K} \leqslant \frac{\sqrt{m}\|\mathbf{y}\|_2}{\gamma m + \lambda_{\min}}$;

(ii) $\|g^o\|_K \leqslant \frac{\sqrt{\lambda_{\max}}\|\mathbf{y}\|_2}{\gamma m + \lambda_{\min}}$;

(iii) $s_K^2\,\|g^o\|_{G_K}^2 - \|g^o\|_K^2 \leqslant \frac{(s_K^2 m - \lambda_{\min})\,\|\mathbf{y}\|_2^2}{(\gamma m + \lambda_{\min})^2}$.

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

$$\|g^o\|_{G_K} \leqslant \sum_{i=1}^{m} |c_i| = \|\mathbf{c}\|_1 \leqslant \sqrt{m}\,\|\mathbf{c}\|_2, \tag{7}$$

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

$$\|\mathbf{c}\|_2 \leqslant \frac{\|\mathbf{y}\|_2}{\gamma m + \lambda_{\min}} \tag{8}$$

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

(ii) By the Representer Theorem, $\|g^o\|_K^2 = \left\langle \sum_{i=1}^m c_i\, K_{x_i}, \sum_{j=1}^m c_j K_{x_j} \right\rangle_K = \sum_{i,j=1}^m c_i\, c_j K(x_i, x_j) = \mathbf{c}^T\, \mathcal{K}[\mathbf{x}]\mathbf{c}$, where $\mathbf{c}^T$ denotes the transpose of the vector $\mathbf{c}$. As $\lambda_{\min}\|\mathbf{c}\|_2^2 \leqslant \mathbf{c}^T\, \mathcal{K}[\mathbf{x}]\mathbf{c} \leqslant \lambda_{\max}\|\mathbf{c}\|_2^2$ [32, p. 21], we have

$$\lambda_{\min}\|\mathbf{c}\|_2^2 \leqslant \|g^o\|_K^2 \leqslant \lambda_{\max}\|\mathbf{c}\|_2^2. \tag{9}$$

Thus by (8), $\|g^o\|_K \leqslant \frac{\sqrt{\lambda_{\max}}\|\mathbf{y}\|_2}{\gamma\, m + \lambda_{\min}}$.

(iii) By (7)–(9), we obtain

$$\begin{aligned} s_K^2\|g^o\|_{G_K}^2 - \|g^o\|_K^2 &\leqslant s_K^2 m\|\mathbf{c}\|_2^2 - \lambda_{\min}\|\mathbf{c}\|_2^2 \leqslant \left(s_K^2 m - \lambda_{\min}\right)\|\mathbf{c}\|_2^2 \\ &\leqslant \frac{(s_K^2\, m - \lambda_{\min})\|\mathbf{y}\|_2^2}{(\gamma m + \lambda_{\min})^2}. \qquad \square \end{aligned}$$

As both $\lambda_{\min}$ and $\lambda_{\max}$ are nonnegative, we can further simplify as follows the upper bounds from Proposition 5.3:

$$\text{(i)}\ \ \|g^o\|_{G_K} \leqslant \frac{\|\mathbf{y}\|_2}{\gamma\sqrt{m}}, \tag{10}$$

$$\text{(ii)}\ \ \|g^o\|_K \leqslant \frac{\sqrt{\lambda_{\max}}\|\mathbf{y}\|_2}{\gamma m}, \tag{11}$$

$$\text{(iii)}\ \ s_K^2\ \|g^o\|_{G_K}^2 - \|g^o\|_K^2 \leqslant \frac{s_K^2\|\mathbf{y}\|_2^2}{\gamma^2 m}. \tag{12}$$

Combining Proposition 5.3 with Theorem 5.2 and inequalities (10)–(12), we shall 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, $\|\mathbf{y}\|_2$, $y_{\max}$, $\lambda_{\min}$, and $\lambda_{\max}$.

