Computer Science and Artificial Intelligence Laboratory

Technical Report

MIT-CSAIL-TR-2005-027
AIM-2005-013
CBCL-248

April 14, 2005

Fast Rates for Regularized Least-squares Algorithm

Andrea Caponnetto, Ernesto De Vito

massachusetts institute of technology, cambridge, ma 02139 usa — www.csail.mit.edu

Abstract

We develop a theoretical analysis of generalization performances of regularized least-squares on reproducing kernel Hilbert spaces for supervised learning. We show that the concept of *effective dimension* of an integral operator plays a central role in the definition of a criterion for the choice of the regularization parameter as a function of the number of samples. In fact a minimax analysis is performed which shows asymptotic optimality of the above mentioned criterion.

This report describes research done at the Center for Biological & Computational Learning, which is in the McGovern Institute for Brain Research at MIT, as well as in the Dept. of Brain & Cognitive Sciences, and which is affiliated with the Computer Sciences & Artificial Intelligence Laboratory (CSAIL).

This research was sponsored by grants from: Office of Naval Research (DARPA) Contract No. MDA972-04-1-0037, Office of Naval Research (DARPA) Contract No. N00014-02-1-0915, National Science Foundation (ITR/SYS) Contract No. IIS-0112991, National Science Foundation (ITR) Contract No. IIS-0209289, National Science Foundation-NIH (CRCNS) Contract No. EIA-0218693, National Science Foundation-NIH (CRCNS) Contract No. EIA-0218506, and National Institutes of Health (Conte) Contract No. 1 P20 MH66239-01A1.

Additional support was provided by: Central Research Institute of Electric Power Industry (CRIEPI), Daimler-Chrysler AG, Compaq/Digital Equipment Corporation, Eastman Kodak Company, Honda R&D Co., Ltd., Industrial Technology Research Institute (ITRI), Komatsu Ltd., Eugene McDermott Foundation, Merrill-Lynch, NEC Fund, Oxygen, Siemens Corporate Research, Inc., Sony, Sumitomo Metal Industries, and Toyota Motor Corporation.

1 Introduction

In this work we investigate the choice of the regularization parameter for the regularized least squares algorithm (RLS) on a reproducing kernel Hilbert space (RKHS) in the regression setting. This problem has already been extensively studied in the statistical learning literature. Probabilistic upper bounds on the excess risk of the empirical estimators are known and usually involve suitable priors on the regression function. In particular we recall that in [11] optimal rates are established assuming some smoothness condition on the regression function. In [3] a covering number technique has been used to obtain explicit bounds expressed in terms of suitable complexity measures of the regression function. In [5],[20], the covering techniques have been replaced by estimates of integral operators through concentration inequalities of vector valued random variables. Although expressed in terms of easily computable quantities the last bounds lack of nearly any information about the actual structure of the kernel in use. We show that such information can be exploited to obtain tighter bounds. The approach we consider is a refinement of the functional analytical techniques presented in [5]. The central concept in this development is the *effective dimension* which, roughly speaking, counts the number of degrees of freedom associated to the kernel and the marginal probability measure for a given condition number. The concept of effective dimension was recently used in [26] and [13] in the analysis of the performances of kernel methods estimators. Indeed in this work we show that effective dimension plays a role in the definition of an effective rule for the choice of the regularization parameter. In fact we prove that this rule is somehow optimal for the minimax problems induced by a certain family of priors.

Since the effective dimension depends on both the kernel and the marginal probability distribution on the input space, our choice for the regularization parameter is not completely distribution independent. In fact the spectrum of the integral operator depends dramatically on the marginal distribution but not on the dimension of the ambient space. These considerations raise the question whether the effective dimension could be estimated by unlabelled data.

The work is organized as follows. In sections 2 we recall very briefly the main concepts of statistical learning theory in the classical setting [4],[9],[16]. In section 3 we fix the notations and define the mathematical objects which will be considered. Furthermore we prove some preliminary results on the structure of RLS estimators and concentration of measure for vector valued random variables. In section 4 we prove the probabilistic upper bound for the excess risk of RLS estimators using the concept of effective dimension. In section 5 we give an explicit rule for the choice of the regularization parameter and compute the corresponding uniform rates of convergence in probability of the actual risk to its minimum. Finally in section 6 using entropy estimates we prove that the rates obtained in the previous section are indeed the optimal ones for the relevant minimax problem.

2 Learning from examples

We briefly recall some basic concepts of statistical learning theory in the regression setting (for details see [23], [9], [17], [2] and references therein).
In the framework of learning from examples there are two sets of variables: the input space X and the output space $Y \subset \mathbb{R}$. The relation between the input $x \in X$ and the output $y \in Y$ is described by a probability distribution $\rho(x,y) = \nu(x)\rho(y|x)$ on $X \times Y$, where ν is the marginal distribution on X and $\rho(\cdot|x)$ is the conditional distribution of y given $x \in X$. The distribution ρ is known only through a sample $\mathbf{z} = (\mathbf{x}, \mathbf{y}) = ((x_1, y_1), \ldots, (x_\ell, y_\ell))$, called *training set*, drawn independently and identically distributed (i.i.d.) according to ρ. Given the sample $\mathbf{z}$, the aim of learning theory is to find a function $f_{\mathbf{z}} : X \to \mathbb{R}$ such that $f_{\mathbf{z}}(x)$ is a good estimate of the output y when a new input x is given. The function $f_{\mathbf{z}}$ is called *estimator* and the map providing $f_{\mathbf{z}}$, for any training set $\mathbf{z}$, is called *learning algorithm.*

Given a function $f : X \to \mathbb{R}$, the ability of f to describe the distribution ρ is measured by its *expected risk* defined as

$$I[f] = \int_{X \times Y} (f(x) - y)^2 \, d\rho(x,y),$$

and the regression function

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

is the minimizer of the expected risk over the space of all the measurable real functions on X. In this sense f_ρ can be seen as the ideal estimator of the distribution probability ρ. However, the regression function cannot be reconstructed exactly since only a finite, possibly small, set of examples $\mathbf{z}$ is given.

To overcome this problem, in the framework of the regularized least squares algorithm, [24], [15], [2], [27], a Hilbert space $\mathcal{H}$ of real functions on X is fixed and the estimator ${f_{\mathbf{z}}}^\lambda$ is defined as the solution of the regularized least squares problem,

$$\min_{f \in \mathcal{H}} \{ \frac{1}{\ell} \sum_{i=1}^{\ell} (f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}{}^2 \}, \tag{1}$$

where λ is a positive parameter to be chosen in order to ensure that the discrepancy

$$I[{f_{\mathbf{z}}}^\lambda] - \inf_{f \in \mathcal{H}} I[f]$$

is small with hight probability. Since ρ is unknown, the above difference is studied by means of a probabilistic bound $B(\lambda, \ell, \eta)$, which is a function depending on the regularization parameter λ, the number ℓ of examples and the confidence level $1 - \eta$, such that

$$\mathbf{P} \left[I[{f_{\mathbf{z}}}^\lambda] - \inf_{f \in \mathcal{H}} I[f] \leq B(\lambda, \ell, \eta) \right] \geq 1 - \eta.$$

In particular, the learning algorithm is *consistent* if it is possible to choose the regularization parameter, as a function of the available data $\lambda = \lambda(\ell, \mathbf{z})$, in such a way that

$$\lim_{\ell \to +\infty} \mathbf{P}\left[I[f_{\mathbf{z}}{}^{\lambda(\ell,\mathbf{z})}] - \inf_{f \in \mathcal{H}} I[f] \geq \epsilon \right] = 0, \tag{2}$$

for every $\epsilon > 0$. The above convergence in probability is usually called *(weak) consistency* of the algorithm (see [7] for a discussion on different types of consistency).

3 Notations, assumptions and preliminary results

We assume that the input space X is a separable metric space and the output space Y is a closed subset of $\mathbb{R}$. We let Z be the product space $X \times Y$, which is a separable metric space.

We denote by ν the marginal distribution on X and by $\rho(y|x)$ the conditional distribution of $y \in Y$ given $x \in X$. Let $L^2(Z, \rho, Y)$ be the Hilbert space of square integrable functions on Z with respect to ρ and we denote by $\|\cdot\|_\rho$ and $\langle \cdot, \cdot \rangle_\rho$ the corresponding norm and scalar product. Similar notation we use for $L^2(X, \rho_X, Y)$. Moreover we assume that ν is *not degenerate*, i.e. all the non-void open subsets of X have a strictly positive measure.

