

The weight-decay technique in learning from data: an optimization point of view

Giorgio Gnecco · Marcello Sanguineti

© Springer-Verlag 2008

Abstract The technique known as “weight decay” in the literature about learning from data is investigated using tools from regularization theory. Weight-decay regularization is compared with Tikhonov’s regularization of the learning problem and with a mixed regularized learning technique. The accuracies of suboptimal solutions to weight-decay learning are estimated for connectionistic models with a-priori fixed numbers of computational units.

Keywords Learning from data · Regularization · Weight decay · Suboptimal solutions · Rates of convergence

1 Introduction

In a variety of applications, an unknown function has to be learned on the basis of a sample of input–output data (Poggio and Smale 2003) (e.g., time-series prediction, system identification, fault diagnosis, weather forecasting, image reconstruction, pattern recognition, development of market models, fitting biological data). The task of supervised learning from data can be mathematically formulated as the minimization

The Authors were partially supported by a PRIN grant from the Italian Ministry for University and Research, project “Models and Algorithms for Robust Network Optimization”.

G. Gnecco · M. Sanguineti (✉)
Department of Communications, Computer, and System Sciences (DIST),
University of Genova, Via Opera Pia 13, 16145 Genova, Italy
e-mail: marcello@dist.unige.it

G. Gnecco
Department of Mathematics (DIMA), University of Genova,
Via Dodecaneso, 35, 16146 Genova, Italy
e-mail: giorgio.gnecco@dist.unige.it

of a functional defined on the basis of the probability distribution generating the data. For a nonempty set $X \subseteq \mathbb{R}^d$ and a joint probability distribution $P(x, y)$ of two random variables $x \in X$ and $y \in \mathbb{R}$, Statistical Learning Theory (Cristianini and Shawe-Taylor 2000; Vapnik 1998) models the learning problem as the minimization of the *expected error functional* $\mathcal{E}(f) \triangleq \int_{X \times \mathbb{R}} (f(x) - y)^2 dP(x, y)$, where $f : X \rightarrow \mathbb{R}$.

In practice, since the probability distribution $P(x, y)$ is usually unknown, the minimization of the expected error functional is replaced by the minimization of the *empirical error functional*, defined for a positive integer m and a data sample $\mathbf{z} \triangleq \{(x_i, y_i) \in X \times \mathbb{R}, i = 1, \dots, m\}$ as

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

So, the learning problems can be modeled as $\min_{f \in \mathcal{H}} \mathcal{E}_{\mathbf{z}}(f)$,¹ where \mathcal{H} is an appropriate space of functions, called *hypothesis space*. Suitable choices of \mathcal{H} allow one to achieve *generalization capability*, i.e., the capability of a model to behave satisfactorily also in the case of data that were not used for learning. To choose a suitable hypothesis space, one can exploit a priori information, which represents one's knowledge on the unknown solution and/or expresses some desired behavior for it.

The problem of approximating a function on the basis of a data sample is often ill-posed (Bertero 1989; Burger and Engl 2000). *Regularization* (Tikhonov and Arsenin 1977) can be used to cope with this drawback. One regularization approach consists in restricting the minimization of the empirical error functional $\mathcal{E}_{\mathbf{z}}$ to a subset M of the hypothesis space \mathcal{H} , called *hypothesis set* and containing only functions with a desired behavior. This form of regularization replaces the original problem with the problem $\min_{f \in M} \mathcal{E}_{\mathbf{z}}(f)$. Another regularization approach consists in minimizing the *regularized empirical error functional*, defined as $\mathcal{E}_{\mathbf{z}}(f) + \gamma \Psi(f)$, where $\Psi : \mathcal{H} \rightarrow \mathbb{R}$ is a functional called *stabilizer* and $\gamma > 0$ is the *regularization parameter*. The corresponding model for the learning problem is $\min_{f \in M} (\mathcal{E}_{\mathbf{z}}(f) + \gamma \Psi(f))$. The parameter γ controls the trade-off between the following two requirements: i) fitting to the data sample (via the value $\mathcal{E}_{\mathbf{z}}(f)$ of the empirical error associated with f); ii) penalizing solutions f that provide a large value of the stabilizer $\Psi(f)$. For certain normed hypothesis spaces \mathcal{H} , the choice $\Psi(\cdot) = \|\cdot\|_{\mathcal{H}}^2$ allows one to enforce certain smoothness properties of the solution (Girosi 1998). So, in this case the parameter γ quantifies the compromise between enforcing closeness to the data sample and avoiding solutions that are not sufficiently smooth. A third regularization possibility consists in combining the two aforesaid methods (i.e., restriction to a hypothesis set and use of a stabilizer).

These three approaches are applications to learning from data of regularization techniques developed during the 1960s for ill-posed inverse problems and known

¹ Whenever we write "min" of a quantity over a set, we implicitly suppose that such a minimum exists. If it does not, we mean that we are interested, for $\varepsilon > 0$, in an ε -near minimum point; in this case, $f_{\varepsilon} \in \mathcal{H} : \mathcal{E}_{\mathbf{z}}(f_{\varepsilon}) < \inf_{f \in \mathcal{H}} \mathcal{E}_{\mathbf{z}}(f) + \varepsilon$.

as *Ivanov's*, *Tikhonov's*, and *Miller's regularizations*, respectively (Bertero 1989, pp. 68–78).

Weight decay (see Bishop 1995 for a overview) was introduced into learning to improve the generalization capability of a model expressed as a linear combination of a set of given basis functions (computational units). Loosely speaking, the method penalizes large values of the coefficients (*weights*) of the linear combination. It can be modeled by adding to the empirical error functional a term given by a squared norm (typically, the Euclidean norm) of the coefficient vector. The name “weight decay” is justified by the fact that, if the empirical error term were not present, the minimization through the gradient descent of the squared norm of the coefficients would lead to their exponential decay (Krogh and Hertz 1992). Training algorithms based on weight decay were investigated, e.g., in Gupta and Lam (1998a,b), Treadgold and Gedeon (1998).

For linear regression problems, the performance of weight decay was theoretically investigated in Krogh and Hertz (1992), where the case of linearization of a nonlinear model was considered, too. As to nonlinear models, a theoretical explanation for the generalization performance of certain neural networks trained through weight decay was given in Bartlett (1998), where binary classification problems were studied using tools from Statistical Learning Theory.

In this work, we investigate the weight-decay learning technique as a way to endow a learning model with generalization capabilities. In the first part of the paper, we compare spectral properties of weight decay with Tikhonov's regularization of the learning problem. In the second part, we investigate the accuracies of certain suboptimal solutions to the same problem.

For certain hypothesis spaces generated by computational units widely used in connectionist models, it is known (Cucker and Smale 2002, p. 42) that the solution to the Tikhonov-regularized learning problem has the form of a linear combination of the m -tuple of the computational units parameterized by the input data vector $\mathbf{x} = (x_1, \dots, x_m)$. The coefficients of the linear combination can be obtained by solving a suitable linear system of equations, and this property can be exploited to develop learning algorithms. In the first part of the paper, in order to study the relationships between weight decay and Tikhonov's regularization, we consider, for weight decay, solutions belonging to the space spanned by the computational units in terms of which the solution to the Tikhonov-regularized learning problem is expressed. To investigate a learning technique that combines the advantages of weight decay and Tikhonov's regularization, we also consider a mixed approach. We compare the three methods in terms of spectral windows (Bertero 1989, Sect. 5.2, pp. 84–88), also called filtering factors, which can be used to evaluate the robustnesses of regularization algorithms against noise in input data.

Then, in the second part of the paper, we investigate the accuracies of suboptimal solutions to weight-decay learning and to the mixed weight-decay/Tikhonov's-regularization learning technique, over hypothesis sets corresponding to models exhibiting the following features: (i) flexible choices of the parameters of the linear combinations and (ii) fewer computational units than the size of the data sample. Indeed, for large data sets, using a number of computational units equal to the number m of data (as required by the analytical expression of the optimal solution) may

lead to very complex models and so may be computationally unfeasible. Moreover, practical applications of such algorithms are limited by the rates of convergence of iterative methods solving the associate systems of equations, as such rates depend on the size of the condition number of the matrices involved therein. For some methods, the computational requirements of solving such systems grow polynomially with the size m of the sample (e.g., for the Gaussian elimination and m large enough, they grow at a rate m^3 Ortega 1990, p. 175). For some data and computational units, keeping the condition number of these matrices small requires a large value of the regularization parameter γ , which may cause a poor fit to the empirical data.

We derive upper bounds on the rates of approximation of the optimal solutions for sequences of suboptimal solutions achievable by minimization over hypothesis sets formed by linear combinations of at most $n < m$ computational units with parameters drawn from the data set. The upper bounds are of the form $A/\sqrt{n} + B/n$, where A and B depend on the properties of the vector $\mathbf{y} = (y_1, \dots, y_m)$ of output data, the properties of the kernel, and the regularization parameter γ .

The paper is organized as follows. In Sect. 2, we shortly review the properties of the hypothesis spaces in which we formulate the learning problem and summarize the definitions and notations used throughout the paper. Section 3 models weight-decay learning as a regularized minimization problem and provides an expression for its solution. In Sect. 4, we state the Tikhonov-regularized learning problem and compare its solution with the solution to weight-decay learning. Section 5 addresses a learning technique resulting from the combination of weight decay and Tikhonov's regularization. In Sect. 6, we investigate the accuracies of suboptimal solutions provided to these regularized learning techniques by connectionistic models characterized by an a-priori fixed number $n < m$ of computational units and a flexible choice of the parameters. Finally, in Sect. 7, some conclusions are drawn.

2 Preliminaries

The hypothesis spaces where we set the learning problems are reproducing Kernel Hilbert spaces (RKHSs), i.e., Hilbert spaces $(\mathcal{H}_K, \langle \cdot, \cdot \rangle_{\mathcal{H}_K})$ formed by functions defined on a nonempty set X such that for every $u \in X$ the evaluation functional \mathcal{F}_u , defined for every $f \in \mathcal{H}_K$ as $\mathcal{F}_u(f) \triangleq f(u)$, is bounded (Aronszajn 1950; Berg et al. 1984; Cucker and Smale 2001). We consider real RKHSs. By the Riesz Representation Theorem (Friedman 1970, p. 200), for every $u \in X$ there exists a unique element $K_u \in (\mathcal{H}_K, \langle \cdot, \cdot \rangle_{\mathcal{H}_K})$, called the *representer* of u , such that for every $f \in \mathcal{H}_K$

$$\mathcal{F}_u(f) = \langle f, K_u \rangle_{\mathcal{H}_K} \quad (1)$$

(called the *reproducing property*).

RKHSs can be characterized in terms of *kernels*. A *positive-semidefinite kernel* is a symmetric function $K : X \times X \rightarrow \mathbb{R}$ such that for all positive integers m , all $(w_1, \dots, w_m) \in \mathbb{R}^m$, and all $(u_1, \dots, u_m) \in X^m$

$$\sum_{i,j=1}^m w_i w_j K(u_i, u_j) \geq 0. \tag{2}$$

If the inequality (2) holds strictly when not all the w_i are zero, then we have a *positive-definite kernel*.² If the function K is also continuous, then it is called a *Mercer kernel*. For a kernel $K : X \times X \rightarrow \mathbb{R}$, a positive integer m , and a vector $\mathbf{x} = (x_1, \dots, x_m) \in X^m$, the $m \times m$ *Gram matrix of the kernel K with respect to \mathbf{x}* is defined as

$$\mathcal{K}[\mathbf{x}]_{i,j} \triangleq K(x_i, x_j).$$

Then, a kernel is positive semidefinite (positive definite) if for every positive integer m its $m \times m$ Gram matrices with respect to every $\mathbf{x} \in X^m$ are positive semidefinite (positive definite, resp.). Every kernel $K : X \times X \rightarrow \mathbb{R}$ generates an RKHS \mathcal{H}_K . Indeed, \mathcal{H}_K can be defined as the completion of the linear span of the set $\{K_u : u \in X\}$ with the inner product $\langle K_u, K_v \rangle_{\mathcal{H}_K} \triangleq K(u, v)$ (see, e.g., [Aronszajn 1950](#) and [Berg et al. 1984](#), p. 81). As in every RKHS \mathcal{H}_K , the inner product and the induced norm can be expressed in terms of the corresponding kernel K , in the following we shall denote them merely by $\langle \cdot, \cdot \rangle_K$ and $\| \cdot \|_K$ instead of $\langle \cdot, \cdot \rangle_{\mathcal{H}_K}$ and $\| \cdot \|_{\mathcal{H}_K}$, respectively.

By the Cauchy-Schwartz inequality, for every $f \in \mathcal{H}_K$ and $u \in X$ one has $|f(u)| = |\langle f, K_u \rangle_K| \leq \|f\|_K \sqrt{K(u, u)} \leq s_K \|f\|_K$, where $s_K = \sup_{u \in X} \sqrt{K(u, u)}$. We assume that s_K is finite, as is the case with commonly encountered choices of kernels ([Schölkopf and Smola 2002](#)). Thus for every kernel K ,

$$\sup_{u \in X} |f(u)| \leq s_K \|f\|_K. \tag{3}$$

As for every $u, v \in X$, $K(u, v)$ is given by the inner product $\langle K_u, K_v \rangle_K$, by the Cauchy-Schwartz inequality we get $|K(u, v)| \leq s_K^2$.

