Analysis and Applications

Discretization Error Analysis for Tikhonov Regularization

Ernesto De Vito

Dipartimento di Matematica, Università di Modena, Via Campi 213/B, 41100 Modena, Italy and I.N.F.N., Sezione di Genova, Via Dodecaneso 33, 16146 Genova, Italy
devito@unimo.it

Lorenzo Rosasco

DISI, Università di Genova Via Dodecaneso 35, 16146 Genova, Italy, and C.B.C.L., Massachusetts Institute of Technology, Bldg. E25-206, 45 Carleton St., Cambridge, MA 02142.
rosasco@disi.unige.it

Andrea Caponnetto

C.B.C.L., Massachusetts Institute of Technology,5Y Bldg. E25-206, 45 Carleton St., Cambridge, MA 02142 and DISI, Università di Genova Via Dodecaneso 35, 16146 Genova, Italy.
caponnet@mit.edu

Received (Day Month Year)
Revised (Day Month Year)

We study the discretization of inverse problems defined by a Carleman operator. In particular we develop a discretization strategy for this class of inverse problems and we give a convergence analysis. Learning from examples as well as the discretization of integral equations can be analysed in our setting.

Keywords: learning from examples; inverse problems, Tikhonov regularization.

Mathematics Subject Classification 2000: 68T05, 68P30

1. Introduction

Let A be a bounded linear operator from a Hilbert space $\mathcal{H}$ into a Hilbert space $\mathcal{G}$ and consider the inverse problem associated with

$$Af = g, \tag{1.1}$$

that is, the problem of finding $f \in \mathcal{H}$ given the datum $g \in \mathcal{G}$ and the model $A : \mathcal{H} \to \mathcal{G}$. Usually the above problem is *ill-posed* [1,2,3] and we can only look for the minimal norm solution of the least-squares problem

$$\min_{f \in \mathcal{H}} \|Af - g\|_{\mathcal{G}}, \tag{1.2}$$

which is called the generalized solution $f^\dagger$. In general g and A are known up to noise, $\|g_{\delta_1} - g\|_{\mathcal{H}} \leq \delta_1$ and $\|A_{\delta_2} - A\|_{\mathcal{L}(\mathcal{H})} \leq \delta_2$, and $f^\dagger$ does not depend continuously on

g and A, so that a regularization procedure is needed to find a stable solution. For example, Tikhonov regularization replaces problem (1.2) with

$$\min_{f\in\mathcal{H}} \left(\|A_{\delta_2} f - g_{\delta_1}\|_{\mathcal{G}}^2 + \lambda \|f\|_{\mathcal{H}}^2 \right). \tag{1.3}$$

The solution of the above problem is now stable with respect to perturbations due to noise [30].

In practice to solve numerically problem (1.3) a suitable discretization is considered and problem (1.1) is replaced with

$$Bf = h \quad B : \mathcal{H} \to \mathcal{Z}, \tag{1.4}$$

where $\mathcal{Z}$ is a finite dimensional subspace of $\mathcal{G}$, B is a bounded linear operator and h is an element of $\mathcal{Z}$. Examples of discretization procedures are degenerate kernel methods, quadrature methods and projection methods (for a review see [3,4,5,2, 6] and references therein). For these methods the convergence of the regularized solution of the discretized problem to the generalized solution of problem (1.1) is controlled by δ_1, δ_2 and the dimension of $\mathcal{Z}$.

In this paper, we develop a framework to deal simultaneously with the perturbation due both to the noise and to the discretization. To this end we directly compare problem (1.1) and problem (1.4) where $\mathcal{Z}$ is not necessarily a subspace of $\mathcal{G}$. Following [7,8], we study

$$B_\delta f = g_\delta$$

regarding the datum g_δ as a perturbation of the exact datum g and the operator B_δ as a perturbation of the exact model A. The critical point in our setting is to give a measure of the discrepancy between g_δ and g, and between B_δ and A, since in general they belong to different spaces. We suggest that the perturbation can be controlled by $\|B_\delta^* g_\delta - A^* g\|_{\mathcal{H}} \leq \delta_1$ and $\|B_\delta^* B_\delta - A^* A\|_{\mathcal{L}(\mathcal{H})} \leq \delta_2$, $\delta = (\delta_1, \delta_2)$. This can be seen, for example, observing that the Tikhonov regularized solution of $B_\delta f = g_\delta$ is

$$f_\delta^\lambda = (B_\delta^* B_\delta + \lambda)^{-1} B_\delta^* g_\delta,$$

so that f_δ^λ depends on $B_\delta^* B_\delta$, which is an operator from $\mathcal{H}$ to $\mathcal{H}$, and on $B_\delta^* g_\delta$, which is an element of $\mathcal{H}$. We note that the output space $\mathcal{Z}$ disappears. We stress that in our approach δ_1 and δ_2 take care of both the noise and the discretization. Our setting appears natural while considering the problem of learning from examples where one has to deal with the stochastic discretization of a linear inverse problem [34]. Nonetheless, we can equally deal with the deterministic discretization of integral equations [6].

