MATHEMATICS OF OPERATIONS RESEARCH
Vol. 33, No. 3, August 2008, pp. 747–756
ISSN 0364-765X | EISSN 1526-5471 | 08 | 3303 | 0747

DOI 10.1287/moor.1080.0317

Approximate Minimization of the Regularized Expected Error over Kernel Models

Věra Kůrková
Institute of Computer Science, Academy of Sciences of the Czech Republic,
Prague 8, Czech Republic, vera@cs.cas.cz, http://www.cs.cas.cz/~vera/

Marcello Sanguineti
Department of Communications, Computer, and System Sciences (DIST), University of Genoa,
16145 Genova, Italy, marcello@dist.unige.it, http://www.dist.unige.it/msanguineti/

Learning from data under constraints on model complexity is studied in terms of rates of approximate minimization of the regularized expected error functional. For kernel models with an increasing number n of kernel functions, upper bounds on such rates are derived. The bounds are of the form $a/n + b/\sqrt{n}$, where a and b depend on the regularization parameter and on properties of the kernel, and of the probability measure defining the expected error. As a special case, estimates of rates of approximate minimization of the regularized empirical error are derived.

Key words: suboptimal solutions; expected error; convex functionals; kernel methods; model complexity; rates of convergence
MSC2000 subject classification: Primary: 68Q32, 58E50, 41A25; secondary: 90C48, 90C99, 46N10, 47B34
OR/MS subject classification: Primary: computers/computer science: artificial intelligence; secondary: programming: infinite dimensional
History: Received January 10, 2007; revised June 30, 2007 and February 15, 2008.

1. Introduction. Finding a function that fits given data amounts to minimizing an error functional. To improve the stability of solutions of such an optimization task, various regularization methods can be used (Bertero [8]). In Tikhonov regularization, a *stabilizer* is added that penalizes undesired properties of the solution. A commonly used stabilizer is the square of the norm on a *reproducing kernel Hilbert space* (*RKHS*). Squares of norms on Hilbert spaces are strictly convex (which guarantees uniqueness of solutions) and include the class of high-frequency filters (Girosi [20]). RHKS' were formally defined by Aronszajn [4], but their theory is based on previous works by Mercer, Schönberg and others (see, e.g., Schönberg [37]).

RKHS' were introduced into data analysis by Parzen and Wahba in the 1960s, as a framework for data smoothing by spline models (see, e.g., Wahba [40]). Since the 1980s, RKHS' have been extensively used in statistics (see, e.g., Berlinet and Thomas-Agnan [7] and the references therein). Kernel spaces also play an important role in machine learning and neurocomputing. A classification algorithm modifying the input-space geometry by using a kernel (called *potential function*) was proposed by Aizerman et al. [2] in 1960s, and later extended by Cortes and Vapnik [14] to the concept of support vector machine (see, e.g., Schölkopf and Smola [36]). Kernels were introduced into neurocomputing by Poggio and Girosi [34] as *kernel networks*, an extension to radial-basis-function networks. For a brief historical outline of RKHS', see Parzen [32].

Most neural-network learning algorithms decrease the *empirical error functional*. Among them, the ones designed to improve generalization decrease this error together with an additional term penalizing undesired solutions. For example, in *weight-decay regularization*, a term bounding from above the kernel norm of the solution is added as a stabilizer (Burger and Neubauer [12]).

A theoretical analysis of regularization in neural-network learning was made by Girosi et al. [21] and Poggio and Girosi [34]. Girosi [20] realized that the high-frequency filters used as stabilizers belong to the class of squares of norms on RKHS'. Thus the *Representer Theorem* (first derived by Kimeldorf and Wahba [25, 26]) describing solutions of regularization problems in RKHS' can also be applied in learning theory to model generalization (see, e.g., the synthesis paper Cucker and Smale [15] or Poggio and Smale [35] and the references therein). The optimal solution characterized by this theorem can be interpreted as an input-output function of a kernel network approximating the data as well as satisfying some global smoothness condition. In particular, for convolution kernels, it can be interpreted as a function computable by a radial-basis-function network. However, the network computing the optimal solution has as many units as the size of the data sample. So, for large data samples, the computation of network parameters requires the numerical solution of large systems of linear equations. This may limit practical applications of algorithms based on this theorem. By contrast, standard neural-network learning algorithms operate on networks with a number of units much smaller than the size of the data sample (see, e.g., Fine [18]).

In this paper, we compare suboptimal solutions achievable by using kernel networks having an a priori bounded number of units with theoretically optimal solutions described by the Representer Theorem. We estimate the speed of convergence of infima of the regularized expected error functional over kernel models with an increasing number of terms to its global minimum described by the Representer Theorem. By applying the results to the discrete case we obtain estimates for empirical error functionals defined by samples of data.

The upper bounds are formulated in terms of the regularization parameter, of the moduli of continuity and convexity of the regularized expected error functional, and of two norms of the function at which the global minimum is achieved: the norm on the ambient RKHS and a certain variational norm tailored to the type of kernel. The variational norm is a critical term for the speed of convergence. We estimate its magnitude by using a suitable integral representation of the optimal solution. Thus we obtain upper bounds of the form $a/n+b/\sqrt{n}$, where n is the number of kernel computational units and a and b depend on the regularization parameter and on some properties of the kernel and the regression function.

The paper is organized as follows. In §2, the notation is defined, and the minimum point of the regularized expected error functional is expressed in terms suitable for estimation of its variational norm. In §3, upper bounds are derived on infima of the regularized expected error over kernel models formed by linear combinations of at most n kernel computational units. In §4, these estimates are applied to the discrete case (corresponding to the empirical error functional) and compared with our earlier estimates from Kůrková and Sanguineti [29]. Section 5 provides a brief summary discussion. For the reader's convenience, we include appendices containing definitions concerning RKHS', convex functionals, and the proof of an auxiliary result on moduli of continuity and convexity of the regularized expected error.

2. Minimization of the regularized expected error. Let X be a *compact* subset of $\mathbb{R}^d$, Y a *bounded* subset of $\mathbb{R}$, and ρ a nondegenerate (i.e., all nonempty open sets have positive measures) *probability measure* on $X\times Y$. In mathematical learning theory and statistics (see, e.g., Cucker and Smale [15], Vapnik [39]), learning from data has been modeled as minimization of the *expected error* (also called *expected risk* or *theoretical error*).

$$\mathscr{E}_\rho(f)=\int_{X\times Y}(f(x)-y)^2\,d\rho, \tag{1}$$

over a suitable set of admissible solutions.

The probability measure ρ induces on X the *marginal probability measure* ρ_X, defined for all subsets $S\subseteq X$ as

$$\rho_X(S)=\rho(\pi^{-1}(X)),$$

where $\pi\colon X\times Y\to X$ denotes the projection from $X\times Y$ to X.

