Available online at www.sciencedirect.com

ScienceDirect

Journal of Complexity 23 (2007) 52–72

Journal of COMPLEXITY

www.elsevier.com/locate/jco

On regularization algorithms in learning theory

Frank Bauer[a], Sergei Pereverzev[b], Lorenzo Rosasco[c, d, *]

[a] *Institute for Mathematical Stochastics, Department of Mathematics, University of Göttingen, Maschmühlenweg 8-10, 37073 Göttingen, Germany*

[b] *Institute for Numerical and Applied Mathematics, Department of Mathematics, University of Göttingen, Lotzestr. 16-18, 37083 Göttingen, Germany*

[c] *Johann Radon Institute for Computational and Applied Mathematics (RICAM), Austrian Academy of Sciences, Altenbergerstrasse 69, A-4040 Linz, Austria*

[d] *DISI, Università di Genova, v. Dodecaneso 35, 16146 Genova, Italy*

Received 19 January 2006; accepted 22 July 2006
Available online 19 October 2006

Abstract

In this paper we discuss a relation between Learning Theory and Regularization of linear ill-posed inverse problems. It is well known that Tikhonov regularization can be profitably used in the context of supervised learning, where it usually goes under the name of regularized least-squares algorithm. Moreover, the gradient descent algorithm was studied recently, which is an analog of Landweber regularization scheme. In this paper we show that a notion of regularization defined according to what is usually done for ill-posed inverse problems allows to derive learning algorithms which are consistent and provide a fast convergence rate. It turns out that for priors expressed in term of variable Hilbert scales in reproducing kernel Hilbert spaces our results for Tikhonov regularization match those in Smale and Zhou [Learning theory estimates via integral operators and their approximations, submitted for publication, retrievable at ⟨http://www.tti-c.org/smale.html⟩, 2005] and improve the results for Landweber iterations obtained in Yao et al. [On early stopping in gradient descent learning, Constructive Approximation (2005), submitted for publication]. The remarkable fact is that our analysis shows that the same properties are shared by a large class of learning algorithms which are essentially all the linear regularization schemes. The concept of operator monotone functions turns out to be an important tool for the analysis.

Keywords: Learning theory; Regularization theory; Non-parametric statistics

[*] Corresponding author. DISI, Università di Genova, v. Dodecaneso 35, 16146 Genova, Italy.
E-mail address: rosasco@disi.unige.it (L. Rosasco).

0885-064X/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.jco.2006.07.001

1. Introduction

In this paper we investigate the theoretical properties of a class of regularization schemes to solve the following regression problem which is relevant to Learning Theory [32,11]. Given a training set $z_i = (x_i, y_i), i = 1, \ldots, n$, drawn i.i.d. according to an unknown probability measure ρ on $X \times Y$, we wish to approximate the regression function

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

We consider approximation schemes in reproducing kernel Hilbert Spaces $\mathcal{H}$ and the quality of the approximation is measured either in the norm in $\mathcal{H}$ or in the norm $\|f\|_\rho = (\int f^2 \, d\rho)^{1/2}$. In the context of Learning Theory the latter is particularly meaningful since weight is put on the points which are most likely to be sampled. Moreover, we are interested in a worst case analysis that is, since an estimator $f_{\mathbf{z}}$ based on $\mathbf{z} = (z_1, \ldots, z_n)$ is a random variable, we look for exponential tail inequalities,

$$P\left[\left\| f_{\mathbf{z}} - f_\rho \right\|_\rho > \varepsilon(n)\tau\right] \leqslant e^{-\tau},$$

where $\varepsilon(n)$ is a positive, decreasing function of the number of samples and $\tau > 0$. To obtain this kind of results, we have to assume some prior on the problem, that is $f_\rho \in \Omega$ for some suitable set Ω (see the discussion in [14]). This is usually done relating the problem to the considered approximation scheme. Following Rosasco et al. [24] we consider a large class of approximation schemes in reproducing kernel Hilbert spaces (RKHS). In this context the prior is usually expressed in terms of some standard Hilbert scale [11].

In this paper we generalize to priors defined in term of *variable* Hilbert scales and refine the analysis in Rosasco et al. [24]. In particular, we can analyze a larger class of algorithms and especially obtain improved probabilistic error estimates. In fact the regularized least-squares algorithm (Tikhonov Regularization), see [27,28,7–9] and reference therein for latest result, and the gradient descent algorithm (Landweber Iteration) in Yao et al. [33] can be treated as special cases of our general analysis. In particular we show that, in the range of prior considered here, our result for Tikhonov regularization match those in Smale and Zhou [27] and improve the results for Landweber iteration obtained in Yao et al. [33] which now share the same rates as Tikhonov regularization. The remarkable fact is that our analysis shows that the same properties are shared by a large class of algorithms which are essentially all the linear regularization algorithms which can be profitably used to solve ill-posed inverse problems [16].

At the same time, this paper is not just a reformulation of the results from the theory of ill-posed problems in the context of Learning Theory. Indeed, standard ill-posed problems theory, as it is presented, for example in Engl et al. [16], is dealing with the situation, when an ill-posed linear operator equation and its perturbed version are considered in some common Hilbert space. The problem of Learning from examples cannot be put in this framework directly, in spite of the fact that under some conditions the regression function can be really considered as a solution of linear ill-posed operator equation (embedding equation). The point is that the sampling operator involved in the discretized or "perturbed" version of this equation acts in Euclidean space, while the operator of the embedding equation is feasible only in an infinite-dimensional functional space. Indeed this is different from the setting in Bissantz et al. [4] where the operator is always assumed to be the same.

The first attempt to resolve this discrepancy has been made in De Vito et al. [13], Yao et al. [33], Rosasco et al. [24], where the estimates of the Lipschitz constants of functions generating regularization methods have been used for obtaining error bounds. But these functions should converge point-wise to the singular function $\sigma \to 1/\sigma$ (see conditions (15), (16)). Therefore, their Lipschitz properties are rather poor. As a result, general error bounds from De Vito et al. [13], Yao et al. [33], Rosasco et al. [24] do not coincide with the estimates [8,27] obtained on the base of meticulous analysis of Tikhonov regularization (particular case of general scheme considered in [24]). In this paper to achieve tight regularization error bound the concept of operator monotone index functions is introduced in the analysis of learning from examples. At first glance it can be viewed as a restriction on the prior, but as we argue in Remark 2, the concept of operator monotonicity covers all types of priors considered so far in Regularization Theory.

In our opinion the approach to the estimation of the regularization error presented in this paper (see Theorem 10) can be also used for obtaining new results in Regularization Theory. In particular, it could be applied to regularized collocation methods. We hope that this idea will be realized in a near future.

Finally, we note that though we mainly discuss a regression setting we can also consider the implication in the context of classification. This is pursued in this paper considering recently proposed assumption [29] on the classification noise. Indeed we can prove classification risk bounds as well as fast rates to Bayes risk.

The plan of the paper is as follows. In Section 2 we present the setting and state the main assumptions. Some background on RKHS is given and the prior on the problem is discussed. In Section 3 we first present the class of algorithms we are going to analyze and then state and prove the main results of the paper.

2. Learning in RKHS

The content of this section is divided as follows. First, we introduce the problem of learning from examples as the problem of approximating a multivariate function from random samples, fix the setting and the notation. Second, we give an account of RKHS since our approximation schemes will be built in such spaces. Third, we discuss the kind of prior assumption we consider on the problem.

2.1. Learning from examples: notation and assumptions

We start giving a brief account of Learning Theory (see [32,11,17,5] and reference therein). We let $Z = X \times Y$ be the sample space, where the input space $X \subset \mathbb{R}^d$ is closed and the output space is $Y \subset \mathbb{R}$. The space Z is endowed with a fixed but unknown probability measure ρ which can be factorized as $\rho(x, y) = \rho_X(x)\rho(y|x)$ where ρ_X is the marginal probability on X and $\rho(y|x)$ is the conditional probability of y given x. A common assumption is $Y = [-B, B]$ for some $B > 0$, here we can assume the weaker conditions considered in De Vito [8], that is for almost all $x \in X$ we assume

$$\int_Y \left(e^{\frac{|y - f_{\mathcal{H}}^{\dagger}(x)|}{M}} - \frac{|y - f_{\mathcal{H}}^{\dagger}(x)|}{M} - 1 \right) d\rho(y|x) \leqslant \frac{\Sigma^2}{2M^2}, \tag{1}$$

where $f_{\mathcal{H}}^{\dagger}$ is an approximation of the regression function (see (5)) and $\Sigma, M \in \mathbb{R}^{+}$. Moreover, we assume

$$\int_Y y^2 \, d\rho(x, y) \leqslant \infty. \tag{2}$$

In this setting, what is given is a training set $\mathbf{z} = (\mathbf{x}, \mathbf{y}) = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ drawn i.i.d. according to ρ and, fixing a loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^{+}$, the goal is to find an estimator $f = f_{\mathbf{z}}$ with a small expected error

$$\mathcal{E}(f) = \int_{X \times Y} \ell(y, f(x)) \, d\rho(x, y).$$

A natural choice for the loss function is the squared loss function $\ell(y, f(x)) = (y - f(x))^2$. In fact the minimizer of $\mathcal{E}(f)$ becomes the regression function

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

where the minimum is taken over the space $L^2(X, \rho_X)$ of square integrable functions with respect to ρ_X. Moreover, we recall that for $f \in L^2(X, \rho_X)$

$$\mathcal{E}(f) = \left\| f - f_\rho \right\|_\rho^2 + \mathcal{E}(f_\rho)$$

so that we can restate the problem as that of approximating the regression function in the norm $\|\cdot\|_\rho = \|\cdot\|_{L^2(X,\rho_X)}$. As we mention in the Introduction we are interested in exponential tail inequalities such that with probability at least $1 - \eta$

$$\left\| f_{\mathbf{z}} - f_\rho \right\|_\rho \leqslant \varepsilon(n) \log \frac{1}{\eta} \tag{3}$$

for some positive decreasing function $\varepsilon(n)$ and $0 < \eta \leqslant 1$. From these kind of results, we can easily obtain bound in expectation

$$\mathbb{E}_{\mathbf{z}} \left[\left\| f_{\mathbf{z}} - f_\rho \right\|_\rho \right] \leqslant \bar{\varepsilon}(n)$$

by standard integration of tail inequalities, that is $\bar{\varepsilon}(n) = \int_0^\infty \exp\{-\frac{t}{\varepsilon(n)}\} \, dt$. Moreover, if $\varepsilon(n)$ decreases fast enough, the Borel–Cantelli Lemma allows to derive almost sure convergence of $\left\| f_{\mathbf{z}} - f_\rho \right\|_\rho \to 0$ as n goes to ∞, namely strong consistency [32,15].