For a positive integer d , by $\| \cdot \|_1$ and $\| \cdot \|_2$ we denote the standard 1-norm and Euclidean norm on \mathbb{R}^d , respectively. An illustrative example of a kernel is the *Gaussian kernel* $K(u, v) = e^{-\rho \|u-v\|_2^2}$ on $\mathbb{R}^d \times \mathbb{R}^d$, where $\rho > 0$. The corresponding RKHS contains all functions obtainable by Gaussian radial-basis-function networks with a fixed “width” equal to ρ . Other well-known examples of kernels are $K(u, v) = e^{-\|u-v\|_2}$, $K(u, v) = \langle u, v \rangle^p$ (*homogeneous polynomial* of degree p), where $\langle \cdot, \cdot \rangle$ is any inner product on \mathbb{R}^d , $K(u, v) = (1 + \langle u, v \rangle)^p$ (*inhomogeneous polynomial* of degree p), and $K(u, v) = (a^2 + \|u - v\|^2)^{-\alpha}$, with $\alpha > 0$ ([Cucker and Smale 2001](#), p. 38). In this paper, we consider positive-definite kernels K ; to fix ideas, one can think of the widespread Gaussian kernel.

One reason for choosing RKHSs as hypothesis spaces is that they allow one to model generalization capability. The main tool to achieve mathematical modeling of generalization consists in exploiting a-priori information about the behavior of

² Some Authors call “positive-definite” the kernels for which (2) is satisfied and “strictly positive-definite” those for which (2) holds strictly when not all the w_i are zero; see, e.g., ([Poggio et al. 2002](#), Definition 2.1).

admissible solutions. Such information can be expressed by the choice of an RKHS over which the empirical error \mathcal{E}_z is minimized, since the norms $\|\cdot\|_K$ on RKHSs defined by a large variety of kernels K often play the role of measures of various types of oscillations of functions in those spaces. Thus, the choice of suitable RKHSs as hypothesis spaces allows one to impose a condition on oscillations of admissible solutions to the learning problem. This can be illustrated on *convolution kernels* $K : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$, i.e., kernels such that $K(u, v) = k(u - v)$ for some $k : \mathbb{R}^d \rightarrow \mathbb{R}$ with a non-negative Fourier transform \tilde{k} . For such kernels, the stabilizer can be expressed as

$$\|f\|_K^2 = (2\pi)^{d/2} \int_{\mathbb{R}^d} \frac{|\tilde{f}(\omega)|^2}{\tilde{k}(\omega)} d\omega, \quad (4)$$

where \tilde{f} is the Fourier transform of f (see [Girosi 1998](#); [Schölkopf and Smola 2002](#), p. 97). For example, the Gaussian kernel is a convolution kernel with a positive Fourier transform. Another example of a convolution kernel with a positive Fourier transform is $K(u, v) = k(u - v)$, where $k(t) = e^{-a\|t\|^2}$. In this case, the rate of decay of $\tilde{k}(\omega)$ is of order $\|\omega\|_2^{-(d+1)}$. In particular, for $d = a = 1$, the norm on the corresponding RKHS is a Sobolev norm ([Gnecco and Sanguinetti 2007](#), Appendix C).

For more examples of kernels and for a discussion of their role in learning theory see, e.g., [Schölkopf and Smola \(2002\)](#). We conclude these preliminaries by introducing some notations used throughout the paper.

For a subset G of a linear space and a positive integer n ,

$$\text{span}_n G \triangleq \left\{ \sum_{i=1}^n w_i g_i : w_i \in \mathbb{R}, \quad g_i \in G \right\}$$

and

$$\text{span } G \triangleq \left\{ \sum_{i=1}^n w_i g_i : w_i \in \mathbb{R}, \quad g_i \in G, \quad n \in \mathbb{N} \right\}$$

are the sets of linear combinations of all n -tuples of elements of G and all linear combinations of elements of G , resp.

We let

$$G_K \triangleq \{K_x : x \in X\}$$

and

$$G_{K_x} \triangleq \{K_{x_1}, \dots, K_{x_m}\}.$$

So, $\text{span}_n G_K$ and $\text{span}_n G_{K_x}$ are the sets of all input/output functions of a computational model with one hidden layer of n computational units computing functions from G_K and G_{K_x} , respectively.

Given a space \mathcal{H} , a set $M \subseteq \mathcal{H}$, and a functional $\Phi : M \rightarrow \mathbb{R}$, we follow standard notation from optimization theory and we denote by

$$(M, \Phi)$$

the problem of minimizing Φ over M . Every $f \in M$ such that $\Phi(f) = \min_{f \in M} \Phi(f)$ is called a *solution* or a *minimum point* of the problem (M, Φ) . Let

$$\operatorname{argmin}(M, \Phi) \triangleq \left\{ f \in M : \Phi(f) = \min_{f \in M} \Phi(f) \right\}$$

be the set of solutions of (M, Φ) . For $\varepsilon > 0$, $\operatorname{argmin}_\varepsilon(M, \Phi)$ is the set of ε -near minimum points of (M, Φ) , i.e.,

$$\operatorname{argmin}_\varepsilon(M, \Phi) \triangleq \left\{ f \in M : \Phi(f) < \inf_{f \in M} \Phi(f) + \varepsilon \right\}$$

A sequence $\{f_n\}$ of elements of M is called a Φ -*minimizing sequence* over M if $\lim_{n \rightarrow \infty} \Phi(f_n) = \inf_{f \in M} \Phi(f)$.

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

A functional Φ is *convex* on a convex set $M \subseteq X$ if for all $h, g \in M$ and all $\lambda \in [0, 1]$, $\Phi(\lambda h + (1 - \lambda)g) \leq \lambda\Phi(h) + (1 - \lambda)\Phi(g)$ and it is *uniformly convex* if there exists a non-negative function $\delta : \mathbb{R}_+ \rightarrow \mathbb{R}_+$, where \mathbb{R}_+ denotes the set of non-negative reals, 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 δ is called a *modulus of convexity* of Φ . [Levitin and Polyak \(1966\)](#)³

3 Solution to the weight-decay learning problem

Given a regularization parameter $\gamma > 0$ and a positive-definite kernel K (e.g., the Gaussian kernel), we define on $\operatorname{span}_n G_K$ the *weight-decay empirical error functional* as

$$\Phi_{WD,\gamma}(f) \triangleq \mathcal{E}_z(f) + \gamma \|\mathbf{c}_f\|_2^2, \tag{5}$$

where for a positive integer n and $\hat{x}_1, \dots, \hat{x}_n \in X$, the components of the vector \mathbf{c}_f are the parameters in the expansion

³ The terminology is not unified: some authors use the term “strictly uniformly convex” instead of “uniformly convex,” while they adopt the term “uniformly convex” for the case where $\delta : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ merely satisfies $\delta(0) = 0$ and $\delta(t_0) > 0$ for some $t_0 > 0$ (see, e.g., [Vladimirov et al. 1978](#) and [Dontchev 1983](#), p. 10).

$$f = \sum_{j=1}^n c_{f,j} K_{\hat{x}_j} \in \text{span}_n G_K. \quad (6)$$

Choosing a positive-definite kernel guarantees that f has a unique representation in terms of $\{c_{f,j}\}$ and $\{\hat{x}_j\}$, thus $\|\mathbf{c}_f\|_2^2$ in (5) is defined unambiguously (otherwise one may choose, among all equivalent representations of f —possibly with different values of n —the infimum of the squared norm $\|\mathbf{c}_f\|_2^2$ of the corresponding coefficient vector \mathbf{c}_f).

The functional (5) is defined on $\text{span } G_K$ but not necessarily on its closure, i.e., the whole \mathcal{H}_K . Indeed, for a continuous kernel it is easy to construct a sequence $\{f_n\}$ such that $f_n \in \text{span}_{2n} G_K$, $\|f_n\|_K \rightarrow 0$ and $\|c_{f_n}\|_2^2 \rightarrow \infty$ as $n \rightarrow \infty$. One example of such a sequence is given by $f_n \triangleq \sum_{i=1}^{2n} (-1)^i K_{\hat{x}_i(n)}$, where, for each n , when i is odd $\hat{x}_i(n)$ and $\hat{x}_{i+1}(n)$ are chosen “sufficiently close” to each other, such that $\|f_n\|_K < \frac{1}{n}$. However, in practice the fact that the functional (5) is not defined on the whole \mathcal{H}_K is not a limitation, as in applications the number of kernel units has to be finite.

The number n of terms in the expression (6) is the dimension of the vector \mathbf{c}_f in the functional (5). In this section, we consider the weight-decay functional corresponding to the choice $n = m$ (i.e., n is equal to the size of the data sample) and $\hat{x}_i = x_i$ for $i = 1, \dots, m$, i.e., we minimize the functional (5) over linear combinations $\text{span}_m G_{K_x}$ of m kernel functions centered at the m input data.

Hence, we model the *weight-decay learning problem* as

$$(\text{span}_m G_{K_x}, \Phi_{WD,\gamma}). \quad (7)$$

The next proposition investigates its solution.

Proposition 1 *Let X be a nonempty set, $K : X \times X \rightarrow \mathbb{R}$ a positive-definite kernel, m a positive integer, $\mathbf{x} = (x_1, \dots, x_m) \in X^m$, $\mathcal{K}[\mathbf{x}]$ the Gram matrix of the kernel K with respect to \mathbf{x} , $\mathbf{y} = (y_1, \dots, y_m) \in \mathbb{R}^m$, and $\gamma > 0$. There exists a unique solution*

$$f_{WD,\gamma} = \sum_{i=1}^m c_{WD,\gamma,i} K_{x_i} \quad (8)$$

to the problem $(\text{span}_m G_{K_x}, \Phi_{WD,\gamma})$, where $\mathbf{c}_{WD,\gamma} = (c_{WD,\gamma,1}, \dots, c_{WD,\gamma,m})$ is the unique solution to the linear system of equations

$$\left(\mathcal{K}[\mathbf{x}] + \gamma m \mathcal{K}^{-1}[\mathbf{x}] \right) \mathbf{c}_{WD,\gamma} = \mathbf{y}. \quad (9)$$

Proof For $f = \sum_{i=1}^m c_{f,i} K_{x_i} \in \text{span}_m G_{K_x}$, the functional (5) can be written as

$$\Phi_{WD,\gamma}(f) = \left\| \frac{\mathcal{K}[\mathbf{x}]\mathbf{c}_f}{\sqrt{m}} - \frac{\mathbf{y}}{\sqrt{m}} \right\|_2^2 + \gamma \|\mathbf{c}_f\|_2^2.$$

According to classical results of regularization theory (see, e.g., [Bertero 1989](#), p. 69 with $L = \mathcal{K}[\mathbf{x}]/\sqrt{m}$, $g = \mathbf{y}/\sqrt{m}$, and $\alpha = \gamma$), there exists a unique solution to (7) and such a solution is a linear combination of functions in $G_{\mathcal{K}_x}$, with coefficients given by

$$\mathbf{c}_{WD,\gamma} = \left(\frac{\mathcal{K}^2[\mathbf{x}]}{m} + \gamma \mathcal{I} \right)^{-1} \frac{\mathcal{K}[\mathbf{x}]}{\sqrt{m}} \frac{\mathbf{y}}{\sqrt{m}} = \left(\mathcal{K}[\mathbf{x}] + \gamma m \mathcal{K}^{-1}[\mathbf{x}] \right)^{-1} \mathbf{y}.$$

□

Proposition 1 expresses the solution to the weight-decay learning problem as a linear combination of kernel functions centered at the data points, with coefficients given by the linear system of Eq. (9). For example, 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 ([Girosi et al. 1995](#)). Numerical issues related to the solution of such a system will be discussed in Sect. 6.

We denote by $\lambda_{\max}(\mathcal{A})$ and $\lambda_{\min}(\mathcal{A})$ the maximum and minimum eigenvalues of a symmetric matrix \mathcal{A} , respectively. To simplify the notation, we write λ_{\max} instead of $\lambda_{\max}(\mathcal{K}[\mathbf{x}])$ and similarly for λ_{\min} . By (9), elementary spectral theory arguments, and some algebra we have the following upper bound on the Euclidean norm of the coefficient vector of the solution:

$$\|\mathbf{c}_{WD,\gamma}\|_2 \leq \max_{\lambda \in [\lambda_{\min}, \lambda_{\max}]} \frac{1}{\lambda + \gamma m \lambda^{-1}} \|\mathbf{y}\|_2 \leq \frac{1}{2\sqrt{\gamma m}} \|\mathbf{y}\|_2.$$

As $\|f_{WD,\gamma}\|_K = \|\mathcal{K}^{\frac{1}{2}}[\mathbf{x}] (\mathcal{K}[\mathbf{x}] + \gamma m \mathcal{K}^{-1}[\mathbf{x}])^{-1} \mathbf{y}\|_2$, by (8), (9), elementary spectral theory arguments, and some algebra, we get the following upper bound on the K -norm of the solution:

$$\|f_{WD,\gamma}\|_K \leq \max_{\lambda \in [\lambda_{\min}, \lambda_{\max}]} \frac{\sqrt{\lambda}}{\lambda + \gamma m \lambda^{-1}} \|\mathbf{y}\|_2 \leq \frac{1}{4} \sqrt[4]{\frac{27}{\gamma m}} \|\mathbf{y}\|_2.$$

Note that the above upper bounds on $\|\mathbf{c}_{WD,\gamma}\|_2$ and $\|f_{WD,\gamma}\|_K$ are proportional to $\|\mathbf{y}\|_2/\sqrt{\gamma m}$ and $\|\mathbf{y}\|_2/\sqrt[4]{\gamma m}$, resp.