The minimum of $\mathscr{E}_\rho$ over the space $\mathscr{L}^2_{\rho_X}(X)$ of ρ_X-square-integrable functions on X is achieved at the *regression function* f_ρ, defined for every $x\in X$ as

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

where $\rho(y\mid x)$ is the *conditional* (*w.r.t.* x) *probability measure* on Y. It is easy to see that $\mathscr{E}_\rho(f)=\|f-f_\rho\|^2_{\mathscr{L}^2_{\rho_X}}+\mathscr{E}_\rho(f_\rho)$ for all $f\in\mathscr{L}^2_{\rho_X}(X)$ (see, e.g., Cucker and Smale [15]). We denote

$$s_Y=\sup\{|y|\mid (y\in Y)(\exists x\in X,\rho(y\mid x)\neq 0)\}.$$

In *Tikhonov's regularization* (Tikhonov [38]), the functional $\mathscr{E}_\rho$ is replaced with the functional $\mathscr{E}_\rho+\gamma\Phi$, where Φ is called a *stabilizer* and γ a *regularization parameter*. Suitable stabilizers are squares of norms on Hilbert spaces. They are strictly convex and thus guarantee the uniqueness of the minimum points of the corresponding regularized functionals. Among such stabilizers, a special role is played by norms on reproducing kernel Hilbert spaces (RKHS') (see Appendix A). They can penalize solutions with undesired high-frequency oscillations, since high-frequency filters of the form

$$\Phi(f)=(2\pi)^{-d}\int_{\mathbb{R}^d}\frac{\hat{f}(s)^2}{\hat{k}(s)}\,ds$$

with $\hat{k}>0$ (where $\hat{k}$ denotes the Fourier transform of k) are squares of norms on RKHS' generated by convolution kernels (Girosi [20]) (see Appendix A).

The expected error regularized by such a stabilizer is the functional $\mathscr{E}_{\rho,\gamma,K}$ on the RKHS induced by the kernel K, defined as

$$\mathscr{E}_{\rho,\gamma,K}(f) = \mathscr{E}_\rho(f) + \gamma\|f\|_K^2 = \int_{X\times Y}(f(x)-y)^2 d\rho + \gamma\|f\|_K^2.$$

Let $L_K\colon \mathscr{L}^2_{\rho_X}(X) \to \mathscr{L}^2_{\rho_X}(X)$ denote the integral operator defined for every $f \in \mathscr{L}^2_{\rho_X}(X)$ as

$$L_K(f)(u) = \int_X f(v)K(u,v)\,d\rho_X.$$

Recall that, for K continuous and X compact, L_K is a compact operator (Friedman [19, pp. 238, 188]). For K symmetric, L_K is self-adjoint, and for K positive-definite, L_K is positive (Friedman [19, pp. 237, 233]). So for a continuous, symmetric, and positive-definite kernel, L_K has an orthonormal family of eigenfunctions $\{\phi_i\}$ with positive eigenvalues $\{\lambda_i\}$. The sequence $\{\lambda_i\}$, ordered in a nonincreasing way, is either finite (when K is degenerate) or convergent to zero. By the Mercer Theorem (see, e.g., Cucker and Smale [15, p. 34]), $K(u,v) = \sum_{i=1}^\infty \lambda_i\phi_i(u)\phi_i(v)$, where the convergence is absolute for all $u, v \in X$ and uniform on $X \times X$ and $\sum_{i=1}^\infty \lambda_i$ is convergent (Cucker and Smale [15, p. 36]). For a function $f = \sum_{i=1}^\infty c_i\,\phi_i \in \mathscr{L}^2_{\rho_X}(X)$, we denote

$$\|f\|_{l^1} = \sum_{i=1}^\infty |c_i|.$$

When K is continuous and X is compact, the regularized expected error functional $\mathscr{E}_{\rho,\gamma,K}$ achieves its minimum over the RKHS $\mathscr{H}_K(X)$ defined by K. The minimum point satisfies

$$f_\gamma = (I + \gamma L_K^{-1})^{-1}(f_\rho) \tag{2}$$

(for the proof, see, e.g., p. 42 of the synthesis work (Cucker and Smale [15])). Note that for a nondegenerate kernel K, the functional L_K^{-1} is defined only on a proper subspace of $\mathscr{L}^2_{\rho_X}(X)$.

The optimal solution f_γ can be alternatively expressed as an image of the regression function f_ρ under an integral operator. Such an expression (which will be needed in Lemma 3.1) uses an operator $K_\gamma\colon X \times X \to \mathbb{R}$ defined as

$$K_\gamma(u,v) = \sum_{i=1}^\infty \frac{\lambda_i}{\lambda_i+\gamma}\phi_i(u)\phi_i(v). \tag{3}$$

The kernel K_γ is positive definite, since the class of symmetric positive definite kernels contains all product kernels and is closed with respect to linear combinations with positive coefficients as well as to pointwise limits (Berg et al. [6]). The following proposition summarizes properties of K_γ and of the integral operator L_{K_γ} defined for every $f \in \mathscr{L}^2_{\rho_X}(X)$ as

$$L_{K_\gamma}(f)(u) = \int_X f(v)K_\gamma(u,v)\,d\rho_X.$$

Proposition 2.1. *Let d be a positive integer, $X \subset \mathbb{R}^d$ compact, $Y \subset \mathbb{R}$ bounded, ρ a nondegenerate probability measure on $X \times Y$, $K\colon X \times X \to \mathbb{R}$ a continuous, positive definite kernel, and $\gamma > 0$. Then*

(i) *K_γ is a continuous positive definite kernel;*
(ii) *$L_{K_\gamma}(\phi_i) = (\lambda_i/(\lambda_i+\gamma))\phi_i$;*
(iii) *$L_{K_\gamma}\colon (\mathscr{L}^2_{\rho_X}(X), \|.\|_{\mathscr{L}^2_{\rho_X}}) \to (\mathscr{H}_K(X), \|.\|_K)$ is a compact operator;*
(iv) *$\mathscr{H}_K(X) = \mathscr{H}_{K_\gamma}(X)$ and $\|.\|_{K_\gamma}^2 = \|.\|_{\mathscr{L}^2_{\rho_X}}^2 + \gamma\|.\|_K^2$.*

Proof. Properties (i) and (ii) follow from the definitions of K_γ and L_{K_γ}.

(iii) Let $f = \sum_{i=1}^\infty c_i\phi_i \in \mathscr{L}^2_{\rho_X}(X)$, so $\sum_{i=1}^\infty c_i^2 < \infty$. Then, $L_{K_\gamma}(f) = \sum_{i=1}^\infty c_i(\lambda_i/(\lambda_i+\gamma))\phi_i$ and $\sum_{i=1}^\infty (c_i\,\lambda_i)^2 \cdot (\lambda_i+\gamma)^{-2}\lambda_i^{-1} \le (\lambda_1/\gamma^2)\sum_{i=1}^\infty c_i^2 < \infty$. As $\mathscr{H}_K(X)$ is a linear subspace of $\mathscr{L}^2_{\rho_X}(X)$ formed by those $f = \sum_{i=1}^\infty c_i\,\phi_i \in \mathscr{L}^2_{\rho_X}(X)$, for which $\sum_{i=1}^\infty c_i^2/\lambda_i < \infty$ (Cucker and Smale [15]), we get $L_{K_\gamma}(f) \in \mathscr{H}_K(X)$.

(iv) By the definition of K_γ and by (A2) from Appendix A, we get $\|f\|_{K_\gamma}^2 = \sum_{i=1}^\infty c_i^2(\lambda_i+\gamma)/\lambda_i = \sum_{i=1}^\infty c_i^2 + \gamma\|f\|_K^2 = \|f\|_{\mathscr{L}^2_{\rho_X}}^2 + \gamma\|f\|_K^2$. So $\|f\|_{K_\gamma} < \infty$ if and only if $\|f\|_K < \infty$. □

Using Proposition 2.1, we reformulate the characterization (2) of the minimum point of the regularized expected error $\mathscr{E}_{\rho,\gamma,K}$ over $\mathscr{H}_K(X)$ in terms of L_{K_γ}.