We assume that the space $\mathcal{H}$ is a reproducing kernel Hilbert space, [1], [18], with a separately continuous kernel $K : X \times X \to \mathbb{R}$ such that

$$\kappa = \sup_{x \in X} K(x, x) < +\infty. \tag{3}$$

As usual $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and $\|\cdot\|_{\mathcal{H}}$ denote the corresponding scalar product and norm. Since K is a separately continuous bounded kernel and X is separable, $\mathcal{H}$ is a real separable Hilbert space whose elements are real continuous functions defined on X, [18]. Moreover, given $x \in X$ the function $K_x = K(\cdot, x)$ belongs to $\mathcal{H}$ and the following *reproducing* property holds

$$f(x) = \langle f, K_x \rangle_{\mathcal{H}} \qquad \forall f \in \mathcal{H} \tag{4}$$

and (3) ensures

$$\|K_x\|_{\mathcal{H}}{}^2 = K(x, x) \leq \kappa. \tag{5}$$

We are ready to state the hypotheses on the probability measure ρ. Firstly we will assume the condition

$$\int_Z y^2 \, d\rho(x, y) < +\infty. \tag{6}$$

Moreover we will assume that some $f_{\mathcal{H}} \in \mathcal{H}$ exists which attains the infimum of the expected risk, that is

$$I[f_{\mathcal{H}}] = \inf_{f \in \mathcal{H}} I[f], \tag{7}$$

and that for some $M > 0$

$$|y - f_{\mathcal{H}}(x)|^2 \leq M \quad \rho - a.s. \tag{8}$$

Let us now review some properties of the operator $A : \mathcal{H} \to L^2(Z, \rho, Y)$ defined as follows

$$(Af)(x, y) = \langle f, K_x \rangle_{\mathcal{H}}.$$

Equation (4) implies that the action of A on an element f is simply

$$(Af)(x, y) = f(x) \quad \forall x \in X,\ f \in \mathcal{H},$$

that is, A is the canonical inclusion of $\mathcal{H}$ into $L^2(Z, \rho, Y)$, where the variable y is *dumb*. However, A changes the norm since $\|f\|_{\mathcal{H}}$ is different from $\|f\|_\rho$. The main properties of the operator A are summarized in the following proposition.

Proposition 1 *The operator A is an injective Hilbert-Schmidt operator from $\mathcal{H}$ into $L^2(Z, \rho, Y)$ and*

$$A^*\phi = \int_Z \phi(x, y) K_x \, d\rho(x, y) \tag{9}$$

$$T := A^*A = \int_X \langle \cdot, K_x \rangle_{\mathcal{H}} K_x \, d\nu(x), \tag{10}$$

where $\phi \in L^2(Z, \rho, Y)$, the first integral converges in norm and the second one in trace norm. In particular T is a trace class injective operator from $\mathcal{H}$ to $\mathcal{H}$.

PROOF. The proof is standard and we report it for completeness.
Since the elements $f \in \mathcal{H}$ are continuous functions, (5) bounds them by

$$|f(x)| = |\langle f, K_x \rangle_{\mathcal{H}}| \leq \|f\|_{\mathcal{H}} \|K_x\|_{\mathcal{H}} \leq \sqrt{\kappa} \|f\|_{\mathcal{H}}.$$

Since ρ is a probability measure, then $f \in L^2(Z, \rho, Y)$ and A is a linear operator from $\mathcal{H}$ to $L^2(Z, \rho, Y)$ with $\|Af\|_\rho \leq \sqrt{\kappa} \|f\|_{\mathcal{H}}$, so that A is bounded.
We now show that A is injective. Let $f \in \mathcal{H}$ and $W = \{x \in X \mid f(x) \neq 0\}$, which is open since f is continuous. Assume now that $Af = 0$, i.e., $f(x) = 0$ for ν-almost all $x \in X$, then W has null measure. Since ν is not degenerate, W is the empty set, i.e., $f(x) = 0$ for all $x \in X$, so that $f = 0$.
We now prove (9). Since K is separately continuous, the map

$$X \ni x \mapsto K_x \in \mathcal{H}$$

is weakly continuous and, since $\mathcal{H}$ is separable, is strongly measurable. Hence, given $\phi \in L^2(Z, \rho, Y)$, the map $(x, y) \mapsto \phi(x, y) K_x$ is measurable from Z to $\mathcal{H}$. Moreover, (5) gives

$$\|\phi(x, y) K_x\|_{\mathcal{H}} \leq |\phi(x, y)| \sqrt{\kappa}$$

for all $x \in X$. Since ρ is finite, ϕ is in $L^1(Z,\rho)$ and, hence, $(x,y) \mapsto \phi(x,y)K_x$ is integrable, as a vector valued map. Finally, for all $f \in \mathcal{H}$,

$$\int_Z \phi(x,y)\,\langle K_x, f\rangle_{\mathcal{H}}\, d\rho(x,y) = \langle \phi, Af\rangle_\rho = \langle A^*\phi, f\rangle_{\mathcal{H}}\,,$$

so (9) holds.
Equation (10) is a consequence of (9),the fact that the integral commutes with the scalar product and the definition of marginal distribution ν.
We now prove that A is a Hilbert-Schmidt operator. Let $(e_n)_{n\in\mathbb{N}}$ be a Hilbert basis of $\mathcal{H}$, which is separable. Since A^*A is a positive operator and $|\langle K_x, e_n\rangle_{\mathcal{H}}|^2$ is a positive function, by monotone convergence theorem, we have that

$$\begin{aligned}
\mathrm{Tr}\,(A^*A) &= \sum_n \int_X |\langle e_n, K_x\rangle_{\mathcal{H}}|^2\, d\nu(x) \\
&= \int_X \sum_n |\langle e_n, K_x\rangle_{\mathcal{H}}|^2\, d\nu(x) \\
&= \int_X \langle K_x, K_x\rangle_{\mathcal{H}}\, d\nu(x) \\
&= \int_X K(x,x)\, d\nu(x) \le \kappa
\end{aligned}$$

and the thesis follows. The properties of T are an easy consequence of the corresponding properties of A. □

The following proposition clarifies the role of the operator A in the context of learning theory. The result is well known in the framework of linear inverse problems, see for example [8]. With slight abuse of notation we denote by y both the variable and the function $(x,y) \mapsto y$, which belongs to $L^2(Z,\rho,Y)$ due to (6).

Proposition 2 *If a minimizer $f_{\mathcal{H}}$ of the expected risk $I[f]$ exists on $\mathcal{H}$, then it is unique, satisfies*

$$Tf_{\mathcal{H}} = A^*y. \tag{11}$$

and

$$I[f] - I[f_{\mathcal{H}}] = \|A(f - f_{\mathcal{H}})\|_\rho{}^2 = \left\|\sqrt{T}(f - f_{\mathcal{H}})\right\|_{\mathcal{H}}{}^2 \tag{12}$$

for all $f \in \mathcal{H}$.
If $\lambda > 0$ a unique minimizer f^λ of the regularized expected risk

$$I[f] + \lambda \|f\|_{\mathcal{H}}{}^2$$

exists and it is given by

$$f^\lambda = (T+\lambda)^{-1}A^*y = (T+\lambda)^{-1}Tf_{\mathcal{H}}. \tag{13}$$

PROOF. Clearly

$$I[f] = \|Af - y\|_\rho^{\,2}$$

for all $f \in \mathcal{H}$. Since $f_{\mathcal{H}}$ is a minimizer, by differentiation we obtain

$$\langle Af, Af_{\mathcal{H}} - y\rangle_\rho = 0 \quad \forall f \in \mathcal{H} \tag{14}$$

and (11) follows. The uniqueness is ensured by the injectivity of T.
Given $f \in \mathcal{H}$

$$\begin{aligned} I[f] - I[f_{\mathcal{H}}] &= \|Af - y\|_\rho^{\,2} - \|Af_{\mathcal{H}} - y\|_\rho^{\,2} \\ &= \|A(f - f_{\mathcal{H}})\|_\rho^{\,2} + 2\langle A(f - f_{\mathcal{H}}), Af_{\mathcal{H}} - y\rangle_\rho \\ &= \|A(f - f_{\mathcal{H}})\|_\rho^{\,2} \end{aligned}$$

since the second term is zero due to (14). Let $A = U\sqrt{T}$ be the polar decomposition. Since U is a partial isometry from the closure of the range of $\sqrt{T}$ onto the closure of the range of A

$$\|A(f - f_{\mathcal{H}})\|_\rho = \left\|\sqrt{T}(f - f_{\mathcal{H}})\right\|_{\mathcal{H}}.$$

Finally, (13) follows taking the derivative be equal to zero. □

Let now $\mathbf{z} = (\mathbf{x}, \mathbf{y}) = ((x_1, y_1), \ldots, (x_\ell, y_\ell)) \in Z^\ell$ be the training set. The above arguments can be repeated replacing the measure ρ with the empirical measure $\rho_{\mathbf{z}} = \frac{1}{\ell}\sum_{i=1}^{\ell} \delta_{(x_i,y_i)}$ where $\delta_{(x,y)}$ is the Dirac measure at point $(x, y) \in Z$. An element $\phi \in L^2(Z, \rho_{\mathbf{z}})$ is completely specified by the vector $\mathbf{w} \in \mathbb{R}^\ell$ given by

$$\mathbf{w}_i = \phi(x_i, y_i)$$

with the condition that $\mathbf{w}_i = \mathbf{w}_j$ whenever $(x_i, y_i) = (x_j, y_j)$. In the following we represent the elements of $L^2(Z, \rho_{\mathbf{z}})$ as vectors in $\mathbb{R}^\ell$ where scalar product is given by

$$\left\langle \mathbf{w}, \mathbf{w}' \right\rangle_{L^2(Z,\rho_{\mathbf{z}})} = \frac{1}{\ell}\sum_{i=1}^{\ell} \mathbf{w}_i \mathbf{w}'_i.$$

We now get a discretized version of A by defining the *sampling operator* [19]

$$A_{\mathbf{z}} : \mathcal{H} \to L^2(Z, \rho_{\mathbf{z}})$$

$$(A_{\mathbf{z}} f)_i = \langle f, K_{x_i}\rangle_{\mathcal{H}} = f(x_i) \quad \forall i = 1, \ldots, \ell.$$

The main properties of the sampling operator are given by the following proposition.