In this paper we search for the estimator $f_{\mathbf{z}}$ in a hypothesis space $\mathcal{H} \subset L^2(X, \rho_X)$ which is a RKHS [26,1]. Before recalling some basic facts on such spaces we discuss some implication of considering approximation schemes in a fixed hypothesis space and in particular in RKHSs. Once we choose $\mathcal{H}$ the best achievable error is clearly

$$\inf_{f \in \mathcal{H}} \mathcal{E}(f). \tag{4}$$

In general, the above error can be bigger than $\mathcal{E}(f_\rho)$ and the existence of an extremal function is not even ensured. Now let $I_K : \mathcal{H} \to L^2(X, \rho_X)$ be the inclusion operator and $P : L^2(X, \rho_X) \to L^2(X, \rho_X)$ the projection on the closure of the range of I_K in $L^2(X, \rho_X)$, Then, as noted in De Vito et al. [13,12], the theory of inverse problems ensures that $Pf_\rho \in R(I_K)$ is a sufficient condition for existence and uniqueness of a minimal norm solution of problem (4) (see [16, Theorem 2.5]).

In fact, such an extremal function, denoted here with $f_{\mathcal{H}}^{\dagger}$ is nothing but the Moore–Penrose (or generalized) solution [1] of the linear embedding equation $I_K f = f_\rho$ since

$$\inf_{f \in \mathcal{H}} \mathcal{E}(f) - \mathcal{E}(f_\rho) = \inf_{f \in \mathcal{H}} \left\| I_K f - f_\rho \right\|_\rho^2 , \tag{5}$$

see [12,13]. As a consequence, rather than studying (3), what we can aim to, if $P f_\rho \in R(I_K)$, are probabilistic bounds on

$$\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_{\mathcal{H}}^{\dagger}) = \left\| f_{\mathbf{z}} - f_{\mathcal{H}}^{\dagger} \right\|_\rho^2 . \tag{6}$$

As we discuss in the following (see Theorem 10) under some more assumption this ensures also a good approximation for f_ρ. For example, if $f_\rho \in \mathcal{H}$ (that is $f_\rho \in R(I_K)$) clearly $f_\rho = f_{\mathcal{H}}^{\dagger}$ (that is $f_\rho = I_K f_{\mathcal{H}}^{\dagger}$).

2.2. *Reproducing Kernel Hilbert spaces and related operators*

A RKHS $\mathcal{H}$ is a Hilbert space of point-wise defined functions which can be completely characterized by a symmetric positive definite function $K : X \times X \to \mathbb{R}$, namely the kernel. If we let $K_x = K(x, \cdot)$, the space $\mathcal{H}$ induced by the kernel K can be built as the completion of the finite linear combinations $f = \sum_{i=1}^{N} c_i K_{x_i}$ with respect to the inner product $\langle K_s, K_x \rangle_{\mathcal{H}} = K(s, x)$. The following reproducing property easily follows $\langle f, K_x \rangle_{\mathcal{H}} = f(x)$, and moreover by Cauchy–Schwartz inequality $\|f\|_\infty \leqslant \sup_{x \in X} \sqrt{K(x,x)}\, \|f\|_{\mathcal{H}}$. In this paper we make the following assumptions [2] on $\mathcal{H}$:

- the kernel is measurable;
- the kernel is bounded, that is

$$\sup_{x \in X} \sqrt{K(x,x)} \leqslant \kappa < \infty . \tag{7}$$

- the space $\mathcal{H}$ is separable.

We now define some operators which will be useful in the following (see [10] for details). We already introduced the inclusion operator $I_K : \mathcal{H} \to L^2(X, \rho_X)$, which is continuous by (7). Moreover, we consider the adjoint operator $I_K^* : L^2(X, \rho_X) \to \mathcal{H}$, the covariance operator $T : \mathcal{H} \to \mathcal{H}$ such that $T = I_K^* I_K$ and the operator $L_K : L^2(X, \rho_X) \to L^2(X, \rho_X)$ such that $L_K = I_K I_K^*$. It can be easily proved that

$$I_K^* = \int_X K_x \, d\rho_X(x), \qquad T = \int_X \langle \cdot, K_x \rangle_{\mathcal{H}} K_x \, d\rho_X(x).$$

The operators T and L_K can be proved to be positive trace class operators (and hence compact). For a function $f \in \mathcal{H}$ we can relate the norm in $\mathcal{H}$ and $L^2(X, \rho_X)$ using T. In fact if we

[1] In Learning Theory $f_{\mathcal{H}}^{\dagger}$ is often called the best in model or the best in the class Bousquet et al. [5].

[2] We note that it is common to assume K to be a Mercer kernel that is a continuous kernel. This assumption, together with compactness of the input space X ensures compactness of the integral operator with kernel K. Under our assumptions it is still possible to prove compactness of the integral operator even when X is not compact [10].

regard $f \in \mathcal{H}$ as a function in $L^2(X, \rho_X)$ we can write

$$\|f\|_\rho = \left\|\sqrt{T} f\right\|_{\mathcal{H}}. \tag{8}$$

This fact can be easily proved recalling that the inclusion operator is continuous and hence admits a polar decomposition $I_K = U\sqrt{T}$, where U is a partial isometry [25].

Finally, replacing ρ_X by the empirical measure $\rho_{\mathbf{x}} = n^{-1} \sum_{i=1}^n \delta_{x_i}$ on a sample $\mathbf{x} = (x_i)_{i=1}^n$ we can define the sampling operator $S_{\mathbf{x}} : \mathcal{H} \to \mathbb{R}^n$ by $(S_{\mathbf{x}} f)_i = f(x_i) = \langle f, K_{x_i}\rangle_{\mathcal{H}}$; $i = 1, \ldots, n$, where the norm $\|\cdot\|_n$ in $\mathbb{R}^n$ is $1/n$ times the euclidean norm. Moreover, we can define $S_{\mathbf{x}}^* : \mathbb{R}^n \to \mathcal{H}$, the empirical covariance operator $T_{\mathbf{x}} : \mathcal{H} \to \mathcal{H}$ such that $T_{\mathbf{x}} = S_{\mathbf{x}}^* S_{\mathbf{x}}$ and the operator $S_{\mathbf{x}} S_{\mathbf{x}}^* : \mathbb{R}^n \to \mathbb{R}^n$. It follows that for $\xi = (\xi_1, \ldots, \xi_n)$

$$S_{\mathbf{x}}^* \xi = \frac{1}{n} \sum_{i=1}^n K_{x_i} \xi_i, \quad T_{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \langle \cdot, K_{x_i}\rangle_{\mathcal{H}} K_{x_i}.$$

Moreover, $S_{\mathbf{x}} S_{\mathbf{x}}^* = n^{-1}\mathbf{K}$ where $\mathbf{K}$ is the kernel matrix such that $(\mathbf{K})_{ij} = K(x_i, x_j)$.

Throughout we indicate with $\|\cdot\|$ the norm in the Banach space $\mathcal{L}(\mathcal{H})$ of bounded linear operators from $\mathcal{H}$ to $\mathcal{H}$.

2.3. A priori assumption on the problem: general source condition

It is well known that to obtain probabilistic bounds such as that in (3) (or rather bounds on (6)) we have to restrict the class of possible probability measures. In Learning Theory this is related to the so-called “no free lunch” Theorem [15] but similar kind of phenomenon occurs in statistics [18] and in regularization of ill-posed inverse problems [16]. Essentially, what happens is that we can always find a solution with convergence guarantees to some prescribed target function but the convergence rates can be arbitrary slow. In our setting this turns into the impossibility to state finite sample bounds holding uniformly with respect to any probability measure ρ.

A standard way to impose restrictions on the class of possible problems is to consider a set of probability measures $\mathcal{M}(\Omega)$ such that the associated regression functions satisfies $f_\rho \in \Omega$. Such a condition is called the prior. The set Ω is usually a compact set determined by smoothness conditions [14]. In the context of RKHSs it is natural to describe the prior in term of the compact operator L_K, considering $f_\rho \in \Omega_{r,R}$ with

$$\Omega_{r,R} = \{f \in L^2(X, \rho_X) : f = L_K^r u, \|u\|_\rho \leqslant R\}. \tag{9}$$

The above condition is often written as $\left\|L_K^{-r} f_\rho\right\|_\rho \leqslant R$ [27]. Note that, when $r = \frac{1}{2}$, such a condition is equivalent to assuming $f_\rho \in \mathcal{H}$ and is independent of the measure ρ, but for arbitrary r it is distribution dependent.