Theorem 5.4. *Let Ω be a nonempty set, $K : \Omega\times\Omega \to \Re$ a kernel, $s_K = \sup_{x\in\Omega}\sqrt{K(x,x)}$, $\gamma > 0$, m a positive integer, $\mathbf{x} = (x_1, \ldots, x_m) \in \Omega^m$, $\mathbf{y} = (y_1, \ldots, y_m) \in \Re^m$, $y_{\max} =$*

$\max\{|y_i| : i = 1, \ldots, m\}$, $g^o = \sum_{i=1}^m c_i\, K_{x_i}$ *the unique solution of* $(\mathcal{H}_K(\Omega), \mathcal{E}_{\gamma,K})$, $\{\varepsilon_n\}$ *a sequence of positive real numbers, and* $\{g_n\}$ *a sequence of* ε_n*-near minimum points of* $(span_n\, G_K, \mathcal{E}_{\gamma,K})$. *Let*

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

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

Then, for every positive integer n the following estimates hold:

(i) $\inf_{g \in span_n\, G_K}\ \mathcal{E}_{\gamma,K}(g) - \mathcal{E}_{\gamma,K}(g^o) \leqslant \frac{\bar{u}}{n} + \frac{\bar{v}}{\sqrt{n}}$;

(ii) $\mathcal{E}_{\gamma,K}(g_n) - \mathcal{E}_K(g^o) \leqslant \frac{\bar{u}}{n} + \frac{\bar{v}}{\sqrt{n}} + \varepsilon_n$;

(iii) $\|g_n - g^o\|_K^2 \leqslant \frac{1}{\gamma}\left(\frac{\bar{u}}{n} + \frac{\bar{v}}{\sqrt{n}} + \varepsilon_n\right)$;

(iv) $\sup_{x \in \Omega} |g_n(x) - g^o(x)|^2 \leqslant \frac{s_K^2}{\gamma}\left(\frac{\bar{u}}{n} + \frac{\bar{v}}{\sqrt{n}} + \varepsilon_n\right)$.

Thus, to obtain a good approximation of the solution of $(\mathcal{H}_K(\Omega), \mathcal{E}_{\gamma,K})$ given by the Representer Theorem by a suboptimal solution computable by a model with at most $n < m$ computational units, both $\frac{\bar{u}}{n}$ and $\frac{\bar{v}}{\sqrt{n}}$ have to be sufficiently small for some n, for which models with n computational units computing functions from G_K are implementable.

6. Estimates for convolution kernels

In this section, we illustrate the estimates given in Theorem 5.4 by examples of RHSH with $\Omega = \Re^d$ and convolution kernels. Let $K(u, v) = \psi(\|u - v\|)$ be a convolution kernel, where $\psi : \Re \to [0, 1]$ is monotonically decreasing and satisfies $\psi(0) = 1$ (this includes the Gaussian kernel). The following corollary estimates rates of convergence of suboptimal solutions for input/output pairs of data $(x_1, y_1), \ldots, (x_m, y_m)$ for which the input data are sufficiently separated so that there exists $t \in [0, 1]$ such that for all distinct $i, j \in \{1, \ldots, m\}$, $\psi(\|x_i - x_j\|) \leqslant t$.

Corollary 6.1. *Let* $K : \Re^d \times \Re^d \to \Re$ *be a kernel such that* $K(s, t) = \psi(\|s - t\|)$ *with* $\psi : \Re \to [0, 1]$ *monotonically decreasing, satisfying* $\psi(0) = 1$, *and such that for all distinct* $i, j \in \{1, \ldots, m\}$, $\psi(\|x_i - x_j\|) \leqslant t$ *for some* $t > 0$. *Let* $\gamma > 0$, m *be a positive integer,* $\mathbf{x} = (x_1, \ldots, x_m) \in \Re^{dm}$, $\mathbf{y} = (y_1, \ldots, y_m) \in \Re^m$, $y_{\max} = \max\{|y_i| : i = 1, \ldots, m\}$, $g^o = \sum_{i=1}^m c_i\, K_{x_i}$ *the unique solution of* $(\mathcal{H}_K(\Re^d), \mathcal{E}_K)$, $\{\varepsilon_n\}$ *a sequence of positive real numbers, and* $\{g_n\}$ *a sequence of* ε_n*-near minimum points of* $(span_n\, G_K, \mathcal{E}_{\gamma,K})$. *Let*

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

$$\hat{v} = \frac{2}{\gamma\sqrt{m}}\|\mathbf{y}\|_2 \left((1+\gamma)\frac{\sqrt{1+(m-1)t}\|\mathbf{y}\|_2}{\gamma m} + y_{\max} \right)$$