Proposition 3 *The sampling operator $A_{\mathbf{z}} : \mathcal{H} \to L^2(Z, \rho_{\mathbf{z}})$ is a finite rank operator and*

$$A_{\mathbf{z}}{}^{*}\mathbf{w} = \frac{1}{\ell}\sum_{i=1}^{\ell} \mathbf{w}_i K_{x_i} \qquad \mathbf{w} \in L^2(Z, \rho_{\mathbf{z}}) \tag{15}$$

$$T_{\mathbf{x}} := A_{\mathbf{z}}{}^{*} A_{\mathbf{z}} = \frac{1}{\ell}\sum_{i=1}^{\ell} \langle \cdot, K_{x_i} \rangle_{\mathcal{H}} K_{x_i}. \tag{16}$$

If $\lambda > 0$ a unique minimizer $f_{\mathbf{z}}{}^{\lambda}$ of the regularized empirical error

$$\frac{1}{\ell}\sum_{i=1}^{\ell} (f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}{}^2$$

exists and is given by

$$f_{\mathbf{z}}^{\lambda} = (T_{\mathbf{x}} + \lambda)^{-1} A_{\mathbf{x}}^{*} \mathbf{y}. \tag{17}$$

PROOF. The content of the proposition is a restatement of Proposition 1 and the fact that the integrals reduce to sums.

Finally we need the following probabilistic inequality due to Pinelis and Sakhanenko [14], [25].

Proposition 4 *Let $(\Omega, \mathcal{F}, P)$ be a probability space and ξ be a random variable on Ω taking value in a real separable Hilbert space $\mathcal{H}$. Assume that there are two positive constants H and σ such that*

$$\|\xi(\omega)\|_{\mathcal{H}} \leq \frac{H}{2} \qquad \text{a.}s \tag{18}$$

$$\mathbb{E}[\|\xi\|_{\mathcal{H}}{}^2] \leq \sigma^2. \tag{19}$$

Let $\ell \in \mathbb{N}$ and $0 < \eta < 1$, then

$$\mathbf{P}^{\ell}\left[(\omega_1, \ldots, \omega_\ell) \in \Omega^{\ell} \mid \left\| \frac{1}{\ell}\sum_{i=1}^{\ell} \xi(\omega_i) - \mathbb{E}[\xi] \right\|_{\mathcal{H}} \leq 2\left(\frac{H}{\ell} + \frac{\sigma}{\sqrt{\ell}}\right)\log\frac{2}{\eta}\right] \geq 1 - \eta. \tag{20}$$

PROOF. It is just a restatement of Th. 3.3.4 of [25], see also [21]. Consider the probability space $(\Omega^{\ell}, \mathcal{F}^{\ell}, P^{\ell})$ and the set of independent random variables with zero mean $\xi_i(\omega_1, \ldots, \omega_\ell) = \xi(\omega_i) - \mathbb{E}[\xi]$ defined on Ω^{ℓ}. The fact that ξ_i are i.i.d and conditions (18), (19) ensure that

$$\|\xi_i\|_{\mathcal{H}} \leq H \qquad \text{a.}s$$
$$\mathbb{E}[\|\xi_i\|_{\mathcal{H}}{}^2] \leq \sigma^2,$$

so that, for all $m \geq 2$ it holds

$$\sum_{i=1}^{\ell} \mathbb{E}[\|\xi_i\|_{\mathcal{H}}^m] \leq \frac{1}{2} m! B^2 H^{m-2},$$

with $B^2 = \ell\sigma^2$. So Th. 3.3.4 of [25] can be applied and it ensures

$$\mathbf{P}\left[\frac{1}{\ell}\left\|\sum_{i=1}^{\ell}(\xi(z_i) - \mathbb{E}[\xi])\right\| \geq \frac{xB}{\ell}\right] \leq 2\exp\left(-\frac{x^2}{2(1+xHB^{-1})}\right).$$

for all $x \geq 0$. Letting $\delta = \frac{xB}{\ell}$, we get the equation

$$\frac{1}{2}(\frac{\ell\delta}{B})^2 \frac{1}{1+\ell\delta HB^{-2}} = \frac{\ell\delta^2\sigma^{-2}}{2(1+\delta H\sigma^{-2})} = \log\frac{2}{\eta},$$

since $B^2 = \ell\sigma^2$. Defining $t = \delta H\sigma^{-2}$

$$\frac{\ell\sigma^2}{2H^2}\frac{t^2}{1+t} = \log\frac{2}{\eta}.$$

The inverse of the function $\frac{t^2}{1+t}$ is the function $g(t) = \frac{1}{2}(t + \sqrt{t^2+4t})$ so

$$\left\|\frac{1}{\ell}\sum_{i=1}^{\ell}\xi(z_i) - \mathbb{E}[\xi]\right\|_{\mathcal{H}} \leq \frac{\sigma^2}{H}\, g\left(\frac{2H^2}{\ell\sigma^2}\log\frac{2}{\eta}\right)$$

with probability greater than $1-\eta$. The thesis follows observing that $g(t) \leq t + \sqrt{t}$ and $2\log\frac{2}{\eta} \geq \sqrt{2\log\frac{2}{\eta}} \geq 1$. □

4 Upper bound

The aim of this section is to give a probabilistic upper bound on the expect risk of the solution given by the regularized least squares algorithm. The bound depends on the number of examples ℓ, the regularization parameter and some prior information on the probability distribution ρ.

In the following, we assume that the space $\mathcal{H}$ and the probability distribution ρ satisfy the assumptions (3), (6), (7) and (8). Set the parameter $\lambda > 0$ we define

(1) the *residual*

$$\mathcal{A}(\lambda) = \left\|f^\lambda - f_{\mathcal{H}}\right\|_\rho^2 = \left\|\sqrt{T}(f^\lambda - f_{\mathcal{H}})\right\|_{\mathcal{H}}^2,$$

where T is given by (10), f^λ by (13) and $f_{\mathcal{H}}$ by (7);

(2) the *reconstruction error*

$$\mathcal{B}(\lambda) = \left\|f^\lambda - f_{\mathcal{H}}\right\|_{\mathcal{H}}^2;$$

(3) the *effective dimension*

$$\mathcal{N}(\lambda) = \mathrm{Tr}[(T+\lambda)^{-1}T],$$

where the trace is finite due to Proposition 1.

In the framework of learning $\mathcal{A}(\lambda)$ is called approximation error, whereas in the framework of approximation theory $\sqrt{\mathcal{B}(\lambda)}$ is the approximation error. To avoid confusion we follow the notation of inverse problems.

We are now ready to state our main result of the section.

Theorem 5 *Let $\mathbf{z} \in Z^\ell$ be a training set drawn i.i.d according to ρ and ${f_\mathbf{z}}^\lambda \in \mathcal{H}$ the corresponding estimator given by (17). With probability greater than $1-\eta$, $0 < \eta < 1$,*

$$I[{f_\mathbf{z}}^\lambda] - I[f_\mathcal{H}] \leq C_\eta \left(\mathcal{A}(\lambda) + \frac{\kappa^2 \mathcal{B}(\lambda)}{\ell^2 \lambda} + \frac{\kappa \mathcal{A}(\lambda)}{\ell \lambda} + \frac{\kappa M}{\ell^2 \lambda} + \frac{M \mathcal{N}(\lambda)}{\ell} \right) \tag{21}$$

provided that

$$\ell \geq \frac{C_\eta \kappa}{2\lambda} \max(\mathcal{N}(\lambda), \sqrt{2/C_\eta}) \tag{22}$$

where $C_\eta = 128 \log^2(8/\eta)$.

PROOF. We split the proof in several steps. Here $\|\cdot\|$ denotes the uniform norm of an operator from $\mathcal{H}$ to $\mathcal{H}$ Let λ, η and ℓ as in the statement of the theorem.
Step 1: Given a training set $\mathbf{z} = (\mathbf{x}, \mathbf{y}) \in Z^\ell$, (12) gives

$$I[{f_\mathbf{z}}^\lambda] - I[f_\mathcal{H}] = \left\| \sqrt{T}({f_\mathbf{z}}^\lambda - f_\mathcal{H}) \right\|_\mathcal{H}^2.$$

As usual,

$${f_\mathbf{z}}^\lambda - f_\mathcal{H} = ({f_\mathbf{z}}^\lambda - f^\lambda) + (f^\lambda - f_\mathcal{H})$$

and (13), (17) give

$$\begin{aligned}
{f_\mathbf{z}}^\lambda - f^\lambda &= \left((T_\mathbf{x}+\lambda)^{-1} {A_\mathbf{z}}^* \mathbf{y}\right) - \left((T+\lambda)^{-1} A^* y\right) \\
&= (T_\mathbf{x}+\lambda)^{-1} \left\{ ({A_\mathbf{z}}^* \mathbf{y} - A^* y) + (T - T_\mathbf{x})(T+\lambda)^{-1} A^* y \right\} \\
(\text{ Eq. (11) }) &= (T_\mathbf{x}+\lambda)^{-1} \left\{ ({A_\mathbf{z}}^* \mathbf{y} - T_\mathbf{x} f_\mathcal{H} + T_\mathbf{x} f_\mathcal{H} - T f_\mathcal{H}) + (T - T_\mathbf{x}) f^\lambda \right\} \\
&= (T_\mathbf{x}+\lambda)^{-1} ({A_\mathbf{z}}^* \mathbf{y} - T_\mathbf{x} f_\mathcal{H}) + (T_\mathbf{x}+\lambda)^{-1} (T - T_\mathbf{x})(f^\lambda - f_\mathcal{H}).
\end{aligned}$$