As noted in De Vito et al. [12,13] the condition $f_\rho \in \Omega_{r,R}$ corresponds to what is called a source condition in the inverse problems literature. In fact if we consider $Pf_\rho \in \Omega_{r,R}$, $r > \frac{1}{2}$, then $Pf_\rho \in R(I_K)$ and we can equivalently consider the prior $f_{\mathcal{H}}^\dagger \in \Omega_{v,R}$ with

$$\Omega_{v,R} = \{f \in \mathcal{H} : f = T^v v, \|v\|_{\mathcal{H}} \leqslant R\}, \tag{10}$$

where $v = r - \frac{1}{2}$ (see, for example, [12, Proposition 3.2]). Recalling that $T = I_K^* I_K$ we see that the above condition is the standard source condition for the linear problem $I_K f = f_\rho$, namely Hölder source condition [16].

Following what is done in inverse problems in this paper we wish to extend the class of possible probability measures $\mathcal{M}(\Omega)$ considering general source condition (see [21] and references therein). We assume throughout that $Pf_\rho \in R(I_K)$ which means that $f_{\mathcal{H}}^\dagger$ exists and solves the normalized embedding equation $Tf = I_K^* f_\rho$. Using the singular value decompositions

$$T = \sum_{i=1}^{\infty} t_i \, \langle \cdot, e_i \rangle_{\mathcal{H}} \, e_i, \quad L_K = \sum_{i=1}^{\infty} t_i \left\langle \cdot, \psi_i \right\rangle_\rho \psi_i,$$

for orthonormal systems $\{e_i\}$ in $\mathcal{H}$ and $\{\psi_i\}$ in $L^2(X, \rho_X)$ and sequence of singular numbers $\kappa^2 \geqslant t_1 \geqslant t_2 \geqslant \cdots \geqslant 0$, one can represent $f_{\mathcal{H}}^\dagger$ in the form

$$f_{\mathcal{H}}^\dagger = \sum_{i=1}^{\infty} \frac{1}{\sqrt{t_i}} \left\langle f_\rho, \psi_i \right\rangle_\rho e_i.$$

Then $f_{\mathcal{H}}^\dagger \in \mathcal{H}$ if and only if

$$\sum_{i=1}^{\infty} \frac{\left\langle f_\rho, \psi_i \right\rangle_\rho^2}{t_i} < \infty,$$

where the above condition is known as Picard's criterion. It provides a zero-smoothness condition on $f_{\mathcal{H}}^\dagger$ (merely $f_{\mathcal{H}}^\dagger \in \mathcal{H}$) and tells us that the Fourier coefficients $\left\langle f_\rho, \psi_i \right\rangle_\rho$ should decay much faster than t_i. Therefore, it seems natural to measure the smoothness of $f_{\mathcal{H}}^\dagger$ by enforcing some faster decay. More precisely, not only Picard's criterion but also the stronger condition

$$\sum_{i=1}^{\infty} \frac{\left\langle f_\rho, \psi_i \right\rangle_\rho^2}{t_i \phi^2(t_i)} < \infty$$

is satisfied, where ϕ is some continuous increasing function defined on the interval $[0, \kappa^2] \supset \{t_i\}$ and such that $\phi(0) = 0$. Then

$$v := \sum_{i=1}^{\infty} \frac{1}{\sqrt{t_i} \phi(t_i)} \left\langle f_\rho, \psi_i \right\rangle_\rho e_i$$

and

$$f_{\mathcal{H}}^\dagger = \sum_{i=1}^{\infty} \phi(t_i) \, \langle v, e_i \rangle \, e_i = \phi(T) v \in \mathcal{H}.$$

Thus, additional smoothness of $f_{\mathcal{H}}^\dagger$ can be expressed as an inclusion

$$f_{\mathcal{H}}^\dagger \in \Omega_{\phi,R} := \{ f \in \mathcal{H} : f = \phi(T) v, \ \|v\|_{\mathcal{H}} \leqslant R \}, \tag{11}$$

that goes usually under the name of *source condition*. The function ϕ is called index function. There is a good reason to further restrict the class of possible index functions. In general, the smoothness expressed through source conditions is not stable with respect to perturbations in the involved operator T. In Learning Theory only the empirical covariance operator $T_{\mathbf{x}}$ is available and it is desirable to control $\phi(T) - \phi(T_{\mathbf{x}})$. This can be achieved by requiring ϕ to be operator monotone. Recall that the function ϕ is operator monotone on $[0, b]$ if for any pair of self-adjoint operators U, V, with spectra in $[0, b]$ if such that $U \leqslant V$ we have $\phi(U) \leqslant \phi(V)$. The partial ordering $B_1 \leqslant B_2$ for self-adjoint operators B_1, B_2 on some Hilbert space $\mathcal{H}$ means that for any $h \in \mathcal{H}$, $\langle B_1 h, h\rangle \leqslant \langle B_2 h, h\rangle$. It follows from the Löwner theorem (see for example [19]) that each operator monotone function on $(0, b)$ admits an analytic continuation in the corresponding strip of the upper half-plane with positive imaginary part. Important implications of the concept of operator monotonicity in the context of regularization can be seen from the following result (see [20,22]).

Theorem 1. *Suppose ψ is an operator monotone index function on* $[0, b]$, *with $b > a$. Then there is a constant $c_\psi < \infty$ depending on $b - a$, such that for any pair B_1, B_2, $\|B_1\|, \|B_2\| \leqslant a$, of non-negative self-adjoint operators on some Hilbert space it holds*

$$\|\psi(B_1) - \psi(B_2)\| \leqslant c_\psi \psi(\|B_1 - B_2\|).$$

Moreover, there is $c > 0$ such that

$$c\frac{\lambda}{\psi(\lambda)} \leqslant \frac{\sigma}{\psi(\sigma)}$$

whenever $0 < \lambda < \sigma \leqslant a < b$.

Thus, operator monotone index functions allow a desired norm estimate for $\phi(T) - \phi(T_{\mathbf{x}})$. Therefore, in the following we consider index functions from the class:

$$\mathcal{F}_C = \left\{\psi : [0, b] \to \mathbb{R}_+, \text{operator monotone}, \psi(0) = 0, \psi(b) \leqslant C, b > \kappa^2\right\}.$$

Note that from the above theorem it follows that an index function $\psi \in \mathcal{F}_C$ cannot converge faster than linearly to 0. To overcome this limitation of the class $\mathcal{F}_C$ we also introduce the class $\mathcal{F}$ of index functions $\phi : [0, \kappa^2] \to \mathbb{R}_+$ which can be split into a part $\psi \in \mathcal{F}_C$ and a monotone Lipschitz part $\vartheta : [0, \kappa^2] \to \mathbb{R}_+$, $\vartheta(0) = 0$, i.e. $\phi(\sigma) = \vartheta(\sigma)\psi(\sigma)$. This splitting is not unique such that we implicitly assume that the Lipschitz constant for ϑ is equal to 1 which means

$$\|\vartheta(T) - \vartheta(T_{\mathbf{x}})\| \leqslant \|T - T_{\mathbf{x}}\|.$$

The fact that an operator-valued function ϑ is Lipschitz continuous if a real function ϑ is Lipschitz continuous follows from Theorem 8.1 in Birman and Solomyak [3].

Remark 2. Observe that for $v \in [0, 1]$ a Hölder-type source condition (10) can be seen as (11) with $\phi(\sigma) = \sigma^v \in \mathcal{F}_C$, $C = b^v$, $b > \kappa^2$ while for $v > 1$ we can write $\phi(\sigma) = \vartheta(\sigma)\psi(\sigma)$ where $\vartheta(\sigma) = \sigma^p / C_1$ and $\psi(\sigma) = C_1 \sigma^{v-p} \in \mathcal{F}_C$, $C = C_1 b^{v-p}$, $b > \kappa^2$, $C_1 = p\kappa^{2(p-1)}$ and $p = [v]$ is an integer part of v. It is clear that the Lipschitz constant for such a $\vartheta(\sigma)$ is equal to 1. At the same time, source conditions (11) with $\phi \in \mathcal{F}$ cover all types of smoothness studied so far in Regularization Theory. For example, $\psi(\sigma) = \sigma^p \log^{-v} 1/\sigma$ with $p = 0, 1, \ldots, v \in [0, 1]$ can be split in a Lipschitz part $\vartheta(\sigma) = \sigma^p$ and an operator monotone part $\psi(\sigma) = \log^{-v} 1/\sigma$.

3. Regularization in learning theory

In this section we first present the class of regularization algorithms we are going to study. Regularization is defined according to what is usual done for ill-posed inverse problems. Second we give the main results of the paper. It turns out that such a notion of regularization allows to derive learning algorithms which are consistent possibly with fast convergence rate. Several corollaries illustrate this fact.

3.1. Regularization algorithms

It is well known that Tikhonov regularization can be profitably used in the context of supervised learning and many theoretical properties have been shown. The question whether other regularization techniques from the theory of ill-posed inverse problems can be valuable in the context of Learning Theory has been considered in Rosasco et al. [24] motivated by some connections between learning and inverse problems [12,13]. In this paper we follow the same approach and provide a refined analysis for algorithms defined by

$$f_{\mathbf{z}}^{\lambda} = g_{\lambda}(T_{\mathbf{x}}) S_{\mathbf{x}}^{*} \mathbf{y}, \tag{12}$$

where the final estimator is defined providing the above scheme with a parameter choice $\lambda_n = \lambda(n, \mathbf{z})$ so that $f_{\mathbf{z}} = f_{\mathbf{z}}^{\lambda_n}$. We show that the following definition characterizes which regularization provide sensible learning algorithms. Interestingly, such a definition is the standard definition characterizing regularization for ill-posed problems [16].

