Learning from data by neural networks with a limited complexity

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

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

Abstract

Learning from data formalized as a minimization of a regularized empirical error is studied in terms of approximate minimization over sets of functions computable by networks with increasing number of hidden units. There are derived upper bounds on speed of convergence of infima achievable over networks with \(n\) hidden units to the global infimum. The bounds are expressed in terms of norms tailored to the type of network units and moduli of continuity of regularized empirical error functionals.

1 Introduction

The basic goal of supervised learning is to adjust network parameters so that the network approximates with a sufficient accuracy a desired functional relationship between inputs and outputs. Typically, such a relationship is not known analytically. Instead, a training set is given consisting of a sample of input/output pairs \(\{(x_i, y_i) \in \mathcal{R}^d \times \mathcal{R}, i = 1, \ldots, m\}\). To be able to utilize a network trained on such a random sample for processing new data, which were not used for training, one needs some global knowledge (“conceptual data”) of the desired input/output function (such as smoothness or lack of high frequency oscillations).

Learning based on both empirical and conceptual data has been modelled [10] as minimization of a functional called regularized empirical error, which is the sum of two functionals \(E_V + \gamma \Psi\). The first one, called the empirical error, is defined as \(E_V(f) = \frac{1}{m} \sum_{i=1}^{m} V(f(x_i), y_i)\), where \(V\) is a loss function, which measures how much is lost if the network computes from an input \(x\) an output \(f(x)\) instead of \(y\). The empirical error \(E_V\) enforces closeness to the sample of empirical data \((x_1, y_1), \ldots, (x_m, y_m)\), while \(\Psi\), called stabilizer, expresses requirements on the global behavior of the desired input/output function, and the regularization parameter \(\gamma\) controls the trade-off between fitting to the empirical and the conceptual data.

It has been argued in [10, p. 219] that “the regularization principles lead to approximation schemes that are equivalent to networks with one layer of hidden units”. Indeed, various versions of the so-called Representer Theorem (see, e.g., [5, p. 42], [9], and [10]) show that for a wide class of stabilizers there exists a unique function minimizing the regularized empirical error and this unique argminimum is of the form of a one-hidden layer network with a linear output and hidden units computing functions corresponding to the type of a stabilizer. In particular, for stabilizers defined by the Gaussian kernel, the Representer Theorem states that the minimum is achieved at a function computable by a Gaussian radial-basis function network with centroids at \(x_1, \ldots, x_m\). The output weights of such a network can be computed by solving a well-posed system of linear equations, depending on the sample size \(m\), the output data \(y_1, \ldots, y_m\), the regularization parameter \(\gamma\), and the stabilizer \(\Psi\). An algorithm constructing such networks has been successfully applied to various pattern recognition tasks (see, e.g., [18]).

A drawback of the algorithm based on this elegant theoretical result is that the number of hidden units in the network is equal to the number of pairs of input/output data. For large data sets, such a network might not be efficiently implementable. Moreover in typical applications of neural networks, either a fixed number of units much smaller than the number of data is chosen before learning or, starting from a small network new hidden units are added till a sufficient performance is achieved. So to model such a typical neural network learning, minimization has to be
considered over sets of functions computable by networks with a limited number of hidden units imposed by feasibility of an implementation.

Motivated by these complexity constraints, we derive an approximate version of the Representer Theorem. We consider minimization of regularized empirical error functionals over admissible sets formed by functions computable by networks with at most \( n \) hidden units. We study such approximate optimization in a more general framework of approximate optimization called the “extended Ritz method”, which exploits variable-basis functions a special case of which are neural networks.

For continuous functionals, we derive upper bounds on rates of convergence of sequences of infima achievable over neural networks with increasing number of hidden units. The bounds are expressed in terms of certain norms, tailored to the type of network units, and moduli of continuity of regularized empirical errors. Our estimates are not merely asymptotic: they hold for every number of hidden units and thus they can be applied to networks with a limited number of hidden units imposed by feasibility of an implementation.

The paper is organized as follows. Section 2 introduces notations and definitions. In section 3, learning from data by neural networks is investigated in the framework of approximate optimization and there are derived estimates of rates of decrease of infima with increasing number of hidden units. In section 4, these estimates are applied to minimization of empirical error functionals regularized using squares of norms in reproducing kernel Hilbert spaces.

