Weight-decay regularization in Reproducing Kernel Hilbert Spaces by variable-basis schemes

GIORGIO GNECCO
Department of Computer and Information Science (DISI)
University of Genoa
Via Dodecaneso, 35, 16146 Genova
ITALY
giorgio.gnecco@dist.unige.it

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

Abstract: The optimization problems associated with various regularization techniques for supervised learning from data (e.g., weight-decay and Tikhonov regularization) are described in the context of Reproducing Kernel Hilbert Spaces. Suboptimal solutions expressed by sparse kernel models with a given upper bound on the number of kernel computational units are investigated. Improvements of some estimates obtained in *Comput. Manag. Sci., vol. 6, pp. 53-79, 2009* are derived. Relationships between sparseness and generalization are discussed.

Key–Words: Learning from data, regularization, weight decay, suboptimal solutions, rates of approximation.

1 Introduction

In supervised learning, an unknown input-output mapping has to be learned on the basis of a sample of input-output data [1]. The problem of approximating a function on the basis of a data sample $\mathbf{z} \stackrel{\triangle}{=} \{(x_i, y_i) \in X \times \mathbb{R}, i = 1, \ldots, m\}$ is often ill-posed [2, 3]. *Regularization* [4] can be used to cope with this drawback.

Among regularization techniques, *weight decay* (see, e.g., [5]) is a learning technique that penalizes large values of the parameters (*weights*) of the model to be learned. For linear regression problems, the performance of weight decay was theoretically investigated in [5], where the case of linearization of a nonlinear model was considered, too. As to nonlinear models, a theoretical motivation of the generalization performance of certain neural networks trained through weight decay was given in [6], where the case of binary classification problems was studied using tools from Statistical Learning Theory.

In this paper, we study the optimization problems associated with the weight-decay and other learning techniques. Each problem is formulated as the minimization of a regularized empirical error functional over a suitable hypothesis space. Then, we compare the solution provided to the learning problem by weight-decay regularization with the solution given by the classical Tikhonov's regularization and a mixed regularization technique (i.e., weight decay combined with Tikhonov's regularization). When one uses hypothesis spaces spanned by kernel functions implemented by computational units widely used in connectionistic models, the solution to the Tikhonov-regularized learning problem has the form of a linear combination of the m-tuple of the kernel functions, parameterized by the input data vector $\mathbf{x} = (x_1, \ldots, x_m)$. 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 order to simplify the analysis and emphasize the relationships between weight decay and Tikhonov's regularization, also for the weight-decay learning problem and the mixed weight-decay/Tikhonov one we consider admissible solutions belonging to linear combinations of kernel functions parameterized by the input data vectors. For these problems one can show [7] that the optimal solutions are obtained by solving systems of linear equations, too.

For large data sets, the use of a number of computational units equal to the number m of data may lead to very complex models and so may be computationally unfeasible. Moreover, practical applications of linear algorithms using m com-

putational units are limited by the rate of convergence of iterative methods solving the systems of linear equations associated with the regularization schemes, 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 data sample (e.g., for the Gaussian elimination and m large enough, they grow at a rate m^3 [8, p. 175]). For some data and kernels, keeping the condition number of these matrices small requires a large value of the regularization parameter $\gamma > 0$, which may cause poor fit to the empirical data.

Motivated by these drawbacks, we also investigate the accuracy of suboptimal solutions to weight-decay learning and to the mixed weight-decay/Tikhonov learning, over hypothesis sets corresponding to models with less computational units than the size of the data sample. 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 $n < m$ kernel functions with parameters drawn from the data set. The upper bounds improve the ones given in [7] for the same regularized learning problems and are of the form $1/n$ (instead of $1/\sqrt{n}$, as in [7]), times a term that depends on the size m of the data sample, properties of the vector $\mathbf{y} = (y_1, \ldots, y_m)$ of output data, properties of the kernel, and the regularization parameter γ.

The final part of the paper discusses some relationships between sparseness and generalization and algorithms that can be used to find sparse suboptimal solutions to regularized learning problems.

2 Notations and definitions

The hypothesis spaces where we set the learning problems are *Reproducing Kernel Hilbert Spaces* (*RKHS*). These can be characterized in terms of *kernels* [9, 10]. A *positive-semidefinite (psd) kernel* is a symmetric function $K : X \times X \to \mathbb{R}$ such that for all positive integers m, all $(w_1, \ldots, w_m) \in \mathbb{R}^m$, and all $(u_1, \ldots, u_m) \in X^m$,

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

In other words, for every m and every $\mathbf{x} = (x_1, \ldots, x_m) \in X^m$, the *Gram matrix of the kernel K with respect to* $\mathbf{x}$, denoted by $\mathcal{K}[\mathbf{x}]$ and defined as $\mathcal{K}[\mathbf{x}]_{i,j} := K(x_i, x_j)$, is positive-semidefinite. If, for all positive integers m, all $(w_1, \ldots, w_m) \in \mathbb{R}^m$, and all $(u_1, \ldots, u_m) \in X^m$ with no repeated entries u_i, the equality in (1) holds only for $w_1 = \ldots = w_m = 0$, then K is called *positive-definite (pd) kernel*. Every psd kernel $K : X \times X \to \mathbb{R}$ generates an RKHS $\mathcal{H}_K(X)$. Indeed, $\mathcal{H}_K(X)$ can be defined as the completion of the linear span of the set $\{K_u : u \in X\}$ with the inner product $\langle K_u, K_v \rangle \stackrel{\triangle}{=} K(u, v)$. In the following, we denote by $\| \cdot \|_K$ the norm on the RKHS $\mathcal{H}_K(X)$.

In this paper we consider pd kernels K; to fix ideas, one can think of the widely-used *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. One reason for choosing RKHSs as hypothesis spaces is that the norms $\| \cdot \|_K$ on RKHSs defined by a large variety of kernels K 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. See, e.g., [11] for details.