Theorem 2.1. *Let d be a positive integer, $X \subset \mathbb{R}^d$ compact, $Y \subset \mathbb{R}$ bounded, ρ a nondegenerate probability measure on $X \times Y$, and $K\colon X \times X \to \mathbb{R}$ a continuous kernel. Then in $\mathscr{H}_K(X)$, $\mathscr{E}_{\rho,\gamma,K}$ has a unique minimum point*

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

In Lemma 3.1, we shall exploit this representation to estimate the norm $\|f_\gamma\|_K$.

3. Rates of approximate minimization of the regularized expected error. Although the regression function f_ρ minimizing $\mathscr{E}_{\rho,K,\gamma}$ may not be in $\mathscr{H}_K(X)$, its image under the operator L_{K_γ} is present because L_{K_γ} maps the whole $\mathscr{L}^2_{\rho_X}(X)$ into its linear subspace $\mathscr{H}_K(X)$. So $L_{K_\gamma}(f_\rho)$ can be approximated in K-norm (hence also uniformly: See (A1) in Appendix A) by a sequence of linear combinations of functions from the set

$$G_K = \{K_x \mid x \in X\},$$

where $K_x\colon X \to \mathbb{R}$ is defined as $K_x(y) = K(x,y)$. However, the sequence of the numbers of terms in the linear combinations might be unbounded, or even when it is bounded, the bound may be too large.

Many neural-network learning algorithms are based on the minimization of a regularized discretized version of the expected error, called *empirical error*. For example, in algorithms with the so-called *weight-decay regularization* (see, e.g., Burger and Neubauer [12]), the hypothesis set is made up of input-output functions of the form $f = \sum_{i=1}^n w_i K(x_i, .)$ and the functional

$$\Phi(f) = \sum_{i=1}^n |w_i|$$

is used as a stabilizer. Since $\|f\|_K \le \sum_{i=1}^n |w_i| K(x_i,x_i) \le s_K \sum_{i=1}^n |w_i|$, such algorithms decrease $\|f\|_K^2$.

To compare the theoretically optimal solution f_γ with suboptimal solutions achievable by using computational models with a reasonably small number of terms, we estimate the speed of convergence of infima of $\mathscr{E}_{\rho,K,\gamma}$ over the sets

$$\operatorname{span}_n G_K = \left\{\sum_{i=1}^n w_i K_{x_i} \mid w_i \in \mathbb{R}, x_i \in X\right\},$$

with increasing n. For positive definite kernels (see Appendix A), the sets G_K are linearly independent, thus the sets $\operatorname{span}_n G_K$ are not convex. Thus, results from convex optimization theory (Borwein and Lewis [11]) cannot be applied and one has to consider infima instead of minima of $\mathscr{E}_{\rho,\gamma,K}$ over $\operatorname{span}_n G_K$.

To estimate the rate of convergence of these infima, we exploit an upper bound from Kůrková and Sanguineti [29, Theorem 4.2] on convergence of approximate infima of continuous functionals; this bound is based on tools from nonlinear approximation theory. It depends on moduli of continuity and convexity (for their definitions, see Appendix B) of the functional to be minimized and on two norms of its minimum point (in our case, f_γ) over the whole space: the norm $\|.\|$ on the ambient space (in our case, $\|.\|_K$) and a certain norm determined by a set G, called *G-variation*. It is defined for a bounded subset of a normed linear space $(X, \|.\|)$ as the Minkowski functional of the closure (with respect to the topology induced by $\|.\|$) of the symmetric convex hull of G, i.e.,

$$\|f\|_G = \inf\left\{c > 0 \;\middle|\; \frac{f}{c} \in \operatorname{cl}\operatorname{conv}(G \cup -G)\right\},$$

(see, e.g, Kůrková and Sanguineti [27] for the properties of $\|.\|_G$). In our case, we use the variation with respect to G_K, i.e., the norm $\|.\|_{G_K}$.

We first estimate moduli of continuity and convexity of $\mathscr{E}_{\rho,\gamma,K}$ on $\mathscr{H}_K(X)$ (for the proof, see Appendix C).

PROPOSITION 3.1. *Let d, n be positive integers, $X \subset \mathbb{R}^d$ compact, $Y \subset \mathbb{R}$ bounded, ρ a nondegenerate probability measure on $X \times Y$, $K\colon X \times X \to \mathbb{R}$ a continuous kernel, $s_K = \sup_{x\in X}\sqrt{K(x,x)}$, and $\gamma > 0$. Then*

(i) *$\mathscr{E}_{\rho,\gamma,K}$ is uniformly convex on $\mathscr{H}_K(X)$ with a modulus of convexity $\delta(t) \le \gamma t^2$;*

(ii) *$\mathscr{E}_{\rho,\gamma,K}$ is continuous on $(\mathscr{H}_K(X), \|.\|_K)$ and for every $g \in \mathscr{H}_K(X)$, the modulus of continuity of $\mathscr{E}_{\rho,\gamma,K}$ at g satisfies $\omega_g(t) \le a_2 t^2 + a_1 t$, where $a_2 = s_K^2 + \gamma$ and $a_1 = 2(\|g\|_K(s_K^2+\gamma) + s_K s_Y)$;*

(iii) *there exists a unique minimum point f_γ of $\mathscr{E}_{\rho,\gamma,K}$ over $\mathscr{H}_K(X)$ and, for every $f \in \mathscr{H}_K(X)$, $\|f - f_\gamma\|_K^2 \le (\mathscr{E}_{\rho,\gamma,K}(f) - \mathscr{E}_{\rho,\gamma,K}(f_\gamma))/\gamma$.*

The regularized expected error $\mathscr{E}_{\rho,\gamma,K}$ is continuous and uniformly convex on $\mathscr{H}_K(X)$. Both its moduli of convexity and continuity are bounded from above by quadratic functions; the function bounding its modulus of continuity ω_g at some $g \in \mathscr{H}_K(X)$ depends on the norm $\|g\|_K$. In particular, for $g = f_\gamma$ we get

$$\omega_{f_\gamma}(t) \le \alpha(t) = (s_K^2+\gamma)t^2 + 2(\|f_\gamma\|_K(s_K^2+\gamma) + s_K s_Y)t. \tag{4}$$

Combining Kůrková and Sanguineti [29, Theorem 4.2] with Proposition 3.1, we obtain the following upper bound on rates of minimization of $\mathscr{E}_{\rho,\gamma,K}$ over $\operatorname{span}_n G_K$. For a functional Φ, a set $\mathscr{H}$, and $\varepsilon > 0$, we denote by $\arg\min_\varepsilon(\mathscr{H}, \Phi)$ the set of all *ε-near minimum points*, i.e.,

$$\arg\min_\varepsilon(\mathscr{H}, \Phi) = \left\{f \in \mathscr{H} \mid \Phi(f) < \inf_{f\in\mathscr{H}} \Phi(f) + \varepsilon\right\}.$$