The inequality $\|f_1 + f_2 + f_3\|_\mathcal{H}^2 \leq 3(\|f_1\|_\mathcal{H}^2 + \|f_2\|_\mathcal{H}^2 + \|f_3\|_\mathcal{H}^2)$ implies

$$I[{f_\mathbf{z}}^\lambda] - I[f_\mathcal{H}] \leq 3 \left(\mathcal{A}(\lambda) + \mathcal{S}_1(\lambda, \mathbf{z}) + \mathcal{S}_2(\lambda, \mathbf{z}) \right) \tag{23}$$

where

$$\mathcal{S}_1(\lambda, \mathbf{z}) = \left\| \sqrt{T} (T_\mathbf{x}+\lambda)^{-1} ({A_\mathbf{z}}^* \mathbf{y} - T_\mathbf{x} f_\mathcal{H}) \right\|_\mathcal{H}^2$$

$$\mathcal{S}_2(\lambda,\mathbf{z}) = \left\|\sqrt{T}(T_{\mathbf{x}}+\lambda)^{-1}(T-T_{\mathbf{x}})(f^\lambda - f_{\mathcal{H}})\right\|_{\mathcal{H}}^2.$$

Step 2: probabilistic bound on $\mathcal{S}_2(\lambda,\mathbf{z})$. Clearly

$$\mathcal{S}_2(\lambda,\mathbf{z}) \leq \left\|\sqrt{T}(T_{\mathbf{x}}+\lambda)^{-1}\right\|_{\mathcal{L}(\mathcal{H})}^2 \left\|(T-T_{\mathbf{x}})(f^\lambda - f_{\mathcal{H}})\right\|_{\mathcal{H}}^2. \tag{24}$$

Step 2.1: probabilistic bound on $\left\|\sqrt{T}(T_{\mathbf{x}}+\lambda)^{-1}\right\|$. Assume that

$$\Theta(\lambda,\mathbf{z}) = \left\|(T+\lambda)^{-1}(T-T_{\mathbf{x}})\right\| \leq \frac{1}{2}, \tag{25}$$

then the Neumann series gives

$$\begin{aligned}\sqrt{T}(T_{\mathbf{x}}+\lambda)^{-1} &= \sqrt{T}(T+\lambda)^{-1}(I-(T+\lambda)^{-1}(T-T_{\mathbf{x}}))^{-1}\\ &= \sqrt{T}(T+\lambda)^{-1}\sum_{n=0}^{+\infty}\left((T+\lambda)^{-1}(T-T_{\mathbf{x}})\right)^n\end{aligned}$$

so that

$$\begin{aligned}\left\|\sqrt{T}(T_{\mathbf{x}}+\lambda)^{-1}\right\| &\leq \left\|\sqrt{T}(T+\lambda)^{-1}\right\| \sum_{n=0}^{+\infty}\left\|(T+\lambda)^{-1}(T-T_{\mathbf{x}})\right\|_{\mathcal{L}(\mathcal{H})}^n\\ &\leq \frac{1}{2\sqrt{\lambda}}\frac{1}{1-\Theta(\lambda,\mathbf{z})},\end{aligned}$$

where, by spectral theorem, $\left\|\sqrt{T}(T+\lambda)^{-1}\right\| \leq \frac{1}{2\sqrt{\lambda}}$. Inequality (25) now gives

$$\left\|\sqrt{T}(T_{\mathbf{x}}+\lambda)^{-1}\right\| \leq \frac{1}{\sqrt{\lambda}}. \tag{26}$$

We claim that (22) implies (25) with probability greater than $1-\eta$. Indeed, let $\mathcal{L}_2(\mathcal{H})$ be the Hilbert space of Hilbert-Schmidt operators on $\mathcal{H}$ (recall that $\langle A,B\rangle_{\mathcal{L}_2(\mathcal{H})} = \mathrm{Tr}[A^*B]$). Let us identify $\mathcal{L}_2(\mathcal{H})$ with $\mathcal{H}\otimes\mathcal{H}$, and let $\xi_1 : X \to \mathcal{L}_2(\mathcal{H})$ be the random variable

$$\xi_1(x) = \langle\cdot, K_x\rangle_{\mathcal{H}}(T+\lambda)^{-1}K_x = K_x \otimes (T+\lambda)^{-1}K_x.$$

Bound (5) and $\left\|(T+\lambda)^{-1}\right\| \leq \frac{1}{\lambda}$ imply

$$\|\xi\|_{\mathcal{H}\otimes\mathcal{H}} = \|K_x\|_{\mathcal{H}}\left\|(T+\lambda)^{-1}K_x\right\|_{\mathcal{H}} \leq \frac{\kappa}{\lambda} = \frac{H_1}{2},$$

and

$$\begin{aligned}\mathbb{E}[\|\xi_1\|_{\mathcal{H}\otimes\mathcal{H}}^2] &= \int_X \|K_x\|_{\mathcal{H}}^2\left\|(T+\lambda)^{-1}K_x\right\|_{\mathcal{H}}^2 d\nu(x)\\ &\leq \kappa\int_X \left\langle(T+\lambda)^{-2}K_x, K_x\right\rangle_{\mathcal{H}} d\nu(x)\\ &= \kappa\,\mathrm{Tr}[(T+\lambda)^{-2}T]\\ &\leq \kappa\left\|(T+\lambda)^{-1}\right\|\mathrm{Tr}[(T+\lambda)^{-1}T]\\ &\leq \frac{\kappa}{\lambda}\mathcal{N}(\lambda) = \sigma_1^2,\end{aligned}$$

Observing that

$$\mathbb{E}[\xi_1] = T(T+\lambda)^{-1} \qquad \frac{1}{\ell}\sum_{i=1}^{\ell}\xi_1(x_i) = (T+\lambda)^{-1}T_{\mathbf{x}},$$

Proposition 4 applied to ξ_1 gives

$$\left\|(T+\lambda)^{-1}T_{\mathbf{x}} - T(T+\lambda)^{-1}\right\|_{\mathcal{L}_2(\mathcal{H})} \leq 2\log(6/\eta)\left(\frac{2\kappa}{\lambda\ell} + \sqrt{\frac{\kappa\mathcal{N}(\lambda)}{\lambda\ell}}\right)$$

with probability greater than $1-\eta/3$. Then for all $\ell \in \mathbb{N}$ satisfying (22)

$$\log(6/\eta)\left(\frac{2\kappa}{\lambda\ell} + \sqrt{\frac{\kappa\mathcal{N}(\lambda)}{\lambda\ell}}\right) \leq \frac{1}{8} + \frac{1}{8} \leq \frac{1}{4}$$

so that

$$\Theta(\lambda, \mathbf{z}) \leq \left\|(T+\lambda)^{-1}T_{\mathbf{x}} - T(T+\lambda)^{-1}\right\|_{\mathcal{L}_2(\mathcal{H})} \leq \frac{1}{2} \tag{27}$$

with probability greater than $1-\eta/3$.

Step 2.2: probabilistic bound on $\left\|(T-T_{\mathbf{x}})(f^{\lambda}-f_{\mathcal{H}})\right\|$. Let $\xi_2 : X \to \mathcal{H}$ be the random variable

$$\xi_2(x) = \left\langle f^{\lambda} - f_{\mathcal{H}}, K_x \right\rangle_{\mathcal{H}} K_x.$$

Bound (5) and the definition of $\mathcal{B}(\lambda)$ give

$$\|\xi_2(x)\|_{\mathcal{H}} \leq {\|K_x\|_{\mathcal{H}}}^2 \left\|f^{\lambda} - f_{\mathcal{H}}\right\|_{\mathcal{H}} \leq \kappa\sqrt{\mathcal{B}(\lambda)} = \frac{H_2}{2},$$

and

$$\begin{aligned}
\mathbb{E}[{\|\xi_2\|_{\mathcal{H}}}^2] &= \int_X {\|K_x\|_{\mathcal{H}}}^2 \left\langle f^{\lambda} - f_{\mathcal{H}}, K_x \right\rangle_{\mathcal{H}}^2 d\nu(x) \\
&\leq \kappa \left\langle T(f^{\lambda} - f_{\mathcal{H}}), f^{\lambda} - f_{\mathcal{H}} \right\rangle_{\mathcal{H}} \\
&= \kappa \left\| \sqrt{T}(f^{\lambda} - f_{\mathcal{H}}) \right\|_{\mathcal{H}}^2 \\
&= \kappa\mathcal{A}(\lambda) = \sigma_2^2.
\end{aligned}$$

Observing that

$$\mathbb{E}[\xi_2] = T(f^{\lambda} - f_{\mathcal{H}}) \qquad \frac{1}{\ell}\sum_{i=1}^{\ell}\xi_2(x_i) = T_{\mathbf{x}}(f^{\lambda} - f_{\mathcal{H}}),$$