For a subset G of a linear space and a positive integer n, we denote by $\operatorname{span} G \stackrel{\triangle}{=} \left\{ \sum_{j=1}^{k} w_j g_j : w_j \in \mathbb{R}, \; g_j \in G, k \in \mathbb{N} \right\}$ and $\operatorname{span}_n G \stackrel{\triangle}{=} \left\{ \sum_{j=1}^{n} w_j g_j : w_j \in \mathbb{R}, \, g_j \in G \right\}$ the sets of all linear combinations of elements of G and of all linear combinations of n-tuples of elements of G, resp. We let $G_K \stackrel{\triangle}{=} \{K_x : x \in X\}$ and, for an input data sample $\mathbf{x}$, $G_{K_\mathbf{x}} \stackrel{\triangle}{=} \{K_{x_1}, \ldots, K_{x_m}\}$. So, $\operatorname{span}_n G_K$ and $\operatorname{span}_n G_{K_\mathbf{x}}$ are the sets of all input/output functions of a computational model with one hidden layer of n computational units implementing functions from G_K and $G_{K_\mathbf{x}}$, resp. In contrast to linear approximation [12], which is also called *fixed-basis approximation* (as the approximating functions belong to a linear subspace generated by the first n elements of a set of functions with a fixed linear ordering), this approximation scheme is sometimes called *variable-basis approximation* [13] or *approximation from a dictionary* [14]. This models, e.g., radial-basis-function (RBF) networks and one-hidden layer perceptrons [15].

Given a linear space $\mathcal{H}$, a set $M \subseteq \mathcal{H}$, and a functional $\Phi : M \to \mathbb{R}$, following standard notation from optimization theory we denote by (M, Φ) the problem of minimizing Φ

over M. Every $f^o \in M$ such that $\Phi(f^o) = \min_{f \in M} \Phi(f)$ is called an *(optimal) solution* or a *minimum point* of the problem (M, Φ). We denote by $\operatorname{argmin}(M, \Phi)$ the set of solutions of (M, Φ), i.e., $\operatorname{argmin}(M, \Phi) \stackrel{\triangle}{=} \{f^o \in M : \Phi(f^o) = \min_{f \in M} \Phi(f)\}$. For $\varepsilon > 0$, $\operatorname{argmin}_\varepsilon(M, \Phi)$ is the set of ε-near minimum points of (M, Φ), i.e., $\operatorname{argmin}_\varepsilon(M, \Phi) \stackrel{\triangle}{=} \{f^o_\varepsilon \in M : \Phi(f^o_\varepsilon) \leq \inf_{f \in M} \Phi(f) + \varepsilon\}$.

Given a normed linear space $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$, let $M \subseteq \mathcal{H}$ and $\Phi : M \to \mathbb{R}$ be a functional. If Φ is continuous and $f \in M$, then the function $\omega_f : [0, +\infty) \to [0, +\infty)$ defined as

$$\omega_f(t) = \sup\{|\Phi(f) - \Phi(g)| : g \in M, \|f - g\|_{\mathcal{X}} \leq t\}$$

is called the *modulus of continuity* of Φ at f. By its definition, $\omega_f(t)$ is a nondecreasing function of t.

For a positive integer d, by $\|\cdot\|_1$ and $\|\cdot\|_2$ we denote the 1-norm and Euclidean norm on $\mathbb{R}^d$, resp.

3 The optimization problem associated with weight decay

In order to deal with the ill-posedness of the learning problem [2, 3], a widely-used regularization approach consists in minimizing the *regularized empirical error functional*, defined for $f \in \mathcal{H}$ as $\mathcal{E}_{\mathbf{z}}(f) + \gamma\, \Psi(f)$, where $\mathcal{E}_{\mathbf{z}}(f) \stackrel{\triangle}{=} \frac{1}{m} \sum_{i=1}^m (f(x_i) - y_i)^2$ is the *empirical error functional*, $\gamma > 0$ is the *regularization parameter*, and $\Psi : \mathcal{H} \to \mathbb{R}$ is a functional called *regularizer*. The corresponding model for the learning problem is $(M, \mathcal{E}_{\mathbf{z}} + \gamma\Psi)$, i.e., $\inf_{f \in M} (\mathcal{E}_{\mathbf{z}}(f) + \gamma\, \Psi(f))$. 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 in correspondence of f); ii) penalizing functions f that give a large value of the regularizer $\Psi(f)$. For certain normed hypothesis spaces $\mathcal{H}$, the choice $\Psi(\cdot) = \|\cdot\|^2_{\mathcal{H}}$ allows one to enforce certain smoothness properties of the solution. In this case, the parameter γ quantifies the compromise between enforcing closeness to the data sample and avoiding solutions that are not sufficiently smooth.

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

$$\Phi_{WD,\gamma}(f) \stackrel{\triangle}{=} \mathcal{E}_{\mathbf{z}}(f) + \gamma\, \|\mathbf{c}_f\|_2^2\,, \tag{2}$$

where the components of the vector $\mathbf{c}_f = (c_{f,1}, \ldots, c_{f,l})^{\mathrm{T}}$ and $\hat{x}_1, \ldots, \hat{x}_l \in X$ are the parameters in the expansion

$$f = \sum_{j=1}^{l} c_{f,j} K_{\hat{x}_j}\,. \tag{3}$$

Choosing a pd kernel guarantees that f has a unique representation of the form (3), thus l and $\|\mathbf{c}_f\|_2^2$ in (2) are defined unambiguously (otherwise one may choose, among all equivalent representations of f - possibly with different values of l - the infimum of the squared norms $\|\mathbf{c}_f\|_2^2$ of the corresponding coefficient vectors $\mathbf{c}_f$). Note that in general the functional (2) cannot be continuously extended on $\mathcal{H}_K(X)$ if K is continuous on $X \times X \subseteq \mathbb{R}^d \times \mathbb{R}^d$. Indeed, by varying the number of kernel units in (2), it is easy to construct a sequence $\{f_{2l}\}$ such that $f_{2l} \in \operatorname{span}_{2l} G_K$, $\|f_{2l}\|_K \to 0$ and $\|\mathbf{c}_{f_{2l}}\|_2^2 \to \infty$ as $l \to \infty$. One example of such a sequence is given by $f_{2l} \stackrel{\triangle}{=} \sum_{j=1}^{2l} (-1)^j K_{\hat{x}_j(l)}$, where, for each l, when j is odd $\hat{x}_j(l)$ and $\hat{x}_{j+1}(l)$ are chosen "sufficiently close" to each other such that $\|f_{2l}\|_K < \frac{1}{l}$.