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

Proof. As $s_K = 1$ and $\lambda_{\max} \leqslant \|\mathcal{K}[\mathbf{x}]\|_1 = \max_{j=1,\ldots,m} \sum_{i=1}^m |K[\mathbf{x}]_{i,j}|$ [32, pp. 6, 21–23], we have $\lambda_{\max} \leqslant 1+(m-1)t$. Hence estimates (i)–(iv) follow from Theorem 5.4 with $\bar{u} = \hat{u}$ and $\bar{v} \leqslant \hat{v}$. □

Bounding from above the right-hand side of the estimates from Corollary 6.1 in terms of the maximum of the absolute values of output data, we obtain the following corollary.

Corollary 6.2. *Let $K : \Re^d \times \Re^d \to \Re$ be a kernel such that $K(s,t) = \psi(\|s-t\|)$ with $\psi : \Re \to [0,1]$ monotonically decreasing, satisfying $\psi(0) = 1$, and such that for all distinct $i, j \in \{1,\ldots,m\}$, $\psi(\|x_i - x_j\|) \leqslant t$ for some $t > 0$. Let $\gamma > 0$, m be a positive integer, $\mathbf{x} = (x_1,\ldots,x_m) \in \Re^{dm}$, $\mathbf{y} = (y_1,\ldots,y_m) \in \Re^m$, $y_{\max} = \max\{|y_i| : i = 1,\ldots,m\}$, $g^o = \sum_{i=1}^m c_i K_{x_i}$ the unique solution of $(\mathcal{H}_K(\Re^d), \mathcal{E}_K)$, $\{\varepsilon_n; n = 1,\ldots,m\}$ positive real numbers, $\{g_n : n = 1,\ldots,m\}$ ε_n-near minimum points of $(span_n G_K, \mathcal{E}_{\gamma,K})$, and $b = \frac{y_{\max}^2}{\gamma}(3\frac{1+\gamma}{\gamma} + 2)$.*

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

Proof. As $\|\mathbf{y}\|_2^2 \leqslant m y_{\max}^2$, we have $\frac{\hat{u}}{n} + \frac{\hat{v}}{\sqrt{n}} \leqslant \frac{1+\gamma}{\gamma^2} y_{\max}^2 \frac{1}{n} + \frac{2}{\gamma} y_{\max}^2 ((1+\gamma)\frac{\sqrt{1+(m-1)t}}{\gamma\sqrt{m}} + 1)\frac{1}{\sqrt{n}}$, which for $t \in [0,1]$ and $n \leqslant m$ is bounded from above by $\frac{1+\gamma}{\gamma} y_{\max}^2 \frac{1}{n} + \frac{2}{\gamma} y_{\max}^2 (\frac{1+\gamma}{\gamma} + 1)\frac{1}{\sqrt{n}} \leqslant 3\frac{1+\gamma}{\gamma^2} y_{\max}^2 + \frac{2}{\gamma} y_{\max}^2 = \frac{y_{\max}^2}{\gamma}(3\frac{1+\gamma}{\gamma} + 2)$. Hence estimates (i)–(iv) follow from Corollary 6.1. □

So, when γ is not too small and $y_{\max}$ is not too large, Corollary 6.2 guarantees a good approximation of the optimal solution by 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* computational units approximates the global minimum over the whole RKHS within $\frac{b}{\sqrt{n}}$, where $b = \frac{y_{\max}^2}{\gamma}(3\frac{1+\gamma}{\gamma} + 2)$.

7. Discussion

We have compared two approaches to learning from data with generalization capability, both modeling learning as a minimization of the empirical error functional with the square loss function regularized by the square of a norm on an RKHS, but differing in the hypothesis sets over which minimization is performed. The first approach, which is based on the Representer Theorem, considers minimization of the regularized empirical error over the whole RKHS, whereas the second one over its subset formed by functions computable by linear combinations of n computational units defined by the kernel.