4 Comparison with the Tikhonov-regularized learning problem

Tikhonov’s regularization in learning from data can be formalized in terms of the following *Tikhonov-regularized empirical error functional*:

$$\Phi_{T,\gamma} \triangleq \mathcal{E}_z(f) + \gamma \|f\|_K^2, \tag{10}$$

where $\|\cdot\|_K$ is the norm on the RKHS \mathcal{H}_K . The squared norm $\|\cdot\|_K^2$ is used as a stabilizer instead of $\|\cdot\|_K$ for technical reasons, as the square of the norm on any Hilbert space is a uniformly convex functional [[Kůrková and Sanguineti 2005](#), Proposition

4.1 (iii)], which implies uniqueness of the solution of the regularized problem (see, e.g., [Dontchev 1983](#), p. 10, [Cucker and Smale 2001](#), pp. 27, 42) and convergence of minimizing sequences to the solution [Levitin and Polyak \(1966\)](#). The corresponding *Tikhonov-regularized learning problem* is

$$(\mathcal{H}_K, \Phi_{T,\gamma}). \quad (11)$$

The existence, the uniqueness, and an explicit formula describing the solution to the learning problem (11) are given by the so-called *Representer Theorem* (see, e.g., ([Cucker and Smale 2001](#), p. 42) and [Poggio and Smale \(2003\)](#); [Girosi et al. \(1995\)](#); [Girosi \(1994\)](#); [Poggio and Girosi \(1990\)](#)).

Theorem 1 (Representer Theorem) *Let X be a nonempty set, $K : X \times X \rightarrow \mathbb{R}$ a kernel, m a positive integer, $\mathbf{x} = (x_1, \dots, x_m) \in X^m$, $\mathcal{K}[\mathbf{x}]$ the Gram matrix of the kernel K with respect to \mathbf{x} , $\mathbf{y} = (y_1, \dots, y_m) \in \mathbb{R}^m$, and $\gamma > 0$. There exists a unique solution $f_{T,\gamma}$ of the problem $(\mathcal{H}_K, \Phi_{T,\gamma})$ and it has the form*

$$f_{T,\gamma} = \sum_{i=1}^m c_{T,\gamma,i} K_{x_i}, \quad (12)$$

where $\mathbf{c}_{T,\gamma} = (c_{T,\gamma,1}, \dots, c_{T,\gamma,m})$ is the unique solution to the linear system of equations

$$(\gamma m \mathcal{I} + \mathcal{K}[\mathbf{x}])\mathbf{c}_{T,\gamma} = \mathbf{y}. \quad (13)$$

The Representer Theorem was first proven in [Kimeldorf and Wahba \(1970\)](#). For Mercer kernels, a proof based on the Mercer Theorem was used in ([Cucker and Smale 2001](#), p. 42). In [Kůrková \(2004\)](#), it was derived from the theory of inverse problems, and a proof using functional derivatives was given in ([Poggio and Smale, 2003](#), pp. 538–539). A weaker form of the result, without a formula for computing the coefficients c_1, \dots, c_m , is sometimes called “Generalized Representer Theorem,” and holds for a stabilizer of the form $\psi(\|\cdot\|_K)$, where $\psi : [0, +\infty) \rightarrow \mathbb{R}$ is a strictly increasing function ([Schölkopf et al. 2001](#)).

The optimal solution to the Tikhonov-regularized learning problem described by Theorem 1 is an element of $\text{span}_m G_{K,\mathbf{x}} \subseteq \text{span}_m G_K$, where $\text{span}_m G_K$ 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 . Numerical issues related to the solution to the linear system of Eq. (13) will be discussed in Sect. 6.

In applications, the number n of computational units is bounded. To compare weight decay with Tikhonov’s regularization, let $n < m$. The following proposition gives an immediate relationship between the weight-decay learning functional and the Tikhonov-regularized one.

Proposition 2 *Let X be a nonempty set, $K : X \times X \rightarrow \mathbb{R}$ a positive-definite kernel, n and m positive integers, $f \in \text{span}_n G_K$, $\mathbf{x} = (x_1, \dots, x_m) \in X^m$, $\mathbf{y} = (y_1, \dots, y_m) \in \mathbb{R}^m$, $\gamma > 0$, and $\bar{\gamma} \triangleq \gamma/(n s_K^2)$. Then*

$$\Phi_{T,\bar{\gamma}}(f) \leq \Phi_{WD,\gamma}(f)$$

Proof For $f = \sum_{j=1}^n c_{f,j} K_{\hat{x}_j} \in \text{span}_n G_K$, $\|f\|_K^2 = \mathbf{c}_f \mathcal{K}[\mathbf{x}] \mathbf{c}_f^\top$, where the superscript “ \top ” denotes transposition. For every $\mathbf{x}, \mathbf{y} \in X$, one has $|K(\mathbf{x}, \mathbf{y})| \leq s_K^2$. So we get $\|f\|_K^2 \leq \|\mathbf{c}_f\|_1^2 s_K^2 \leq \|\mathbf{c}_f\|_2^2 n s_K^2$. \square

By Proposition 2, for every $f \in \text{span}_n G_K$ the value of the weight-decay functional with parameter γ is an upper bound on the value of the Tikhonov-regularized functional with parameter $\bar{\gamma} \triangleq \gamma/(n s_K^2)$.

As, according to the Representer Theorem, the solution to the Tikhonov-regularized learning problem $(\mathcal{H}_K, \Phi_{T,\gamma})$ belongs to $\text{span}_m G_{K_X}$, one can restate such a problem as $(\text{span}_m G_{K_X}, \Phi_{T,\gamma})$ and compare its solution with the solution to the weight-decay learning problem $(\text{span}_m G_{K_X}, \Phi_{WD,\gamma})$.

The comparison between the optimal values $\mathbf{c}_{WD,\gamma}$ and $\mathbf{c}_{T,\gamma}$ of the coefficient vectors of the solutions to problems $(\text{span}_m G_{K_X}, \Phi_{WD,\gamma})$ and $(\text{span}_m G_{K_X}, \Phi_{T,\gamma})$, resp., is particularly interesting when \mathbf{y} is an eigenvector \mathbf{y}_λ of the matrix $\mathcal{K}[\mathbf{x}]$ associated with the eigenvalue λ . In this case, the equations $\mathbf{c}_{WD,\gamma} = (\mathcal{K}[\mathbf{x}] + \gamma m \mathcal{K}^{-1}[\mathbf{x}])^{-1} \mathbf{y}$ and $\mathbf{c}_{T,\gamma} = (\mathcal{K}[\mathbf{x}] + \gamma m \mathcal{I})^{-1} \mathbf{y}$ give

$$\mathbf{c}_{WD,\gamma} = \frac{1}{\lambda + \gamma m \lambda^{-1}} \mathbf{y}_\lambda = \frac{1}{\lambda} \frac{\lambda}{\lambda + \gamma m \lambda^{-1}} \mathbf{y}_\lambda = \frac{1}{\lambda} W_{WD,\gamma}(\lambda) \mathbf{y}_\lambda, \tag{14}$$

where $W_{WD,\gamma}(\lambda) \triangleq \frac{\lambda}{\lambda + \gamma m \lambda^{-1}}$, and

$$\mathbf{c}_{T,\gamma} = \frac{1}{\lambda + \gamma m} \mathbf{y}_\lambda = \frac{1}{\lambda} \frac{\lambda}{\lambda + \gamma m} \mathbf{y}_\lambda = \frac{1}{\lambda} W_{T,\gamma}(\lambda) \mathbf{y}_\lambda, \tag{15}$$

where $W_{T,\gamma}(\lambda) \triangleq \frac{\lambda}{\lambda + \gamma m}$.

Following the theory of spectral windows for inverse problems (Bertero 1989, Sect. V.C, pp. 84–88) $W_{WD,\gamma}$ and $W_{T,\gamma}$ play the roles of *filtering factors*, which allow one to compare the robustness of different regularization approaches with respect to noise in the input data. $W_{WD,\gamma}$ and $W_{T,\gamma}$ take on small values in the neighborhood of $\lambda = 0$, hence they filter out spectral components that are more sensitive to noise. Equations (14) and (15) express the coefficients of the optimal solutions as products of $1/\lambda$, which corresponds to the absence of regularization, times a filtering factor, which corresponds to the regularization term. Elementary calculations show that the two filtering factors $W_{WD,\gamma}(\lambda)$ and $W_{T,\gamma}(\lambda)$ have the same asymptotic behavior:

$$\lim_{\lambda \rightarrow 0} W_{WD,\gamma}(\lambda) = \lim_{\lambda \rightarrow 0} W_{T,\gamma}(\lambda) = 0$$

and

$$\lim_{\lambda \rightarrow +\infty} W_{WD,\gamma}(\lambda) = \lim_{\lambda \rightarrow +\infty} W_{T,\gamma}(\lambda) = 1.$$

Moreover, $W_{WD,\gamma}(\lambda) < W_{T,\gamma}(\lambda)$ if $\lambda < 1$, $W_{WD,\gamma}(\lambda) = W_{T,\gamma}(\lambda)$ if $\lambda = 1$, and $W_{WD,\gamma}(\lambda) > W_{T,\gamma}(\lambda)$ if $\lambda > 1$.

5 A mixed regularized learning technique

By considering, for $\gamma_T, \gamma_{WD} > 0$, the minimization of the *mixed regularized functional*

$$\Phi_{WDT, \gamma_T, \gamma_{WD}}(f) \triangleq \mathcal{E}_z(f) + \gamma_T \|f\|_K^2 + \gamma_{WD} \|\mathbf{c}_f\|_2^2,$$

it is possible to combine weight decay and Tikhonov's regularization. Within the context of learning in Sobolev spaces and with an (ideally) infinite number of samples, an analysis of such an approach was made in [Burger and Neubauer \(2002\)](#). For simplicity and without loss of generality, we let $\gamma_T = \gamma_{WD} = \frac{\gamma}{2}$ and define the *mixed-regularized learning problem*

$$\Phi_{WDT, \frac{\gamma}{2}}(f) \triangleq \Phi_{WDT, \frac{\gamma}{2}, \frac{\gamma}{2}}(f).$$

The next proposition investigates the problem $(\text{span}_m G_{K_x}, \Phi_{WDT, \frac{\gamma}{2}})$ and gives a formula for its solution.

Proposition 3 *Let X be a nonempty set, $K : X \times X \rightarrow \mathbb{R}$ a positive-definite kernel, m a positive integer, $\mathbf{x} = (x_1, \dots, x_m) \in X^m$, $\mathcal{K}[\mathbf{x}]$ the Gram matrix of the kernel K with respect to \mathbf{x} , $\mathbf{y} = (y_1, \dots, y_m) \in \mathbb{R}^m$, and $\gamma > 0$. There exists a unique solution*

$$f_{WDT, \frac{\gamma}{2}} = \sum_{i=1}^m c_{WDT, \frac{\gamma}{2}, i} K_{x_i} \quad (16)$$

to the problem $(\text{span}_m G_{K_x}, \Phi_{WDT, \frac{\gamma}{2}})$, where the coefficient vector $\mathbf{c}_{WDT, \frac{\gamma}{2}} = (c_{WDT, \frac{\gamma}{2}, 1}, \dots, c_{WDT, \frac{\gamma}{2}, m})$ is the unique solution to the linear system of equations

$$\left(\mathcal{K}[\mathbf{x}] + \frac{\gamma}{2} m (\mathcal{I} + \mathcal{K}^{-1}[\mathbf{x}]) \right) \mathbf{c}_{WDT, \frac{\gamma}{2}} = \mathbf{y}. \quad (17)$$

Proof For $f \in \text{span}_m G_{K_x}$,

$$\begin{aligned} \Phi_{WDT, \gamma_T, \gamma_{WD}}(f) &= \left\| \frac{\mathcal{K}[\mathbf{x}]\mathbf{c}_f}{\sqrt{m}} - \frac{\mathbf{y}}{\sqrt{m}} \right\|_2^2 + \gamma_T \|\mathcal{K}^{\frac{1}{2}}[\mathbf{x}]\mathbf{c}_f\|_2^2 + \gamma_{WD} \|\mathbf{c}_f\|_2^2 \\ &= \left\| \frac{\mathcal{K}[\mathbf{x}]\mathbf{c}_f}{\sqrt{m}} - \frac{\mathbf{y}}{\sqrt{m}} \right\|_2^2 + \gamma_T \left\| \left(\mathcal{K}[\mathbf{x}] + \frac{\gamma_{WD}}{\gamma_T} \mathcal{I} \right)^{\frac{1}{2}} \mathbf{c}_f \right\|_2^2. \end{aligned}$$

By the extension of Tikhonov's regularization from ([Bertero 1989](#), p. 79) setting $\gamma_T = \gamma_{WD} \triangleq \frac{\gamma}{2}$ allows the coefficients of the minimizer over $\text{span}_m G_{K_x}$ of $\Phi_{WDT, \frac{\gamma}{2}}$

to be given by

$$\begin{aligned} \mathbf{c}_{f_{WDT, \frac{\gamma}{2}}} &= \left(\frac{\mathcal{K}^2[\mathbf{x}]}{m} + \frac{\gamma}{2} (\mathcal{K}[\mathbf{x}] + \mathcal{I}) \right)^{-1} \frac{\mathcal{K}[\mathbf{x}]}{\sqrt{m}} \frac{\mathbf{y}}{\sqrt{m}} \\ &= \left(\mathcal{K}[\mathbf{x}] + \frac{\gamma}{2} m (\mathcal{I} + \mathcal{K}^{-1}[\mathbf{x}]) \right)^{-1} \mathbf{y}. \end{aligned}$$

□

Similarly to Proposition 1 and Theorem 1, Proposition 3 expresses the solution to the mixed learning problem as a linear combination of kernel functions centered in the data points, with the coefficients given by the solution to the linear system of Eq. (17). Numerical issues related to the solution of such a system will be discussed in Sect. 6.