The number l of terms in the expression (3) is equal to the dimension of the vector $\mathbf{c}_f$ in (2). In the following, we consider the weight-decay functional corresponding to the choice $l = m$ (i.e., l equal to the size of the data sample) and $\hat{x}_j = x_j$ for $j = 1, \ldots, m$. In other words, we investigate the minimization of the functional (2) over linear combinations $\operatorname{span}_m G_{K_{\mathbf{x}}}$ of m kernel functions centered at the m input data. Hence, we model the *weight-decay learning problem* as

$$(\operatorname{span}_m G_{K_{\mathbf{x}}}, \Phi_{WD,\gamma})\,. \tag{4}$$

The next proposition, from [7], expresses the solution to the weight-decay learning problem as a linear combination of kernel functions centered at the input data points, with coefficients obtained by solving a linear system of equations. Its proof given in [7] is based on the theory of regularization of inverse problems; for completeness of exposition, here we sketch a simpler proof.

Proposition 1 *Let X be a nonempty set, $K : X \times X \to \mathbb{R}$ a pd kernel, m a positive integer, $\mathbf{x} = (x_1, \ldots, x_m) \in X^m$ with no repeated entries x_i, $\mathbf{y} = (y_1, \ldots, y_m)^{\mathrm{T}} \in \mathbb{R}^m$, and $\gamma > 0$. Then there exists a unique solution*

$$f^o_{WD,\gamma} = \sum_{j=1}^{m} c^o_{WD,\gamma,j} K_{x_j} \tag{5}$$

to the problem $(\mathrm{span}_m\, G_{K_\mathbf{x}}, \Phi_{WD,\gamma})$, *where* $\mathbf{c}^o_{WD,\gamma} = (c^o_{WD,\gamma,1}, \ldots, c^o_{WD,\gamma,m})^{\mathrm{T}}$ *is the unique solution to the linear system of equations*

$$\left(\mathcal{K}[\mathbf{x}] + \gamma\, m \mathcal{K}^{-1}[\mathbf{x}]\right)\, \mathbf{c}^o_{WD,\gamma} = \mathbf{y}\,. \tag{6}$$

Proof. Let $F = \{\mathbf{e}_1, \ldots, \mathbf{e}_m\}$ be the canonical orthonormal basis of $\mathbb{R}^m$ and define the m-variable function $\tilde{\Phi}_{WD,\gamma} : \mathbb{R}^m \to \mathbb{R}$ such that, for every $f = \sum_{j=1}^m c_{f,j} K_{x_j} \in \mathrm{span}\, G_{K_\mathbf{x}}$, $\tilde{\Phi}_{WD,\gamma}(\mathbf{c}_f) \stackrel{\triangle}{=} \Phi_{WD,\gamma}(f)$. The problems $(\mathrm{span}\, G_{K_\mathbf{x}}, \Phi_{WD,\gamma})$ and $(\mathrm{span}\, F, \tilde{\Phi}_{WD,\gamma})$ are clearly equivalent, and straightforward computations show that

$$\begin{aligned}\Phi_{WD,\gamma}(\mathbf{c}_f) = \mathbf{c}_f^{\mathrm{T}} \left(\tfrac{1}{m}\mathcal{K}^2[\mathbf{x}] + \gamma\mathcal{I}\right) \mathbf{c}_f \\ -\tfrac{2}{m}\mathbf{c}_f^{\mathrm{T}}\mathcal{K}[\mathbf{x}]\mathbf{y} + \tfrac{1}{m}\mathbf{y}^{\mathrm{T}}\mathbf{y}\,.\end{aligned} \tag{7}$$

Then (6) follows by minimizing (7) w.r.t. $\mathbf{c}_f$. □

4 Comparison with Tikhonov regularization

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

$$\Phi_{T,\gamma} \stackrel{\triangle}{=} \mathcal{E}_\mathbf{z}(f) + \gamma\, \|f\|_K^2\,. \tag{8}$$

The corresponding *Tikhonov-regularized learning problem* is

$$(\mathcal{H}_K(X), \Phi_{T,\gamma})\,. \tag{9}$$

Existence, uniqueness, and an explicit formula describing the solution to the learning problem (9) are given by the so-called *Representer Theorem* (see, e.g., [16, p. 42]). For a nonempty set X, $K : X \times X \to \mathbb{R}$ a psd kernel, m a positive integer, $\mathbf{x} = (x_1, \ldots, x_m) \in X^m$, $\mathbf{y} = (y_1, \ldots, y_m)^{\mathrm{T}} \in \mathbb{R}^m$, and $\gamma > 0$, the Representer Theorem states that there exists a unique solution $f^o_{T,\gamma}$ to the problem $(\mathcal{H}_K(X), \Phi_{T,\gamma})$ and it has the form

$$f^o_{T,\gamma} = \sum_{j=1}^m c^o_{T,\gamma,j} K_{x_j}, \tag{10}$$

where $\mathbf{c}^o_{T,\gamma} = (c^o_{T,\gamma,1}, \ldots, c^o_{T,\gamma,m})^{\mathrm{T}}$ is the unique solution to the linear system of equations

$$(\mathcal{K}[\mathbf{x}] + \gamma\, m\, \mathcal{I})\mathbf{c}^o_{T,\gamma} = \mathbf{y}\,. \tag{11}$$

The optimal solution to the Tikhonov-regularized learning problem described by (10) is an element of $\mathrm{span}_m\, G_{K,\mathbf{x}} \subseteq \mathrm{span}_m\, G_K$. Since, by the Representer Theorem, the solution to the Tikhonov-regularized learning problem $(\mathcal{H}_K(X), \Phi_{T,\gamma})$ belongs to $\mathrm{span}_m\, G_{K_\mathbf{x}}$, one can restate such a problem as $(\mathrm{span}_m\, G_{K_\mathbf{x}}, \Phi_{T,\gamma})$ and compare its solution with the solution to the weight-decay learning problem $(\mathrm{span}_m\, G_{K_\mathbf{x}}, \Phi_{WD,\gamma})$. This is done in [7] in terms of *spectral windows*.