Definition 1 (*Regularization*). We say that a family $g_{\lambda} : [0, \kappa^2] \to \mathbb{R}$, $0 < \lambda \leqslant \kappa^2$, is regularization if the following conditions hold

- There exists a constant D such that

$$\sup_{0<\sigma\leqslant\kappa^2} |\sigma g_{\lambda}(\sigma)| \leqslant D. \tag{13}$$

- There exists a constant B such that

$$\sup_{0<\sigma\leqslant\kappa^2} |g_{\lambda}(\sigma)| \leqslant \frac{B}{\lambda}. \tag{14}$$

- There exists a constant γ such that

$$\sup_{0<\sigma\leqslant\kappa^2} |1 - g_{\lambda}(\sigma)\sigma| \leqslant \gamma. \tag{15}$$

- The qualification of the regularization g_{λ} is the maximal v such that

$$\sup_{0<\sigma\leqslant\kappa^2} |1 - g_{\lambda}(\sigma)\sigma| \sigma^{v} \leqslant \gamma_{v} \lambda^{v}, \tag{16}$$

  where γ_v does not depend on λ.

The above condition are standard in the theory of inverse problems and, as shown in Theorem 10, are also sufficient to obtain consistent learning schemes. In Rosasco et al. [24] an extra condition was required on g_{λ}, namely a Lipschitz condition. Here, we show that at least in the considered range of prior such a condition can be dropped and the conditions considered for inverse problems are sufficient to learning. We give some examples which will be discussed in the following (see [16] for details and [24] for more discussion in the context of learning).

Example 3 (*Tikhonov*). The choice $g_\lambda(\sigma) = \frac{1}{\sigma+\lambda}$ corresponds to Tikhonov regularization or the regularized least squares algorithm. In this case we have $B = D = \gamma = 1$. The qualification of the method is 1 and $\gamma_\nu = 1$.

Example 4 (*Landweber iteration*). We assume for simplicity that $\kappa = 1$. Then Landweber iteration is defined by $g_t(\sigma) = \sum_{i=0}^{t-1}(1-\sigma)^i$ where we identify $\lambda = t^{-1}, t \in \mathbb{N}$. This corresponds to the gradient descent algorithm in Yao et al. [33] with constant step-size. In this case we have $B = D = \gamma = 1$. Any $\nu \in [0, \infty)$ can be considered as qualification of the method and $\gamma_\nu = 1$ if $0 < \nu \leqslant 1$ and $\gamma_\nu = \nu^\nu$ otherwise.

Example 5 (*Spectral cut-off*). A classical regularization algorithms for ill-posed inverse problems is spectral cut-off or truncated singular value decomposition (TSVD) defined by

$$g_\lambda = \begin{cases} \dfrac{1}{\sigma}, & \sigma \geqslant \lambda, \\ 0, & \sigma < \lambda. \end{cases}$$

Up-to our knowledge this method is not used in Learning Theory and could not be treated in the analysis of Rosasco et al. [24]. In this case we have $B = D = \gamma = 1$. The qualification of the method is arbitrary and $\gamma_\nu = 1$.

Example 6 (*Accelerated Landweber iteration*). Finally, we consider a class of methods called Accelerated Landweber or Semiiterative regularization. Here, again assume for simplicity that $\kappa = 1$ and identify $\lambda = t^{-2}, t \in \mathbb{N}$. Such methods are defined by $g_t(\sigma) = p_{t-1}(\sigma)$ where p_{t-1} is a polynomial of degree $t-1$. In this case $D = \gamma = 1$, $B = 2$. The so-called ν-method falls into this class of schemes. Though they usually have finite qualification the advantage of this iterative algorithms is that they require a number of iteration which is considerably smaller than Landweber iteration (see [16, Chapter 6]).

We end this section discussing the important interplay between qualification and a source condition. To this aim we need the following definition from Mathé and Pereverzev [21].

Definition 2. We say that the qualification ν_0 covers ϕ, if there is $c > 0$ such that

$$c\frac{\lambda^{\nu_0}}{\phi(\lambda)} \leqslant \inf_{\lambda \leqslant \sigma \leqslant \kappa^2} \frac{\sigma^{\nu_0}}{\phi(\sigma)}, \tag{17}$$

where $0 < \lambda \leqslant \kappa^2$.

The following important result is a restatement of Proposition 3 in Pereverzev [21].

Proposition 7. *Let ϕ be a non-decreasing index function and let g_λ be a regularization with qualification which covers ϕ. Then the following inequality holds true*:

$$\sup_{0<\sigma\leqslant\kappa^2} |1 - g_\lambda(\sigma)\sigma|\phi(\sigma) \leqslant c_g\phi(\lambda), \quad c_g = \frac{\gamma_\nu}{c},$$

where c is a constant from (17).

Remark 8. The index functions $\phi \in \mathcal{F}$ are covered by regularization with infinite qualification such as spectral cut-off or Landweber iteration. Moreover, from Theorem 1 above it follows that the index functions $\phi \in \mathcal{F}_C$ are covered by the qualification of Tikhonov regularization. Note also that if the function $\sigma \to \sigma^\nu/\phi(\sigma)$ is increasing then (17) is certainly satisfied with $c = 1$.

3.2. Main result

The following result provides us with error estimates for a fixed value of the regularization parameter λ. In order to give the proof we need the following Lemma whose proof is postponed to Section 3.4.

Lemma 9. *Let Assumption* (1) *hold and* κ *as in* (7). *For* $0 < \eta \leqslant 1$ *and* $n \in \mathbb{N}$ *let*

$$G_\eta = \left\{ \mathbf{z} \in Z^n : \left\| T_{\mathbf{x}} f_{\mathcal{H}}^\dagger - S_{\mathbf{x}}^* \mathbf{y} \right\|_{\mathcal{H}} \leqslant \delta_1, \, \|T - T_{\mathbf{x}}\| \leqslant \delta_2 \right\},$$

with

$$\delta_1 := \delta_1(n, \eta) = 2\left(\frac{\kappa M}{n} + \frac{\kappa \Sigma}{\sqrt{n}} \right) \log \frac{4}{\eta},$$

$$\delta_2 := \delta_2(n, \eta) = \frac{1}{\sqrt{n}} 2\sqrt{2}\kappa^2 \log \frac{4}{\eta}.$$

Then

$$P\left[G_\eta\right] \geqslant 1 - \eta.$$

The above result provides us with the probabilistic perturbation measures which quantify the effect of random sampling. We are now ready to state the following theorem.

Theorem 10. *Let* $\lambda \in (0, 1]$. *Assume that* (1) *and* (2) *hold. Moreover, assume that* $Pf_\rho \in R(I_K)$ *and* $f_{\mathcal{H}}^\dagger \in \Omega_{\phi,R}$. *We let* $f_{\mathbf{z}}^\lambda$ *as in* (12), *satisfying Definition* 1 *and assume that the regularization has a qualification which covers* $\phi(\sigma)\sqrt{\sigma}$. *If*

$$\lambda \geqslant \frac{1}{\sqrt{n}} 2\sqrt{2}\kappa^2 \log \frac{4}{\eta} \tag{18}$$

for $0 < \eta < 1$ *then with probability at least* $1 - \eta$

$$\left\| f_{\mathbf{z}}^\lambda - f_{\mathcal{H}}^\dagger \right\|_\rho \leqslant \left(C_1 \phi(\lambda)\sqrt{\lambda} + C_2 \frac{1}{\sqrt{\lambda n}} \right) \log \frac{4}{\eta}, \tag{19}$$

where $C_1 = 2(1 + c_\psi)c_g R$ *and* $C_2 = \left((1 + c_g)\gamma C R 2\sqrt{2}\kappa^2 + \left(\sqrt{DB} + B\right)\left(\kappa\Sigma + \frac{M}{\sqrt{2}\kappa}\right)\right)$.

Moreover, with probability at least $1 - \eta$

$$\left\| f_{\mathbf{z}}^\lambda - f_{\mathcal{H}}^\dagger \right\|_{\mathcal{H}} \leqslant \left(C_3 \phi(\lambda) + C_4 \frac{1}{\lambda\sqrt{n}} \right) \log \frac{4}{\eta}, \tag{20}$$

where $C_3 = (1 + c_\psi)c_g R$ *and* $C_4 = \left(\gamma C R 2\sqrt{2}\kappa^2 + B\left(\kappa\Sigma + \frac{M}{\sqrt{2}\kappa}\right)\right)$.

Proof. Let δ_1, δ_2 and G_η as in Lemma 9. Then from this lemma we know that

$$P\left[G_\eta\right] \geqslant 1-\eta. \tag{21}$$

Moreover, we let

$$r_\lambda(\sigma) = 1 - \sigma g_\lambda(\sigma). \tag{22}$$

We consider the following decomposition into two terms:

$$\begin{aligned} f_{\mathcal{H}}^\dagger - f_{\mathbf{z}}^\lambda &= f_{\mathcal{H}}^\dagger - g_\lambda(T_{\mathbf{x}})S_{\mathbf{x}}^*\mathbf{y} \\ &= (f_{\mathcal{H}}^\dagger - g_\lambda(T_{\mathbf{x}})T_{\mathbf{x}}f_{\mathcal{H}}^\dagger) + (g_\lambda(T_{\mathbf{x}})T_{\mathbf{x}}f_{\mathcal{H}}^\dagger - g_\lambda(T_{\mathbf{x}})S_{\mathbf{x}}^*\mathbf{y}). \end{aligned} \tag{23}$$

The idea is then to separately bound each term both in the norm in $\mathcal{H}$ and in $L^2(X, \rho_X)$.