Proposition 4 applied to ξ_2 gives

$$\left\|(T-T_{\mathbf{x}})(f^{\lambda}-f_{\mathcal{H}})\right\|_{\mathcal{H}} \leq 2\log(6/\eta)\left(\frac{2\kappa\sqrt{\mathcal{B}(\lambda)}}{\ell} + \sqrt{\frac{\kappa\mathcal{A}(\lambda)}{\ell}}\right). \tag{28}$$

with probability greater than $1-\eta/3$. Replacing (26), (28) in (24), for all $\ell \in \mathbb{N}$ satisfying (22) it holds

$$\mathcal{S}_2(\lambda, \mathbf{z}) \leq 8 \log^2(6/\eta) \left(\frac{4\kappa^2 \mathcal{B}(\lambda)}{\ell^2 \lambda} + \frac{\kappa \mathcal{A}(\lambda)}{\ell \lambda} \right) \tag{29}$$

with probability greater than $1-2\eta/3$.
Step 3: probabilistic bound on $\mathcal{S}_1(\lambda, \mathbf{z})$. Clearly

$$\mathcal{S}_1(\lambda, \mathbf{z}) \leq \left\| \sqrt{T}(T_{\mathbf{x}} + \lambda)^{-1}(T+\lambda)^{\frac{1}{2}} \right\|^2_{\mathcal{L}(\mathcal{H})} \left\| (T+\lambda)^{-\frac{1}{2}} \left(A_{\mathbf{z}}{}^* \mathbf{y} - T_{\mathbf{x}} f_{\mathcal{H}} \right) \right\|^2_{\mathcal{H}}. \tag{30}$$

Step 3.1: bound on $\left\| \sqrt{T}(T_{\mathbf{x}} + \lambda)^{-1}(T+\lambda)^{\frac{1}{2}} \right\|$. Clearly,

$$\sqrt{T}(T_{\mathbf{x}} + \lambda)^{-1}(T+\lambda)^{\frac{1}{2}} = \sqrt{T}(T+\lambda)^{-\frac{1}{2}} \left\{ I - (T+\lambda)^{-\frac{1}{2}}(T - T_{\mathbf{x}})(T+\lambda)^{-\frac{1}{2}} \right\}^{-1}.$$

Spectral theorem ensures that $\left\| \sqrt{T}(T+\lambda)^{-\frac{1}{2}} \right\| \leq 1$ so, reasoning as in Step 2.1,

$$\left\| \sqrt{T}(T_{\mathbf{x}} + \lambda)^{-1}(T+\lambda)^{\frac{1}{2}} \right\| \leq 2 \tag{31}$$

provided that

$$\left\| (T+\lambda)^{-\frac{1}{2}}(T - T_{\mathbf{x}})(T+\lambda)^{-\frac{1}{2}} \right\| \leq \frac{1}{2}. \tag{32}$$

If $B = (T+\lambda)^{-\frac{1}{2}}(T - T_{\mathbf{x}})(T+\lambda)^{-\frac{1}{2}}$, then

$$\begin{aligned}
\|B\|^2_{\mathcal{L}_2(\mathcal{H})} &= \operatorname{Tr}\left((T+\lambda)^{-1}(T - T_{\mathbf{x}})(T+\lambda)^{-1}(T - T_{\mathbf{x}}) \right) \\
&= \left\langle (T+\lambda)^{-1}(T - T_{\mathbf{x}}), \left((T+\lambda)^{-1}(T - T_{\mathbf{x}}) \right)^* \right\rangle_{\mathcal{L}_2(\mathcal{H})} \\
&\leq \left\| (T+\lambda)^{-1}(T - T_{\mathbf{x}}) \right\|_{\mathcal{L}_2(\mathcal{H})} \left\| \left((T+\lambda)^{-1}(T - T_{\mathbf{x}}) \right)^* \right\|_{\mathcal{L}_2(\mathcal{H})} \\
&= \left\| (T+\lambda)^{-1}(T - T_{\mathbf{x}}) \right\|^2_{\mathcal{L}_2(\mathcal{H})},
\end{aligned}$$

and, for all $\ell \in \mathbb{N}$ satisfying (22), (27) ensures that (32) holds with probability $1-2\eta/3$.
Step 3.2: bound on $\left\| (T+\lambda)^{-\frac{1}{2}} \left(A_{\mathbf{z}}{}^* \mathbf{y} - T_{\mathbf{x}} f_{\mathcal{H}} \right) \right\|_{\mathcal{H}}$. Let $\xi_3 : X \times Y \to \mathcal{H}$ be the random variable

$$\xi_3(x, y) = (T+\lambda)^{-\frac{1}{2}} K_x \left(y - f_{\mathcal{H}}(x) \right).$$

The definition of M gives

$$\|\xi_3(x, y)\|_{\mathcal{H}} \leq \left\| (T+\lambda)^{-\frac{1}{2}} \right\|_{\mathcal{H}} \|K_x\|_{\mathcal{H}} \sqrt{M} \leq \sqrt{\frac{\kappa M}{\lambda}} = \frac{H_3}{2}$$

almost surely, and

$$\begin{aligned}\mathbb{E}[\|\xi_3\|_{\mathcal{H}}{}^2] &= \int_{X\times Y} (y - f_{\mathcal{H}}(x))^2 \left\|(T+\lambda)^{-\frac{1}{2}} K_x\right\|_{\mathcal{H}}{}^2 d\nu(x) \\ &\leq M \int_X \left\langle (T+\lambda)^{-1} K_x, K_x \right\rangle_{\mathcal{H}} d\nu(x) \\ &= M \operatorname{Tr}[(T+\lambda)^{-1} T] = M\mathcal{N}(\lambda) = \sigma_3^2.\end{aligned}$$

Equation (11) gives

$$\mathbb{E}[\xi_3] = (T+\lambda)^{-\frac{1}{2}} (A^* y - T f_{\mathcal{H}}) = 0,$$

so Proposition 4 applied to ξ_3 ensures

$$\left\|(T+\lambda)^{-\frac{1}{2}} (A_{\mathbf{z}}{}^* \mathbf{y} - T_{\mathbf{x}} f_{\mathcal{H}})\right\|_{\mathcal{H}} \leq 2\log(6/\eta) \left(\frac{2}{\ell}\sqrt{\frac{\kappa M}{\lambda}} + \sqrt{\frac{M\mathcal{N}(\lambda)}{\ell}}\right) \tag{33}$$

with probability greater than $1 - \eta/3$. Replacing (31), (33) in (30)

$$\mathcal{S}_1(\lambda, \mathbf{z}) \leq 32 \log^2(6/\eta) \left(\frac{4\kappa M}{\ell^2 \lambda} + \frac{M\mathcal{N}(\lambda)}{\ell}\right). \tag{34}$$

with probability greater than $1 - \eta$.
Replacing bounds (29), (34) in (23),

$$I[f_{\mathbf{z}}{}^{\lambda}] - I[f_{\mathcal{H}}] \leq 3\mathcal{A}(\lambda) + 8\log^2(6/\eta) \left(\frac{4\kappa^2 \mathcal{B}(\lambda)}{\ell^2 \lambda} + \frac{\kappa \mathcal{A}(\lambda)}{\ell\lambda} + \frac{16\kappa M}{\ell^2\lambda} + \frac{4M\mathcal{N}(\lambda)}{\ell}\right)$$

and (21) follows by bounding the numerical constants with 128.

5 A priori regularization parameter choice

In this section we discuss the choice of the parameter $\lambda = \lambda_\ell$ as a function of the number of examples ℓ in such a way to obtain a maximal rate of convergence.

The following lemma studies the dependence of $\mathcal{A}(\lambda)$, $\mathcal{B}(\lambda)$ and $\mathcal{N}(\lambda)$ on λ. We let N be the dimension of $\mathcal{H}$ (possibly $N = +\infty$) and

$$T = \sum_{n=1}^{N} t_n \langle \cdot, e_n \rangle e_n$$

be the spectral decomposition of T with $0 < t_{n+1} \leq t_n$ and $(e_n)_{n=1}^N$ be a basis of $\mathcal{H}$.

Lemma 6 *With the above notations,*

$$\lim_{\lambda \to 0} \mathcal{N}(\lambda) = N, \tag{35}$$

in particular if $N = +\infty$ and $t_n = O(n^{-b})$ for some $b > 1$

$$\mathcal{N}(\lambda) = O(\lambda^{-\frac{1}{b}}). \tag{36}$$

Moreover, if $f_{\mathcal{H}} \in \text{Im } T^a$ *with* $0 \leq a \leq 1/2$*, then*

$$\mathcal{A}(\lambda) \leq \lambda^{2a+1} \left\| T^{-a} f_{\mathcal{H}} \right\|_{\mathcal{H}}^2 \tag{37}$$

and

$$\mathcal{B}(\lambda) \leq \lambda^{2a} \left\| T^{-a} f_{\mathcal{H}} \right\|_{\mathcal{H}}^2. \tag{38}$$

PROOF. Firstly we study $\mathcal{N}(\lambda)$. Since

$$\mathcal{N}(\lambda) = \sum_{n=1}^{N} \frac{t_n}{t_n + \lambda} \tag{39}$$

clearly, $\lim_{\lambda \to 0} \mathcal{N}(\lambda) = N$. Assume now that $N = +\infty$ and $t_n = O(n^{-b})$ with $b > 1$, in fact since, by eq 39, $\mathcal{N}(\lambda)$ is an increasing function of t_n for every n, without loss of generality we can consider the case $t_n = n^{-b}$. The sequence $(t_n)_{n \in \mathbb{N}}$ is strictly positive and decreasing then, by integral test, $\mathcal{N}(\lambda)$ has the same behavior of

$$\begin{aligned}
\mathcal{M}(\lambda) &= \int_1^{\infty} \frac{1}{1 + t^b \lambda} \, dt \\
(\tau^b = t^b \lambda) &= \lambda^{-\frac{1}{b}} \int_{\lambda^{\frac{1}{b}}}^{+\infty} \frac{1}{1 + \tau^b} \, d\tau \\
&\leq \lambda^{-\frac{1}{b}} \int_0^{+\infty} \frac{1}{1 + \tau^b}, \, d\tau
\end{aligned}$$

so that $\mathcal{N}(\lambda) = O(\lambda^{-\frac{1}{b}})$.