5 Combining weight decay and Tikhonov regularization

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

$$\Phi_{WDT,\gamma_T,\gamma_{WD}}(f) \stackrel{\triangle}{=} \mathcal{E}_\mathbf{z}(f) + \gamma_T\, \|f\|_K^2 + \gamma_{WD}\, \|\mathbf{c}_f\|_2^2$$

it is possible to combine weight decay and Tikhonov regularization. For simplicity and without loss of generality, we take $\gamma_T = \gamma_{WD} = \gamma/2$ and we define the *mixed regularized learning problem* $(\mathrm{span}_m\, G_{K_\mathbf{x}}, \Phi_{WDT,\gamma/2})$, where

$$\Phi_{WDT,\gamma/2}(f) \stackrel{\triangle}{=} \Phi_{WDT,\gamma/2,\gamma/2}(f)\,.$$

The next proposition, also taken from [7], investigates the problem $(\mathrm{span}_m\, G_{K_\mathbf{x}}, \Phi_{WDT,\gamma/2})$ and gives a formula for its solution. Its proof is similar to that of Proposition 1, so is omitted.

Proposition 2 *Let X be a nonempty set, $K : X \times X \to \mathbb{R}$ a pd kernel, m a positive integer, $\mathbf{x} = (x_1, \ldots, x_m) \in X^m$ with no repeated entries x_i, $\mathbf{y} = (y_1, \ldots, y_m)^{\mathrm{T}} \in \mathbb{R}^m$, and $\gamma > 0$. Then there exists a unique solution*

$$f^o_{WDT,\gamma/2} = \sum_{j=1}^m c^o_{WDT,\gamma/2,j} K_{x_j} \tag{12}$$

to the problem $(\mathrm{span}_m\, G_{K_\mathbf{x}}, \Phi_{WDT,\gamma/2})$, *where* $\mathbf{c}^o_{WDT,\gamma/2} = (c^o_{WDT,\gamma/2,1}, \ldots, c^o_{WDT,\gamma/2,m})^{\mathrm{T}}$ *is the unique solution to the linear system of equations*

$$\left(\mathcal{K}[\mathbf{x}] + \frac{\gamma}{2}\, m\, (\mathcal{I} + \mathcal{K}^{-1}[\mathbf{x}]\,)\right)\, \mathbf{c}^o_{WDT,\gamma/2} = \mathbf{y}\,. \tag{13}$$

Similarly to Proposition 1, Proposition 2 expresses the solution to the mixed regularized learning problem as a linear combination of kernel functions centered at the data points, with

coefficients obtained by solving the linear system of equations (13). The expression $\mathbf{c}^o_{f_{WDT,\gamma/2}} = \left(\mathcal{K}[\mathbf{x}] + \frac{\gamma}{2}\, m\left(\mathcal{I} + \mathcal{K}^{-1}[\mathbf{x}]\right)\right)^{-1} \mathbf{y}$ can be compared in terms of spectral windows with the expressions $\mathbf{c}^o_{f_{WD,\gamma}} = \left(\mathcal{K}[\mathbf{x}] + \gamma\, m\mathcal{K}^{-1}[\mathbf{x}]\right)\right)^{-1} \mathbf{y}$ and $\mathbf{c}^o_{f_{T,\gamma}} = \left(\mathcal{K}[\mathbf{x}] + \gamma\, m\,\mathcal{I}\right)^{-1} \mathbf{y}$ for the coefficients of the solutions to the weight-decay and the Tikhonov-regularized learning problems, respectively [7].

6 Accuracy of suboptimal solutions

The expressions (6), (11), and (13) for the coefficients of the linear combinations providing the solutions to the respective problems require to solve linear systems of equations, so, in principle, they can be used to design linear learning algorithms. However, their applications are limited by the rates of convergence of iterative methods solving linear systems of equations.

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 $\mathrm{cond}(\mathcal{A}) = \|\mathcal{A}\|\,\|\mathcal{A}^{-1}\|$, where $\|\mathcal{A}\|$ denotes the norm of $\mathcal{A}$ as a linear operator on $(\mathbb{R}^m, \|\cdot\|)$. For a symmetric matrix $\mathcal{A}$, we denote by $\lambda_{\max}(\mathcal{A})$ and $\lambda_{\min}(\mathcal{A})$ its maximum and minimum eigenvalues, respectively. It is easy to check that for every norm $\|\cdot\|$ on $\mathbb{R}^m$ and every $m \times m$ symmetric nonsingular matrix $\mathcal{A}$, $\mathrm{cond}(\mathcal{A}) \geq \frac{|\lambda_{\max}(\mathcal{A})|}{|\lambda_{\min}(\mathcal{A})|}$ and $\mathrm{cond}_2(\mathcal{A}) = \frac{|\lambda_{\max}(\mathcal{A})|}{|\lambda_{\min}(\mathcal{A})|}$, where $\mathrm{cond}_2(\mathcal{A})$ denotes the condition number of $\mathcal{A}$ with respect to the $\|\cdot\|_2$-norm on $\mathbb{R}^m$ [8, p. 35]. To simplify the notation, we write $\lambda_{\max}$ instead of $\lambda_{\max}(\mathcal{K}[\mathbf{x}])$ and similarly for $\lambda_{\min}$.

For pd kernels and every $\mathbf{x}$ with no repeated entries x_i, the matrix $\mathcal{K}[\mathbf{x}]$ is positive definite, so all its eigenvalues are positive. By simple algebraic manipulations and spectral theory, for the condition numbers of the matrices involved in the solutions of the linear systems of equations (6), (11), and (13), simple calculations give

$$\mathrm{cond}_2(\mathcal{K}[\mathbf{x}] + \gamma\, m\,\mathcal{K}^{-1}[\mathbf{x}]) \leq \mathrm{cond}_2(\mathcal{K}[\mathbf{x}])\,, \quad (14)$$

$$\mathrm{cond}_2(\mathcal{K}[\mathbf{x}] + \gamma\, m\mathcal{I}) = \mathrm{cond}_2(\mathcal{K}[\mathbf{x}])\,, \quad (15)$$

$$\mathrm{cond}_2(\mathcal{K}[\mathbf{x}] + \gamma\, m\mathcal{I}) \leq 1 + \frac{\lambda_{\max}}{\gamma\, m}\,, \quad (16)$$