Recall that we denote by $\lambda_{\max}(\mathcal{A})$ and $\lambda_{\min}(\mathcal{A})$ the maximum and minimum eigenvalues of a symmetric matrix \mathcal{A} , respectively, and that, to simplify the notation, we write λ_{\max} instead of $\lambda_{\max}(\mathcal{K}[\mathbf{x}])$ and similarly for λ_{\min} . By (17), elementary spectral theory arguments, and some algebra we obtain the following upper bound on the Euclidean norm of the coefficient vector of the solution:

$$\|\mathbf{c}_{f_{WDT, \frac{\gamma}{2}}}\|_2 \leq \max_{\lambda \in [\lambda_{\min}, \lambda_{\max}]} \frac{1}{\lambda + \frac{\gamma}{2} m(1 + \lambda^{-1})} \|\mathbf{y}\|_2 \leq \frac{1}{2\sqrt{\frac{\gamma}{2} m} + \frac{\gamma}{2} m} \|\mathbf{y}\|_2.$$

As $\|f_{WDT, \frac{\gamma}{2}}\|_K = \|\mathcal{K}^{\frac{1}{2}}[\mathbf{x}] (\mathcal{K}[\mathbf{x}] + \frac{\gamma}{2} m(\mathcal{I} + \mathcal{K}^{-1}[\mathbf{x}]))^{-1} \mathbf{y}\|_2$, by (16), (17), elementary spectral theory arguments, and some algebra, we get the following upper bound on the K -norm of the solution:

$$\begin{aligned} \|f_{WDT, \frac{\gamma}{2}}\|_K &\leq \max_{\lambda \in [\lambda_{\min}, \lambda_{\max}]} \frac{\sqrt{\lambda}}{\lambda + \frac{\gamma}{2} m(1 + \lambda^{-1})} \|\mathbf{y}\|_2 \\ &\leq \frac{\sqrt{\lambda^*}}{\lambda^* + \frac{\gamma}{2} m(1 + \lambda^{*-1})} \|\mathbf{y}\|_2, \end{aligned}$$

where $\lambda^* \triangleq \frac{\frac{\gamma}{2} m + \sqrt{(\frac{\gamma}{2} m)^2 + 12 \frac{\gamma}{2} m}}{2}$ is the value that maximizes $\frac{\sqrt{\lambda}}{\lambda + \frac{\gamma}{2} m(1 + \lambda^{-1})}$ over $[0, +\infty)$.

The expression $\mathbf{c}_{f_{WDT, \frac{\gamma}{2}}} = (\mathcal{K}[\mathbf{x}] + \frac{\gamma}{2} m (\mathcal{I} + \mathcal{K}^{-1}[\mathbf{x}]))^{-1} \mathbf{y}$ has to be compared with the expressions $\mathbf{c}_{f_{WD, \gamma}} = (\mathcal{K}[\mathbf{x}] + (\gamma m \mathcal{K}^{-1}[\mathbf{x}]))^{-1} \mathbf{y}$ and $\mathbf{c}_{f_{T, \gamma}} = (\gamma m \mathcal{I} + \mathcal{K}[\mathbf{x}])^{-1} \mathbf{y}$ for the coefficients of the solutions to the weight-decay and Tikhonov-regularized learning problems, respectively. An interesting comparison can be made in terms of filtering factors, when \mathbf{y}_λ is an eigenvector of $\mathcal{K}[\mathbf{x}]$ associated with the

eigenvalue λ . In this case, we get

$$\begin{aligned} \mathbf{c}_{f_{WDT, \frac{\gamma}{2}}} &= \frac{1}{\lambda + \frac{\gamma}{2} m (1 + \lambda^{-1})} \mathbf{y}^\lambda = \frac{1}{\lambda} \frac{\lambda}{\lambda + \frac{\gamma}{2} m (1 + \lambda^{-1})} \mathbf{y}^\lambda \\ &= \frac{1}{\lambda} W_{WDT, \frac{\gamma}{2}}(\lambda) \mathbf{y}^\lambda, \end{aligned}$$

where $W_{WDT, \frac{\gamma}{2}}(\lambda) \triangleq \frac{\lambda}{\lambda + \frac{\gamma}{2} m (1 + \lambda^{-1})}$ is the corresponding filtering factor.

Table 1 summarizes the solutions and the filtering factors for weight-decay learning, Tikhonov's-regularization learning, and the mixed weight-decay/Tikhonov's-regularization learning.

For a comparison with the previous two filtering factors, the behaviors of $W_{WD, \gamma}(\lambda)$, $W_{T, \gamma}(\lambda)$, and $W_{WDT, \frac{\gamma}{2}}(\lambda)$ are reported in Fig. 1. Note that the graph of $W_{WDT, \frac{\gamma}{2}}(\lambda)$ lies between those of $W_{T, \gamma}(\lambda)$ and $W_{WD, \gamma}(\lambda)$. The three curves intersect at $\lambda = 1$ (independently of the value of γm); hence, for an eigenvector associated with $\lambda = 1$ the three regularization terms are equal. A significant fact is that “small” values of λ in both $W_{WD, \gamma}(\lambda)$ and $W_{WDT, \frac{\gamma}{2}}(\lambda)$ are filtered out more than in $W_{T, \gamma}(\lambda)$ (indeed, computing the derivatives of the three filtering factors we obtain $W'_{WD, \gamma}(0) = W'_{WDT, \frac{\gamma}{2}}(0) = 0$, whereas $W'_{T, \gamma}(0) = \frac{1}{\gamma m}$). This gives a theoretical motivation for the use of weight decay (or weight decay combined with Tikhonov's regularization) in learning from data. Indeed, also for numerical reasons one may wish to have an upper bound on the size of the coefficients $c_{f,i}$ of the learned regressor $f = \sum_{i=1}^m c_{f,i} K_{x_i}$. However, when one has Tikhonov's regularization, the filtering factor $W_{T, \gamma}(\lambda)$ may not sufficiently penalize small values of λ . Thus, the learned regressor $f_{T, \gamma}$ corresponding to Tikhonov's regularization may have a large value of $\|c_f\|_2^2$ although it may have, for sufficiently large values of the regularization parameter, a small value of $\|f\|_K^2$.

In Fig. 2, we show that for different values of the parameters γ_1 and γ_2 it is possible to make the graphs of $W_{T, \gamma_1}(\lambda)$ and $W_{WDT, \frac{\gamma_2}{2}}(\lambda)$ “very similar” to each other, for sufficiently large values of λ , still preserving a more desirable smoothing behavior of $W_{WDT, \frac{\gamma_2}{2}}(\lambda)$ for sufficiently small λ . It is likely that the location of the threshold can be controlled more easily by allowing for different values of γ_T and γ_{WD} in the mixed regularized functional.

Finally, it is interesting to observe that, although the actual eigenvalues of $\mathcal{K}[\mathbf{x}]$ depend on the random extraction of data, the filtering factors $W_{WD, \gamma}(\lambda)$, $W_{T, \gamma}(\lambda)$, and $W_{WDT, \frac{\gamma}{2}}(\lambda)$ depend only on the product γm , i.e. the product between the regularization parameter and the number of samples.

6 Suboptimal solutions to weight-decay learning

Proposition 1, Theorem 1, and Proposition 3 give explicit formulas for the solutions to the learning problem regularized via weight decay, Tikhonov's regularization, and the mixed weight decay/Tikhonov approach, respectively. The expressions (9), (13), and (17) for the coefficients of the linear combinations providing the solutions to the

Table 1 The learning problems considered in this paper

Learning technique	Functional	Minimization problem	Linear system of equations	Filtering factor
WD	$\Phi_{WD,\gamma}(f) = \mathcal{E}_z(f) + \gamma \ \mathbf{e}_f\ _2^2$	$(\text{span}_m G_{K_X}, \Phi_{WD,\gamma})$	$(\mathcal{K}[\mathbf{X}] + \gamma m \mathcal{K}^{-1}[\mathbf{X}]) \mathbf{c} = \mathbf{y}$	$\frac{\lambda}{\lambda + \gamma m \lambda^{-1}}$
Tikhonov	$\Phi_{T,\gamma}(f) = \mathcal{E}_z(f) + \gamma \ f\ _K^2$	$(\mathcal{H}_K, \Phi_{T,\gamma})$	$(\gamma m \mathcal{I} + \mathcal{K}[\mathbf{X}]) \mathbf{c} = \mathbf{y}$	$\frac{\lambda}{\lambda + \gamma m}$
WD/Tikhonov	$\Phi_{WDT,\frac{\gamma}{2}}(f) = \mathcal{E}_z(f) + \frac{\gamma}{2} (\ \mathbf{e}_f\ _2^2 + \ f\ _K^2)$	$(\text{span}_m G_{K_X}, \Phi_{WDT,\frac{\gamma}{2}})$	$(\mathcal{K}[\mathbf{X}] + \frac{\gamma}{2} m (\mathcal{I} + \mathcal{K}^{-1}[\mathbf{X}])) \mathbf{c} = \mathbf{y}$	$\frac{\lambda}{\lambda + \frac{\gamma}{2} m (1 + \lambda^{-1})}$

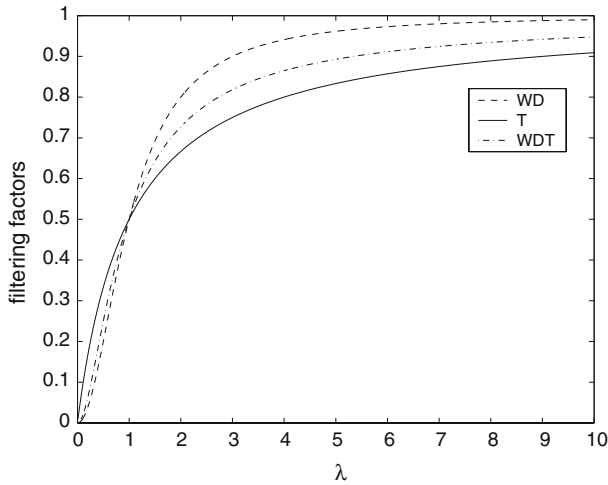


Fig. 1 Plots of the three filtering factors $W_{WD,\gamma}(\lambda)$, $W_{T,\gamma}(\lambda)$ and $W_{WDT,\frac{\gamma}{2}}(\lambda)$ for $\gamma m = 1$. Similar plots are obtained for the other values of γm

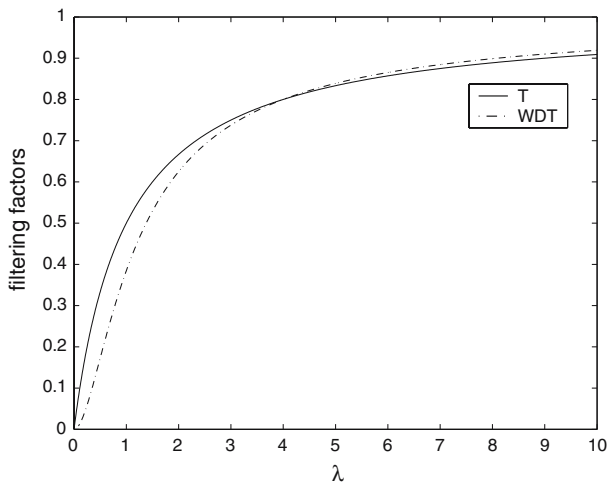


Fig. 2 Plots of the filtering factors $W_{T,\gamma_1}(\lambda)$ and $W_{WDT,\frac{\gamma_2}{2}}(\lambda)$ for $\gamma_1 m = 1$ and $\gamma_2 m = 1.6$

respective problems require one to solve linear systems of equations, so in principle they can be used to design learning algorithms. For algorithms based on the Representer Theorem, see, e.g., (Cucker and Smale, 2001, p. 42) and (Poggio and Smale, 2003, pp. 538-539). As discussed in Kůrková and Sanguineti (2005), their applications are limited by the rates of convergence of iterative methods solving the linear system of Eq. (13), which depend on the size of the condition number of its matrix. Similar drawbacks hold for the linear systems in (9) and (17).