The results about $\mathcal{A}(\lambda)$ and $\mathcal{B}(\lambda)$ are standard [10,8]. We give the proof for completeness. Equations (13) gives

$$\begin{aligned}
\mathcal{B}(\lambda) &= \left\| (T + \lambda)^{-1} T f_{\mathcal{H}} - f_{\mathcal{H}} \right\|_{\mathcal{H}}^2 \\
&= \left\| \lambda (T + \lambda)^{-1} f_{\mathcal{H}} \right\|_{\mathcal{H}}^2 \\
&= \sum_{n=1}^{N} (\frac{\lambda}{t_n + \lambda})^2 |\langle f_{\mathcal{H}}, e_n \rangle_{\mathcal{H}}|^2.
\end{aligned} \tag{40}$$

Assume now that $f_{\mathcal{H}} \in \text{Im } T^a$ with $0 \leq a \leq 1/2$. Since the function x^a is concave on $]0, +\infty[$

$$(\frac{t_n}{\lambda})^a \leq 1 + a \frac{t_n}{\lambda} \leq 1 + \frac{t_n}{\lambda},$$

so that

$$\frac{\lambda}{t_n + \lambda} = \frac{1}{1 + \frac{t_n}{\lambda}} \leq \frac{\lambda^a}{t_n^a},$$

then, replacing the above inequality in (40)

$$\mathcal{B}(\lambda) \leq \lambda^{2a} \left\| T^{-a} f_{\mathcal{H}} \right\|_{\mathcal{H}}^2.$$

Reasoning as above we obtain

$$\begin{aligned}\mathcal{A}(\lambda) &= \left\| \sqrt{T}\left((T+\lambda)^{-1}Tf_{\mathcal{H}} - f_{\mathcal{H}}\right)\right\|_{\mathcal{H}}^{2} \\ &= \sum_{n=1}^{N} t_n \left(\frac{\lambda}{t_n+\lambda}\right)^2 |\langle f_{\mathcal{H}}, e_n\rangle_{\mathcal{H}}|^2 \\ &\leq \sum_{n=1}^{N} t_n \left(\frac{\lambda^{a+\frac{1}{2}}}{t_n^{a+\frac{1}{2}}}\right)^2 |\langle f_{\mathcal{H}}, e_n\rangle_{\mathcal{H}}|^2 \\ &= \lambda^{2a+1} \sum_{n=1}^{N} \frac{1}{t_n^{2a}} |\langle f_{\mathcal{H}}, e_n\rangle_{\mathcal{H}}|^2 = \lambda^{2a+1} \left\|T^{-a} f_{\mathcal{H}}\right\|_{\mathcal{H}}^{2}\end{aligned}$$

where the inequality follows since the function $x^{a+\frac{1}{2}}$ is concave.

The following theorem analyzes the a priori choice for the regularization parameter and the corresponding rate of convergence to the regression function in a suitable prior class. We use the stochastic order symbol O_P defined, [22],by the equivalence

$$X_n = O_P(k_n) \Leftrightarrow \lim_{T\to\infty} \limsup_{n\to\infty} \mathbf{P}\left[|X_n| > Tk_n\right] = 0,$$

where $(X_n)_{n\in\mathbb{N}}$ is a sequence of random variables and $(k_n)_{n\in\mathbb{N}}$ a sequence of positive numbers.

Theorem 7 *Assume $f_{\mathcal{H}} \in \operatorname{Im} T^a$ with $0 < a \leq 1/2$ and for $\ell \in \mathbb{N}$ let λ_ℓ the unique solution of the equation*

$$\ell\lambda_\ell^c = \mathcal{N}(\lambda_\ell), \tag{41}$$

where $c = 2a+1$. Then, with the assumptions of Theorem 5

$$if \begin{cases} N < +\infty \\ N = +\infty, \quad t_n = O(n^{-b}) \end{cases} \quad then \quad I[f_{\mathbf{z}}{}^{\lambda_\ell}] - I[f_{\mathcal{H}}] = \begin{cases} O_P\left(\ell^{-1}\right) \\ O_P\left(\ell^{-\frac{bc}{cb+1}}\right) \end{cases}$$

where $b > 1$.

PROOF.

First of all, from representation (39), we note that $\mathcal{N}(\lambda)$ is a positive, non-increasing continuous function of λ. Then it is clear that equation (41) has a unique non-negative solution λ_ℓ, and $\lim_{\ell\to\infty} \lambda_\ell = 0$.

Now let us fix $\eta \in (0,1)$, and note that since by assumption $c > 1$, there exists $\ell(\eta) \in \mathbb{N}$ such that $\ell > \ell(\eta)$ implies

$$\ell\lambda_\ell = \lambda_\ell^{1-c}\mathcal{N}(\lambda_\ell) \geq \frac{C_\eta \kappa}{2} \max(\mathcal{N}(\lambda_\ell), \sqrt{2/C_\eta}) \tag{42}$$

where we followed the notation in Theorem 5. Since the inequality above is equivalent to the constraint (22), Theorem 5 can be applied, giving

$$\mathbf{P}\left[|X_\ell| > C_\eta \left(\mathcal{A}(\lambda_\ell) + \frac{\kappa^2 \mathcal{B}(\lambda_\ell)}{\ell^2 \lambda_\ell} + \frac{\kappa \mathcal{A}(\lambda_\ell)}{\ell \lambda_\ell} + \frac{\kappa M}{\ell^2 \lambda_\ell} + \frac{M\mathcal{N}(\lambda_\ell)}{\ell}\right)\right] \leq \eta, \ \forall \ell > \ell(\eta)$$

where we introduced the sequence of random variables $X_\ell = I[{f_{\mathbf{z}}}^{\lambda_\ell}] - I[f_{\mathcal{H}}]$.

We can simplify the form of the upper bound above recalling that by Lemma 6, $\mathcal{B}(\lambda_\ell) = O(1)$, which implies $\mathcal{B}(\lambda_\ell)\ell^{-2}\lambda_\ell^{-1} = O(\ell^{-2}\lambda_\ell^{-1})$ such that the fourth term in the sum above is not asymptotically smaller than the second one. Reasoning in a similar way, from inequality (42) we see that $\ell^{-1}\lambda_\ell^{-1} = O(1)$ such that $\mathcal{A}(\lambda_\ell)\ell^{-1}\lambda_\ell^{-1} = O(\mathcal{A}(\lambda_\ell))$ and $\ell^{-2}\lambda_\ell^{-1} = O(\mathcal{N}(\lambda_\ell)\ell^{-1})$, which means that the first and fifth terms are not asymptotically smaller than the third and fourth ones respectively. These arguments on asymptotic orders of the terms in the probabilistic bound above lead to the conclusion that a positive constant C' and natural number $\ell'(\eta)$ exist such that

$$\mathbf{P}\left[|X_\ell| > C_\eta C' \left(\mathcal{A}(\lambda_\ell) + \frac{\mathcal{N}(\lambda_\ell)}{\ell}\right)\right] \leq \eta, \ \forall \ell > \ell'(\eta). \tag{43}$$

The bound above can be restated in terms of stochastic order symbols in fact, since η was arbitrarily chosen in the interval $(0, 1)$, from the definition of the constants C_η it follows that

$$\mathbf{P}\left[|X_\ell| > T \left(\mathcal{A}(\lambda_\ell) + \frac{\mathcal{N}(\lambda_\ell)}{\ell}\right)\right] \leq 8e^{-\sqrt{T/128C'}}, \ \forall T > 128C' \log^2 8 \ \ \forall \ell > \ell''(T), \tag{44}$$

and from the definition of stochastic order symbol this implies

$$I[{f_{\mathbf{z}}}^{\lambda_\ell}] - I[f_{\mathcal{H}}] = O_P\left(\mathcal{A}(\lambda_\ell) + \frac{\mathcal{N}(\lambda_\ell)}{\ell}\right). \tag{45}$$

In order to complete the proof we now have to estimate the asymptotic order of the sequence of positive numbers appearing on the r.h.s. of the equality above. This can be accomplished bounding by $O(\lambda_\ell^c)$ both the approximation term $\mathcal{A}(\lambda_\ell)$ and the sampling term $\mathcal{N}(\lambda_\ell)/\ell$. The first estimate is obtained by inequality (37) in Lemma 6, the second one by the very definition of λ_ℓ, equation (41). Then we can furthermore simplify the previous expression for the stochastic order of $I[{f_{\mathbf{z}}}^{\lambda_\ell}] - I[f_{\mathcal{H}}]$, in fact we have

$$I[{f_{\mathbf{z}}}^{\lambda_\ell}] - I[f_{\mathcal{H}}] = O_P\left(\lambda_\ell^c\right). \tag{46}$$

Finally let us consider separately the two cases in text of the theorem. If $N < +\infty$, by equation (35) in Lemma 6 and equation (41), it is clear that $\lambda_\ell^c = O(\ell^{-1})$, which proves the first statement of the theorem. If instead $N = +\infty$ and $t_n = O(n^{-b})$ for some $b > 1$ this time from equality (36) in Lemma 6 and equation (41) it follows

$$\lambda_\ell^c \ell = O(\lambda_\ell^{-\frac{1}{b}}),$$

which implies $\lambda_\ell^c = O(\ell^{-\frac{bc}{bc+1}})$, proving the second part of the theorem.