$$\mathrm{cond}_2(\mathcal{K}[\mathbf{x}] + \frac{\gamma}{2}\, m(\mathcal{I} + \mathcal{K}^{-1}[\mathbf{x}])) \leq \mathrm{cond}_2(\mathcal{K}[\mathbf{x}])\,, \quad (17)$$

and

$$\mathrm{cond}_2(\mathcal{K}[\mathbf{x}] + \tfrac{\gamma}{2}\, m(\mathcal{I} + \mathcal{K}^{-1}[\mathbf{x}])) \leq \frac{\lambda_{\max}}{\frac{\gamma}{2}\, m} + \frac{\frac{\gamma}{2}\, m\,(\lambda_{\min}+1)}{\lambda_{\min}\,(\lambda_{\min}+\frac{\gamma}{2}\, m\,)}\,. \quad (18)$$

By equations (14), (15), and (17), when $\mathrm{cond}_2(\mathcal{K}[\mathbf{x}])$ is sufficiently small, good conditioning of the respective matrices is guaranteed 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. On the other hand, the regularization parameter γ can always be chosen "large enough" such that $\mathrm{cond}_2(\mathcal{K}[\mathbf{x}] + \gamma\, m\,\mathcal{K}^{-1}[\mathbf{x}])$, $\mathrm{cond}_2(\mathcal{K}[\mathbf{x}] + \gamma\, m\mathcal{I})$, and $\mathrm{cond}_2(\mathcal{K}[\mathbf{x}] + \frac{\gamma}{2}\, m(\mathcal{I} + \mathcal{K}^{-1}[\mathbf{x}]))$ are close to $\mathrm{cond}_2(\mathcal{K}[\mathbf{x}])$, 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 it cannot be too large. The problem of choosing the regularization parameter in Tikhonov and other regularization techniques is studied, e.g., in [17].

Summing up, when a small condition number of the matrices in (6), (11), and (13) and a good fit to the empirical data cannot be simultaneously guaranteed, one has to consider other learning techniques. In particular, one may be interested in suboptimal solutions depending on a number of computational units smaller than m, as they have smaller memory requirements, better interpretability, and in some cases are easier to find than the optimal ones (e.g., through the so-called *greedy algorithms* [18]). For instance, in contrast to the respective optimal solutions, which are linear combinations of $K_{x_1}, \ldots, K_{x_m}$ determined by the sample $\mathbf{x} = (x_1, \ldots, x_m)$ of input data, one may search for suboptimal solutions formed by linear combinations of $n < m$ such functions, or that depend on *arbitrary n-tuples* of elements of $G_K \stackrel{\triangle}{=} \{K_x : x \in X\}$.

In the following, we investigate the accuracy of suboptimal solutions over $(\mathrm{span}_n\, G_{K_{\mathbf{x}}}, \Phi_{WD,\gamma})$ and $(\mathrm{span}_n\, G_{K_{\mathbf{x}}}, \Phi_{WDT,\gamma/2})$ to the solutions $f^o_{WD,\gamma}$ $f^o_{WDT,\gamma/2}$ provided by Propositions 1 and 2, resp.

The next theorem improves [7, Theorem 4] and estimates for increasing values of $n < m$ the rates of approximation of $f^o_{WD,\gamma}$, which can be obtained by suboptimal solutions to the problem $(\mathrm{span}_n\, G_{K_{\mathbf{x}}}, \Phi_{WD,\gamma})$. Note that, differently from [7, Theorem 4], we do not require the condition $s_K = \sup_{x \in X} \sqrt{K(x,x)} < +\infty$.

Theorem 3 *Let X be a nonempty set, $K : X \times X \to \mathbb{R}$ a pd kernel, $\mathbf{z}$ a data sample of size m*

with no repeated entries x_i, $\lambda_{\min}$ and $\lambda_{\max}$ the minimum and the maximum eigenvalues of $\mathcal{K}[\mathbf{x}]$, resp., and $\gamma > 0$. Let $\alpha(t) \stackrel{\triangle}{=} \left(\frac{\lambda_{\max}^2}{m} + \gamma\right) t^2$, and $\Delta_{WD,\gamma} \stackrel{\triangle}{=} \|\mathbf{c}^o_{WD,\gamma}\|_1^2 - \|\mathbf{c}^o_{WD,\gamma}\|_2^2$. Then the following holds.

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

$$\inf_{f \in \mathrm{span}_n G_{K_\mathbf{x}}} \Phi_{WD,\gamma}(f) - \Phi_{WD,\gamma}(f^o_{WD,\gamma}) \leq \alpha\left(\sqrt{\frac{\Delta_{WD,\gamma}}{n}}\right).$$

(ii) Let $\varepsilon_n > 0$ and $f_n \in \mathrm{argmin}_{\varepsilon_n}(\mathrm{span}_n G_{K_\mathbf{x}}, \Phi_{WD,\gamma})$. Then $\|f_n - f^o_{WD,\gamma}\|_K^2 \leq \frac{\lambda_{\max}}{\frac{\lambda_{\min}^2}{m}+\gamma}\left[\alpha\left(\sqrt{\frac{\Delta_{WD,\gamma}}{n}}\right) + \varepsilon_n\right]$.