We have derived upper bounds on the errors of approximation of the optimal solution by the suboptimal ones obtainable with n increasing. We have shown that when the absolute values of output data are not too large and the regularization parameter is not too small, suboptimal solutions approximate the optimal one within an accuracy $\frac{c}{\sqrt{n}}$ with c moderate. In such cases, algorithms operating on models with n computational units can approximate the optimal solution quite well. Hence, when the solution of the system of linear equations described in the Representer Theorem is not computationally feasible or when the system is ill-conditioned, models with bounded complexity provide a useful and quite accurate alternative to the learning algorithms based on the Representer Theorem. For convolution kernels on $\Re^d \times \Re^d$ the upper bounds from Corollaries 6.1 and 6.2 do not depend on the number d of variables, so the approximation of the optimal solution by such models does not exhibit the curse of dimensionality [4].

Minimization over a set of parameters of a chosen model is a nonlinear programming problem [35, p. 1489], which can be solved by iterative methods such as gradient descent [7, pp. 103–106,173–174] (possibly with additive stochastic terms to avoid local minima, due to the nonconvexity of $\mathcal{E}_{\gamma,K}$ as a function of the parameters), genetic algorithms [21], and simulated annealing [1].

Acknowledgements

The authors thank GIORGIO GNECCO, Martin Holeňa, Paul Kainen, Per Kullstam, and Andrew Vogt for fruitful comments and discussions.

References

[1] E. Aarts, J. Korst, Simulated Annealing and Boltzmann Machines: A Stochastic Approach to Combinatorial Optimization and Neural Computing, Wiley, New York, 1989.
[2] N. Aronszajn, Theory of reproducing kernels, Trans. Amer. Math. J. 68 (1950) 337–404.
[3] A.R. Barron, Universal approximation bounds for superpositions of a sigmoidal function, IEEE Trans. Inform. Theory 39 (1993) 930–945.
[4] R. Bellman, Dynamic Programming, Princeton University Press, Princeton, NJ, 1957.
[5] C. Berg, J.P.R. Christensen, P. Ressel, Harmonic Analysis on Semigroups, Springer, New York, 1984.
[6] M. Bertero, Linear inverse and ill-posed problems, Adv. Electron. Electron Phys. 75 (1989) 1–120.
[7] D.P. Bertsekas, Nonlinear Programming, Athena Scientific, Belmont, MA, 1999.
[8] C. Cortes, V. Vapnik, Support vector networks, Mach. Learning 20 (1995) 1–25.
[9] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods, Cambridge University Press, Cambridge, 2000.