THEOREM 3.1. *Let d, n be positive integers, $X \subset \mathbb{R}^d$ compact, $Y \subset \mathbb{R}$ bounded, ρ a nondegenerate probability measure on $X \times Y$, $K\colon X \times X \to \mathbb{R}$ a continuous kernel, $s_K = \sup_{x\in X}\sqrt{K(x,x)}$, $\gamma > 0$, f_γ the minimum point of $\mathscr{E}_{\rho,\gamma,K}$ over $\mathscr{H}_K(X)$, $\{\varepsilon_n\} \subset \mathbb{R}_+$, and $f_n \in \arg\min_{\varepsilon_n}(\operatorname{span}_n G_K, \mathscr{E}_{\rho,\gamma,K})$. Then,*

(i) $\inf_{f\in\operatorname{span}_n G_K} \mathscr{E}_{\rho,\gamma,K}(f) - \mathscr{E}_{\rho,\gamma,K}(f_\gamma) \le \alpha(\sqrt{((s_K\|f_\gamma\|_{G_K})^2 - \|f_\gamma\|_K^2)/n})$;

(ii) $\|f_n - f_\gamma\|_K^2 \le (1/\gamma)(\alpha(\sqrt{((s_K\|f_\gamma\|_{G_K})^2 - \|f_\gamma\|_K^2)/n}) + \varepsilon_n)$;

(iii) $\sup_{x\in X}|f_n(x) - f_\gamma(x)|^2 \le (s_K^2/\gamma)(\alpha(\sqrt{((s_K\|f_\gamma\|_{G_K})^2 - \|f_\gamma\|_K^2)/n}) + \varepsilon_n)$.

The next lemma shows that, with increasing γ, the K-norm of the regularized solution f_γ decreases at least as fast as $s_Y/2\sqrt{\gamma}$.

LEMMA 3.1. *Let d be a positive integer, $X \subset \mathbb{R}^d$ compact, $Y \subset \mathbb{R}$ bounded, and $\gamma > 0$. Then, for every nondegenerate probability measure ρ on $X \times Y$ and every continuous positive semidefinite kernel $K\colon X \times X \to \mathbb{R}$, the unique minimum point f_γ of $\mathscr{E}_{\rho,\gamma,K}$ over $\mathscr{H}_K(X)$ satisfies*

$$\|f_\gamma\|_K \le \frac{s_Y}{2\sqrt{\gamma}}.$$

PROOF. Let $f_\rho = \sum_{i=1}^\infty c_i\phi_i$. Then, by Theorem 2.1, $f_\gamma = \sum_{i=1}^\infty (c_i\lambda_i/(\lambda_i+\gamma))\phi_i$ and so $\|f_\gamma\|_K^2 = \sum_{i=1}^\infty (c_i\lambda_i/(\lambda_i+\gamma))^2(1/\lambda_i) = \sum_{i=1}^\infty c_i^2\lambda_i/(\lambda_i+\gamma)^2$. It is easy to check that, for all i, $\lambda_i/(\lambda_i+\gamma)^2 \le 1/4\gamma$, and so $\|f_\gamma\|_K^2 = \sum_{i=1}^\infty c_i^2\lambda_i/(\lambda_i+\gamma)^2 \le (1/4\gamma)\sum_{i=1}^\infty c_i^2 = (1/4\gamma)\|f_\rho\|_{\mathscr{L}^2_{\rho_X}}^2 \le s_Y^2/4\gamma$. □

Lemma 3.1 and Equation (4) imply the following estimate of the modulus of continuity ω_{f_γ} of $\mathscr{E}_{\rho,\gamma,K}$ at f_γ formulated in terms of s_K, s_Y, and the regularization parameter γ:

$$\omega_{f_\gamma}(t) \le (s_K^2+\gamma)t^2 + s_Y\left(\frac{s_K^2}{\sqrt{\gamma}} + 2s_K + \sqrt{\gamma}\right)t. \tag{5}$$

To estimate $\|f_\gamma\|_{G_K}$ we use an extension of Kůrková and Kainen [30, Theorem 2.2] stating that if f can be expressed as

$$f(v) = \int_X h(u)K(u,v)\,d\lambda(u) = L_K(h), \tag{6}$$

with $K\colon X \times X \to \mathbb{R}$ continuous, X compact, and $h \in \mathscr{L}^1_\lambda(X)$, where λ is the Lebesgue measure, then

$$\|f\|_{G_K} \le \|h\|_{\mathscr{L}^1_\lambda}. \tag{7}$$

In the next proposition, using the integral representation $f_\gamma = L_{K_\gamma}(f_\rho)$, we express f_γ as an element of the range of L_K.

PROPOSITION 3.2. *Let d be a positive integer, $X \subset \mathbb{R}^d$ compact, $Y \subset \mathbb{R}$ bounded, $K\colon X \times X \to \mathbb{R}$ a continuous kernel, $\{\lambda_i\}$ the sequence of the eigenvalues of L_K, $\gamma > 0$, and f_γ be the unique minimum point of $\mathscr{E}_{\rho,\gamma,K}$ over $\mathscr{H}_K(X)$. Then, for every nondegenerate probability measure ρ on $X \times Y$,*

(i) *there exists $\bar{f}_\rho \in \mathscr{L}^2_{\rho_X}(X)$ such that $f_\gamma = L_K(\bar{f}_\rho)$;*

(ii) $\|f_\gamma\|_{G_K} \le \sum_{i=1}^\infty |c_i|1/(\lambda_i+\gamma)$, *where $f_\rho = \sum_{i=1}^\infty c_i\phi_i$;*

(iii) $\|f_\gamma\|_{G_K} \le (1/\gamma)\|f_\rho\|_{l^1}$.

PROOF. (i) Let $f_\rho = \sum_{i=1}^\infty c_i\phi_i$ be the representation of f_ρ as an element of $\mathscr{L}^2_{\rho_X}(X)$. Define $\bar{f}_\rho = \sum_{i=1}^\infty c_i \cdot (\lambda_i+\gamma)^{-1}\phi_i$. By Theorem 2.1, $f_\gamma = L_{K_\gamma}(f_\rho) = \sum_{i=1}^\infty c_i(\lambda_i/(\lambda_i+\gamma))\phi_i = \sum_{i=1}^\infty \lambda_i(c_i/(\lambda_i+\gamma)\phi_i) = L_K(\bar{f}_\rho)$.

(ii) By (i), $f_\gamma(u) = \int_X \bar{f}_\rho(v)K(u,v)\,d\rho_X$. Inspection of the proof of Kůrková and Kainen [30, Theorem 2.2] shows that Equation (6) implies Equation (7) also when λ is replaced with any nondegenerate probability measure. Thus, by continuity of K, we get $\|f_\gamma\|_{G_K} \le \|\bar{f}_\rho\|_{\mathscr{L}^1_{\rho_X}} \le \sum_{i=1}^\infty |c_i| \cdot 1/(\lambda_i+\gamma)$.

(iii) For all i $1/(\lambda_i+\gamma) \le 1/\gamma$, so by (ii) we have $\|f_\gamma\|_{G_K} \le (1/\gamma)\sum_{i=1}^\infty |c_i| = (1/\gamma)\|f_\rho\|_{l^1}$. □

Proposition 3.2(iii) gives an estimate of $\|f_\gamma\|_{G_K}$ in terms of the regularization parameter γ and the l^1-norm of the regression function f_ρ. Combining this estimate with Theorem 3.1 and the upper bound Equation (5) on the modulus of continuity ω_{f_γ}, we get the next upper bound on the speed of convergence of approximate infima of the regularized expected error $\mathscr{E}_{\rho,\gamma,K}$ over $\operatorname{span}_n G_K$.