Proof. (i) Since $\mathbf{c}^o_{WD,\gamma}$ is optimal for the problem $(\mathrm{span}\, F, \tilde{\Phi}_{WD,\gamma})$, by the first-order optimality condition for unconstrained optimization and the specific form of $\tilde{\Phi}_{WD,\gamma}$ it follows $\tilde{\Phi}_{WD,\gamma}(\mathbf{c}_f) = \tilde{\Phi}_{WD,\gamma}(\mathbf{c}^o_{WD,\gamma}) + \frac{1}{2} H(\mathbf{c}^o_{WD,\gamma})\|\mathbf{c}_f - \mathbf{c}^o_{WD,\gamma}\|_2^2$, where the Hessian $H(\mathbf{c}^o_{WD,\gamma})$ has the expression $H(\mathbf{c}^o_{WD,\gamma}) = 2\left(\frac{\mathcal{K}^2[\mathbf{x}]}{m} + \gamma\mathcal{I}\right)$ and is positive definite. So the function $\alpha(t) \stackrel{\triangle}{=} (\frac{\lambda_{\max}^2}{m}+\gamma)t^2$ is the modulus of continuity of $\tilde{\Phi}_{WD,\gamma}$ on $(\mathbb{R}^m, \|\cdot\|_2)$. Let $F = \{\mathbf{e}_1, \ldots, \mathbf{e}_m\}$ be the canonical orthonormal basis of $\mathbb{R}^m$. Then by the definitions of $\tilde{\Phi}_{WD,\gamma}$ and of modulus of continuity, we have

$$\begin{aligned}
&\inf_{f \in \mathrm{span}_n G_{K_\mathbf{x}}} \Phi_{WD,\gamma}(f) - \Phi_{WD,\gamma}(f^o_{WD,\gamma}) \\
&= \inf_{\mathbf{c}_f \in \mathrm{span}_n F} \tilde{\Phi}_{WD,\gamma}(\mathbf{c}_f) - \tilde{\Phi}_{WD,\gamma}(\mathbf{c}^o_{WD,\gamma}) \\
&\leq \alpha\left(\inf_{\mathbf{c}_f \in \mathrm{span}_n F} \|\mathbf{c}^o_{WD,\gamma} - \mathbf{c}_f\|_2\right).
\end{aligned}$$

The estimate $\inf_{\mathbf{c}_f \in \mathrm{span}_n F} \|\mathbf{c}^o_{WD,\gamma} - \mathbf{c}_f\|_2 \leq \sqrt{\frac{\Delta_{WD,\gamma}}{n}}$ follows by applying Theorem 5 in the Appendix.

(ii) By the proof of (i) and the definition of f_n, we get

$$\left(\frac{\lambda_{\min}^2}{m} + \gamma\right)\|\mathbf{c}_f - \mathbf{c}^o_{WD,\gamma}\|_2^2 \leq \alpha\left(\sqrt{\frac{\Delta_{WD,\gamma}}{n}}\right) + \varepsilon_n. \tag{19}$$

The estimate is obtained by applying (19) and $\|f_n - f^o_{WD,\gamma}\|_K^2 \leq \lambda_{\max}\|\mathbf{c}_f - \mathbf{c}^o_{WD,\gamma}\|_2^2$ (which follows easily by the definition of the norm in an RKHS). □

The next theorem improves [7, Theorem 4] and estimates for increasing values of $n < m$ the rates of approximation of $f^o_{WDT,\gamma/2}$, which can be obtained by suboptimal solutions to the problem $(\mathrm{span}_n G_{K_\mathbf{x}}, \Phi_{WDT,\gamma/2})$. The proof is similar to that of Theorem 3, so it is omitted.

Theorem 4 *Let X be a nonempty set, $K : X \times X \to \mathbb{R}$ a pd kernel, $\mathbf{z}$ a data sample of size m with no repeated entries x_i, $\lambda_{\min}$ and $\lambda_{\max}$ the minimum and the maximum eigenvalues of $\mathcal{K}[\mathbf{x}]$, resp., and $\gamma > 0$. Let $\beta(t) \stackrel{\triangle}{=} \left(\frac{\lambda_{\max}^2}{m} + \frac{\gamma}{2}(\lambda_{\max}+1)\right)t^2$, and $\Delta_{WDT,\gamma/2} \stackrel{\triangle}{=} \|\mathbf{c}^o_{WDT,\gamma/2}\|_1^2 - \|\mathbf{c}^o_{WDT,\gamma/2}\|_2^2$. Then the following hold.*

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

$$\inf_{f \in \mathrm{span}_n G_{K_\mathbf{x}}} \Phi_{WDT,\gamma/2}(f) - \Phi_{WDT,\gamma/2}(f^o_{WDT,\gamma/2}) \leq \beta\left(\sqrt{\frac{\Delta_{WDT,\gamma/2}}{n}}\right).$$

(ii) Let $\varepsilon_n > 0$ and $f_n \in \mathrm{argmin}_{\varepsilon_n}(\mathrm{span}_n G_{K_\mathbf{x}}, \Phi_{WDT,\gamma/2})$. Then

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

7 Sparseness and generalization

Besides advantages in terms of memory requirements and, in some cases, computational requirements, a sparse suboptimal solution to a learning problem stated on a RKHS should have good generalization properties. If an a-priori upper bound on the $\|\cdot\|_K$-norms of suboptimal solutions is known, then one can study the associated estimation errors following the approach of [19, Chapters 4 and 7] (see also [20]), which employs upper bounds on Rademacher's complexity for balls in RKHSs, or [21, Section 6], which uses other measures of complexity for such balls.

It is worth remarking that there are situations in which sparseness itself enforces good generalization capability (quantitatively expressed by suitable bounds from statistical learning theory). For example, in the context of binary classification problems, [22, Theorem 4] gives a lower

bound on the probability of finding, after a training process from m data samples, a sparse kernel-based binary classifier with a small generalization error, provided that a (not necessarily sparse) kernel-based binary classifier that correctly classifies all the training set exists and has a large margin. The main tools used there to obtain such a result are *Littlestone-Warmuth's compression lemma* [23] and an extension to kernel-based binary classifiers of classical *Novikoff's mistake bound* for perceptron learning [24]. Related results for regression problems were given in [25, Theorem 3].

8 On algorithms for sparse suboptimal solutions

Various algorithms have been proposed in the literature to find sparse suboptimal solutions to approximation and optimization problems. In the following, we report a discussion from [20] on those useful for the regularization techniques considered in this paper.

The context common to all such algorithms is the following. Given a (typically redundant) set D of functions, called *dictionary*, which are elements of a finite- or infinite-dimensional Hilbert space $\mathcal{H}$, for a "small" positive integer n one aims to find an accurate suboptimal solution f_n^s from $\mathrm{span}_n D$ to a function approximation problem, or, more generally, to a functional optimization problem.