The paper is organized as follows. In Section 2 we discuss the main example we have in mind, i.e. learning from examples. In particular we recall a recently proposed formalization of learning as an inverse problem. In Section 3 we develop the general setting and give the main results. In Section 4 we specialize to linear problems induced by a Carleman operator giving an unifying framework for both integral

equations and approximation problems in reproducing kernel Hilbert spaces. In particular we provide an estimate of the perturbation δ in two different settings. In Section 4.2 the discrete data are deterministically given. As a simple example we consider the problem of computing the derivative of a function g when a finite set of samples $y_i = g(x_i)$ is given. In Section 4.3 we come back to learning theory considering the discrete data as random variables and obtaining a probabilistic bound on the discrete regularized solution.

2. An Inverse Problem Perspective on Learning Theory

In order to motivate the need to extend the discretization procedure usually considered in the theory of inverse problems to the scheme discussed in the introduction, we give a brief account of the theory of learning from examples. For sake of clarity we consider only the regularized least-squares algorithm in the regression setting with quadratic loss function.

The theory of learning from examples was developed in the last two decades as a mathematical model for learning in brain and cognitive science (for an account of learning theory and its applications see [9,10,11,12] and references therein). As noticed by many authors, the theory is strongly related with function approximation [13,14] and nonparametric regression [15]. It was recently shown [34] that the problem of learning can also be reformulated as an inverse problem. In this section, we review this connection.

The following ingredients define the mathematical setting of learning theory where we focus on the regularized least-squares algorithm, see [13,19,20,17] and references therein.

- The **sample space** $Z = X \times Y$ where the input space X is a closed subset of $\mathbb{R}^d$ and the output space Y is a bounded subset of $\mathbb{R}$.
- The **probability distribution** $\rho(x, y) = \rho(y|x)\rho_X(x)$ on the sample space Z. The measure ρ is fixed but unknown. We denote by $L^2(X, \rho_X)$ the Hilbert space of functions $f : X \to \mathbb{R}$, which are square integrable with respect to ρ_X and by $\|f\|_\rho$ the corresponding norm.
- The **regression function**

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

  (the integral is finite and f_ρ is in $L^2(X, \rho_X)$ since Y is bounded).
- The **hypothesis space** $\mathcal{H}$, which is a (separable) reproducing kernel Hilbert space [21] with a measurable kernel $K : X \times X \to \mathbb{R}$ bounded by

$$K(x, x) \leq \kappa^2. \tag{2.1}$$

- The **training sets** $\mathbf{z} = \{(x_1, y_1), \cdots, (x_n, y_n)\} \in Z^n$ where the n examples $(x_i, y_i) \in Z$ are drawn *i.i.d.* according to ρ (that is, Z^n is endowed with the probability distribution ρ^n).

- The **regularized least-squares algorithm**. Given $n \in \mathbb{N}$, for all training sets $\mathbf{z} \in Z^n$, the estimator $f_{\mathbf{z},n}$ is defined as

$$f_{\mathbf{z},n} = \operatorname{argmin}_{f \in \mathcal{H}} \left(\frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda_{\mathbf{z},n} \|f\|_{\mathcal{H}}^2 \right).$$

  where $\lambda_{\mathbf{z},n} > 0$ is the regularization parameter depending on n and $\mathbf{z}$.

We recall [21] that the elements of $\mathcal{H}$ are functions $f : X \to \mathbb{R}$ such that

$$f(x) = \langle f, K_x \rangle_{\mathcal{H}} \qquad x \in X,\ f \in \mathcal{H}, \tag{2.2}$$

where $K_x \in \mathcal{H}$ is the function $K_x(t) = K(t, x)$. Moreover, since the above functional is coercive and strictly convex the solution $f_{\mathbf{z},n}$ exists and is unique [22]. Clearly $f_{\mathbf{z},n}$ is a random variable on Z^n taking values in $\mathcal{H}$. One of the goals of learning theory is the study of the *generalization properties* of the algorithm when the number n of examples increases. Working with the squared loss, this amounts to give a probabilistic upper bound on

$$\mathbb{P}_{\mathbf{z} \sim \rho^n} \left[\|f_{\mathbf{z},n} - f_\rho\|_\rho^2 \geq \inf_{f \in \mathcal{H}} \|f - f_\rho\|_\rho^2 + \epsilon \right]$$

for all $\epsilon > 0$ [11,15,17].