6 Asymptotic Optimality

In this section we prove that the asymptotic power rates obtained are indeed optimal. The cited power rates were derived assuming that the eigenvalues of the operator T fulfill the power upper bound $t_n = O(n^{-b})$.

On the other hand the main result of this section relies on a power lower bound for the same eigenvalues, $t_n = \Omega(n^{-b})$. It is shown that under this assumption a lower bound of the same form as the previous result can be established for the asymptotic rate of convergence for the risk.

These two results allow to conclude that in the case $t_n = \Theta(n^{-b})$, that is if the eigenvalues of T have a power asymptotic growth, then the asymptotic rate of convergence for the risk of RLS estimators obtained by the described choice for the regularization parameter, is optimal conditionally to the marginal distribution ρ_X and the prior $f_\rho \in \mathcal{F}$.

We now state and prove the just introduced theorem. Let us first name $\mathcal{M}(\mathcal{F}, \rho_X)$ the set of probability distribution ρ on $X \times Y$ which realize the marginal probability distribution ρ_X and compatible with the prior condition $f_\rho \in \mathcal{F}$. The theorem is based on a recent result, [6], expressed in terms of the *tight packing numbers* of the prior class $\mathcal{F} \subseteq L_2(X, \rho_X)$

$$\bar{\mathcal{N}}(\mathcal{F}, \rho_X, \delta, c_0, c_1) := \sup\{k |\ \exists\ (g_i)_{i=1}^k \in \mathcal{F}^k,\ \text{ s.t. } \forall i \neq j\ \ c_0\delta \leq \|g_i - g_j\|_\nu \leq c_1\delta\}$$

where $0 < c_0 \leq c_1 < \infty$ are two fixed real numbers. The main lower bound in ([6] Theorem 3.1) can be restated in the following form

Theorem 8 *Define $\bar{\mathcal{N}}(\epsilon) := \bar{\mathcal{N}}(\mathcal{F}, \rho_X, 2\sqrt{\epsilon}/c_0, c_0, c_1)$. Suppose that for $\epsilon > 0$ the net of functions $(g_i)_{i=1}^{\bar{\mathcal{N}}(\epsilon)}$ (occurring in the definition of tight packing numbers) satisfies $\|g_i\|_{\mathcal{C}(X)} \leq 1/4$ for $i = 1, \cdots, \bar{\mathcal{N}}(\epsilon)$. Let $\bar{\eta} := \mathrm{e}^{-3/\mathrm{e}}$, then for all $\ell \in \mathbb{N}$*

$$\inf_{f_\cdot} \sup_{\rho \in \mathcal{M}(\mathcal{F}, \rho_X)} \mathbf{P}_{\mathbf{z} \sim \rho^\ell} [I[f_{\mathbf{z}}] \geq I[f_\rho] + \epsilon] \geq \min(\frac{1}{2}, \bar{\eta}\sqrt{\bar{\mathcal{N}}(\epsilon)} \exp(-8\ell\epsilon c_1^2/c_0^2))$$

where the infimum is over the set of all the learning maps $\mathbf{z} \to f_{\mathbf{z}}$.

We now specialize this general result to our framework, that is we consider the two parameters family of prior classes

$$\mathcal{F}(a, R) := \{f \in \mathcal{H} |\ T^{-a} f \in \mathcal{H} \ \text{ with } \left\|T^{-a} f\right\|_{\mathcal{H}} \leq R\} \tag{47}$$

for $0 \leq a \leq 1/2$ and $R > 0$.

The lower bound on the asymptotic rate of risk for general learning maps under these priors and fixed marginal distribution ρ_X, can be stated as follows

Theorem 9 *Let us assume that* $\dim \mathcal{H} = +\infty$. *Moreover let the eigenvalues* $(t_n)_{n=1}^{+\infty}$ *of the operator* T *fulfill the condition* $t_n = \Omega(n^{-b})$ *for some* $b > 0$. *Then for every* $0 \leq a \leq 1/2$ *and* $R > 0$ *there exist* $\bar{\ell}$ *and* $C > 0$ *such that for every* $\ell > \bar{\ell}$

$$\inf_{f_\cdot} \sup_{\rho \in \mathcal{M}(\mathcal{F}, \rho_X)} \mathbf{P}_{\mathbf{z} \sim \rho^\ell} \left[I[f_{\mathbf{z}}] \geq I[f_\rho] + C\ell^{-\frac{cb}{cb+1}} \right] \geq \bar{\eta},$$

where $\mathcal{F} := A\mathcal{F}(a, R)$, $c := 2a + 1$ *and* $\bar{\eta} := \mathrm{e}^{-3/\mathrm{e}}$.

The proof of Theorem (9) relies on the following Lemma regarding packing numbers over sets of binary strings endowed with the Hamming distance

$$\mathrm{d_H}(\sigma, \sigma') := |\{1 \leq i \leq K \ \text{ s.t. } \sigma_i \neq \sigma'_i\}|,$$

where we adopted the notation $\sigma := (\sigma_i)_{i=1}^K \in \{-1, +1\}^K$. The proof relies on a standard concentration of measure result, Hoeffding's inequality [12] .

Lemma 10 *For every* $K > 17$ *there exists a subset* L *of the cube* $\{-1, +1\}^K$ *such that*

$$\mathrm{d_H}(\sigma, \sigma') > \frac{K}{2} \qquad \forall\ \sigma, \sigma' \in L, \quad \sigma \neq \sigma'$$

and $|L| \geq B^K$ *where* $B := \exp(1/24)$.

PROOF. If $\sigma := (\sigma_i)_{i=1}^K$ is a random point on the cube, then the components σ_i are independent random variables distributed according to the measure $1/2(\delta_{-1} + \delta_{+1})$. Let σ and σ' be independent random points on the cube, then note that

$$\mathrm{d_H}(\sigma, \sigma') = \sum_{i=1}^K |\sigma_i - \sigma'_i| = \sum_{i=1}^K \theta_i$$

where θ_i are independent random variables distributed according to the measure $1/2(\delta_0 + \delta_2)$. It follows that Hoeffding's inequality can be applied to the random variable $\mathrm{d_H}(\sigma, \sigma')$, yielding for every $\delta > 0$

$$\mathbf{P}\left[|\mathrm{d_H}(\sigma, \sigma') - K| \geq \delta \right] \leq 2\exp(-\frac{\delta^2}{2K}).$$

Setting $\delta = K/2$ in the inequality above, we obtain

$$\mathbf{P}\left[\mathrm{d_H}(\sigma, \sigma') \leq \frac{K}{2} \right] \leq 2\exp(-\frac{K}{8}). \tag{48}$$

Now draw $m := \ulcorner B^K \urcorner$ (where $\ulcorner x \urcorner$ is the lowest integer greater than x) independent random points $\sigma^{(j)}$ $(j = 1, \cdots, m)$ on the cube. From inequality (48), by union bound it holds

$$\mathbf{P}\left[\exists\ 1 \le j,k \le m,\ j \ne k, \quad \text{with } \mathrm{d_H}(\sigma^{(j)},\sigma^{(k)}) \le \frac{K}{2}\right]$$
$$\le (m^2-m)\exp(-\frac{K}{8}) \le (B^{2K}+B^K+1)B^{-3K} < 1,$$

where we used the inequality $x \le \ulcorner x \urcorner \le x+1$ and the assumption $K > 17$. The last inequality implies that at least one subset of the cube with the required properties exists.

We are now ready to prove the main result of this section.

PROOF. [Theorem (9)] The proof consists in showing a large subset $A\mathcal{L}_\epsilon$ of the the prior class $\mathcal{F} := A\mathcal{F}(a,R)$ defined in (47) whose tight packing numbers can be easily estimated. Then Theorem (8) will be applied directly to $A\mathcal{L}_\epsilon$ yielding the claimed lower bound.

Firstly define $R' := \min\{R, \kappa^{-c}/4\}$, obviously

$$\mathcal{F}(a,R') \subseteq \mathcal{F}(a,R).$$

Moreover for every f in $\mathcal{F}(a,R')$ it holds

$$\|f\|_{\mathcal{C}(X)} \le \kappa\|f\|_{\mathcal{H}} \le \kappa\,\|T^a\|\,\left\|T^{-a}f\right\|_{\mathcal{H}} \le \kappa^c R' \le \frac{1}{4} \tag{49}$$

where we used the fact that $\|T\| \le \kappa^2$. This bound will be useful when applying Theorem (8) to the subset $A\mathcal{L}_\epsilon$ of $A\mathcal{F}(a,R')$.