THEOREM 3.2. *Let d, n be positive integers, $X \subset \mathbb{R}^d$ compact, $Y \subset \mathbb{R}$ bounded, $K\colon X \times X \to \mathbb{R}$ a continuous kernel, $s_K = \sup_{x\in X}\sqrt{K(x,x)}$, $\gamma > 0$, f_γ the unique minimum point of $\mathscr{E}_{\rho,\gamma,K}$ over $\mathscr{H}_K(X)$, $\{\varepsilon_n\} \subset \mathbb{R}_+$, $f_n \in \arg\min_{\varepsilon_n}(\operatorname{span}_n G_K, \mathscr{E}_{\rho,\gamma,K})$, $b = (s_K^2+\gamma)(s_K/\gamma)^2$, and $c = (s_K s_Y/\gamma)(s_K^2/\sqrt{\gamma} + 2s_K + \sqrt{\gamma})$. Then, for every*

nondegenerate probability measure ρ on $X \times Y$, the following estimates hold:

(i) $\inf_{f \in \mathrm{span}_n G_K} \mathscr{E}_{\rho,\gamma,K}(f) - \mathscr{E}_{\rho,\gamma,K}(f_\gamma) \le (b/n)\|f_\rho\|_{l^1}^2 + (c/\sqrt{n})\|f_\rho\|_{l^1}$;

(ii) $\|f_n - f_\gamma\|_K^2 \le (1/\gamma)((b/n)\|f_\rho\|_{l^1}^2 + (c/\sqrt{n})\|f_\rho\|_{l^1} + \varepsilon_n)$;

(iii) $\sup_{x\in X} |f_n(x) - f_\gamma(x)|^2 \le (s_K^2/\gamma)((b/n)\|f_\rho\|_{l^1}^2 + (c/\sqrt{n})\|f_\rho\|_{l^1} + \varepsilon_n)$.

Proof. Theorem 3.1 implies

$$\inf_{f \in \mathrm{span}_n G_K} \mathscr{E}_{\rho,\gamma,K}(f) - \mathscr{E}_{\rho,\gamma,K}(f_\gamma) \le \alpha\left(\frac{s_K\|f_\gamma\|_{G_K}}{\sqrt{n}}\right), \tag{8}$$

$$\|f_n - f_\gamma\|_K^2 \le \frac{1}{\gamma}\left(\alpha\left(\frac{s_K\|f_\gamma\|_{G_K}}{\sqrt{n}}\right) + \varepsilon_n\right), \tag{9}$$

and

$$\sup_{x\in X} |f_n(x) - f_\gamma(x)|^2 \le \frac{s_K^2}{\gamma}\left(\alpha\left(\frac{s_K\|f_\gamma\|_{G_K}}{\sqrt{n}}\right) + \varepsilon_n\right). \tag{10}$$

By the estimate Equation (5) of the modulus of continuity ω_{f_γ} and the estimate of $\|f_\gamma\|_{G_K}$ from Proposition 3.2(iii), we get

$$\begin{aligned}\alpha\left(\frac{s_K\|f_\gamma\|_{G_K}}{\sqrt{n}}\right) &\le (s_K^2+\gamma)\left(\frac{s_K\|f_\rho\|_{l^1}}{\gamma\sqrt{n}}\right)^2 + s_Y\left(\frac{s_K^2}{\sqrt{\gamma}} + 2s_K + \sqrt{\gamma}\right)\frac{s_K\|f_\rho\|_{l^1}}{\gamma\sqrt{n}}\\ &= \frac{s_K^2+\gamma}{n}\left(\frac{s_K}{\gamma}\right)^2\|f_\rho\|_{l^1}^2 + \frac{s_K s_Y}{\gamma\sqrt{n}}\left(\frac{s_K^2}{\sqrt{\gamma}} + 2s_K + \sqrt{\gamma}\right)\|f_\rho\|_{l^1}.\end{aligned} \tag{11}$$

The bounds (i), (ii), and (iii) follow from Equation (11) combined with Equations (8), (9), and (10), respectively. □

By Theorem 3.2, the larger γ, the faster the convergence. For a fixed γ, the estimates imply fast convergence for probability measures ρ with small l^1-norms of the regression functions f_ρ.

4. Approximate minimization of the regularized empirical error. Let d be a positive integer, $X \subset \mathbb{R}^d$, $Y \subset \mathbb{R}$, and $z = \{(x_i, y_i) \subset X \times Y | i = 1, \dots, m\}$ a sample of data. By $\mathscr{E}_z$ we denote the *empirical error functional*, defined as

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

The empirical error functional is a special case of the expected error; it is determined by the discrete probability measure ρ on $X \times Y$ satisfying $\rho(x_i, y_i) = 1/m$ and otherwise $\rho(x, y) = 0$. For such ρ, $\mathscr{E}_z = \mathscr{E}_\rho$. It is easy to check that, in this case, the regression function f_ρ satisfies for all $i = 1, \dots, m$, $f_\rho(x_i) = y_i$, $(\mathscr{L}^2_{\rho_X}(X), \|.\|_{\mathscr{L}^2_{\rho_X}}) = (\mathbb{R}^m, \|.\|_{2,m})$, where $\|u\|_{2,m}^2 = (1/m)\sum_{i=1}^m u_i^2$, and $\|f_\rho\|_{l^1} = (1/m)\|y\|_1$, where $y = (y_1, \dots, y_m)$.

For a kernel K and a regularization parameter $\gamma > 0$, we denote the *regularized empirical error functional* by

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

Recall that the unique minimum point f_γ of $\mathscr{E}_{z,\gamma,K}$ over $\mathscr{H}_K(X)$ satisfies (see, e.g., Cucker and Smale [15, p. 42])

$$f_\gamma = \sum_{i=1}^{m} c_i K_{x_i}, \tag{12}$$

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

$$(\gamma m\mathcal{I} + \mathscr{K}[x])c = y, \tag{13}$$

$\mathcal{I}$ is the $m \times m$ identity matrix and $\mathscr{K}[x]_{ij} = K(x_i, x_j)$ is the $m \times m$ *Gram matrix of the kernel K with respect to* $x = (x_1, \dots, x_m)$ and $y = (y_1, \dots, y_m)$.

As a corollary of Theorem 3.2, we obtain the following estimates for minimization of $\mathscr{E}_{z,\gamma,K}$ over $\mathrm{span}_n G_K$.