We start dealing with the first term. Using (11) and (22) we can write

$$\begin{aligned} f_{\mathcal{H}}^\dagger - g_\lambda(T_{\mathbf{x}})T_{\mathbf{x}}f_{\mathcal{H}}^\dagger &= (I - g_\lambda(T_{\mathbf{x}})T_{\mathbf{x}})\phi(T)v \\ &= r_\lambda(T_{\mathbf{x}})\phi(T_{\mathbf{x}})v + r_\lambda(T_{\mathbf{x}})(\phi(T) - \phi(T_{\mathbf{x}}))v \\ &= r_\lambda(T_{\mathbf{x}})\phi(T_{\mathbf{x}})v + r_\lambda(T_{\mathbf{x}})\vartheta(T_{\mathbf{x}})(\psi(T) - \psi(T_{\mathbf{x}}))v \\ &\quad + r_\lambda(T_{\mathbf{x}})(\vartheta(T) - \vartheta(T_{\mathbf{x}}))\psi(T)v. \end{aligned} \tag{24}$$

When considering the norm in $\mathcal{H}$ we know that Proposition 7 applies since ϕ (as well as ϑ) is covered by the qualification of g_λ. The fact that ϑ is covered by the qualification of g_λ can be seen from the following chain of inequalities:

$$\begin{aligned} \inf_{\lambda\leqslant\sigma\leqslant\kappa^2} \frac{\sigma^{v_0}}{\vartheta(\sigma)} &= \inf_{\lambda\leqslant\sigma\leqslant\kappa^2} \frac{\sigma^{v_0}\psi(\sigma)}{\vartheta(\sigma)\psi(\sigma)} \geqslant \psi(\lambda) \inf_{\lambda\leqslant\sigma\leqslant\kappa^2} \frac{\sigma^{v_0}}{\phi(\sigma)} \\ &\geqslant c\psi(\lambda)\frac{\lambda^{v_0}}{\phi(\lambda)} = c\frac{\lambda^{v_0}}{\vartheta(\lambda)}, \end{aligned}$$

where we rely on the fact that $\phi(\lambda) = \psi(\lambda)\vartheta(\lambda)$ is covered by the qualification of g_λ, and an operator monotone index function $\psi(\lambda)$ is non-decreasing. Then we can use (16), (15), (11) and Theorem 1 to get the bound

$$\left\| f_{\mathcal{H}}^\dagger - g_\lambda(T_{\mathbf{x}})T_{\mathbf{x}}f_{\mathcal{H}}^\dagger \right\|_{\mathcal{H}} \leqslant c_g R\phi(\lambda) + c_g c_\psi R\vartheta(\lambda)\psi(\|T - T_{\mathbf{x}}\|) + \gamma C R \|T - T_{\mathbf{x}}\|$$

and for $\mathbf{z} \in G_\eta$ we have

$$\left\| f_{\mathcal{H}}^\dagger - g_\lambda(T_{\mathbf{x}})T_{\mathbf{x}}f_{\mathcal{H}}^\dagger \right\|_{\mathcal{H}} \leqslant (1 + c_\psi)c_g R\phi(\lambda) + \gamma C R\delta_2, \tag{25}$$

where we used (18) to have $\vartheta(\lambda)\psi(\|T - T_{\mathbf{x}}\|) \leqslant \vartheta(\lambda)\psi(\delta_2) \leqslant \vartheta(\lambda)\psi(\lambda) = \phi(\lambda)$. Some more reasoning is needed to get the bound in $L^2(X, \rho_X)$. To this aim in place of (24) we consider

$$\sqrt{T}(f_{\mathcal{H}}^\dagger - g_\lambda(T_{\mathbf{x}})T_{\mathbf{x}}f_{\mathcal{H}}^\dagger) = (\sqrt{T} - \sqrt{T_{\mathbf{x}}})(I - g_\lambda(T_{\mathbf{x}})T_{\mathbf{x}})f_{\mathcal{H}}^\dagger + \sqrt{T_{\mathbf{x}}}(I - g_\lambda(T_{\mathbf{x}})T_{\mathbf{x}})f_{\mathcal{H}}^\dagger. \tag{26}$$

The first addend is easy to bound since from Condition (18) and operator monotonicity of $\psi(\sigma) = \sqrt{\sigma}$ we get

$$\left\| \sqrt{T} - \sqrt{T_{\mathbf{x}}} \right\| \leqslant \sqrt{\|T - T_{\mathbf{x}}\|} \leqslant \sqrt{\delta_2} \leqslant \sqrt{\lambda} \tag{27}$$

for $\mathbf{z} \in G_\eta$. Then from the above inequality and from (25) we get

$$\left\| (\sqrt{T} - \sqrt{T_{\mathbf{x}}})(I - g_\lambda(T_{\mathbf{x}})T_{\mathbf{x}}) f_{\mathcal{H}}^\dagger \right\|_{\mathcal{H}} \leqslant (1 + c_\psi) c_g R \phi(\lambda) \sqrt{\lambda} + \gamma C R \sqrt{\lambda} \delta_2. \tag{28}$$

On the other hand, the second addend can be further decomposed using (11)

$$\begin{aligned} \sqrt{T_{\mathbf{x}}}(I - g_\lambda(T_{\mathbf{x}})T_{\mathbf{x}})\phi(T)v = & \sqrt{T_{\mathbf{x}}} r_\lambda(T_{\mathbf{x}})\phi(T_{\mathbf{x}})v \\ & + \sqrt{T_{\mathbf{x}}} r_\lambda(T_{\mathbf{x}})\vartheta(T_{\mathbf{x}})(\psi(T) - \psi(T_{\mathbf{x}}))v \\ & + \sqrt{T_{\mathbf{x}}} r_\lambda(T_{\mathbf{x}})(\vartheta(T) - \vartheta(T_{\mathbf{x}}))\psi(T)v. \end{aligned}$$

Using (16), (15), (11) and Theorem 1 we get for $\mathbf{z} \in G_\eta$

$$\left\| \sqrt{T_{\mathbf{x}}}(I - g_\lambda(T_{\mathbf{x}})T_{\mathbf{x}}) f_{\mathcal{H}}^\dagger \right\|_{\mathcal{H}} \leqslant (1 + c_\psi) c_g R \phi(\lambda) \sqrt{\lambda} + c_g \gamma C R \sqrt{\lambda} \delta_2,$$

where again we used (18) to have $\psi(\|T - T_{\mathbf{x}}\|) \leqslant \psi(\delta_2) \leqslant \psi(\lambda)$. Now we can put the above inequality and (28) together to obtain the following bound in the ρ-norm:

$$\left\| \sqrt{T}(f_{\mathcal{H}}^\dagger - g_\lambda(T_{\mathbf{x}})T_{\mathbf{x}} f_{\mathcal{H}}^\dagger) \right\|_{\mathcal{H}} \leqslant 2(1 + c_\psi) c_g R \phi(\lambda) \sqrt{\lambda} + (1 + c_g) \gamma C R \sqrt{\lambda} \delta_2. \tag{29}$$

We are now ready to consider the second term in (23). If we consider the norm in $\mathcal{H}$ we can write

$$g_\lambda(T_{\mathbf{x}})T_{\mathbf{x}} f_{\mathcal{H}}^\dagger - g_\lambda(T_{\mathbf{x}})S_{\mathbf{x}}^* \mathbf{y} = g_\lambda(T_{\mathbf{x}})(T_{\mathbf{x}} f_{\mathcal{H}}^\dagger - S_{\mathbf{x}}^* \mathbf{y})$$

and for $\mathbf{z} \in G_\eta$ then condition (14) immediately yields

$$\left\| g_\lambda(T_{\mathbf{x}})T_{\mathbf{x}} f_{\mathcal{H}}^\dagger - g_\lambda(T_{\mathbf{x}})S_{\mathbf{x}}^* \mathbf{y} \right\|_{\mathcal{H}} \leqslant \frac{B}{\lambda} \delta_1. \tag{30}$$

Moreover, when considering the norm in $L^2(X, \rho_X)$ we simply have

$$\begin{aligned} \sqrt{T}(g_\lambda(T_{\mathbf{x}})T_{\mathbf{x}} f_{\mathcal{H}}^\dagger - g_\lambda(T_{\mathbf{x}})S_{\mathbf{x}}^* \mathbf{y}) = & \sqrt{T_{\mathbf{x}}} g_\lambda(T_{\mathbf{x}})(T_{\mathbf{x}} f_{\mathcal{H}}^\dagger - S_{\mathbf{x}}^* \mathbf{y}) \\ & + (\sqrt{T} - \sqrt{T_{\mathbf{x}}}) g_\lambda(T_{\mathbf{x}})(T_{\mathbf{x}} f_{\mathcal{H}}^\dagger - S_{\mathbf{x}}^* \mathbf{y}). \end{aligned} \tag{31}$$

It is easy to show that

$$\left\| \sqrt{T_{\mathbf{x}}} g_\lambda(T_{\mathbf{x}}) \right\| \leqslant \frac{\sqrt{DB}}{\sqrt{\lambda}}$$

in fact $\forall h \in \mathcal{H}$ from Cauchy–Schwartz inequality we have

$$\begin{aligned} \left| \left\langle \sqrt{T_{\mathbf{x}}} g_\lambda(T_{\mathbf{x}})h, \sqrt{T_{\mathbf{x}}} g_\lambda(T_{\mathbf{x}})h \right\rangle_{\mathcal{H}} \right| = & \left| \langle g_\lambda(T_{\mathbf{x}})h, T_{\mathbf{x}} g_\lambda(T_{\mathbf{x}})h \rangle \right| \\ \leqslant & \|g_\lambda(T_{\mathbf{x}})h\|_{\mathcal{H}} \|T_{\mathbf{x}} g_\lambda(T_{\mathbf{x}})h\|_{\mathcal{H}} \\ \leqslant & D \frac{B}{\lambda} \|h\|_{\mathcal{H}}^2, \end{aligned}$$

where we used (13) and (14). We can use the definition of δ_1 with the above inequality to bound the first addend in (31) and the definition of δ_1 with inequalities (27), (14) to bound the second addend in (31). Then, using (18), we have $\sqrt{\delta_2} \leqslant \sqrt{\lambda}$ so that

$$\left\| \sqrt{T}(g_\lambda(T_{\mathbf{x}})T_{\mathbf{x}} f_{\mathcal{H}}^\dagger - g_\lambda(T_{\mathbf{x}})S_{\mathbf{x}}^* \mathbf{y}) \right\|_{\mathcal{H}} \leqslant \frac{\sqrt{DB}}{\sqrt{\lambda}}\delta_1 + \sqrt{\delta_2}\frac{B}{\lambda}\delta_1 \leqslant \frac{(\sqrt{DB}+B)}{\sqrt{\lambda}}\delta_1 \tag{32}$$

for $\mathbf{z} \in G_\eta$. We now are in the position to derive the desired bounds.