Let us recall that

$$T = \sum_{n=1}^{+\infty} t_n \langle \cdot, e_n\rangle\, e_n$$

is the spectral decomposition of T with $0 < t_{n+1} \le t_n$ and $(e_n)_{n=1}^{+\infty}$ is a basis of $\mathcal{H}$. Then by definition (47) it is clear that

$$\mathcal{F}(a,R') = \{\sum_{n=1}^{+\infty} c_n e_n |\ \sum_{n=1}^{+\infty} c_n^2 t_n^{-2a} \le R'^2\}. \tag{50}$$

Since by assumption $t_n = \Omega(n^{-b})$, it holds

$$\exists\ N', C' > 0 \qquad \text{s.t. } \forall n \ge N' \quad t_n \ge C' n^{-b}. \tag{51}$$

Now let us set the constants $f := R'^{\frac{2}{bc}} C'^{-\frac{1}{b}}$, $\bar\epsilon := (f/(2N'+34))^{bc}$ and define the family of sets

$$\mathcal{D}_\epsilon := \{n \in \mathbb{N} |\ N' \leq n \leq f\epsilon^{-\frac{1}{bc}}\}, \qquad 0 < \epsilon < \bar{\epsilon}. \tag{52}$$

From this definition and property (51) it follows that

$$t_n^c \geq \epsilon R'^{-2}, \qquad \forall\ n \in \mathcal{D}_\epsilon. \tag{53}$$

Furthermore, due to the constraint for the allowed values of ϵ in (52), the cardinalities of the sets $\mathcal{D}_\epsilon$ fulfill the inequality

$$|\mathcal{D}_\epsilon| > \max\{\frac{f}{2}\epsilon^{-\frac{1}{bc}}, 17\}. \tag{54}$$

From the family of inequalities (53) and equation (50) it follows

$$\mathcal{F}(a, R') \supseteq \mathcal{C}_\epsilon := \{\sum_{n\in\mathcal{D}_\epsilon} \sqrt{\frac{\epsilon}{|\mathcal{D}_\epsilon| t_n}}\sigma_n e_n |\ (\sigma_n)_{n\in\mathcal{D}_\epsilon} \in \{-1, +1\}^{|\mathcal{D}_\epsilon|}\}.$$

Since by inequality (54), $|\mathcal{D}_\epsilon| > 17$, we can apply Lemma (10) to the binary cube $\{-1, +1\}^{|\mathcal{D}_\epsilon|}$. This ensures us that a subset L_ϵ of the cube exists with cardinality $|L_\epsilon| > B^{|\mathcal{D}_\epsilon|}$ for some constant B, and such that for all distinct σ and σ' in L_ϵ it holds

$$\mathrm{d_H}(\sigma, \sigma') > \frac{|\mathcal{D}_\epsilon|}{2}. \tag{55}$$

We are finally ready to define the set of functions $A\mathcal{L}_\epsilon$ to which Theorem (8) will be applied. In fact we have the following straightforward chain of inclusions

$$\mathcal{F}(a, R') \supseteq \mathcal{C}_\epsilon \supseteq \mathcal{L}_\epsilon := \{\sum_{n\in\mathcal{D}_\epsilon} \sqrt{\frac{\epsilon}{|\mathcal{D}_\epsilon| t_n}}\sigma_n e_n |\ (\sigma_n)_{n\in\mathcal{D}_\epsilon} \in L_\epsilon\}.$$

Let us first notice that since $\|Af\|_\nu = \left\|T^{1/2} f\right\|_{\mathcal{H}}$, by inequalities (55) it follows that for all distinct g and g' in $A\mathcal{L}_\epsilon$ it holds

$$\frac{\epsilon}{2} \leq \left\|g - g'\right\|_\nu^{\ 2} \leq 2\epsilon.$$

Then letting $c_1^2 = 2$ and $c_0^2 = 1/2$ in the definition of tight packing numbers for the class $A\mathcal{L}_\epsilon$, it is clear that $A\mathcal{L}_\epsilon$ itself is a maximal $\sqrt{\epsilon}$-net of functions. Moreover from (49) $\|g\|_{\mathcal{C}(X)} \leq 1/4$ for every $g \in A\mathcal{L}_\epsilon$, then we can apply Theorem (8) obtaining for all $\ell \in \mathbb{N}$

$$\begin{aligned}\inf_{f.} \sup_{\rho\in\mathcal{M}(\mathcal{F},\rho_X)} \mathbf{P}_{\mathbf{z}\sim\rho^\ell}\left[I[f_{\mathbf{z}}] \geq I[f_\rho] + \frac{\epsilon}{8}\right] &\geq \inf_{f.} \sup_{\rho\in\mathcal{M}(\mathcal{L},\rho_X)} \mathbf{P}_{\mathbf{z}\sim\rho^\ell}\left[I[f_{\mathbf{z}}] \geq I[f_\rho] + \frac{\epsilon}{8}\right] \\ &\geq \min(\frac{1}{2}, \bar{\eta}\sqrt{\bar{\mathcal{N}}(\frac{\epsilon}{8})}\exp(-32\ell\epsilon)),\end{aligned}$$

where $\mathcal{F} := A\mathcal{F}(a, R)$, $\mathcal{L} := A\mathcal{L}_\epsilon$ and by inequality (54) it holds

$$\bar{\mathcal{N}}(\frac{\epsilon}{8}) := \bar{\mathcal{N}}(A\mathcal{L}_\epsilon, \rho_X, \sqrt{\epsilon}, \sqrt{1/2}, \sqrt{2}) = |L_\epsilon| \geq B^{|\mathcal{D}_\epsilon|} > B^{\frac{f}{2}\epsilon^{-\frac{1}{bc}}}.$$

The result claimed by the Theorem follows substituting in the inequalities above ϵ with the expression

$$\epsilon(\ell) = 8C\,\mathrm{e}^{-\frac{bc}{bc+1}},$$

where

$$C := \frac{1}{8}\left(\frac{f}{128}\log B\right)^{\frac{bc}{bc+1}},$$

and

$$\ell \geq \bar{\ell} := \left(\frac{\bar{\epsilon}}{8C}\right)^{-\frac{bc+1}{bc}},$$

in order to enforce the constraint $\epsilon < \bar{\epsilon}$ and the condition

$$\sqrt{\bar{\mathcal{N}}\left(\frac{\epsilon(\ell)}{8}\right)}\exp(-32\ell\epsilon(\ell)) \geq 1.$$

Acknowledgments

We would like to thank T. Poggio, A. Rakhlin, L. Rosasco and S. Smale for useful discussions and suggestions.

References

[1] N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68:337–404, 1950.

[2] F. Cucker and S. Smale. On the mathematical foundations of learning. *Bull. Amer. Math. Soc. (N.S.)*, 39(1):1–49 (electronic), 2002.

[3] Felipe Cucker and Steve Smale. Best choices for regularization parameters in learning theory: on the bias-variance problem. *Found. Comput. Math.*, 2(4):413–428, 2002.

[4] Felipe Cucker and Steve Smale. On the mathematical foundations of learning. *Bull. Amer. Math. Soc. (N.S.)*, 39(1):1–49 (electronic), 2002.

[5] E. De Vito, A. Caponnetto, and L. Rosasco. Model selection for regularized least-squares algorithm in learning theory. *Foundations of Computational Mathematics*, 5(1):59–85, February 2005.

[6] R. DeVore, G. Kerkyacharian, D. Picard, and V. Temlyakov. Mathematical methods for supervised learning. Technical report, Industrial Mathematics Institute, University of South Carolina, 2004.

[7] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Number 31 in Applications of mathematics. Springer, New York, 1996.

[8] H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of inverse problems*, volume 375 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1996.

[9] T. Evgeniou, M. Pontil, and T. Poggio. Regularization networks and support vector machines. *Adv. Comp. Math.*, 13:1–50, 2000.

[10] C. W. Groetsch. *The theory of Tikhonov regularization for Fredholm equations of the first kind*, volume 105 of *Research Notes in Mathematics*. Pitman (Advanced Publishing Program), Boston, MA, 1984.

[11] L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A Distribution-free Theory of Non-parametric Regression*. Springer Series in Statistics, 2002.

[12] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

[13] S. Mendelson. On the performance of kernel classes. *Journal of Machine Learning Research*, 4:759–771, 2003.

[14] I. F. Pinelis and A. I. Sakhanenko. Remarks on inequalities for probabilities of large deviations. *Theory Probab. Appl.*, 30(1):143–148, 1985.

[15] T. Poggio and F. Girosi. A theory of networks for approximation and learning. In C. Lau, editor, *Foundation of Neural Networks*, pages 91–106. IEEE Press, Piscataway, N.J., 1992.

[16] Tomaso Poggio and Steve Smale. The mathematics of learning: dealing with data. *Notices Amer. Math. Soc.*, 50(5):537–544, 2003.

[17] B. Schölkopf and A.J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

[18] L. Schwartz. Sous-espaces hilbertiens d'espaces vectoriels topologiques et noyaux associés (noyaux reproduisants). *J. Analyse Math.*, 13:115–256, 1964.

[19] S. Smale and D. Zhou. Shannon sampling and function reconstruction from point values. *Bull. Amer. Math. Soc. (N.S.)*, 41(3):279–305 (electronic), 2004.

[20] S. Smale and D. Zhou. Shannon sampling II : Connections to learning theory. *submitted to Appl. Comput. Harmonic Anal.*, 2004.

[21] I. Steinwart. Sparseness of support vector machines. *Journal of Machine Learning Research*, 4:1071–1105, 2003.

[22] S. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, 2000.

[23] V. N. Vapnik. *Statistical learning theory*. Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons Inc., New York, 1998. A Wiley-Interscience Publication.

[24] G. Wahba. *Spline models for observational data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1990.

[25] V. Yurinsky. *Sums and Gaussian vectors*, volume 1617 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1995.

[26] T. Zhang. Effective dimension and generalization of kernel learning. *NIPS 2002*, pages 454–461.

[27] T. Zhang. Leave-one-out bounds for kenrel methods. *Neural Computation*, 13:1397–1437, 2003.