Corollary 4.1. *Let d, m, n be positive integers, $X \subset \mathbb{R}^d$, $Y \subset \mathbb{R}$, K: $X \times X \to \mathbb{R}$ a kernel, $s_K = \sup_{x\in X}\sqrt{K(x,x)}$, $\gamma > 0$, $z = \{(x_i, y_i) \in X \times Y, i = 1, \dots, m\}$, $y_{\max} = \max\{|y_i| \mid i = 1, \dots, m\}$, f_γ be the unique solution of $(\mathscr{H}_K(X), \mathscr{E}_{z,\gamma,K})$, $\{\varepsilon_n\}$ a sequence of positive real numbers, and $\{f_n\}$ a sequence of ε_n-near minimum*

points of $(span_n G_K, \mathscr{E}_{z,\gamma,K})$, $b = s_K^2(s_K^2+\gamma)/\gamma^2$ *and* $c = y_{\max}(s_K/\gamma)(s_K^2/\sqrt{\gamma}+2s_K+\sqrt{\gamma})$. *Then*

(i) $\inf_{f\in \mathrm{span}_n G_K} \mathscr{E}_{z,\gamma,K}(f) - \mathscr{E}_{z,\gamma,K}(f_\gamma) \le b\|y\|_1^2/m^2 n + c\|y\|_1/m\sqrt{n}$;

(ii) $\|f_n - f_\gamma\|_K^2 \le (1/\gamma)(b\|y\|_1^2/m^2 n + c\|y\|_1/m\sqrt{n}+\varepsilon_n)$;

(iii) $\sup_{x\in X}|f_n(x)-f_\gamma(x)|^2 \le (s_K^2/\gamma)(b\|y\|_1^2/m^2 n + c\|y\|_1/m\sqrt{n}+\varepsilon_n)$.

In Kůrková and Sanguineti [29], we derived estimates for minimization of the regularized empirical error directly from properties of $\mathscr{E}_{z,\gamma,K}$. In Kůrková and Sanguineti [29, Theorem 5.4], we obtained the following upper bounds:

$$\inf_{f\in \mathrm{span}_n G_K} \mathscr{E}_{z,\gamma,K}(f) - \mathscr{E}_{z,\gamma,K}(f_\gamma) \le \frac{b'}{n} + \frac{c'}{\sqrt{n}}, \tag{14}$$

$$\|f_n - f_\gamma\|_K^2 \le \frac{1}{\gamma}\left(\frac{b'}{n} + \frac{c'}{\sqrt{n}} + \varepsilon_n\right), \tag{15}$$

$$\sup_{x\in X}|f_n(x)-f_\gamma(x)|^2 \le \frac{s_K^2}{\gamma}\left(\frac{b'}{n} + \frac{c'}{\sqrt{n}} + \varepsilon_n\right), \tag{16}$$

where

$$b' = (s_K^2+\gamma)\left(\frac{s_K\|y\|_2}{\gamma\sqrt{m}}\right)^2, \tag{17}$$

$$c' = 2\left((s_K^2+\gamma)\frac{\sqrt{\lambda_{\max}}\|y\|_2}{\gamma m} + y_{\max}s_K\right)\frac{s_K\|y\|_2}{\gamma\sqrt{m}}, \tag{18}$$

and $\lambda_{\max}$ is the maximum eigenvalue of the Gram matrix $\mathscr{K}[x]$.

In Table 1, the upper bounds derived in this paper as a special case of those obtained for the expected error are compared with estimates derived in Kůrková and Sanguineti [29] directly for the empirical error.

Note that, if γ is not too small, then, for data for which $\|y\|_1$ is close to $\sqrt{m}\|y\|_2$, the coefficient $(s_K^2+\gamma)(s_K\|y\|_1/\gamma m)^2$ of the term $1/n$ is close to the coefficient $(s_K^2+\gamma)(s_K\|y\|_2/\gamma\sqrt{m})^2$. So, in this case, the results from Kůrková and Sanguineti [29, Theorem 3.1] and Corollary 4.1 are quite similar. However, in this case, the coefficients of the term $1/\sqrt{n}$ differ. For $\lambda_{\max} \le \gamma/4$, the estimate

$$\left(\frac{s_K^2}{\gamma}2\sqrt{\lambda_{\max}} + 2s_K + 2\sqrt{\lambda_{\max}}\right)\frac{s_K\, y_{\max}\,\|y\|_1}{\gamma m}, \tag{19}$$

of the coefficient $2((s_K^2+\gamma)\sqrt{\lambda_{\max}}\|y\|_2/\gamma m + y_{\max}s_K)s_K\|y\|_2/\gamma\sqrt{m}$ at the term $1/\sqrt{n}$ from Kůrková and Sanguineti [29, Theorem 3.1] is better than the coefficient

$$\left(\frac{s_K^2}{\sqrt{\gamma}} + 2s_K + \sqrt{\gamma}\right)\frac{s_K\,\|y\|_1}{\gamma m}\, y_{\max} \tag{20}$$

from Corollary 4.1.

5. Discussion. The Representer Theorem describes theoretically the optimal solution of a learning task with a generalization capability modeled by using a global condition on the solution, which can be expressed in terms of a kernel norm. We have compared such an optimal solution with suboptimal ones, which can be achieved by various neural-network algorithms operating on networks with a limited number n of computational units. We have estimated the speeds of convergence of such suboptimal solutions.

TABLE 1. Upper bounds from Kůrková and Sanguineti [29, Theorem 3.1] and Corollary 4.1.

	Upper bounds from Corollary 4.1	Upper bounds from Kůrková and Sanguineti [29, Theorem 3.1]
$\frac{1}{n}$	$(s_K^2+\gamma)\left(\frac{s_K\|y\|_1}{\gamma m}\right)^2$	$(s_K^2+\gamma)\left(\frac{s_K\|y\|_2}{\gamma\sqrt{m}}\right)^2$
$\frac{1}{\sqrt{n}}$	$\left(\frac{s_K^2}{\sqrt{\gamma}} + 2s_K + \sqrt{\gamma}\right)\frac{s_K\|y\|_1}{\gamma m}\, y_{\max}$	$2\left((s_K^2+\gamma)\frac{\sqrt{\lambda_{\max}}\|y\|_2}{\gamma m} + y_{\max}s_K\right)\frac{s_K\|y\|_2}{\gamma\sqrt{m}}$

By contrast to studies (e.g., Burger and Neubauer [12]) that give only asymptotic results of the form $\mathcal{O}(1/\sqrt{n})$, we have obtained upper bounds that hold for all n. Moreover, we have specified characteristics of the data (properties of the regression function) and the kernels on which the estimates depend.

Our estimates have been derived using a result from nonlinear approximation theory developed by Maurey (reported in Pisier [33, Lemma 2, p. V.2]), Jones [24, p. 611], and Barron [5, p. 934, Lemma 1]. As this result gives a bound on the worst-case error (it holds for all functions with the same variation norm), our estimates may not be tight: For some regression functions, faster rates might hold.

Because Maurey-Jones-Barron's theorem is not constructive, it does not suggest a method to build suitable networks. In practical applications, network parameters are searched for by algorithms based on gradient descent with stochastic perturbations (Bertsekas [9, pp. 38–40, 103–104]), genetic algorithms (Goldberg [22]), simulated annealing (Aarts and Korst [1]), global stochastic optimization based on Monte Carlo or quasi-Monte Carlo methods (Yin [41]), and Basis Pursuit (Chen et al. [13]). Among learning algorithms explicitly developed for neural networks, see, e.g., Alessandri et al. [3], Bertsekas and Tsitsiklis [10], Grippo [23], and the references therein; for algorithms implementing weight-decay, see Burger and Neubauer [12] and Chen et al. [13].

Appendix A. Reproducing kernel Hilbert spaces (RKHS'). For RKHS' we refer the reader to Aronszajn [4], Berg et al. [6]. For their role in statistics and in learning theory, see, e.g., Berlinet and Thomas-Agnan [7] and Schölkopf and Smola [36], respectively. Here we just recall basic concepts and definitions.