2 Preliminaries

By a normed linear space \((X, \| \cdot \|)\) we mean a real normed linear space and \(\mathcal{R}\) denotes the set of real numbers. A functional \(\Phi : X \rightarrow \mathcal{R}\) 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 \(\Phi\) at \(f\) is a function \(\omega : \mathcal{R}_+ \rightarrow \mathcal{R}_+\) defined as

\[
\omega(\alpha) = \sup \{ |\Phi(f) - \Phi(g)| : \|f - g\| \leq \alpha \}.
\]

\(\Phi\) is convex over a convex set \(M \subseteq X\) if for all \(h, g \in M\) and all \(\lambda \in [0, 1]\),

\[
\Phi(\lambda h + (1 - \lambda)g) \leq \lambda \Phi(h) + (1 - \lambda)\Phi(g).
\]

\(\Phi\) is strictly uniformly convex on a convex set \(M \subseteq X\) if there exists a function \(\delta : \mathcal{R}_+ \rightarrow \mathcal{R}_+\), such that \(\delta(0) = 0\), \(\delta(t) > 0\) for all \(t > 0\) and for all \(h, g \in M\) and all \(\lambda \in [0, 1]\),

\[
\Phi(\lambda h + (1 - \lambda)g) \leq \lambda \Phi(h) + (1 - \lambda)\Phi(g) - \lambda(1 - \lambda)\delta(\|h - g\|).
\]

Any such function \(\delta\) is called a modulus of convexity of \(\Phi\) [6, Chapter 1].

Using standard notation (see, e.g., [7]), we denote by \((M, \Phi)\) the problem of infimizing a functional \(\Phi\) over a subset \(M\) of \(X\). \(M\) is called the set of admissible solutions or admissible set. By

\[
\arg\min (M, \Phi) = \{g \in M : \Phi(g) = \inf_{g \in M} \Phi(g)\}
\]

is denoted the set of argminima of the problem \((M, \Phi)\) and for any \(\varepsilon > 0\),

\[
\arg\min_{\varepsilon} (M, \Phi) = \{g \in M : \Phi(g) < \inf_{g \in M} \Phi(g) + \varepsilon\}
\]

is the set of its \(\varepsilon\)-near argminima. A sequence \(\{g_n\}\) of elements of \(M\) is called \(\Phi\)-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 \((M, \Phi)\) with \(M\) non-empty, there always exists a minimizing sequence. The problem \((M, \Phi)\) is Tychonov well-posed if it has a unique global minimum in \(M\) to which every minimizing sequence converges [7, p. 1]. The modulus of Tychonov well-posedness of \((M, \Phi)\) at an argminimum \(g^o\) is a function \(\xi : \mathcal{R}_+ \rightarrow \mathcal{R}_+\) such that for every \(\alpha \in \mathcal{R}_+\),

\[
\xi(\alpha) = \inf_{g \in M \cap S_\alpha(g^o)} \Phi(g) - \Phi(g^o).
\]

Note that the modulus of Tychonov well-posedness is defined for any problem that has an argminimum, even when such a problem is not Tychonov well-posed.

The closure of \(M\) is

\[
\overline{M} = \{f \in X : (\forall \varepsilon > 0) (B_\varepsilon(f) \cap M) \neq \emptyset\}.
\]

The Minkowski functional of \(M \subseteq X\) denoted by \(p_M\), is defined for every \(f \in X\) as

\[
p_M(f) = \inf\{\lambda \in \mathcal{R}_+ : f/\lambda \in M\}.
\]

3 Learning from data as approximate optimization

Let \(\{(x_i, y_i) \in \mathcal{R}^d \times \mathcal{R}, i = 1, \ldots, m\}\) be a finite set of input/output pairs of data. Learning from such data can be formalized as minimization of the empirical error functional, defined as

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

where \(V : \mathcal{R}^2 \rightarrow \mathcal{R}\) is a loss function. The most common loss function is the square loss defined as

\[
V(f((x), y) = (f(x) - y)^2.
\]
There have also been used other loss functions such as the absolute value loss
\[ V(f(x), y) = |f(x) - y| \]
and Vapnik’s ε-insensitive loss
\[ V(f(x), y) = \max(|f(x) - y|, \varepsilon, 0). \]

Tychonov regularization replaces the functional \( E_V \) with
\[ E_{V, \gamma, \Phi} = E_V + \gamma \Psi, \]
where \( \Psi \) is a functional called stabilizer and \( \gamma \) is a regularization parameter [19]. Various stabilizers have been used to penalize high oscillations of admissible functions. An important class of such functionals are squares of norms in certain Hilbert spaces called reproducing kernel Hilbert spaces (see next section). Another important type of stabilizers has the form
\[ \Phi(f) = \int_{\mathcal{H}} \frac{\tilde{f}(s)}{F(s)} ds, \]
where \( \tilde{f} \) denotes the Fourier transform and \( F \) is a function such that \( \frac{1}{F} \) is a high-frequency filter (i.e., \( \lim_{|s| \to \infty} \tilde{F}(s) = 0 \)).