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

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

where $\|\mathcal{A}\|$ denotes the norm of \mathcal{A} as a linear operator on $(\mathbb{R}^m, \|\cdot\|)$. It is easy to check that for every norm $\|\cdot\|$ on \mathbb{R}^m and every $m \times m$ nonsingular matrix \mathcal{A} , $\text{cond}(\mathcal{A}) \geq \frac{|\lambda_{\max}(\mathcal{A})|}{|\lambda_{\min}(\mathcal{A})|}$ (we denote by $\lambda_{\max}(\mathcal{A})$ and $\lambda_{\min}(\mathcal{A})$ the maximum and minimum eigenvalues of a symmetric matrix \mathcal{A} , respectively), and for every symmetric nonsingular $m \times m$ matrix \mathcal{A} , $\text{cond}_2(\mathcal{A}) = \frac{|\lambda_{\max}(\mathcal{A})|}{|\lambda_{\min}(\mathcal{A})|}$, where $\text{cond}_2(\mathcal{A})$ denotes the condition number of \mathcal{A} with respect to the $\|\cdot\|_2$ -norm on \mathbb{R}^m (Ortega 1990, p. 35).

As we consider positive-definite kernels, the matrix $\mathcal{K}[\mathbf{x}]$ is positive definite, so all its eigenvalues are positive (Ortega 1990, p. 7). By simple algebraic manipulations and spectral theory, for the condition numbers of the matrices involved in the solutions of the linear systems (9), (13), and (17), we get (recall that, to simplify the notation, we denote by λ_{\min} and λ_{\max} the minimum and maximum eigenvalues of the Gram matrix $\mathcal{K}[\mathbf{x}]$, respectively)

$$\begin{aligned} \text{cond}_2(\mathcal{K}[\mathbf{x}] + \gamma m \mathcal{K}^{-1}[\mathbf{x}]) &\leq \frac{\max_{\lambda \in [\lambda_{\min}, \lambda_{\max}]} \left\{ \lambda + \frac{\gamma m}{\lambda} \right\}}{\min_{\lambda \in [\lambda_{\min}, \lambda_{\max}]} \left\{ \lambda + \frac{\gamma m}{\lambda} \right\}} \\ &\leq \frac{\lambda_{\max} + \frac{\gamma m}{\lambda_{\min}}}{\lambda_{\min} + \frac{\gamma m}{\lambda_{\max}}} = \frac{\lambda_{\max}}{\lambda_{\min}} \frac{\lambda_{\min} \lambda_{\max} + \gamma m}{\lambda_{\max} \lambda_{\min} + \gamma m} \\ &= \text{cond}_2(\mathcal{K}[\mathbf{x}]), \end{aligned} \tag{18}$$

$$\text{cond}_2(\gamma m \mathcal{I} + \mathcal{K}[\mathbf{x}]) = \frac{\gamma m + \lambda_{\max}}{\gamma m + \lambda_{\min}} \leq \frac{\lambda_{\max}}{\lambda_{\min}} = \text{cond}_2(\mathcal{K}[\mathbf{x}]), \tag{19}$$

$$\text{cond}_2(\gamma m \mathcal{I} + \mathcal{K}[\mathbf{x}]) = \frac{\gamma m}{\gamma m + \lambda_{\min}} + \frac{\lambda_{\max}}{\gamma m + \lambda_{\min}} \leq 1 + \frac{\lambda_{\max}}{\gamma m}, \tag{20}$$

$$\begin{aligned} \text{cond}_2(\mathcal{K}[\mathbf{x}] + \frac{\gamma}{2} m (\mathcal{I} + \mathcal{K}^{-1}[\mathbf{x}])) &\leq \frac{\lambda_{\max} + \frac{\gamma}{2} m \left(1 + \frac{1}{\lambda_{\min}} \right)}{\lambda_{\min} + \frac{\gamma}{2} m \left(1 + \frac{1}{\lambda_{\max}} \right)} \\ &= \frac{\lambda_{\max}}{\lambda_{\min}} \frac{\lambda_{\min} \lambda_{\max} + \frac{\gamma}{2} m (\lambda_{\min} + 1)}{\lambda_{\max} \lambda_{\min} + \frac{\gamma}{2} m (\lambda_{\max} + 1)} \\ &\leq \text{cond}_2(\mathcal{K}[\mathbf{x}]), \end{aligned} \tag{21}$$

and

$$\begin{aligned} \text{cond}_2(\mathcal{K}[\mathbf{x}] + \frac{\gamma}{2} m (\mathcal{I} + \mathcal{K}^{-1}[\mathbf{x}])) &\leq \frac{\lambda_{\max}}{\lambda_{\min} + \frac{\gamma}{2} m \left(1 + \frac{1}{\lambda_{\max}} \right)} \\ &\quad + \frac{\frac{\gamma}{2} m \left(1 + \frac{1}{\lambda_{\min}} \right)}{\lambda_{\min} + \frac{\gamma}{2} m \left(1 + \frac{1}{\lambda_{\max}} \right)} \end{aligned}$$

$$\begin{aligned}
&\leq \frac{\lambda_{\max}}{\frac{\gamma}{2} m} + \frac{\lambda_{\max}}{\lambda_{\min}} \frac{\frac{\gamma}{2} m (\lambda_{\min} + 1)}{\lambda_{\min} \lambda_{\max} + \frac{\gamma}{2} m (\lambda_{\max} + 1)} \\
&\leq \frac{\lambda_{\max}}{\frac{\gamma}{2} m} + \frac{\lambda_{\max}}{\lambda_{\min}} \frac{\frac{\gamma}{2} m (\lambda_{\min} + 1)}{\lambda_{\min} \lambda_{\max} + \frac{\gamma}{2} m \lambda_{\max}} \\
&= \frac{\lambda_{\max}}{\frac{\gamma}{2} m} + \frac{\frac{\gamma}{2} m (\lambda_{\min} + 1)}{\lambda_{\min} (\lambda_{\min} + \frac{\gamma}{2} m)}. \tag{22}
\end{aligned}$$

When $\text{cond}_2(\mathcal{K}[\mathbf{x}])$ is sufficiently small, Eq. (18), (19), and (21) guarantee good conditioning of the respective matrices, for every value of γ . However, for large values of the size m of the data sample, the matrix $\mathcal{K}[\mathbf{x}]$ might be ill-conditioned. For example, when the data are uniformly distributed over an interval, the probability that $\mathcal{K}[\mathbf{x}]$ is ill-conditioned increases with m (see Cuesta-Albertos and Wschebor 2003, Theorem 2.2 and Demmel 1987, Theorem 5.1). On the other hand, as

$$\lim_{\gamma \rightarrow \infty} \left(1 + \frac{\lambda_{\max}}{\gamma m} \right) = 1$$

and

$$\lim_{\gamma \rightarrow \infty} \left(\frac{\lambda_{\max}}{\frac{\gamma}{2} m} + \frac{\frac{\gamma}{2} m (\lambda_{\min} + 1)}{\lambda_{\min} (\lambda_{\min} + \frac{\gamma}{2} m)} \right) = 1 + \frac{1}{\lambda_{\min}},$$

Equations (18), (20), and (22) guarantee that the regularization parameter γ can be chosen such that $\text{cond}_2(\mathcal{K}[\mathbf{x}] + \gamma m \mathcal{K}^{-1}[\mathbf{x}])$, $\text{cond}_2(\gamma m \mathcal{I} + \mathcal{K}[\mathbf{x}])$, and $\text{cond}_2(\mathcal{K}[\mathbf{x}] + \frac{\gamma}{2} m (\mathcal{I} + \mathcal{K}^{-1}[\mathbf{x}]))$ are close to $\text{cond}_2(\mathcal{K}[\mathbf{x}])$, 1, and $1 + \frac{1}{\lambda_{\min}}$, resp. Unfortunately, good conditioning of the matrices is not the only requirement for γ , as its value must also allow a good fit to the empirical data and thus cannot be too large. The problem of choosing γ in order to minimize the expected error was investigated in Cucker and Smale (2002).

When a value of γ guaranteeing both a small condition number and a good fit to the empirical data cannot be found, one has to develop algorithms for learning from data that differ from the one based on Proposition 1, Theorem 1, and Proposition 3. A variety of learning algorithms have been developed in the field of neurocomputing. Typically, such algorithms operate on connectionistic models with fewer computational units than the size m of the data sample used as a training set. 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 number m of data. Moreover, the values of the computational units' parameters (e.g., the centers in the case of RBF networks) are not necessarily set equal to the input vectors from the data sample, but are often optimized during learning.

In contrast with the respective optimal solutions, which are linear combinations of K_{x_1}, \dots, K_{x_m} determined by the sample $\mathbf{x} = (x_1, \dots, x_m)$ of input data, suboptimal solutions are formed by linear combinations of $n < m$ such functions, or they belong to arbitrary n -tuples of elements of $G_K \triangleq \{K_x : x \in X\}$. In applications, a proper n -tuple together with coefficients of the linear combination can be adjusted by a

suitable nonlinear programming algorithm, such as gradient descent (Bertsekas 1999, pp. 103–106, 173–174) (possibly with additive stochastic terms to avoid local minima), genetic algorithms (Goldberg 1989), and simulated annealing (Aarts and Korst 1989).

Estimates of the rates of approximation, by suboptimal solutions to the problems $(\text{span}_n G_{K_x}, \Phi_{T,\gamma})$, of the solution $f_{T,\gamma}$ provided by the Representer Theorem for the Tikhonov-regularized learning problem $(\mathcal{H}_K, \Phi_{T,\gamma})$ were derived in Kůrková and Sanguineti (2005). As, for every positive integer n , $\text{span}_n G_{K_x} \subseteq \text{span}_n G_K$, the estimates also hold for $(\text{span}_n G_K, \Phi_{T,\gamma})$. In the following, we investigate the accuracies of suboptimal solutions over $(\text{span}_n G_{K_x}, \Phi_{WD,\gamma})$ and $(\text{span}_n G_{K_x}, \Phi_{WDT,\frac{\gamma}{2}})$ to the solutions $f_{WD,\gamma}$, $f_{WDT,\frac{\gamma}{2}}$ provided by Propositions 1 and 3, resp. It is known that, in general, weight decay with the $\|\cdot\|_2$ norm does not guarantee the sparseness of the optimal solution (better sparseness properties can be obtained with the $\|\cdot\|_1$ norm Bishop 2006); this motivates our investigations on sparse suboptimal solutions.

To estimate the rates of approximation of $f_{WD,\gamma}$ and $f_{WDT,\frac{\gamma}{2}}$, which can be obtained by suboptimal solutions to the problems $(\text{span}_n G_{K_x}, \Phi_{WD,\gamma})$ and $(\text{span}_n G_{K_x}, \Phi_{WDT,\frac{\gamma}{2}})$, respectively, we shall exploit the following theorem from Kůrková and Sanguineti (2001). It is a special case of a reformulation (given in Kůrková 1997, see also Kůrková et al. 1998) of a result on the approximation of elements in the closure of the convex hull of a set, by n -tuples of its elements. Given an orthonormal basis F of a Hilbert space X , by $\|\cdot\|_{1,F}$ we denote the 1-norm with respect to F .

Theorem 2 (Kůrková and Sanguineti 2001, Theorem 2) *Let $(\mathcal{X}, \|\cdot\|)$ be a Hilbert space and F its orthonormal basis. For every $\phi \in \mathcal{X}$ and every positive integer n ,*

$$\|\phi - \text{span}_n F\| \leq \sqrt{\frac{\|\phi\|_{1,F}^2 - \|\phi\|^2}{n}}.$$

We start by estimating, for increasing values of the number $n < m$ of elements in the linear combinations, the rates of approximation of $f_{WD,\gamma}$ by suboptimal solutions to the problems $(\text{span}_n G_{K_x}, \Phi_{WD,\gamma})$. For $f = \sum_{i=1}^n c_{f,i} K_{x_i} \in \text{span}_n G_{K_x}$, we let

$$\tilde{\Phi}_{WD,\gamma}(\mathbf{c}_f) \triangleq \Phi_{WD,\gamma} \left(\sum_{i=1}^n c_{f,i} K_{x_i} \right).$$

So, $\tilde{\Phi}_{WD,\gamma}$ is an n -variable function obtained from the functional $\Phi_{WD,\gamma}$ for $f = \sum_{i=1}^n c_{f,i} K_{x_i}$.

The next proposition states the continuity and convexity properties of $\tilde{\Phi}_{WD,\gamma}$.