Recalling (21) and (23), we can put (25) and (30) together to get with probability at least $1-\eta$,

$$\left\| f_{\mathbf{z}}^\lambda - f_{\mathcal{H}}^\dagger \right\|_{\mathcal{H}} \leqslant (1+c_\psi)c_g R\phi(\lambda) + \gamma C R \delta_2 + \frac{B}{\lambda}\delta_1.$$

We can then simplify the above bound. In fact $\delta_2 \leqslant \delta_2/\lambda$ since $\lambda \leqslant 1$ so that

$$\gamma C R \delta_2 \leqslant \log\frac{4}{\eta}\gamma C R 2\sqrt{2}\kappa^2 \frac{1}{\lambda\sqrt{n}}.$$

Moreover, from the explicit expression of δ_1, using (18) and $\lambda \leqslant 1$ it is easy to prove that

$$\frac{B}{\lambda}\delta_1 \leqslant \log\frac{4}{\eta} B\left(\kappa\Sigma + \frac{M}{\sqrt{2}\kappa}\right)\frac{1}{\lambda\sqrt{n}}.$$

Putting everything together we have (20) in fact

$$\left\| f_{\mathbf{z}}^\lambda - f_{\mathcal{H}}^\dagger \right\|_{\mathcal{H}} \leqslant \left(C_3\phi(\lambda) + C_4\frac{1}{\lambda\sqrt{n}} \right)\log\frac{4}{\eta},$$

where $C_3 = (1+c_\psi)c_g R$ and $C_4 = (\gamma C R 2\sqrt{2}\kappa^2 + B(\kappa\Sigma + \frac{M}{\sqrt{2}\kappa}))$.

Similarly, we can use Eq. (8) to write

$$\left\| f_{\mathbf{z}}^\lambda - f_{\mathcal{H}}^\dagger \right\|_\rho = \left\| \sqrt{T}(f_{\mathbf{z}}^\lambda - f_{\mathcal{H}}^\dagger) \right\|_{\mathcal{H}}$$

and from (29) and (32) we get with probability at least $1-\eta$

$$\left\| \sqrt{T}(f_{\mathbf{z}}^\lambda - f_{\mathcal{H}}^\dagger) \right\|_{\mathcal{H}} \leqslant 2(1+c_\psi)c_g R\phi(\lambda)\sqrt{\lambda} + (1+c_g)\gamma C R\sqrt{\lambda}\delta_2 + \frac{(\sqrt{DB}+B)}{\sqrt{\lambda}}\delta_1$$

which can be further simplified as above to get (19). □

Remark 11 (*Assumptions on the regularization parameter*). A condition similar to (18) has been considered in Smale and Zhou [27] and Caponnetto De Vito [7,8]. It simply indicates the range of regularization parameters, for which the error estimates (19) and (20) are non-trivial. For example, if λ does not satisfy (18) then right-hand side of (20) becomes larger than a fixed constant $C_4/(2\sqrt{2}\kappa^2)$, which is not reasonable. Thus, condition (18) is not restrictive at all. In fact it is automatically satisfied for the best a priori choice of the regularization parameter (see Theorem 14) balancing the values of the terms in the estimates (19) and (20). Finally, the condition $\lambda < 1$ is considered only to simplify the results and can be replaced by $\lambda < a$ for some positive constant a (and in particular for $a = \kappa$) that would eventually appear in the bound.

Remark 12 (*Assumption on the best in the model*). If $\mathcal{H}$ is dense in $L^2(X, \rho_X)$ or $f_\rho \in \mathcal{H}$ clearly we can replace $f_{\mathcal{H}}^{\dagger}$ with f_ρ since $\mathcal{E}(f_{\mathcal{H}}^{\dagger}) = \inf_{f \in \mathcal{H}} \mathcal{E}(f) = \mathcal{E}(f_\rho)$.

A drawback in our approach is that we have to assume the existence of $f_{\mathcal{H}}^{\dagger}$. Though this assumption is necessary to study result in the $\mathcal{H}$-norm it can be relaxed when looking for bounds in $L^2(X, \rho_X)$. In fact, as discussed in De Vito et al. [12,13], Yao et al. [33] if $f_{\mathcal{H}}^{\dagger}$ does not exist we can still consider

$$\mathcal{E}(f_{\mathbf{z}}) - \inf_{\mathcal{H}} \mathcal{E}(f) = \left\| f_{\mathbf{z}} - P f_\rho \right\|_\rho^2$$

in place of (6). For this kind of prior (only for Hölder source condition) the best results were obtained in Smale and Zhou [27] for Tikhonov regularization. The result on Landweber iteration in Yao et al. [33] also cover this case though the dependence on the number of examples is worse than for Tikhonov. Results for general regularization schemes were obtained in Rosasco et al. [24] requiring the regularization g_λ to be Lipschitz, but the dependence on the number of examples was again spoiled.

Remark 13 (*Bounds uniform w.r.t. λ*). Inspecting the proof of the above theorem we see that the family of good training sets such that the bounds hold with high probability do not depend on the value of the regularization parameter. This turns out to be useful to define a data-driven strategy for the choice of λ.

From the above results we can immediately derive a data independent (a priori) parameter choice $\lambda_n = \lambda(n)$. Next theorems show the error bounds obtained providing the one parameter family of algorithms in (12) with such a regularization parameter choice.

Theorem 14. *We let $\Theta(\lambda) = \phi(\lambda)\lambda$. Under the same assumptions of Theorem* 10 *we choose*

$$\lambda_n = \Theta^{-1}(n^{-\frac{1}{2}}) \tag{33}$$

and let $f_{\mathbf{z}} = f_{\mathbf{z}}^{\lambda_n}$. Then for $0 < \eta < 1$ and $n \in \mathbb{N}$ such that

$$\Theta^{-1}(n^{-\frac{1}{2}}) n^{\frac{1}{2}} \geqslant 2\sqrt{2}\kappa^2 \log \frac{4}{\eta} \tag{34}$$

the following bound holds with probability at least $1 - \eta$:

$$\left\| f_{\mathbf{z}} - f_{\mathcal{H}}^{\dagger} \right\|_\rho \leqslant (C_1 + C_2)\phi(\Theta^{-1}(n^{-\frac{1}{2}}))\sqrt{\Theta^{-1}(n^{-\frac{1}{2}})} \log \frac{4}{\eta},$$

with C_1 and C_2 as in Theorem 10. *Moreover, with probability at least $1 - \eta$*

$$\left\| f_{\mathbf{z}} - f_{\mathcal{H}}^{\dagger} \right\|_{\mathcal{H}} \leqslant (C_3 + C_4)\phi(\Theta^{-1}(n^{-\frac{1}{2}})) \log \frac{4}{\eta},$$

with C_3 and C_4 as in Theorem 10.

Proof. If we choose λ_n as in (33) then for n such that (34) holds we have that condition (18) is verified and we can apply the bounds of Theorem 10 to λ_n. The results easily follow noting that the proposed parameter choice is the one balancing the two terms in (19) in fact the following equation is verified for $\lambda = \lambda_n$:

$$\phi(\lambda)\sqrt{\lambda} = \frac{1}{\sqrt{\lambda n}}$$

($\phi(\lambda) = \lambda^{-1}n^{-1/2}$ for the $\mathcal{H}$-norm). □

Several corollaries easily follow. The following result considers the stochastic order [30] of convergence with respect to the ρ-norm.

Corollary 15. *Under the same assumptions of Theorem* 14 *if λ_n is chosen according to* (33) *and $f_{\mathbf{z}} = f_{\mathbf{z}}^{\lambda_n}$ then*

$$\lim_{A\to\infty} \limsup_{n\to\infty} \sup_{\rho\in\mathcal{M}(\Omega_{\phi,R})} P\left[\left\| f_{\mathbf{z}} - f_{\mathcal{H}}^{\dagger}\right\|_{\rho} > A a_n\right] = 0$$

for $a_n = \phi(\Theta^{-1}(n^{-\frac{1}{2}}))\sqrt{\Theta^{-1}(n^{-\frac{1}{2}})}$.

Proof. We let $A = (C_3 + C_4)\log\frac{4}{\eta}$ and solve with respect to η to get

$$\eta_A = 4e^{-\frac{A}{C_3+C_4}}.$$

Then we know from Theorem 14 that for n such that (34) holds

$$P\left[\left\| f_{\mathbf{z}} - f_{\mathcal{H}}^{\dagger}\right\|_{\rho} > A\phi(\Theta^{-1}(n^{-\frac{1}{2}})\sqrt{\Theta^{-1}(n^{-\frac{1}{2}})}\right] \leqslant \eta_A$$

and clearly

$$\limsup_{n\to\infty} \sup_{\rho\in\mathcal{M}(\Omega_{\phi,R})} P\left[\left\| f_{\mathbf{z}} - f_{\mathcal{H}}^{\dagger}\right\|_{\rho} > A\phi(\Theta^{-1}(n^{-\frac{1}{2}})\sqrt{\Theta^{-1}(n^{-\frac{1}{2}})}\right] \leqslant \eta_A.$$

The theorem is proved since $\eta_A \to 0$ as $A \to \infty$. □

Remark 16 (*Kernel independent lower bounds*). Up-to our knowledge no minimax lower bounds exist for the class of priors considered here. In fact in Caponnetto and De Vito [7,8] lower bounds are presented for $\rho \in \mathcal{M}(\Omega_{r,R})$, that is Hölder source condition, and considering the case when the eigenvalues of T have a polynomial decay $t_i \propto i^{-b}$, $b > 1$. In this case lower rate $a_n = n^{-\frac{rb}{2rb+1}}$, $1/2 < r \leqslant 1$ are shown to be optimal. Here, we do not make any assumption on the kernel and, in this sense, our results are kernel independent. This situation can be thought of as the limit case when $b = 1$. As it can be seen from next corollary we share the same dependence on the smoothness index r.