A Reproducing Kernel Hilbert Space (RKHS) is a Hilbert space X formed by functions defined on a nonempty set X such that, for every $u \in X$, the evaluation functional $\mathcal{F}_u$, defined for any $f \in X$ as $\mathcal{F}_u(f) = f(u)$, is bounded (Aronszajn [4], Berg et al. [6], Cucker and Smale [15]).

RKHS' can be characterized in terms of kernels, which are *symmetric positive semidefinite* functions $K\colon X \times X \to \mathbb{R}$, i.e., functions satisfying, for all positive integers m, all $(w_1, \ldots, w_m) \in \mathbb{R}^m$ and all $(u_1, \ldots, u_m) \in X^m$,

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

Every kernel $K\colon X \times X \to \mathbb{R}$ generates an RKHS $\mathcal{H}_K(X)$ that is the completion of the linear span of the set $\{K_u \mid u \in X\}$, with the inner product defined as $\langle K_u, K_v \rangle_K = K(u, v)$ and the induced norm $\|\cdot\|_K$ (see, e.g., Aronszajn [4] and Berg et al. [6, p. 81]).

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

$$\sup_{u \in X} |f(u)| \leq s_K \|f\|_K. \tag{A1}$$

The role of $\|.\|_K^2$ as a stabilizer can be illustrated by two examples of classes of kernels. The first is formed by *Mercer kernels*, i.e., continuous, symmetric, and positive definite functions $K\colon X \times X \to \mathbb{R}$, where $X \subset \mathbb{R}^d$ is compact. For a Mercer kernel K, $\|f\|_K^2$ can be expressed using eigenvectors and eigenvalues of the compact linear operator $L_K\colon \mathcal{L}_2(X) \to \mathcal{C}(X)$, defined, for every $f \in \mathcal{L}_2(X)$, as $L_K(f)(x) = \int_X K(x,u) f(u)\, du$, where $\mathcal{L}_2(X)$ and $\mathcal{C}(X)$ denote the spaces of square integrable and of continuous functions on X, respectively. By the Mercer Theorem (Cucker and Smale [15, p. 34]),

$$\|f\|_K^2 = \sum_{i=1}^{\infty} \frac{c_i^2}{\lambda_i}, \tag{A2}$$

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 $\mathcal{H}_K(X)$ formed by the eigenvectors of L_K.

Note that the sequence $\{\lambda_i\}$ is either finite or convergent to zero (for K smooth enough, the convergence to zero is rather fast (Dunford and Schwartz [17, p. 1119])). Thus, the stabilizer $\|.\|_K^2$ penalizes functions for which the sequences of coefficients $\{c_i\}$ do not converge to zero sufficiently quickly. So the functional $\|.\|_K^2$ plays the role of a high-frequency filter.

The second class of kernels illustrating the role of $\|.\|_K^2$ as a stabilizer consists of *convolution kernels*, i.e., kernels defined on $\mathbb{R}^d \times \mathbb{R}^d$ such that $K(x,y) = k(x-y)$, for which the Fourier transform $\tilde{k}$ of k is positive. For such kernels, the value of the stabilizer at any $f \in \mathcal{H}_K(X)$ can be expressed as

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

(see, e.g., Girosi [20] and Schölkopf and Smola [36, p. 97]). The function $1/\tilde{k}$ plays a role analogous to that of the sequence $\{1/\lambda_i\}$ in the case of a Mercer kernel. For example, the Gaussian kernel is a convolution kernel with a positive Fourier transform.

Appendix B. Convex functionals. For concepts related to optimization and convexity, we refer readers to Borwein and Lewis [11].

A *modulus of continuity* of a functional $\Phi\colon (\mathcal{H}, \|.\|) \to \mathbb{R}$ at $f \in \mathcal{H}$ is defined, for every $t > 0$, as $\omega_f(t) = \sup\{|\Phi(f) - \Phi(g)|\colon \|f - g\| \le t\}$.

A functional Φ is *convex* on a convex set $M \subseteq \mathcal{H}$ if for all $h, g \in M$ and all $\lambda \in [0,1]$, one has $\Phi(\lambda h + (1-\lambda)g) \le \lambda\Phi(h) + (1-\lambda)\Phi(g)$ and Φ is *uniformly convex* if there exists a function $\delta\colon \mathbb{R}_+ \to \mathbb{R}_+$ such that $\delta(0) = 0$, for all $t > 0$, $\delta(t) > 0$, and for all $h, g \in M$ and all $\lambda \in [0,1]$, $\Phi(\lambda h + (1-\lambda)g) \le \lambda\Phi(h) + (1-\lambda)\Phi(g) - \lambda(1-\lambda)\delta(\|h - g\|)$. Any such function δ is called a *modulus of convexity* of Φ.

Appendix C. Proof of Proposition 3.1. This proof follows steps similar to those of the proof of properties of the regularized empirical error from Kůrková and Sanguineti [29, Proposition 5.1].

(i) It is easy to check that the sum of a convex functional and a uniformly convex functional with modulus of convexity δ is a uniformly convex functional with the same modulus δ, and that the square of the norm on a Hilbert space is a uniformly convex functional with the modulus of convexity bounded from above by t^2 (see, e.g., Kůrková and Sanguineti [28, Proposition 2.1(iv)]). The convexity of $\mathcal{E}_\rho$ can be easily proved by using the Fubini Theorem (Friedman [19, p. 85]).

(ii) We first estimate the modulus of continuity of $\mathcal{E}_\rho$. By the Fubini Theorem (Friedman [19, p. 85]) and (A1), for every $f, g \in \mathcal{H}_K(X)$, we have

$$\begin{aligned}
|\mathcal{E}_\rho(f) - \mathcal{E}_\rho(g)| &\le \int_{X\times Y} |(f(x) - g(x))\,(f(x) + g(x) - 2y)|\, d\rho \\
&\le s_K \|f - g\|_K \int_{X\times Y} |(f(x) + g(x) - 2y)|\, d\rho \\
&\le s_K \|f - g\|_K \int_X \int_Y (|f(x) + g(x)| + |2y|)\, d\rho(y \mid x)\, d\rho_X \\
&\le s_K \|f - g\|_K \left(\int_X (s_K \|f - g\|_K + 2 s_K \|g\|_K)\, d\rho_X + \int_Y |2y|\, d\rho(y \mid x) \right). \qquad \text{(C1)}
\end{aligned}$$

As $\int_Y |2y| d\rho(y \mid x) \le 2 s_Y$, we get

$$|\mathcal{E}_\rho(f) - \mathcal{E}_\rho(g)| \le s_K \|f - g\|_K (s_K \|f - g\|_K + 2 s_K \|g\|_K + 2 s_Y).$$

So $\mathcal{E}_\rho$ is continuous on $(\mathcal{H}_K(X), \|.\|_K)$, and its modulus of continuity at g is bounded from above by $s_K^2 t^2 + 2 s_K (s_K \|g\|_K + s_Y) t$.

Since

$$\|f\|_K^2 - \|g\|_K^2 = \langle f - g, f + g \rangle_K \le \|f - g\|_K (\|f - g\|_K + 2\|g\|_K),$$

the modulus of continuity of $\gamma \|.\|_K^2$ at g is bounded from above by $\gamma t^2 + 2\|g\|_K \gamma t$.