Proposition 4 *Let X be a nonempty set, $K : X \times X \rightarrow \mathbb{R}$ a positive-definite kernel, m a positive integer, $\mathbf{x} = (x_1, \dots, x_m) \in X^m$, $\mathbf{y} = (y_1, \dots, y_m) \in \mathbb{R}^m$, $|\mathbf{y}|_{\max} \triangleq \max\{|y_i| : i = 1, \dots, m\}$, $\gamma > 0$, and λ_{\max} the maximum eigenvalue of the Gram matrix $\mathcal{K}[\mathbf{x}]$. Then for every positive integer $n \leq m$*

- (i) *the function $\tilde{\Phi}_{WD,\gamma}$ is uniformly convex on $\text{span}_n G_{K_x}$, with a modulus of convexity γt^2 ;*

- (ii) for every $f = \sum_{i=1}^n c_{f,i} K_{x_i} \in \text{span}_n G_{K_x}$, the function $\tilde{\Phi}_{WD,\gamma}$ is continuous, with a modulus of continuity bounded from above by the function $\alpha_{\mathbf{c}_f}(t) \triangleq a_2 t^2 + a_1 t$, where $a_1 \triangleq 2(\sqrt{\lambda_{\max}} \|f\|_K s_K^2 + \sqrt{\lambda_{\max}} |y|_{\max} s_K + \gamma \|\mathbf{c}_f\|_2)$ and $a_2 \triangleq \lambda_{\max} s_K^2 + \gamma$.

Proof (i) For $f = \sum_{i=1}^n c_{f,i} K_{x_i}$, the functional \mathcal{E}_z becomes a function of \mathbf{c}_f , since the kernel functions are fixed; we denote it by $\tilde{\mathcal{E}}_z(\mathbf{c}_f)$. As $\tilde{\mathcal{E}}_z$ is convex and $\gamma \|\cdot\|_2^2$ is uniformly convex with a modulus of convexity $\delta(t) = \gamma t^2$ (see, e.g., (Kůrková and Sanguineti, 2005, Proposition 2.1 (ii))), the function $\tilde{\Phi}_{T,\gamma}(\cdot) = \tilde{\mathcal{E}}_z(\cdot) + \gamma \|\cdot\|_2^2$ is uniformly convex, too, with a modulus of convexity $\delta(t) = \gamma t^2$ (see, e.g., (Kůrková and Sanguineti, 2005, Proposition 2.1 (i))).

- (ii) From the definition of $\Phi_{WD,\gamma}$ we derive

$$|\Phi_{WD,\gamma}(f) - \Phi_{WD,\gamma}(g)| \leq |\mathcal{E}_z(f) - \mathcal{E}_z(g)| + \gamma \left| \|\mathbf{c}_f\|_2^2 - \|\mathbf{c}_g\|_2^2 \right|. \tag{23}$$

Let $f, g \in \text{span}_n G_{K_x}$ be such that $f = \sum_{i=1}^m c_{f,i} K_{x_i}$ and $g = \sum_{i=1}^m c_{g,i} K_{x_i}$ with $\|\mathbf{c}_f - \mathbf{c}_g\|_2 < t$, and take $t > 0$. Then $\|f - g\|_K \leq \|\mathcal{K}^{\frac{1}{2}}[\mathbf{x}](\mathbf{c}_f - \mathbf{c}_g)\|_2 \leq \sqrt{\lambda_{\max}} \|\mathbf{c}_f - \mathbf{c}_g\|_2 < \sqrt{\lambda_{\max}} t$. For every $u > 0$ and f, g such that $\|f - g\|_K \leq u$,

$$\begin{aligned} |\mathcal{E}_z(f) - \mathcal{E}_z(g)| &\leq \frac{1}{m} \left| \sum_{i=1}^m (f(\mathbf{x}_i) - y_i)^2 - (g(\mathbf{x}_i) - y_i)^2 \right| \\ &= \frac{1}{m} \left| \sum_{i=1}^m (f(\mathbf{x}_i) - g(\mathbf{x}_i)) (f(\mathbf{x}_i) + g(\mathbf{x}_i) - 2y_i) \right| \\ &\leq \sup_{\mathbf{x} \in X} |f(\mathbf{x}) - g(\mathbf{x})| \left(\sup_{\mathbf{x} \in X} |f(\mathbf{x}) + g(\mathbf{x})| + 2|y|_{\max} \right) \\ &\leq u s_K (s_K \|f + g\|_K + 2|y|_{\max}) \\ &\leq u s_K (2 s_K \|f\|_K + u s_K + 2|y|_{\max}) \\ &= 2(\|f\|_K s_K^2 + |y|_{\max} s_K) u + s_K^2 u^2. \end{aligned} \tag{24}$$

So, for $\|\mathbf{c}_f - \mathbf{c}_g\|_2 < t$, we get

$$|\tilde{\mathcal{E}}_z(\mathbf{c}_f) - \tilde{\mathcal{E}}_z(\mathbf{c}_g)| < 2(\|f\|_K s_K^2 + |y|_{\max} s_K) \sqrt{\lambda_{\max}} t + s_K^2 \lambda_{\max} t^2. \tag{25}$$

For $\|\mathbf{c}_f - \mathbf{c}_g\|_2 < t$, the regularization term gives

$$\begin{aligned} \gamma |\|\mathbf{c}_f\|_2^2 - \|\mathbf{c}_g\|_2^2| &= \gamma (\|\mathbf{c}_f\|_2 + \|\mathbf{c}_g\|_2) \|\mathbf{c}_f\|_2 - \|\mathbf{c}_g\|_2 \\ &\leq \gamma (2\|\mathbf{c}_f\|_2 + \|\mathbf{c}_f - \mathbf{c}_g\|_2) \|\mathbf{c}_f - \mathbf{c}_g\|_2 \\ &< \gamma (2\|\mathbf{c}_f\|_2 + t) t. \end{aligned} \tag{26}$$

The statement follows from (23), (25), and (26). □

The next theorem estimates, for increasing values of $n < m$, the rates of approximation of $f_{WD,\gamma}$, which can be obtained by suboptimal solutions to the problems $(\text{span}_n G_{K_X}, \Phi_{WD,\gamma})$.

Theorem 3 *Let X be a nonempty set, $K : X \times X \rightarrow \mathbb{R}$ a positive definite kernel, $s_K = \sup_{x \in X} \sqrt{K(x, x)}$, \mathbf{z} a data sample of size m , $|y|_{\max} \triangleq \max\{|y_i| : i = 1, \dots, m\}$, λ_{\min} and λ_{\max} the minimum and the maximum eigenvalues of the Gram matrix $\mathcal{K}[\mathbf{x}]$, resp., and $\gamma > 0$. Let $b_1 \triangleq 2(\sqrt{\lambda_{\max}} \|f_{WD,\gamma}\|_K s_K^2 + \sqrt{\lambda_{\max}} |y|_{\max} s_K + \gamma \|\mathbf{c}_{WD,\gamma}\|_2)$, $b_2 \triangleq \lambda_{\max} s_K^2 + \gamma$, $\alpha(t) \triangleq b_2 t^2 + b_1 t$, and $\Delta_{WD,\gamma} \triangleq \|\mathbf{c}_{WD,\gamma}\|_1^2 - \|\mathbf{c}_{WD,\gamma}\|_2^2$.*

(i) *For every positive integer $n < m$*

$$\inf_{f \in \text{span}_n G_{K_X}} \Phi_{WD,\gamma}(f) - \Phi_{WD,\gamma}(f_{WD,\gamma}) \leq \alpha\left(\sqrt{\frac{\Delta_{WD,\gamma}}{n}}\right).$$

(ii) *Let $\varepsilon_n \geq 0$ and $f_n \in \text{argmin}_{\varepsilon_n}(\text{span}_n G_{K_X}, \Phi_{WD,\gamma})$. Then*

$$\|f_n - f_{WD,\gamma}\|_K^2 \leq \frac{\lambda_{\max}}{\gamma} \left[\alpha\left(\sqrt{\frac{\Delta_{WD,\gamma}}{n}}\right) + \varepsilon_n \right].$$

Proof (i) Let $f_{WD,\gamma} \triangleq \sum_{i=1}^m c_{WD,\gamma,i} K_{x_i} \in \text{span}_m G_{K_X}$. According to Theorem 2 with $(\mathcal{X}, \|\cdot\|) = (\mathbb{R}^m, \|\cdot\|_2)$ and $F = \{\mathbf{e}_1, \dots, \mathbf{e}_m\}$ (i.e., the canonical orthonormal basis of \mathbb{R}^m), there exists $\hat{\mathbf{c}}_{WD,\gamma} \in \text{span}_n \{\mathbf{e}_1, \dots, \mathbf{e}_m\}$ such that

$$\|\mathbf{c}_{WD,\gamma} - \hat{\mathbf{c}}_{WD,\gamma}\|_2 \leq \sqrt{\frac{\|\mathbf{c}_{WD,\gamma}\|_1^2 - \|\mathbf{c}_{WD,\gamma}\|_2^2}{n}}. \tag{27}$$

Let $\hat{f}_{WD,\gamma} \triangleq \sum_{i=1}^m \hat{c}_{WD,\gamma,i} K_{x_i} \in \text{span}_m G_{K_X}$. By (27) and Proposition 4 (ii), we get

$$\begin{aligned} \Phi_{WD,\gamma}(\hat{f}_{WD,\gamma}) - \Phi_{WD,\gamma}(f_{WD,\gamma}) &= \tilde{\Phi}_{WD,\gamma}(\hat{\mathbf{c}}_{WD,\gamma}) - \tilde{\Phi}_{WD,\gamma}(\mathbf{c}_{WD,\gamma}) \\ &\leq \alpha(\|\mathbf{c}_{WD,\gamma} - \hat{\mathbf{c}}_{WD,\gamma}\|_2) \leq \alpha\left(\sqrt{\frac{\Delta_{WD,\gamma}}{n}}\right) \end{aligned}$$

(ii) By Proposition 4 (i) and the properties of the modulus of convexity (see, e.g., (Kůrková and Sanguineti, 2005, Proposition 2.1 (iii))), we obtain

$$\begin{aligned} \gamma \|\mathbf{c}_{f_n} - \mathbf{c}_{f_{WD,\gamma}}\|^2 &\leq \tilde{\Phi}_{WD,\gamma}(\mathbf{c}_{f_n}) - \tilde{\Phi}_{WD,\gamma}(\mathbf{c}_{f_{WD,\gamma}}) \\ &= \Phi_{WD,\gamma}(f_n) - \Phi_{WD,\gamma}(f_{WD,\gamma}) \end{aligned} \tag{28}$$

For every $f, g \in \text{span}_m G_{K_X}$ such that $f = \sum_{i=1}^m c_{f,i} K_{x_i}$ and $g = \sum_{i=1}^m c_{g,i} K_{x_i}$ we have $\|f - g\|_K \leq \|\mathcal{K}^{\frac{1}{2}}[\mathbf{x}](\mathbf{c}_f - \mathbf{c}_g)\|_2 \leq \sqrt{\lambda_{\max}} \|\mathbf{c}_f - \mathbf{c}_g\|_2$. So, $\|f_n - f_{WD,\gamma}\|_K^2 \leq \lambda_{\max} \|\mathbf{c}_{f_n} - \mathbf{c}_{f_{WD,\gamma}}\|_2^2$ and we conclude by (i) and (28). \square

By $\mathbf{c}_{WD,\gamma} = (\mathcal{K}[\mathbf{x}] + \gamma m\mathcal{K}^{-1}[\mathbf{x}])^{-1} \mathbf{y}$, the definition of $\|\cdot\|_1$ norm for a matrix (Golub and Loan (1996), Chap. 2.3) spectral theory, and simple calculations we get the following upper bound on $\Delta_{WD,\gamma}$:

$$\begin{aligned} \Delta_{WD,\gamma} &\leq \left\| (\mathcal{K}[\mathbf{x}] + \gamma m\mathcal{K}^{-1})^{-1} \right\|_1^2 \|\mathbf{y}\|_1^2 - \min_{\lambda \in [\lambda_{\min}, \lambda_{\max}]} \frac{1}{(\lambda + \gamma m\lambda^{-1})^2} \|\mathbf{y}\|_2^2 \\ &= \left\| (\mathcal{K}[\mathbf{x}] + \gamma m\mathcal{K}^{-1})^{-1} \right\|_1^2 \|\mathbf{y}\|_1^2 - \min_{\lambda \in \{\lambda_{\min}, \lambda_{\max}\}} \frac{1}{(\lambda + \gamma m\lambda^{-1})^2} \|\mathbf{y}\|_2^2. \end{aligned}$$

Finally, we investigate the accuracy of suboptimal solutions to the mixed weight-decay/Tikhonov learning problem. For $f = \sum_{i=1}^n c_{f,i} K_{x_i} \in \text{span}_n G_{K_x}$, we let

$$\tilde{\Phi}_{WD T, \frac{\gamma}{2}}(\mathbf{c}_f) \triangleq \Phi_{WD T, \frac{\gamma}{2}} \left(\sum_{i=1}^n c_{f,i} K_{x_i} \right).$$

So, $\tilde{\Phi}_{WD T, \frac{\gamma}{2}}$ is a n -variable function obtained from the functional $\Phi_{WD T, \frac{\gamma}{2}}$ in correspondence of $f = \sum_{i=1}^n c_{f,i} K_{x_i}$.