The following result considers the case of Hölder source conditions, that is the case when Condition (11) reduces to (10). Recalling the equivalence between (9) an (10) we state the following result considering $v = r - \frac{1}{2}$ to have an easier comparison with previous results.

Corollary 17. *Under the same assumption of Theorem* 14 *let* $\phi(\sigma) = \sigma^{v}$, $v = r - \frac{1}{2}$. *Now choose* λ_n *as in* (33) *and let* $f_{\mathbf{z}} = f_{\mathbf{z}}^{\lambda_n}$. *Then for* $0 < \eta < 1$ *and*

$$n > \left(2\sqrt{2}\kappa^2 \log \frac{4}{\eta}\right)^{\frac{4r+2}{2r+3}} \tag{35}$$

the following bounds hold with probability at least $1 - \eta$:

$$\left\| f_{\mathbf{z}} - f_{\mathcal{H}}^{\dagger} \right\|_{\rho} \leqslant (C_1 + C_2) n^{-\frac{r}{2r+1}} \log \frac{4}{\eta},$$

with C_1 *and* C_2 *as in Theorem* 10 *and*

$$\left\| f_{\mathbf{z}} - f_{\mathcal{H}}^{\dagger} \right\|_{\mathcal{H}} \leqslant (C_3 + C_4) n^{-\frac{r-1/2}{2r+1}} \log \frac{4}{\eta},$$

with C_3 *and* C_4 as in *Theorem* 10.

Proof. By a simple computation we have $\lambda_n = \Theta^{-1}(n^{-1/2}) = n^{-\frac{1}{2r+1}}$. Moreover, Condition (34) can now be written explicitly as in (35). The proof follows plugging the explicit form of ϕ and λ_n in the bounds of Theorem 14. □

Remark 18. Clearly if in place of $Pf_\rho \in \Omega_{r,R}$ we take $f_\rho \in \Omega_{r,R}$ with $r > \frac{1}{2}$ then $f_\rho \in \mathcal{H}$ and we can replace $f_{\mathcal{H}}^{\dagger}$ with f_ρ since $\inf_{f \in \mathcal{H}} \mathcal{E}(f) = \mathcal{E}(f_\rho)$.

In particular, we discuss the bounds corresponding to the examples of regularization algorithms discussed in Section 3.1 and for the sake of clarity we restrict ourselves to polynomial source condition and $\mathcal{H}$ dense.

Tikhonov regularization: In the considered range of prior ($r > \frac{1}{2}$) the above results match those obtained in Smale and Zhou [27] for Tikhonov regularization. We observe that this kind of regularization suffers from a saturation effect and the results no longer improve after a certain regularity level, $r = 1$ (or $r = \frac{3}{2}$ for the $\mathcal{H}$-norm) is reached. This is a well-known fact in the theory of inverse problems.

Landweber iteration: In the considered range of prior ($r > \frac{1}{2}$) the above results improve on those obtained in Smale and Zhou [33] for gradient descent learning. Moreover, as pointed out in Yao et al. [33] such an algorithm does not suffer from saturation and the rate can be extremely good if the regression function is regular enough (that is if r is big enough) though the constant gets worse.

Spectral cut-off regularization: The spectral cut-off regularization does not suffer from the saturation phenomenon and moreover the constant does not change with the regularity of the solution, allowing extremely good theoretical properties. Note that such an algorithm is computationally feasible if one can compute the SVD of the kernel matrix **K**.

Accelerated Landweber iteration: The semiiterative methods though suffering from a saturation effect may have some advantage on Landweber iteration from the computational point of view. In fact recalling that we can identify $\lambda = t^{-2}$ it is easy to see that they require the square root of the number of iterations required by Landweber iteration to get the same convergence rate.

Remark 19. Note that, though assuming that $f_{\mathcal{H}}^{\dagger}$ exists, we improve on the result in Rosasco et al. [24] and show that in the considered range of prior we can drop the Lipschitz assumption on

g_λ and obtain the same dependence on the number of examples n and on the confidence level η for all regularization g_λ satisfying Definition 1. This class of algorithms includes all the methods considered in Rosasco et al. [24] and in general all the linear regularization algorithms to solve ill-posed inverse problems. The key to avoid the Lipschitz assumption on g_λ is exploiting the stability of the source condition w.r.t. to operator perturbation.

3.3. Regularization for binary classification: risk bounds and Bayes consistency

We briefly discuss the performance of the proposed class of algorithms in the context of binary classification [6], that is when $Y = \{-1, 1\}$. The problem is that of discriminating the elements of two classes and as usual we can take sign $f_{\mathbf{z}}^{\lambda}$ as our decision rule. In this case some natural error measures can be considered. The *risk* or *misclassification error* is defined as

$$R(f) = \rho_Z(\{(x, y) \in Z \mid \operatorname{sign} f(x) \neq y\}),$$

whose minimizer is the *Bayes rule* sign f_ρ. The quantity we aim to control is the *excess risk*

$$R(f_{\mathbf{z}}) - R(f_\rho).$$

Moreover, as proposed in Smale and Zhou [27] it is interesting to consider

$$\left\| \operatorname{sign} f_{\mathbf{z}} - \operatorname{sign} f_\rho \right\|_\rho .$$

To obtain bounds on the above quantities the idea is to relate them to $\left\| f_{\mathbf{z}} - f_\rho \right\|_\rho$. A straightforward result can be obtained recalling that

$$R(f_{\mathbf{z}}) - R(f_\rho) \leqslant \left\| f_{\mathbf{z}} - f_\rho \right\|_\rho$$

see Bartlett et al. [2], Yao et al. [33]. Anyway it is interesting to consider the case when some extra information is available on the noise affecting the problem. This can be done considering Tsybakov noise condition

$$\rho_X(\{x \in X : |f_\rho(x)| \leqslant L\}) \leqslant B_q L^q \quad \forall L \in [0, 1], \tag{36}$$

where $q \in [0, \infty]$ [29]. As shown in Proposition 6.2 in Yao et al. [33] (see also [2]) the following inequalities hold for $\alpha = \frac{q}{q+1}$:

$$R(f_{\mathbf{z}}) - R(f_\rho) \leqslant 4c_\alpha \left\| f_{\mathbf{z}} - f_\rho \right\|_\rho^{\frac{2}{2-\alpha}},$$

$$\left\| \operatorname{sign} f_{\mathbf{z}} - \operatorname{sign} f_\rho \right\|_\rho \leqslant 4c_\alpha \left\| f_{\mathbf{z}} - f_\rho \right\|_\rho^{\frac{\alpha}{2-\alpha}},$$

with $c_\alpha = B_q + 1$.

A direct application of Theorem 14 immediately leads to the following result

Corollary 20. *Assume that $\mathcal{H}$ is dense in $L^2(X, \rho_X)$ and that the same assumptions of Theorem* 14 *hold. Choose λ_n according to* (33) *and let $f_{\mathbf{z}} = f_{\mathbf{z}}^{\lambda_n}$. Then for* $0 < \eta < 1$ *and n satisfying* (34) *the following bounds hold with probability at least* $1-\eta$:

$$R(f_{\mathbf{z}})-R(f_\rho)\leqslant 4c_\alpha\left((C_1+C_2)\phi(\Theta^{-1}(n^{-\frac{1}{2}}))\sqrt{\Theta^{-1}(n^{-\frac{1}{2}})}\log\frac{4}{\eta}\right)^{\frac{2}{2-\alpha}},$$

$$\left\|\operatorname{sign} f_{\mathbf{z}}-\operatorname{sign} f_\rho\right\|_\rho\leqslant 4c_\alpha\left((C_1+C_2)\phi(\Theta^{-1}(n^{-\frac{1}{2}}))\sqrt{\Theta^{-1}(n^{-\frac{1}{2}})}\log\frac{4}{\eta}\right)^{\frac{\alpha}{2-\alpha}},$$

with C_1,C_2,C_3 *and* C_4 *given in Theorem* 10.

Corollary 17 shows that for polynomial source conditions this means all the proposed algorithms achieve risk bounds on $R(f_{\mathbf{z}})-R(f_\rho)$ of order $n^{-\frac{2r}{(2r+1)(2-\alpha)}}$ if n is big enough (satisfying (35)). In other words the algorithms we propose are Bayes consistent with fast rates of convergence.

3.4. Probabilistic estimates

In our setting the perturbation measure due to random sampling are expressed by the quantities $\left\|T_{\mathbf{x}}f^\dagger_{\mathcal{H}}-S^*_{\mathbf{x}}\mathbf{y}\right\|_{\mathcal{H}}$ and $\|T-T_{\mathbf{x}}\|_{\mathcal{L}(\mathcal{H})}$ which are clearly random variables. Lemma 9 gives suitable probabilistic estimates. Its proof is trivially obtained by the following propositions.