[10] F. Cucker, S. Smale, On the mathematical foundations of learning, Bull. Amer. Math. Soc. 39 (2001) 1–49.
[11] F. Cucker, S. Smale, Best choices for regularization parameters in learning theory: on the bias-variance problem, Found. Comput. Math. 2 (2002) 413–428.
[12] J.A. Cuesta-Albertos, M. Wschebor, Some remarks on the condition number of a real random square matrix, J. Complexity 19 (2003) 548–554.
[13] J. Demmel, The geometry of ill-conditioning, J. Complexity 3 (1987) 201–229.
[14] A.L. Dontchev, Perturbations, Approximations, and Sensitivity Analysis of Optimal Control Systems, Lecture Notes in Control and Information Sciences, vol. 52, Springer, Berlin, Heidelberg, 1983.
[15] A.L. Dontchev, T. Zolezzi, Well-Posed Optimization Problems, Lecture Notes in Mathemathics, vol. 1543, Springer, Berlin, Heidelberg, 1993.
[18] F. Girosi, Regularization theory, Radial Basis Functions and networks, in: V. Cherkassky, J.H. Friedman, H. Wechsler (Eds.), From Statistics to Neural Networks, Theory and Pattern Recognition Applications—Subseries F, Computer and Systems Sciences, Springer, Berlin, 1994, , pp. 166–187.
[19] F. Girosi, An equivalence between sparse approximation and support vector machines, Neural Comput. 10 (1998) 1455–1480.
[20] F. Girosi, M. Jones, T. Poggio, Regularization theory and neural networks architectures, Neural Comput. 7 (1995) 219–269.
[21] D.E. Goldberg, Genetic Algorithms in Search, Optimization, and Machine Learning, Addison-Wesley, Reading, MA, 1989.
[22] L.K. Jones, A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training, Ann. Statist. 20 (1992) 608–613.
[23] G.S. Kimeldorf, G. Wahba, A correspondence between Bayesian estimation on stochastic processes and smoothing by splines, Ann. Math. Statist. 41 (1970) 495–502.
[24] V. Kůrková, Dimension-independent rates of approximation by neural networks, in: K. Warwick, M. Kárný (Eds.), Computer-Intensive Methods in Control and Signal Processing: The Curse of Dimensionality, Birkhäuser, Boston, 1997, , pp. 261–270.
[25] V. Kůrková, High-dimensional approximation by neural networks, in: J. Stuykens et al. (Eds.), Advances in Learning Theory: Methods, Models and Applications, IOS Press, Amsterdam, 2003, , pp. 69–88.
[26] V. Kůrková, Learning from data as an inverse problem, in: J. Antoch (Ed.), COMPSTAT 2004—Proceedings in Computational Statistics, Physica-Verlag, Springer, Heidelberg, 2004, , pp. 1377–1384.
[27] V. Kůrková, M. Sanguineti, Bounds on rates of variable-basis and neural-network approximation, IEEE Trans. Inform. Theory 47 (2001) 2659–2665.
[28] V. Kůrková, M. Sanguineti, Comparison of worst case errors in linear and neural network approximation, IEEE Trans. Inform. Theory 48 (2002) 264–275.
[29] V. Kůrková, M. Sanguineti, Error estimates for approximate optimization by the extended Ritz method, SIAM J. Optim., to appear.
[30] V. Kůrková, P. Savický, K. Hlaváčková, Representations and rates of approximation of real–valued Boolean functions by neural networks, Neural Networks 11 (1998) 651–659.
[31] E.S. Levitin, B.T. Polyak, Convergence of minimizing sequences in conditional extremum problems, Sov. Math. Dokl. 7 (1996) 764–767 (English translation: Dokl. Akad. Nauk SSSR 168 (1966) 997–1000).
[32] J.M. Ortega, Numerical Analysis: A Second Course, SIAM, Philadelphia, 1990 (reprint of the 1972 edition by Academic Press, New York).
[33] E. Parzen, An approach to time series analysis, Ann. Math. Statist. 32 (1961) 951–989.
[34] G. Pisier, Remarques sur un résultat non publié de B. Maurey, Séminaire d'Analyse Fonctionnelle, 1980–81, Exposé no. V, École Polytechnique, Centre de Mathématiques, Palaiseau, France, pp. V.1–V.12.
[35] T. Poggio, F. Girosi, Networks for approximation and learning, Proc. IEEE 78 (1990) 1481–1497.
[36] T. Poggio, F. Girosi, Regularization algorithms for learning that are equivalent to multilayer networks, Science 247 (1990) 978–982.
[37] T. Poggio, S. Smale, The mathematics of learning: dealing with data, Notices Amer. Math. Soc. 50 (2003) 536–544.
[39] B. Schölkopf, R. Herbrich, A.J. Smola, R.C. Williamson. A generalized Representer Theorem, Lecture Notes in Artificial Intelligence (Proceedings of the Annual Conference on Computational Learning Theory—COLT), Springer, London, 2001, pp. 416–426.

[40] B. Schölkopf, A.J. Smola, Learning With Kernels—Support Vector Machines Regularization, Optimization and Beyond, MIT Press, Cambridge, MA, 2002.
[41] I.J. Schönberg, Metric spaces and completely monotone functions, Ann. Math. 39 (1938) 811–841.
[42] A.J. Smola, B. Schölkopf, Sparse greedy matrix approximation for machine learning, in: Proceedings of the 17th International Conference on Machine Learning, Morgan Kaufmann, San Francisco, CA, 2000, pp. 911–918.
[43] A.N. Tikhonov, Solutions of incorrectly formulated problems and the regularization method, Soviet Math. Dokl. 4 (1963) 1035–1038.
[44] A.N. Tikhonov, V.Y. Arsenin, Solutions of Ill-posed Problems, W.H. Winston, Washington, DC, 1977.
[45] V. Vapnik, Statistical Learning Theory, Wiley, Chichester, GB, 1998.
[46] A.A. Vladimirov, Yu.E. Nesterov, Yu.N. Chekanov, On uniformly convex functionals, Vestnik Moskov. Univ. Ser. 15 Vychisl. Mat. Kibernet. 3 (1978) 12–23 (English translation: Moscow Univ. Comput. Math. Cybern. (1979) 10–21).
[47] G. Wahba, Splines Models for Observational Data, CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 59. SIAM, Philadelphia, PA, 1990.