The next proposition states the continuity and convexity properties of $\tilde{\Phi}_{WD T, \frac{\gamma}{2}}$.

Proposition 5 *Let X be a nonempty set, $K : X \times X \rightarrow \mathbb{R}$ a positive-definite kernel, m a positive integer, $\mathbf{x} = (x_1, \dots, x_m) \in X^m$, $\mathbf{y} = (y_1, \dots, y_m) \in \mathbb{R}^m$, $|y|_{\max} \triangleq \max\{|y_i| : i = 1, \dots, m\}$, $\gamma > 0$, and λ_{\min} , λ_{\max} the minimum and maximum eigenvalues of the Gram matrix $\mathcal{K}[\mathbf{x}]$, respectively. Then for every positive integer $n \leq m$*

(i) *the function $\tilde{\Phi}_{WD T, \frac{\gamma}{2}}$ is uniformly convex on $\text{span}_n G_{K_x}$, with a modulus of convexity $\frac{\gamma}{2} (\lambda_{\min} + 1) t^2$;*

(ii) *for every $f = \sum_{i=1}^n c_{f,i} K_{x_i} \in \text{span}_n G_{K_x}$, the function $\tilde{\Phi}_{WD T, \frac{\gamma}{2}}$ is continuous, with a modulus of continuity bounded from above by the function $\beta_{\mathbf{c}_f}(t) \triangleq a_2 t^2 + a_1 t$, where $a_1 \triangleq 2(\sqrt{\lambda_{\max}} \|f\|_K s_K^2 + \sqrt{\lambda_{\max}} |y|_{\max} s_K + \frac{\gamma}{2} (\lambda_{\max} + 1) \|\mathbf{c}_f\|_2)$ and $a_2 \triangleq \lambda_{\max} s_K^2 + \frac{\gamma}{2} (\lambda_{\max} + 1)$.*

Proof (i) We first show that the modulus of convexity of the function $\|\mathcal{A}\mathbf{x}\|_2^2$ is bounded from above by the quadratic function $\lambda_{\min}^2(\mathcal{A}) t^2$, where \mathcal{A} is an $m \times m$ positive-definite matrix and $\lambda_{\min}(\mathcal{A})$ is its minimum eigenvalue. By the definition of uniform convexity, denoting by $\langle \cdot, \cdot \rangle$ the inner product in \mathbb{R}^d , for every $s \in [0, 1]$ and every $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ we

get

$$\begin{aligned} \|\mathcal{A}(s\mathbf{x} + (1-s)\mathbf{y})\|_2^2 &= \langle s\mathcal{A}\mathbf{x} + (1-s)\mathcal{A}\mathbf{y}, s\mathcal{A}\mathbf{x} + (1-s)\mathcal{A}\mathbf{y} \rangle \\ &= s^2 \langle \mathcal{A}\mathbf{x}, \mathcal{A}\mathbf{x} \rangle + (1-s)^2 \langle \mathcal{A}\mathbf{y}, \mathcal{A}\mathbf{y} \rangle + 2s(1-s) \langle \mathcal{A}\mathbf{x}, \mathcal{A}\mathbf{y} \rangle \\ &= s \langle \mathcal{A}\mathbf{x}, \mathcal{A}\mathbf{x} \rangle - s(1-s) \langle \mathcal{A}\mathbf{x}, \mathcal{A}\mathbf{x} \rangle + (1-s) \langle \mathcal{A}\mathbf{y}, \mathcal{A}\mathbf{y} \rangle \\ &\quad - s(1-s) \langle \mathcal{A}\mathbf{y}, \mathcal{A}\mathbf{y} \rangle + 2s(1-s) \langle \mathcal{A}\mathbf{x}, \mathcal{A}\mathbf{y} \rangle \\ &= s \langle \mathcal{A}\mathbf{x}, \mathcal{A}\mathbf{x} \rangle + (1-s) \langle \mathcal{A}\mathbf{y}, \mathcal{A}\mathbf{y} \rangle - s(1-s) \langle \mathcal{A}(\mathbf{x}-\mathbf{y}), \mathcal{A}(\mathbf{x}-\mathbf{y}) \rangle \\ &\leq s \langle \mathcal{A}\mathbf{x}, \mathcal{A}\mathbf{x} \rangle + (1-s) \langle \mathcal{A}\mathbf{y}, \mathcal{A}\mathbf{y} \rangle - s(1-s) \lambda_{\min}^2 \|\mathbf{x}-\mathbf{y}\|_2^2. \end{aligned}$$

Hence, $\frac{\gamma}{2} \|(\mathcal{K}[\mathbf{x}] + \mathcal{I})^{\frac{1}{2}}(\cdot)\|_2^2$ is uniformly convex, with a modulus of convexity $\delta(t) = \frac{\gamma}{2}(\lambda_{\min} + 1)t^2$. Let $\tilde{\mathcal{E}}_{\mathbf{z}}(\mathbf{c}_f) \triangleq \mathcal{E}_{\mathbf{z}}(f)$, considered for $f = \sum_{i=1}^n c_{f,i} K_{x_i}$ as a function of \mathbf{c}_f (since the kernel functions are fixed). Then also $\tilde{\mathcal{E}}_{\mathbf{z}}(\cdot) + \frac{\gamma}{2} \|(\mathcal{K}^{\frac{1}{2}}[\mathbf{x}])(\cdot)\|_2^2 + \|\cdot\|_2^2$ is uniformly convex with modulus of convexity $\delta(t) = \frac{\gamma}{2}(\lambda_{\min} + 1)t^2$ (see, e.g., (Kůrková and Sanguineti, 2005, Proposition 2.1 (i))).

(ii) By the definition of $\Phi_{WDT, \frac{\gamma}{2}}$ we have

$$\begin{aligned} |\Phi_{WDT, \frac{\gamma}{2}}(f) - \Phi_{WDT, \frac{\gamma}{2}}(g)| &\leq |\mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}_{\mathbf{z}}(g)| \\ &\quad + \frac{\gamma}{2} \left| \|(\mathcal{K}[\mathbf{x}] + \mathcal{I})^{\frac{1}{2}}\mathbf{c}_f\|_2^2 - \|(\mathcal{K}[\mathbf{x}] + \mathcal{I})^{\frac{1}{2}}\mathbf{c}_g\|_2^2 \right|. \end{aligned} \tag{29}$$

Let $f, g \in \text{span}_n G_{K_x}$ be such that $f = \sum_{i=1}^n c_{f,i} K_{x_i}$ and $g = \sum_{i=1}^n c_{g,i} K_{x_i}$ and take $t > 0$ such that $\|\mathbf{c}_f - \mathbf{c}_g\|_2 < t$. Then $\|f - g\|_K \leq \|(\mathcal{K}^{\frac{1}{2}}[\mathbf{x}])(\mathbf{c}_f - \mathbf{c}_g)\|_2 \leq \sqrt{\lambda_{\max}} \|\mathbf{c}_f - \mathbf{c}_g\|_2 < \sqrt{\lambda_{\max}} t$.

For $\|\mathbf{c}_f - \mathbf{c}_g\|_2 < t$, the regularization term gives

$$\begin{aligned} \frac{\gamma}{2} \left| \|(\mathcal{K}[\mathbf{x}] + \mathcal{I})^{\frac{1}{2}}\mathbf{c}_f\|_2^2 - \|(\mathcal{K}[\mathbf{x}] + \mathcal{I})^{\frac{1}{2}}\mathbf{c}_g\|_2^2 \right| &\leq \frac{\gamma}{2} \left(2\|(\mathcal{K}[\mathbf{x}] + \mathcal{I})^{\frac{1}{2}}\mathbf{c}_f\|_2 \right. \\ &\quad \left. + \|(\mathcal{K}[\mathbf{x}] + \mathcal{I})^{\frac{1}{2}}(\mathbf{c}_f - \mathbf{c}_g)\|_2 \right) \|(\mathcal{K}[\mathbf{x}] + \mathcal{I})^{\frac{1}{2}}(\mathbf{c}_f - \mathbf{c}_g)\|_2 \\ &< \frac{\gamma}{2} \left(2\sqrt{\lambda_{\max} + 1}\|\mathbf{c}_f\|_2 + \sqrt{\lambda_{\max} + 1}t \right) \sqrt{\lambda_{\max} + 1}t. \end{aligned} \tag{30}$$

The statement follows from (24), (29), and (30). □

The next theorem estimates, for increasing values of $n < m$, the rates of approximation of $f_{WDT, \frac{\gamma}{2}}$, which can be obtained by suboptimal solutions to the problems $(\text{span}_n G_{K_x}, \Phi_{WDT, \frac{\gamma}{2}})$.

Theorem 4 *Let X be a nonempty set, $K : X \times X \rightarrow \mathbb{R}$ a positive definite kernel, $s_K = \sup_{x \in X} \sqrt{K(x, x)}$, \mathbf{z} a data sample of size m , $|y|_{\max} \triangleq \max\{|y_i| : i = 1, \dots, m\}$, λ_{\min} and λ_{\max} the minimum and maximum eigenvalues of the Gram matrix $\mathcal{K}[\mathbf{x}]$, resp., and $\gamma > 0$. Moreover, let $c_1 \triangleq 2(\sqrt{\lambda_{\max}} \|f_{WDT, \frac{\gamma}{2}}\|_K s_K^2 + \sqrt{\lambda_{\max}} |y|_{\max} s_K +$*

$\frac{\gamma}{2}(\lambda_{\max} + 1)\|\mathbf{c}_{f_{WDT, \frac{\gamma}{2}}}\|_2$, $c_2 \triangleq \lambda_{\max} s_K^2 + \frac{\gamma}{2}(\lambda_{\max} + 1)$, $\beta(t) \triangleq c_1 t + c_2 t^2$, and $\Delta_{WDT, \frac{\gamma}{2}} \triangleq \|\mathbf{c}_{WDT, \frac{\gamma}{2}}\|_1^2 - \|\mathbf{c}_{WDT, \frac{\gamma}{2}}\|_2^2$.

(i) For every positive integer $n < m$

$$\inf_{f \in \text{span}_n G_{K_x}} \Phi_{WDT, \frac{\gamma}{2}}(f) - \Phi_{WDT, \frac{\gamma}{2}}(f_{WDT, \frac{\gamma}{2}}) \leq \beta \left(\sqrt{\frac{\Delta_{WDT, \frac{\gamma}{2}}}{n}} \right)$$

(ii) Let $\varepsilon_n \geq 0$ and $f_n \in \text{argmin}_{\varepsilon_n}(\text{span}_n G_{K_x}, \Phi_{WDT, \frac{\gamma}{2}})$. Then

$$\|f_n - f_{WDT, \frac{\gamma}{2}}\|_K^2 \leq 2 \frac{\lambda_{\max}}{\gamma(\lambda_{\min} + 1)} \left[\beta \left(\sqrt{\frac{\Delta_{WDT, \frac{\gamma}{2}}}{n}} \right) + \varepsilon_n \right].$$

Proof (i) Let $f_{WDT, \frac{\gamma}{2}} \triangleq \sum_{i=1}^m c_{WDT, \frac{\gamma}{2}, i} K_{x_i} \in \text{span}_m G_{K_x}$. By Theorem 2 with $(\mathcal{X}, \|\cdot\|) = (\mathbb{R}^m, \|\cdot\|_2)$ and $F = \{\mathbf{e}_1, \dots, \mathbf{e}_m\}$ (i.e., the canonical orthonormal basis of \mathbb{R}^m), there exists $\hat{\mathbf{c}}_{WDT, \frac{\gamma}{2}} \in \text{span}_n \{\mathbf{e}_1, \dots, \mathbf{e}_m\}$ such that

$$\|\mathbf{c}_{WDT, \frac{\gamma}{2}} - \hat{\mathbf{c}}_{WDT, \frac{\gamma}{2}}\|_2 \leq \sqrt{\frac{\|\mathbf{c}_{WDT, \frac{\gamma}{2}}\|_1^2 - \|\mathbf{c}_{WDT, \frac{\gamma}{2}}\|_2^2}{n}}. \tag{31}$$

Let $\hat{f}_{WDT, \frac{\gamma}{2}} \triangleq \sum_{i=1}^m \hat{c}_{WDT, \frac{\gamma}{2}, i} K_{x_i} \in \text{span}_m G_{K_x}$. By (31) and Proposition 5 (ii), we have

$$\begin{aligned} \Phi_{WDT, \frac{\gamma}{2}}(\hat{f}_{WDT, \frac{\gamma}{2}}) - \Phi_{WDT, \frac{\gamma}{2}}(f_{WDT, \frac{\gamma}{2}}) &= \tilde{\Phi}_{WDT, \frac{\gamma}{2}}(\hat{\mathbf{c}}_{WDT, \frac{\gamma}{2}}) \\ &\quad - \tilde{\Phi}_{WDT, \frac{\gamma}{2}}(\mathbf{c}_{WDT, \frac{\gamma}{2}}) \\ &\leq \beta \left(\|\mathbf{c}_{WDT, \frac{\gamma}{2}} - \hat{\mathbf{c}}_{WDT, \frac{\gamma}{2}}\|_2 \right) \\ &\leq \beta \left(\sqrt{\frac{\Delta_{WDT, \frac{\gamma}{2}}}{n}} \right) \end{aligned}$$