Minimization of functionals over sets of functions computable by neural networks can be studied in a more general framework of approximate optimization by the so-called “extended Ritz method” (see [16], [20]), which approximates a general admissible set of functions by its intersections with sets of the form
\[ \text{span}_n \mathcal{G} = \left\{ \sum_{i=1}^{n} w_i g_i : w_i \in \mathcal{R}, g_i \in \mathcal{G} \right\}, \]
where \( \mathcal{G} \) is a set of functions. This approximation scheme is called the variable-basis approximation [14, 15]. It includes splines with free nodes, trigonometric polynomials with free frequencies and feedforward neural networks (in such a case, \( \mathcal{G} \) is the set of functions computable by a given type of hidden unit).

Rates of approximation by this variable-basis scheme have been extensively studied (see, e.g., [2], [9], [15], [10], [13]). Their estimates have been used to describe sets of functions that can be approximated within a given accuracy by networks whose number of hidden units grows polynomially with the number of network inputs.

A useful tool for estimating rates of neural network approximation is Maurey-Jones-Barron’s theorem [17], [11],[2], which we present here in terms of a norm tailored the type of hidden units (see, e.g., [13]). For a subset \( \mathcal{G} \) of a normed linear space \((X, ||||), \) \( \mathcal{G} \)-variation, denoted by \( ||||_{\mathcal{G}} \), is defined as the Minkowski functional of the set \( cl \text{conv} (\mathcal{G} \cup -\mathcal{G}) \) [12], i.e.,
\[ ||f||_{\mathcal{G}} = \inf \{ c > 0 : f/c \in cl \text{conv} (\mathcal{G} \cup -\mathcal{G}) \}. \]

Theorem 3.1 (Maurey-Jones-Barron) Let \((X, ||||)\) be a Hilbert space, \( \mathcal{G} \) its bounded subset and \( s_{\mathcal{G}} = \sup_{g \in \mathcal{G}} ||g||. \) Then for every \( f \in X \) and every positive integer \( n \) we have
\[ ||f - \text{span}_n \mathcal{G}|| \leq \sqrt{\frac{s_{\mathcal{G}} ||f||_{\mathcal{G}}^2 - ||f||^2}{n}}. \]

Thus, any function in a ball of radius \( r \) in \( \mathcal{G} \)-variation can be approximated by a neural network with \( n \) hidden units computing functions from \( \mathcal{G} \) within accuracy \( r/\sqrt{n} \). Note that this estimate holds for functions of any number of variables and so it allows to describe sets of functions that can be approximated without the curse of dimensionality.

For \( \mathcal{G} \) corresponding to Heaviside perceptrons, \( \mathcal{G} \)-variation is a generalization of the concept of total variation that has been studied in integration theory. \( \mathcal{G} \)-variation is also a generalization of the notion of \( l_1 \)-norm as for \( \mathcal{G} \) an orthonormal basis of a separable Hilbert space, it is equal to the \( l_1 \)-norm with respect to \( \mathcal{G} \), defined for every \( f \in X \) as \( ||f||_{l_1, \mathcal{G}} = \sum_{g \in \mathcal{G}} |f \cdot g| \) [14]. For other properties of \( \mathcal{G} \)-variation see, e.g., [14] and [15].

The following theorem from [16] estimates speed of convergence of approximate infima and argminima over nested families of neural networks with increasing number of hidden units to the global infima and argminima. The estimates are formulated in terms of the number \( n \) of hidden units, \( \mathcal{G} \)-variation (for \( \mathcal{G} \) corresponding to the type of functions computed by these units), modulus of continuity of the functional to be minimized, and modulus of variation is also a generalization of the notion of \( l_1 \)-variation that has been studied in integration theory.