Thus, $\mathcal{E}_{\rho,\gamma,K}$ is continuous on $(\mathcal{H}_K(X), \|.\|_K)$ with the modulus of continuity at g bounded from above by $(s_K^2 + \gamma) t^2 + 2((s_K^2 + \gamma)\|g\|_K + s_K s_Y) t$.

(iii) The existence of a unique minimum point f_γ is guaranteed by Theorem 2.1. By Dontchev [16, p. 10] (see also Kůrková and Sanguineti [29, Proposition 4.1(ii)]), for every functional Φ that is uniformly convex over a convex set M with a modulus of convexity δ and a minimum point f^o, for any $f \in M$ we have $\delta(\|f - f^o\|) \le \Phi(f) - \Phi(f^o)$. Hence, for every $f \in \mathcal{H}_K(X)$ we get $\gamma \|f - f_\gamma\|_K^2 \le |\mathcal{E}_{\rho,\gamma,K}(f) - \mathcal{E}_{\rho,\gamma,K}(f_\gamma)|$.

Acknowledgments. The authors were partially supported by the 2007–2009 Scientific Agreement among the University of Genoa, the National Research Council of Italy, and the Academy of Sciences of the Czech Republic, Project "Learning from Data by Neural Networks and Kernel Methods as Approximate Optimization." The first author was partially supported by GA ČR projects 201/05/0557 and 201/08/1744 and by the Institutional Research Plan AV0Z10300504. The second author was partially supported by a PRIN grant from the Italian Ministry of University and Research, Project "Models and Algorithms for Robust Network Optimization."

References

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

[2] Aizerman, M. A., E. M. Braverman, L. I. Rozonoer. 1964. Theoretical foundations of the potential function method in pattern recognition learning. *Automation Remote Control* **25** 821–837.

[3] Alessandri, A., M. Sanguineti, M. Maggiore. 2002. Optimization-based learning with bounded error for feedforward neural networks. *IEEE Trans. Neural Networks* **13**(2) 261–273.
[4] Aronszajn, N. 1950. Theory of reproducing kernels. *Trans. Amer. Math. Soc.* **68**(3) 337–404.
[5] Barron, A. R. 1993. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inform. Theory* **39**(33) 930–945.
[6] Berg, C., J. P. R. Christensen, P. Ressel. 1984. *Harmonic Analysis on Semigroups*. Springer-Verlag, New York.
[7] Berlinet, A., C. Thomas-Agnan. 2004. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer.
[8] Bertero, M. 1989. Linear inverse and ill-posed problems. *Adv. Electronics Electron Phys.* **75** 1–120.
[9] Bertsekas, D. P. 1999. *Nonlinear Programming*. Athena Scientific, Belmont, MA.
[10] Bertsekas, D. P., J. N. Tsitsiklis. 1996. *Neuro-Dynamic Programming*. Athena Scientific, Belmont, MA.
[11] Borwein, J., A. Lewis. 2000. *Convex Analysis and Nonlinear Optimization*. *CMS Books in Mathematics*. Springer, New York.
[12] Burger, M., A. Neubauer. 2003. Analysis of Tikhonov regularization for function approximation by neural networks. *Neural Networks* **16** 79–90.
[13] Chen, S. S., D. L. Donoho, M. A. Saunders. 1998. Atomic decomposition by Basis Pursuit. *SIAM J. Sci. Comput.* **20** 33–61.
[14] Cortes, C., V. Vapnik. 1995. Support-vector networks. *Machine Learn.* **20** 273–297.
[15] Cucker, F., S. Smale. 2002. On the mathematical foundations of learning. *Bull. AMS* **39** 1–49.
[16] Dontchev, A. L. 1983. *Perturbations, Approximations and Sensitivity Analysis of Optimal Control Systems. Lecture Notes in Control and Information Sciences*, Vol. 52. Springer-Verlag, Berlin.
[17] Dunford, N., J. T. Schwartz. 1963. *Linear Operators. Part II: Spectral Theory*. Interscience Publishers, New York.
[18] Fine, T. L. 1999. *Feedforward Neural Network Methodology*. Springer.
[19] Friedman, A. 1982. *Modern Analysis*. Dover, New York.
[20] Girosi, F. 1998. An equivalence between sparse approximation and support vector machines. *Neural Comput.* **10** 1455–1480.
[21] Girosi, F., M. Jones, T. Poggio. 1995. Regularization theory and neural networks architectures. *Neural Comput.* **7** 219–269.
[22] Goldberg, D. E. 1989. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley,
[23] Grippo, L. 2000. Convergent on-line algorithms for supervised learning in neural networks. *IEEE Trans. Neural Networks* **11**(6) 1284–1299.
[24] Jones, L. K. 1992. A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training. *Ann. Statist.* **20**(1) 608–613.
[25] Kimeldorf, G. S., G. Wahba. 1970. A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Ann. Math. Statist.* **41**(2) 495–502.
[26] Kimeldorf, G. S., G. Wahba. 1971. Some results on Tchebyceffian spline functions. *J. Math. Anal. Appl.* **33**(1) 82–95.
[27] Kůrková, V., M. Sanguineti. 2002. Comparison of worst-case errors in linear and neural network approximation. *IEEE Trans. Inform. Theory* **48**(1) 264–275.
[28] Kůrková, V., M. Sanguineti. 2005a. Error estimates for approximate optimization by the extended Ritz method. *SIAM J. Optim.* **15**(2) 261–287.
[29] Kůrková, V., M. Sanguineti. 2005b. Learning with generalization capability by kernel methods of bounded complexity. *J. Complexity* **21** 350–367.
[30] Kůrková, V., P. C. Kainen, V. Kreinovich. 1997. Estimates of the number of hidden units and variation with respect to half-spaces. *Neural Networks* **10**(6) 1061–1068.
[31] Parzen, E. 1961. An approach to time series analysis. *Ann. Math. Statist.* **32** 951–989.
[32] Parzen, E. 2006. Reproducing kernel hilbert spaces outline. http://www.stat.tamu.edu/~jianhua/rkhs/Manny.pdf.
[33] Pisier, G. 1980-81. Remarques sur un resultat non publié de B. Maurey. *Seminaire d'Analyse Fonctionelle* **I(12)**. École Polytechnique, Centre de Mathématiques, Palaiseau, France.
[34] Poggio, T., F. Girosi. 1990. Networks for approximation and learning. *Proc. IEEE* **78**(9) 1481–1497.
[35] Poggio, T., S. Smale. 2003. The mathematics of learning: Dealing with data. *Notices of the AMS* **50**(5) 537–544.
[36] Schölkopf, B., A. J. Smola. 2002. *Learning with Kernels*. The MIT Press, Cambridge, MA.
[37] Schönberg, I. J. 1938. Metric spaces and completely monotone functions. *Ann. Math.* **39**(4) 811–841.
[38] Tikhonov, A. N. 1963. Solutions of incorrectly formulated problems and the regularization method. *Soviet Math. Dokl.* **4** 1035–1038.
[39] Vapnik, V. N. 1998. *Statistical Learning Theory*. John Wiley & Sons, New York.
[40] Wahba, G. 1990. Splines models for observational data. CBMS-NSF Regional Conference Series in Applied Mathematics **59**. SIAM, Philadelphia.
[41] Yin, G. 1963. Rates of convergence for a class of global stochastic optimization algorithms. *SIAM J. Optim.* **10** 99–120.