Proposition 21. *If Assumption* (1) *holds then for all* $n\in\mathbb{N}$ *and* $0<\eta<1$

$$P\left[\left\|T_{\mathbf{x}}f^\dagger_{\mathcal{H}}-S^*_{\mathbf{x}}\mathbf{y}\right\|_{\mathcal{H}}\leqslant 2\left(\frac{\kappa M}{n}+\frac{\kappa\Sigma}{\sqrt{n}}\right)\log\frac{2}{\eta}\right]\geqslant 1-\eta.$$

Proposition 22. *Recalling* $\kappa=\sup_{x\in X}\|K_x\|_{\mathcal{H}}$, *we have for all* $n\in\mathbb{N}$ *and* $0<\eta<1$,

$$P\left[\|T-T_{\mathbf{x}}\|\leqslant\frac{1}{\sqrt{n}}2\sqrt{2}\kappa^2\sqrt{\log\frac{2}{\eta}}\right]\geqslant 1-\eta.$$

The latter proposition was proved in De Vito et al. [13]. The proof of the first estimate is a simple application of the following concentration result for Hilbert space valued random variable used in Caponnetto and De Vito [7] and based on the results in Pinelis and Sakhanenko [23].

Proposition 23. *Let* $(\Omega,\mathcal{B},P)$ *be a probability space and* ξ *a random variable on* Ω *with values in a real separable Hilbert space* $\mathcal{K}$. *Assume there are two constants* H, σ *such that*

$$\mathbb{E}\left[\|\xi-\mathbb{E}[\xi]\|^m_{\mathcal{K}}\right]\leqslant\tfrac{1}{2}m!\sigma^2H^{m-2}\quad\forall m\geqslant 2 \tag{37}$$

then, for all $n\in\mathbb{N}$ *and* $0<\eta<1$,

$$P\left[\|\xi-\mathbb{E}[\xi]\|_{\mathcal{K}}\leqslant 2\left(\frac{H}{n}+\frac{\sigma}{\sqrt{n}}\right)\log\frac{2}{\eta}\right]\geqslant 1-\eta.$$

We can now give the proof of Proposition 21.

Proof. We consider the random variable $\xi:Z\to\mathcal{H}$ defined by

$$\xi=K_x(y-f^\dagger_{\mathcal{H}}(x))$$

with values in the reproducing kernel Hilbert space $\mathcal{H}$. It easy to prove that ξ is a zero mean random variable, in fact

$$\begin{aligned}\mathbb{E}\left[\xi\right] &= \int_{X\times Y} K_x y - K_x \left\langle f_{\mathcal{H}}^{\dagger}, K_x\right\rangle_{\mathcal{H}} d\rho(x,y) \\ &= \int_X d\rho_X(x) K_x \left(\int_Y y\, d\rho(y|x)\right) - \int_X \left\langle f_{\mathcal{H}}^{\dagger}, K_x\right\rangle_{\mathcal{H}} K_x\, d\rho_X(x) \\ &= I_K^* f_\rho - T f_{\mathcal{H}}^{\dagger}.\end{aligned}$$

Recalling (5) we see a standard results in the theory of inverse problems ensures that $T f_{\mathcal{H}}^{\dagger} = I_K^* f_\rho$ (see [16, Theorem 2.6]) so that the above mean is zero. Moreover, Assumption (1) ensures (see, for example, [31])

$$\int_Y (y - f_{\mathcal{H}}^{\dagger}(x))^m (d\rho(y|x)) \leqslant \frac{1}{2} m! \Sigma^2 M^{m-2} \quad \forall m \geqslant 2$$

so that

$$\begin{aligned}\mathbb{E}\left[\|\xi\|_{\mathcal{H}}^m\right] &= \int_{X\times Y} \left(\left\langle K_x(y - f_{\mathcal{H}}^{\dagger}(x)), K_x(y - f_{\mathcal{H}}^{\dagger}(x))\right\rangle_{\mathcal{H}}\right)^{\frac{m}{2}} d\rho(x,y) \\ &= \int_X d\rho_X(x) K(x,x) \int_Y (y - f_{\mathcal{H}}^{\dagger}(x))^2 (d\rho(y|x)) \\ &\leqslant \kappa^m \frac{1}{2} m! \Sigma^2 M^{m-2} \leqslant \frac{1}{2} m! (\kappa\Sigma)^2 (\kappa M)^{m-2}.\end{aligned}$$

The proof follows applying Proposition 23 with $H = \kappa M$ and $\sigma = \kappa\Sigma$. □

Acknowledgments

The third author is grateful to A. Caponnetto and E. De Vito for useful discussions and suggestions. This work was partially written when the first and third author visited Johann Radon Institute for Computational and Applied Mathematics (RICAM) within the framework of Radon Semester 2005. The support of RICAM is gratefully acknowledged. The first author is financed by the Graduiertenkolleg 1023, University of Göttingen and the third author is funded by the FIRB Project ASTAA and the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778.

References

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

[2] P.L. Bartlett, M.J. Jordan, J.D. McAuliffe, Convexity, classification, and risk bounds, Technical Report 638, Department of Statistics, U.C. Berkeley, 2003.

[3] M.S. Birman, M. Solomyak, Double operators integrals in Hilbert spaces, Integral Equations Oper. Theory (2003) 131–168.

[4] N. Bissantz, T. Hohage, A, Munk, F. Ruymgaart, Convergence rates of general regularization methods for statistical inverse problems and applications, preprint, 2006.

[5] O. Bousquet, S. Boucheron, G. Lugosi, Introduction to Statistical Learning Theory, Lectures Notes in Artificial Intelligence, vol. 3176, Springer, Heidelberg, Germany, Wiley-Interscience Publication, 2004, pp. 169–207.

[6] O. Bousquet, S. Boucheron, G. Lugosi, Theory of classification: a survey of recent advances, ESAIM Probab. Statist. (2004), to appear.

[7] A. Caponnetto, E. De Vito, Fast rates for regularized least-squares algorithm, Technical Report CBCL Paper 248/AI Memo 2005-033, Massachusetts Institute of Technology, Cambridge, MA, 2005.

[8] A. Caponnetto, E. De Vito, Optimal rates for regularized least-squares algorithm, Foundation Comput. Math. (2005), accepted for publication.

[9] A. Caponnetto, L. Rosasco, E. De Vito, A. Verri, Empirical effective dimensions and fast rates for regularized least-squares algorithm, Technical Report CBCL Paper 252/AI Memo 2005-019, Massachusetts Institute of Technology, Cambridge, MA, 2005.

[10] C. Carmeli, E. De Vito, A. Toigo, Reproducing kernel Hilbert spaces and mercer theorem, eprint arXiv: math/0504071. Available at ⟨http://arxiv.org⟩, 2005.

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

[12] E. De Vito, L. Rosasco, A. Caponnetto, Discretization error analysis for tikhonov regularization, Anal. Appl. (2005), to appear.

[13] E. De Vito, L. Rosasco, A. Caponnetto, U. De Giovannini, F. Odone, Learning from examples as an inverse problem, J. Mach. Learning Res. 6 (2005) 883–904.

[14] R. DeVore, G. Kerkyacharian, D. Picard, V. Temlyakov, On mathematical methods of learning, Technical Report 2004:10, Industrial Mathematics Institute, Department of Mathematics University of South Carolina, retrievable at ⟨http://www.math/sc/edu/imip/04papers/0410.ps⟩, 2004.

[15] L. Devroye, L. Györfi, G. Lugosi, A Probabilistic Theory of Pattern Recognition, Applications of Mathematics, vol. 31, Springer, New York, 1996.

[16] H.W. Engl, M. Hanke, A. Neubauer, Regularization of Inverse Problems, Mathematics and its Applications, vol. 375, Kluwer Academic Publishers Group, Dordrecht, 1996.

[17] T. Evgeniou, M. Pontil, T. Poggio, Regularization networks and support vector machines, Adv. Comp. Math. 13 (2000) 1–50.

[18] L. Györfi, M. Kohler, A. Krzyzak, H. Walk, A Distribution-free Theory of Non-parametric Regression, Springer Series in Statistics, New York, 1996.

[19] F. Hansen, Operator inequalities associated to Jensen's inequality, survey of "Classical Inequalities", pp. 67–i98.

[20] P. Mathé, S. Pereverzev, Moduli of continuity for operator monotone functions, Numer. Funct. Anal. Optim. 23 (2002) 623–631.

[21] P. Mathé, S. Pereverzev, Geometry of linear ill-posed problems in variable Hilbert scale, Inverse Problems 19 (2003) 789–803.

[22] P. Mathé, S. Pereverzev, Regularization of some linear ill-posed problems with discretized random noisy data, Math. Comput. (2005), accepted for publication.

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

[24] L. Rosasco, E. De Vito, A. Verri, Spectral methods for regularization in learning theory, Technical Report DISI-TR-05-18, DISI, Universitá degli Studi di Genova, Italy, retrievable at ⟨http://www.disi.unige.it/person/RosascoL⟩, 2005.

[25] W. Rudin, Functional Analysis, International Series in Pure and Applied Mathematics, McGraw-Hill, Princeton, 1991.

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

[27] S. Smale, D. Zhou, Learning theory estimates via integral operators and their approximations, submitted for publication, retrievable at ⟨http://www.tti-c.org/smale.html⟩, 2005.

[28] S. Smale, D. Zhou, Shannon sampling ii: connections to learning theory, retrievable at ⟨http://www.tti-c.org/smale.html⟩, 2005, to appear.

[29] A.B. Tsybakov, Optimal aggregation of classifiers in statistical learning, Ann. Statist. 32 (2004) 135–166.

[30] S.A. van de Geer, Empirical Process in M-Estimation, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge, 2000.

[31] S.W. van de Vaart, J.A. Wellner, Weak Convergence and Empirical Process Theory, Springer Series in Statistics, New York, 1996, Springer, New York, 1996.

[32] V.N. Vapnik, Statistical learning theory, Adaptive and Learning Systems for Signal Processing, Communications, and Control, Wiley, New York, a Wiley-Interscience Publication, 1998.

[33] Y. Yao, L. Rosasco, A. Caponnetto, On early stopping in gradient descent learning, Constructive Approximation (2005), submitted for publication.