Theorem 3.2 Let \((X, ||||)\) be a Hilbert space, \( M \) and \( G \) be its subsets, \( s_{\mathcal{G}} = \sup_{g \in \mathcal{G}} ||g||. \) Closed, convex, and 0 \( \in \text{int} M. \) Let \( \Phi : X \to (-\infty, +\infty) \) be a functional, \( g^0 \in \text{argmin}_{\Phi} (M, \Phi). \) \( \Phi \) is continuous at \( g^0 \) with a modulus of continuity \( \omega, \) and let \( \{\varepsilon_n\} \) be a sequence of positive reals such that \( \lim_{n \to \infty} \varepsilon_n = 0 \) and \( g_n \in \text{argmin}_{\varepsilon_n} (M \cap \text{span}_n G, \Phi) \). Then for every integer \( n \) the following hold:
(i) \( \inf_{\varepsilon_n \neq 0} \text{span}_n G \Phi(g) - \Phi(g^0) \leq \omega \left( 1 + c||g^0|| \right) \sqrt{\frac{\sup_{\varepsilon_n \neq 0} ||f||_{\mathcal{G}}^2 - ||f||^2}{n}}; \)
(ii) if \( ||f||_{\mathcal{G}} < \infty, \) then \( \{g_n\} \) is a \( \Phi \)-minimizing sequence over \( M \) and \( \Phi(g_n) - \Phi(g^0) \leq \omega \left( 1 + c||g^0|| \right) \sqrt{\frac{\sup_{\varepsilon_n \neq 0} ||f||_{\mathcal{G}}^2 - ||f||^2}{n}} + \varepsilon_n; \)
(iii) if \( \xi \) is the modulus of Tychonov well-posedness of \((M, \Phi)\) at \( g^0, \) then \( \xi(||g_n - g^0||) \leq \omega \left( 1 + c||g^0|| \right) \sqrt{\frac{\sup_{\varepsilon_n \neq 0} ||f||_{\mathcal{G}}^2 - ||f||^2}{n}} + \varepsilon_n, \) where \( c \) is the Lipschitz constant of \( \Phi \) on \( X. \)
Theorem 3.2 estimates speed of convergence of the sequence \( \{g_n\} \) of approximate solutions of the problem \((M, \Phi)\) obtainable by networks with \( n \) hidden units. Moreover, it estimates speed of convergence of infima of \( \Phi \) over \( M \cap \text{span}_n G \) to the infimum over the whole set \( M \). Inspection of these estimates shows that if the solution of the optimization problem \((M, \Phi)\) is in the ball of a radius \( r \) in \( G \)-variation, then rates of approximate optimization are bounded from above by \( \omega(1 + c||g^*||r/\sqrt{n}) \). Thus when with increasing number of variables one keeps a fixed bound on \( G \)-variations of global argminima (or even on all admissible functions), the curse of dimensionality in approximate optimization by neural networks with increasing number of hidden units can be avoided. Note that for \( G \) corresponding to neural network computational units, balls in \( G \)-variation are not shrinking with increasing dimension [13].

As the estimates from Theorem 3.2 are not merely asymptotic, they allow one to estimate the “quality” of approximate solutions achievable over neural networks with any number of hidden units.

4 An approximate version of the Representer Theorem

In this section, we apply tools derived in Section 3 to learning from data by neural networks with a fixed number of hidden units.

A widespread model of such learning is in terms of minimization of a regularized empirical error functional with square loss function and stabilizers of the form \( \| \cdot \|^2_K \), where \( \| \cdot \|_K \) is the norm on a reproducing kernel Hilbert space (RKHS). A RKHS \( H_K(\Omega) \) is a Hilbert space of functions defined on a set \( \Omega \) such that for every \( x \in \Omega \) the evaluation functional \( F_x \), defined for any \( f \in H_K(\Omega) \) as \( F_x(f) = f(x) \), is bounded.

Recall that a mapping \( K : \Omega \times \Omega \rightarrow \mathcal{R} \) is positive semidefinite if for all positive integers \( m \), all \((w_1, \ldots, w_m) \in \mathcal{R}^m \), and all \((x_1, \ldots, x_m) \in \Omega \), we have

\[
\sum_{i,j=1}^m w_i w_j K(x_i, x_j) \geq 0.
\]

To every RKHS one can associate a unique symmetric, positive semidefinite mapping \( K : \Omega \times \Omega \rightarrow \mathcal{R} \) called kernel [1], [4, p. 81], such that for every \( f \in H_K(\Omega) \),

\[
f(x) = \langle f, K(x, \cdot) \rangle_K.
\]