When no structure is imposed on the dictionary D, the problem of finding the best approximation of a function $f \in \mathcal{H}$ from $\mathrm{span}_n D$ is NP-hard [26]. However, the problem may drastically simplify when the elements of the dictionary have a suitable structure. The simplest situation arises when they are orthogonal. The case of a dictionary with nearly-orthogonal elements, i.e., a dictionary with small *coherence* [14], stays halfway between these two extremes and provides a computationally tractable problem, for which constructive approximation results are available [27, 28, 14, 29]. As in our context $D = G_{K_{\mathbf{x}}}$, we may want to choose a kernel K such that the dictionary $G_{K_{\mathbf{x}}}$ has small coherence. If this is not possible, then, as in [30], one may consider as dictionary a suitable subset of $G_{K_{\mathbf{x}}}$ with a small coherence.

In the remaining of this section, we shall discuss three families of algorithms to derive sparse suboptimal solutions.

Greedy algorithms. Starting from an initial sparse suboptimal solution f_n^s with a small n (usually $n = 0$ and $f_0^s = 0$), typically greedy algorithms obtain inductively an $(n+1)$-term suboptimal solution f_{n+1}^s as a linear combination of the n-term one f_n^s and a new element from the dictionary. So, a sequence of low-dimensional optimization problems has to be solved. Depending on how such problems are defined, different kinds of greedy algorithms are obtained; see, e.g., [31, 32, 33]. These algorithms are particularly suitable to derive sparse suboptimal solutions to Tikhonov regularization.

Remarkable properties were proven for *Matching Pursuit* and *Orthogonal Matching Pursuit*; their kernel versions, known as *Kernel Matching Pursuits*, are studied, e.g., in [34, 25]. Given $f \in \mathcal{H}$ and provided that the positive integer n and the dictionary D are suitably chosen, the n-term approximations found by these two algorithms are only a well-defined factor $C(n) > 0$ worse than the best approximation of f in terms of any n elements of D (see [28, Theorem 2.1], [29, Theorem 2.6], [14, Featured Theorem 3], and [27, Theorem 3.5] for some values of $C(n)$ and estimates on the number of the iterations).

Algorithms based on low-rank approximation of the Gram matrix. Another possibility consists in replacing the Gram matrix by a low-rank approximation, by using greedy algorithms and/or randomization techniques [35, 36]. Low-rank approximations can be used, e.g., to find a sparse suboptimal solution to Tikhonov regularization [35].

Algorithms based on convex formulations of the problem. Also when the functional to be minimized is convex and defined on a convex set, when suboptimal solutions in $\mathrm{span}_n D$ are searched for, the corresponding optimization problem may be not convex any more. Then one may consider a related convex optimization problem with sparse optimal solutions, for which efficient convex optimization algorithms can be exploited. In linear regression, e.g., adding an upper bound on the l_1-norm of the coefficients instead of their l_2-norm (or an l_1 penalization term instead of an l_2 one) is known to enforce the sparseness of the solution. This is called the *Least Absolute Shrinkage and Selection Operator* (*LASSO*) problem [37], for which kernel versions have been proposed in the literature [38]. In [39], an algorithm well-suited to LASSO was proposed, which

shows how the degree of sparseness of its solution is controlled by varying the regularization parameter. Another computationally promising technique to solve LASSO is based on the operator-splitting approach studied in [40]. Some limitations of LASSO have been overcome by its extension called *elastic net* [41], where the l_1 and l_2 penalization terms are simultaneously present. In [41], it was shown that the elastic net can be considered as a LASSO on an extended artificial data set, so that the algorithms from [39] can be still applied. A kernel version of the elastic net was developed in [42].

9 Conclusions

In a variety of applications, an unknown function has to be learned on the basis of a sample of input-output data. Usually such a problem is ill-posed, unless a-priori knowledge is incorporated into the learning model. This can be achieved by regularization techniques.

Representer Theorems in Reproducing Kernel Hilbert Spaces (RKHSs) describe the optimal solutions to various regularized learning problems. For data samples of size m, their solutions are expressed as linear combinations of m computational units determined by the kind of hypothesis space, and for which the optimal coefficients can be obtained by solving certain linear systems of equations. However, solving such systems may be computationally demanding for large data sets and may suffer from ill-conditioning.

We have investigated the accuracies of suboptimal solutions obtainable by arbitrary n-tuples of computational units, with $n < m$. In particular, we have investigated the learning technique known as "weight decay" and we have compared it with learning techniques based on Tikhonov regularization and on the combination of the latter with weight decay.

The upper bounds that we have obtained in Theorems 3 and 4 exhibit a common feature: they are of the form A/n, where $A > 0$ depends on properties of the output data vector $\mathbf{y}$ and of the regularized functional. Thus, in the presence of large data samples and when algorithms based on Representer Theorems suffer from ill-conditioning, algorithms operating on models with $n < m$ computational units can provide useful alternatives, as sparse models can approximate the optimal solutions quite well. The estimates in this paper improve the ones given in [7] for the same regularized learning problems.

Acknowledgement

The authors were partially supported by a grant "Progetti di Ricerca di Ateneo 2008" of the University of Genova, project "Solution of Functional Optimization Problems by Nonlinear Approximators and Learning from Data".

Appendix

In the proofs of Theorems 3 and 4 we exploit the following result from [43]. It is a special case of a reformulation (given in [44]; see also [45]) 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 separable Hilbert space $\mathcal{H}$, by $\|\cdot\|_{1,F}$ we denote the 1-norm with respect to F, defined for every $\phi \in \mathcal{H}$ as $\|\phi\|_{1,F} = \sum_{f\in F} |\langle \phi, f\rangle_{\mathcal{H}}|$. This is a norm on the set $\{\phi \in \mathcal{H} : \|\phi\|_{1,F} < \infty\}$.