(ii) By Proposition 5 (i) and the properties of the modulus of convexity (see, e.g., (Kůrková and Sanguineti, 2005, Proposition 2.1 (iii))), we get

$$\begin{aligned} \frac{\gamma}{2}(\lambda_{\min} + 1)\|\mathbf{c}_{f_h} - \mathbf{c}_{f_{WDT, \frac{\gamma}{2}}}\|_2^2 &\leq \tilde{\Phi}_{WDT, \frac{\gamma}{2}}(\mathbf{c}_{f_h}) - \tilde{\Phi}_{WDT, \frac{\gamma}{2}}(\mathbf{c}_{f_{WDT, \frac{\gamma}{2}}}) \\ &= \Phi_{WDT, \frac{\gamma}{2}}(f_h) - \Phi_{WDT, \frac{\gamma}{2}}(f_{WDT, \frac{\gamma}{2}}). \end{aligned} \tag{32}$$

For every $f, g \in \text{span}_m G_{K_x}$ such that $f = \sum_{i=1}^m c_{f,i} K_{x_i}$ and $g = \sum_{i=1}^m c_{g,i} K_{x_i}$ we get $\|f - g\|_K \leq \|\mathcal{K}^{\frac{1}{2}}[\mathbf{x}](\mathbf{c}_f - \mathbf{c}_g)\|_2 \leq \sqrt{\lambda_{\max}} \|\mathbf{c}_f - \mathbf{c}_g\|_2$. So, $\|f_n - f_{WDT, \frac{\gamma}{2}}\|_K^2 \leq \lambda_{\max} \|\mathbf{c}_{f_n} - \mathbf{c}_{f_{WDT, \frac{\gamma}{2}}}\|_2^2$ and we conclude by (i) and (32). \square

By $\mathbf{c}_{WDT, \frac{\gamma}{2}} = (\mathcal{K}[\mathbf{x}] + \frac{\gamma}{2} m(\mathcal{I} + \mathcal{K}^{-1}))^{-1}$, elementary spectral theory arguments, the definition of the $\|\cdot\|_1$ norm for a matrix (Golub and Loan 1996, Chap. 2.3), and some algebra we have the following upper bound on $\Delta_{WDT, \frac{\gamma}{2}}$:

$$\begin{aligned} \Delta_{WDT, \frac{\gamma}{2}} &\leq \left\| \left(\mathcal{K}[\mathbf{x}] + \frac{\gamma}{2} m(\mathcal{I} + \mathcal{K}^{-1}) \right)^{-1} \right\|_1^2 \|\mathbf{y}\|_1^2 \\ &\quad - \min_{\lambda \in [\lambda_{\min}, \lambda_{\max}]} \frac{1}{\left(\lambda + \frac{\gamma}{2} m(1 + \lambda^{-1}) \right)^2} \|\mathbf{y}\|_2^2 \\ &= \left\| \left(\mathcal{K}[\mathbf{x}] + \frac{\gamma}{2} m(\mathcal{I} + \mathcal{K}^{-1}) \right)^{-1} \right\|_1^2 \|\mathbf{y}\|_1^2 \\ &\quad - \min_{\lambda \in \{\lambda_{\min}, \lambda_{\max}\}} \frac{1}{\left(\lambda + \frac{\gamma}{2} m(1 + \lambda^{-1}) \right)^2} \|\mathbf{y}\|_2^2. \end{aligned}$$

7 Conclusions

We have investigated, from the point of view of regularization, the learning technique known as “weight decay” and we have compared it with learning techniques based on Tikhonov’s regularization and the combination of the latter with weight decay. We have modeled the corresponding learning tasks as minimizations of error functionals over certain hypothesis spaces (or some subsets of theirs), which allow one to model generalization capabilities.

For data samples of size m , we have expressed the solutions to these learning problems as linear combinations of m computational units determined by the kinds of hypothesis space, for which the coefficients can be obtained by solving suitable linear systems of equations.

Motivated by the efficiency of implementation and by favorable computational requirements, we have investigated the accuracies of suboptimal solutions obtainable by $n < m$ computational units and their rates of approximation of the optimal solutions to the respective learning problems. We have shown that suboptimal solutions approximate the optimal ones with an error bounded from above by a quantity of the form $A/\sqrt{n} + B/n$, where A and B depend on the properties of the output data vector \mathbf{y} , the properties of the kernel, and the regularization parameter γ . So, when solving the systems of linear equations is not computationally feasible or when the systems are ill-conditioned, algorithms operating on models with $n < m$ computational units provide useful alternatives, as they can approximate the optimal solutions quite well. The minimizations required by these algorithms entail nonlinear programming problems (Poggio and Girosi 1990, p. 1489), which can be solved by iterative methods such as gradient descent (Bertsekas 1999, pp. 103–106, 173–174) (possibly with additive stochastic terms to avoid local minima), genetic algorithms (Goldberg 1989), and simulated annealing (Aarts and Korst 1989).

Suboptimal solutions to learning by weight decay and Tikhonov’s regularization were investigated in Burger and Neubauer (2002) for function approximation by neural networks in Sobolev spaces, with an ideally infinite number of samples. For Tikhonov’s

regularization, an extension of this analysis to a finite number of samples was made in Hofinger and Pillichshammer (2005), where also an algorithm to train neural networks was discussed (an extension of this algorithm was studied in Hofinger (2006)).

We have modeled weight-decay learning as the regularized optimization problems $(\text{span}_m G_{K_x}, \Phi_{WD,\gamma})$ and $(\text{span}_n G_{K_x}, \Phi_{WD,\gamma})$, with $n < m$. Then, our analysis leaves the more general problems $(\text{span}_m G_K, \Phi_{WD,\gamma})$, $(\text{span}_n G_K, \Phi_{WD,\gamma})$ and $(\text{span} G_K, \Phi_{WD,\gamma})$ open. In particular, note that, since the regularization term in (5) is not of the form $\psi(\|f\|_K)$ for a strictly increasing function ψ , at least one of the hypotheses of the so-called “Generalized Representer Theorem” (Schölkopf et al., 2001, Theorem 4) does not hold. A further analysis is required to study the behaviors of the optimal solutions to $(\text{span} G_K, \Phi_{WD,\gamma})$ and $(\text{span}_n G_K, \Phi_{WD,\gamma})$, for a given upper bound $n < m$ on the number of computational units, and to investigate whether some variation of the Representer Theorem holds also for these two problems.

References

- Aarts E, Korst J (1989) Simulated annealing and Boltzmann machines: a stochastic approach to combinatorial optimization and neural computing. Wiley,
- Aronszajn N (1950) Theory of reproducing kernels. *Trans AMS* 68:337–404
- Bartlett PL (1998) The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE Trans Inf Theory* 44(2):525–536
- Berg C, Christensen JPR, Ressel P (1984) Harmonic analysis on semigroups. Springer, New York
- Bertero M (1989) Linear inverse and ill-posed problems. *Adv Electron Electron Phys* 75:1–120
- Bertsekas DP (1999) Nonlinear programming. Athena Scientific, Belmont
- Bishop C (1995) Neural networks for pattern recognition. Oxford University Press, London
- Bishop C (2006) Pattern recognition and machine learning. Springer, Heidelberg
- Burger M, Engl H (2000) Training neural networks with noisy data as an ill-posed problem. *Adv Comput Math* 13:335–354
- Burger M, Neubauer A (2002) Analysis of Tikhonov regularization for function approximation by neural networks. *Neural Netw* 16:79–90
- Cristianini N, Shawe-Taylor J (2000) An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press, London
- Cucker F, Smale S (2001) On the mathematical foundations of learning. *Bull AMS* 39:1–49
- Cucker F, Smale S (2002) Best choices for regularization parameters in learning theory: on the bias-variance problem. *Found Comput Math* 2:413–428
- Cuesta-Albertos JA, Wschebor M (2003) Some remarks on the condition number of a real random square matrix. *J Compl* 19:548–554
- Demmel J (1987) The geometry of ill-conditioning. *J Compl* 3:201–229
- Dontchev AL (1983) Perturbations, approximations and sensitivity analysis of optimal control systems. *Lecture Notes in Control and Information Sciences*, vol 52. Springer, Berlin
- Friedman A (1970) Foundations of Modern Analysis. Holt, Rinehart, and Winston, New York
- Girosi F, Jones M, Poggio T (1995) Regularization theory and neural networks architectures. *Neural Comput* 7:219–269
- Girosi F (1998) An equivalence between sparse approximation and support vector machines. *Neural Comput* 10:1455–1480
- Girosi F (1994) Regularization theory, radial basis functions and networks. In: Cherkassky JHFV, Wechsler H (eds) From Statistics to Neural Networks. Theory and pattern recognition applications, ser. NATO ASI Series F, Computer and Systems Sciences, Springer, Berlin, pp 166–187
- Gnecco G, Sanguineti M (2007) Accuracy of suboptimal solutions to kernel principal component analysis. *Comput Optim Appl*. doi:10.1007/s10589-007-9108-y
- Goldberg DE (1989) Genetic algorithms in search, optimization, and machine learning. Addison-Wesley, Reading
- Golub GH, Loan CFV (1996) Matrix computations. John Hopkins University Press, London

- Gupta A, Lam M (1998) The weight decay backpropagation for generalizations with missing values. *Ann Oper Res* 78:165–187
- Gupta A, Lam M (1998) Weight decay backpropagation for noisy data. *Neural Netw* 11:1127–1138
- Hofinger A (2006) Nonlinear function approximation: computing smooth solutions with an adaptive greedy algorithm. *J Approxim Theory* 143:159–175
- Hofinger A, Pillichshammer F (2005) Learning a function from noisy samples at a finite sparse set of points. J. Kepler University, Linz, Technical Report, SFB F013
- Kimeldorf GS, Wahba G (1970) A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Ann Math Stat* 41:495–502
- Krogh A, Hertz JA (1992) A simple weight decay can improve generalization. In: *Advances in neural information processing systems*, vol. 4. Morgan Kaufmann Pub., pp 950–957
- Kůrková V (1997) Dimension-independent rates of approximation by neural networks. In: Warwick K, Kárný M (eds) *Computer-intensive methods in control and signal processing. The curse of dimensionality*. Birkhäuser, Boston, pp 261–270
- Kůrková V (2004) Learning from data as an inverse problem. In: Antoch J (ed) *COMPSTAT 2004—proceedings in computational statistics*. Physica-Verlag/Springer, Heidelberg, pp 1377–1384
- Kůrková V, Sanguinetti M (2001) Bounds on rates of variable-basis and neural-network approximation. *IEEE Trans Inf Theory* 47:2659–2665
- Kůrková V, Sanguinetti M (2005) Error estimates for approximate optimization by the extended Ritz method. *SIAM J Optim* 15:461–487
- Kůrková V, Sanguinetti M (2005) Learning with generalization capability by kernel methods of bounded complexity. *J Compl* 21:350–367
- Kůrková V, Savický P, Hlaváčková K (1998) Representations and rates of approximation of real-valued Boolean functions by neural networks. *Neural Netw* 11:651–659
- Levitin ES, Polyak BT (1966) Convergence of minimizing sequences in conditional extremum problems. *Dokl Akad Nauk SSSR* 168(5):764–767
- Ortega JM (1990) *Numerical analysis: a second course*. SIAM, Philadelphia
- Poggio T, Girosi F (1990) Networks for approximation and learning. *Proc IEEE* 78:1481–1497
- Poggio T, Smale S (2003) The mathematics of learning: dealing with data. *Notices AMS* 50:536–544
- Poggio T, Mukherjee S, Rifkin R, Rakhlin A, Verri A (2002) “b”. In: Winkler J, Niranjana M (eds) *Uncertainty in Geometric Computations*. Kluwer, Dordrecht, pp 131–141
- Schölkopf B, Smola AJ (2002) *Learning with kernels—support vector machines, regularization, optimization, and beyond*. MIT Press, Cambridge
- Schölkopf B, Herbrich R, Smola AJ, Williamson RC (2001) A generalized representer theorem. In: *Proceedings of COLT’01, Lecture Notes in Artificial Intelligence*. Springer, Heidelberg, pp 416–424
- Tikhonov AN, Arsenin VY (1977) *Solutions of ill-posed problems*. W.H. Winston, Washington
- Treadgold NK, Gedeon TD (1998) Simulated annealing and weight decay in adaptive learning: the SARPROP algorithm. *IEEE Trans Neural Netw* 9(4):662–668
- Vapnik VN (1998) *Statistical learning theory*. Wiley, New York
- Vladimirov AA, Nesterov YE, Chekanov YN (1978) On uniformly convex functionals. *Vestnik Moskovskogo Universiteta. Seriya 15—Vychislitel’naya Matematika i Kibernetika*, vol 3, pp 12–23 (English translation: *Moscow University Computational Mathematics and Cybernetics*, pp 10–21, 1979)