A kernel \( K : \Omega \times \Omega \rightarrow \mathcal{R} \) is called Mercer kernel if \( \Omega \) is compact and \( K \) is symmetric, continuous and positive semidefinite. By [5, p. 35], for \( \Omega \) compact and \( K \) a Mercer kernel, the space \( H_K(\Omega) \) is contained in the space \( C(\Omega) \) of continuous functions on \( \Omega \). We denote by \( \| \cdot \|_C \) the supremum norm on \( C(\Omega) \). A kernel \( K \) is called convolution or translation invariant if there exists a function \( k : \Omega \rightarrow \mathcal{R} \) such that \( K(x, y) = k(x - y) \). Using \( \| \cdot \|^2_K \) as a stabilizer, the regularized empirical error functional with the square loss function has the form

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

The solution of the problem of minimizing \( \mathcal{E}_{\gamma, K} \) over the whole space \( H_K(\Omega) \) is described by the following theorem, called the “Representer Theorem” [5, p. 42] (there are also other versions for more general loss functions and other stabilizers [8, 9, 10]).

**Theorem 4.1** Let \( \Omega \subset \mathcal{R}^d \) be compact, \( K : \Omega \times \Omega \rightarrow \mathcal{R} \) be a Mercer kernel, \( H_K(\Omega) \) be the RKHS defined by \( K \).

Then there exists a unique argminimum of the problem \((H_K(\Omega), \mathcal{E}_{\gamma, K})\) of the form

\[
g^\gamma(x) = \sum_{i=1}^m a_i K(x, x_i),
\]

where \( a = (a_1, \ldots, a_m) \) is the unique solution of the well-posed linear system

\[
(K(x) + \gamma \mathbf{I}) a = y,
\]

where \( I \) is the \( m \times m \) identity matrix, and \( K(x) \) is the \( m \times m \) matrix defined as \( K(x)_{i,j} = K(x_i, x_j) \).

Notice that the unique argminimum described in Theorem 4.1 has the form of a variable-basis function from \( \text{span}_m G_K \), where \( G_K = \{ K(x, \cdot) : x \in \Omega \} \). In particular, for the convolution kernel defined by the Gaussian function, the argminimum can be computed by a Gaussian radial-basis function network with \( m \) hidden units. However for large values of \( m \), such a network might not be implementable.

Taking advantage of Theorem 3.2 we can derive an approximate version of the Representer Theorem which estimates “quality” of approximate solutions obtained by minimization of \( \mathcal{E}_{\gamma, K} \) over sets of functions computable by \( \text{span}_n G \), possibly with \( n \ll m \). We first state basic properties of the functional \( \mathcal{E}_{\gamma, K} \) (for their proofs see [16]).

**Proposition 4.2** Let \( m, d \) be positive integers, \( \Omega \) be a compact subset of \( \mathcal{R}^d \), \( K : \Omega \times \Omega \rightarrow \mathcal{R} \) be a Mercer kernel, \( \gamma > 0 \) and \( \{ (x_1, y_1), \ldots, (x_m, y_m) \} \subset (\Omega \times \mathcal{R})^m \). Then

(i) \( \mathcal{E}_{\gamma, K} \) is strictly uniformly convex on \( H_K(\Omega) \) with modulus of convexity \( \delta(t) = t^2 \).

(ii) at every \( f \in H_K(\Omega) \), \( \mathcal{E}_{\gamma, K} \) is continuous with a modulus of continuity bounded from above by

\[
\alpha(t) = a_2 t^2 + a_1 t,
\]
where $a_1 = 2(m||f||_K c_K + mb\sqrt{c_K} + \gamma||f||_K)$, $a_2 = m c_K + \gamma$ and $b = \max\{|y_i|: i = 1, \ldots, m\}$.

(iii) for $M \subset \mathcal{H}_K(\Omega)$ closed convex and bounded or $M = \mathcal{H}_K(\Omega)$, the problem $(M, \mathcal{E}_{\gamma,K})$ has a unique argminima $g^\ast$.

(iv) for $M \subset \mathcal{H}_K(\Omega)$ closed convex and bounded or $M = \mathcal{H}_K(\Omega)$, any $g^\ast \in \text{argmin}(M, \mathcal{E}_{\gamma,K})$ and $f \in M$,

$$||f - g^\ast||_K^2 \leq |\mathcal{E}_{\gamma,K}(f) - \mathcal{E}_{\gamma,K}(g^\ast)|.$$

Combining Theorem 3.2 with Proposition 4.2 we derive the following estimate of rates of convergence of approximate minima achievable over sets $\text{span}_n G$ to the global minimum over the whole space (for the details of the proof see [15]).