Theorem 5 [43, Theorem 2] *Let $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ be a separable Hilbert space and F its orthonormal basis. For every $\phi \in \mathcal{H}$ and every positive integer n,*

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

References:

[1] Poggio, T., Smale, S.: The mathematics of learning: Dealing with data. Notices of the AMS **50** (2003) 536–544

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

[3] Burger, M., Engl, H.: Training neural networks with noisy data as an ill-posed problem. Advances in Computational Mathematics **13** (2000) 335–354

[4] Tikhonov, A.N., Arsenin, V.Y.: Solutions of Ill-posed Problems. W.H. Winston, Washington, D.C. (1977)

[5] Krogh, A., Hertz, J.A.: A simple weight decay can improve generalization. In: Advances in Neural Information Processing Systems. Volume 4., Morgan Kaufmann Pub. (1992) 950–957

[6] Bartlett, P.L.: 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. on Information Theory **44**(2) (1998) 525–536

[7] Gnecco, G., Sanguineti, M.: The weight-decay technique in learning from data: An optimization point of view. Computational Management Science **6** (2009) 53–79

[8] Ortega, J.M.: Numerical Analysis: A Second Course. SIAM, Philadelphia (1990)

[9] Aronszajn, N.: Theory of reproducing kernels. Trans. of AMS **68** (1950) 337–404

[10] Taouali, O., Saidi, N., Messaoud, H.: Identification of non linear MISO process using RKHS and Volterra models. WSEAS Transactions on Systems **8** (2009) 723–732

[11] Kůrková, V., Sanguineti, M.: Learning with generalization capability by kernel methods of bounded complexity. J. of Complexity **21** (2005) 350–367

[12] Singer, I.: Best Approximation in Normed Linear Spaces by Elements of Linear Subspaces. Springer-Verlag, Berlin Heidelberg (1970)

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

[14] Gribonval, R., Vandergheynst, P.: On the exponential convergence of matching pursuits in quasi-incoherent dictionaries. IEEE Trans. on Information Theory **52** (2006) 255–261

[15] Popescu, M.C., Balas, V.E., Perescu-Popescu, L., Mastorakis, N.: Multilayer perceptrons and neural networks. WSEAS Transactions on Circuits and Systems **8** (2009) 579–588

[16] Cucker, F., Smale, S.: On the mathematical foundations of learning. Bulletin of AMS **39** (2001) 1–49

[17] Hämarik, U., Raus, T.: Choice of the regularization parameter in ill-posed problems with rough estimate of the noise level of data. WSEAS Transactions on Mathematics **4** (2005) 76–81

[18] Zhang, T.: Sequential greedy approximation for certain convex optimization problems. IEEE Trans. on Information Theory **49** (2003) 682–691

[19] Shawe-Taylor, J., Cristianini, N.: Kernel Methods for Pattern Analysis. Cambridge University Press (2004)

[20] Gnecco, G., Sanguineti, M.: Regularization techniques and suboptimal solutions to optimization problems in learning from data. Neural Computation (to appear)

[21] Evgeniou, T., Pontil, M., Poggio, T.: Regularization networks and support vector machines. Advances in Computational Mathematics **13** (2000) 1–50

[22] Graepel, T., Herbrich, R.: From margin to sparsity. Advances in Neural Information System Processing **13** (2001) 210–216

[23] Littlestone, N., Warmuth, M.: Relating data compression and learnability. Technical report, University of California, Santa Cruz (1986)

[24] Novikoff, A.: On convergence proofs for perceptrons. In: Proceeding of the Symp. on the Mathematical Theory of Automata. Volume 12. (1962) 615–622

[25] Hussain, Z., Shawe-Taylor, J.: Theory of matching pursuit. In Koller, D., Schuurmans, D., Bengio, Y., Bottou, L., eds.: Advances in Neural Information Processing Systems 21 (Proc. 22nd Annual Conf. on Neural Information Processing Systems), MIT Press (2009)

[26] Davis, G., Mallat, S., Avellaneda, M.: Adaptive greedy approximations. Constructive Approximation **13** (1997) 57–98

[27] Das, A., Kempe, D.: Algorithms for subset selection in linear regression. In: Proc. 40th Annual ACM Symp. on Theory of Computing, ACM, New York (2008) 45–54

[28] Gilbert, A.C., Muthukrishnan, S., Strauss, M.J.: Approximation of functions over redundant dictionaries using coherence. In: Proc. 14th Annual ACM-SIAM Symp. on Discrete Algorithms. (2003)

[29] Tropp, J.A.: Greed is good: Algorithmic results for sparse approximation. IEEE Trans. on Information Theory **50** (2004) 2231–2242

[30] Honeine, P., Richard, C., Bermudez, J.C.M.: On-line nonlinear sparse approximation of functions. In: Proc. IEEE Int. Symp. on Information Theory. (2007) 956–960

[31] Zhang, X., Polycarpou, M.M., Parisini, T.: A robust detection and isolation scheme for abrupt and incipient faults in nonlinear systems. IEEE Trans. on Automatic Control **47** (2002) 576–593

[32] Zhang, X., Polycarpou, M.M., Parisini, T.: Robust fault isolation of a class of nonlinear input-output systems. Int. J. of Control **74** (2001) 1295–1310

[33] Zhang, T.: Sequential greedy approximation for certain convex optimization problems. IEEE Trans. on Information Theory **49** (2003) 682–691

[34] Vincent, P., Bengio, Y.: Kernel Matching Pursuit. Machine Learning **48** (2002) 165–187

[35] Smola, A.J., Schölkopf, B.: Sparse greedy matrix approximation for machine learning. In: Proc. 17th Int. Conf. on Machine Learning, Morgan Kaufmann (2000) 911–918

[36] Drineas, P., Mahoney, M.W.: On the Nyström method for approximating a Gram matrix for improved kernel-based learning. J. of Machine Learning Research **6** (2005) 2153–2175

[37] Tibshirani, R.: Regression shrinkage and selection via the LASSO. J. of the Royal Statistical Society, Series B **58** (1996) 267–288

[38] Roth, V.: The generalized Lasso. IEEE Trans. on Neural Networks **15** (2004) 16–28

[39] Efron, B., Hastie, T., Johnstone, L., Tibshirani, R.: Least angle regression. Annals of Statistics **32** (2004) 407–499

[40] Hale, E.T., Yin, W., Zhang, Y.: Fixed-point continuation for ℓ_1-minimization: Methodology and convergence. SIAM J. on Optimization **19** (2008) 1107–1130

[41] Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. of the Royal Statistical Society B **67** (2005) 301–320

[42] De Mol, C., De Vito, E., Rosasco, L.: Elastic-net regularization in learning theory. J. of Complexity **25** (2009) 201–230

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

[44] Kůrková, V.: 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 (1997) 261–270

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