**Theorem 4.3** Let $\Omega \subset \mathcal{R}^d$ be compact, $K : \Omega \times \Omega \rightarrow \mathcal{R}$ be a Mercer kernel, $s_K = \sup_{x \in \Omega} \sqrt{K(x,x)}$.

$(\mathcal{H}_K(\Omega), ||.||_K)$ be the RKHS defined by $K$, $\mathcal{E} : \mathcal{H}_K(\Omega) \rightarrow \mathcal{R}$ be the empirical error $\mathcal{E}(f) = \frac{1}{m} \sum_{i=1}^{m} |f(x_i) - y_i|^2$, where $(x_1, \ldots, x_m) \in \Omega^m$, $(y_1, \ldots, y_m) \in \mathcal{R}^m$, $\gamma > 0$, $g^\ast(x) = \sum_{i=1}^{m} a_i K(x,x_i)$ be the unique minimum of $(\mathcal{H}_K(\Omega), \mathcal{E}, K)$ given by the Representer Theorem, $\{\varepsilon_n\}$ is a sequence of positive reals such that $\lim_{n \rightarrow \infty} \varepsilon_n = 0$ and $\{g_n\}$ be a sequence of $\varepsilon_n$-minimisers of problems $\text{span}_n G_K, \mathcal{E}_{\gamma,K}$. Then for every positive integer $n$ the following hold:

(i) $\inf_{g \in \text{span}_n G_K} \mathcal{E}_{\gamma,K}(g) - \mathcal{E}_{\gamma,K}(g^\ast) \leq \alpha \left( \sqrt{\frac{\text{tr}(g^\ast) ||g^\ast||_K^2 - ||g^\ast||_K^2}{n}} \right) + \varepsilon_n$;

(ii) if $||g^\ast||_G < \infty$, then $\{g_n\}$ is a $\mathcal{E}_{\gamma,K}$-minimising sequence over $\mathcal{H}_K(\Omega)$ and $\mathcal{E}_{\gamma,K}(g_n) - \mathcal{E}_{\gamma,K}(g^\ast) \leq \alpha \sqrt{\frac{\text{tr}(g^\ast) ||g^\ast||_K^2 - ||g^\ast||_K^2}{n}} + \varepsilon_n$;

(iii) $\|g_n - g^\ast\|_K^2 \leq \alpha \left( \sqrt{\frac{\text{tr}(g^\ast) ||g^\ast||_K^2 - ||g^\ast||_K^2}{n}} + \varepsilon_n \right)$;

(iv) $\|g_n - g^\ast\|_K^2 \leq \sqrt{c_K} \left( \alpha \sqrt{\frac{\text{tr}(g^\ast) ||g^\ast||_K^2 - ||g^\ast||_K^2}{n}} + \varepsilon_n \right)$,

where $c_K = \sup_{x,y \in \Omega} |K(x,y)|$, $\alpha(t) = a_2 t^2 + a_1 t$, $a_1 = 2(m||f||_K c_K + mb\sqrt{c_K} + \gamma||f||_K)$, $a_2 = m c_K + \gamma$ and $b = \max\{|y_i|: i = 1, \ldots, m\}$.

For any Mercer kernel $K$ (in particular for the Gaussian kernel), Theorem 4.3 can be applied to estimate rates of approximate optimization of the regularized empirical error functional $\mathcal{E}_{\gamma,K}$ over sets $\text{span}_n G_K$, corresponding to networks with $r$ hidden units. As the result is not only asymptotic, it holds for any $n$, in particular for $n$ much smaller that the size $m$ of the data sample. The dominant term in these estimates is $\frac{||(g^\ast)||_G^2}{n}$. Thus if the global minimum $g^\ast$ described by the Representer Theorem is in a ball of $G_K$-variation of a fixed radius $r$, then the estimates converge rather quickly (as $r/n$).

It follows from the definition of $G_K$-variation and Theorem 4.1 that $||g^\ast||_G \leq \sum_{i=1}^{m} |s_i|$, where $\alpha = (\alpha_1, \ldots, \alpha_m)$ is the unique solution of the well-posed linear system $(K(x) + \gamma m I)\alpha = y$, $I$ is the $m \times m$ identity matrix, and $K(x)$ is the $m \times m$ matrix defined as $K(x)_{i,j} = K(x_i, x_j)$.

**References**