We now rewrite the above problem in the framework given in the introduction, for a wider discussion see [34].
If $I_K : \mathcal{H} \to L^2(X, \rho_X)$ is the inclusion operator, which is continuous by (2.1), the least-squares problem associated with the linear problem

$$I_K f = f_\rho \tag{2.3}$$

is

$$\inf_{f \in \mathcal{H}} \|I_K f - f_\rho\|_\rho^2 = \inf_{f \in \mathcal{H}} \|f - f_\rho\|_\rho^2,$$

so that the problem of finding the best estimator in $\mathcal{H}$ is equivalent to solving problem (2.3) in the least-squares sense. In particular, the generalized solution $f^\dagger$ exists if and only if $P f_\rho \in \mathcal{H}$ where P is the projection onto the closure of $\mathcal{H}$ in $L^2(X, \rho_X)$ (in learning theory $f^\dagger$ is usually denoted by $f_{\mathcal{H}}$ [10].) Moreover, the definition of projection P gives

$$\|f - f_\rho\|_\rho^2 - \inf_{f \in \mathcal{H}} \|f - f_\rho\|_\rho^2 = \|I_K f - P f_\rho\|_\rho^2 \tag{2.4}$$

for all $f \in \mathcal{H}$, which is the square of the residual of f in the framework of inverse problems [2].
Finally, if $\mathbf{z} = (\mathbf{x}, \mathbf{y})$ with $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$, let $S_{\mathbf{x},n} : \mathcal{H} \to \mathbb{R}^n$ be the sampling operator [23,17]

$$(S_{\mathbf{x},n} f)_i = f(x_i) = \langle f, K_{x_i} \rangle_{\mathcal{H}} \qquad i = 1, \ldots, n.$$

where $\|\cdot\|_n$ is $1/n$ times the euclidean norm in $\mathbb{R}^n$. Then

$$\frac{1}{n}\sum_{i=1}^{n}(f(x_i)-y_i)^2+\lambda\|f\|_{\mathcal{H}}^2=\|S_{\mathbf{x},n}f-\mathbf{y}\|_n^2+\lambda\|f\|_{\mathcal{H}}^2,$$

and the estimator $f_{\mathbf{z},n}$ is the Tikhonov regularization of the linear inverse problem

$$S_{\mathbf{x},n}f=\mathbf{y} \tag{2.5}$$

with the choice $\lambda=\lambda_{\mathbf{z},n}$.

The problem of learning can be seen as the problem of solving the exact problem (2.3) (in the least-squares sense) when a discretized problem $S_{\mathbf{x},n}f=\mathbf{y}$ is randomly given. As suggested in the introduction, the convergence of the regularized solution of the discretized problem can be controlled by

$$\left\|I_K^*I_K-S_{\mathbf{x},n}^*S_{\mathbf{x},n}\right\|_{\mathcal{L}(\mathcal{H})}\leq\delta_1\qquad\|I_K^*f_\rho-S_{\mathbf{x},n}\mathbf{y}\|_{\mathcal{H}}\leq\delta_2.$$

Clearly, since both $S_{\mathbf{x},n}$ and $\mathbf{y}$ are random variables, the above bounds are to be considered in a probabilistic sense. As a consequence of the theory we develop in the following section, we will prove in Section 4.3 that, for a suitable a priori choice of the parameter $\lambda=\lambda_n$,

$$\mathbb{P}_{\mathbf{z}\sim\rho^n}\left[\|f_{\mathbf{z},n}-Pf_\rho\|_\rho^2>\inf_{f\in\mathcal{H}}\|f-f_\rho\|_\rho^2+(C\epsilon+R)^2n^{-\frac{r}{r+1}}\right]\leq 2\left(e^{-C_1\epsilon^2}+e^{-C_1\epsilon^2}\right)$$

where $0<r\leq 1$ and C, C_1, C_2, R are constants.

We end the section, observing that, the basic goal of learning is to control $\|I_Kf_{\mathbf{z},n}-Pf_\rho\|_\rho^2$, which is the square of the residual of the solution. However, if $f^\dagger=f_{\mathcal{H}}$ exists, it is also of interest [17] to bound the reconstruction error $\|f_{\mathbf{z},n}-f_{\mathcal{H}}\|_{\mathcal{H}}$, which is standard both in inverse problems and in approximation theory. In the following we treat both errors.

3. Error estimates for Tikhonov regularization

In this section, first we describe and briefly discuss the general setting, then we state and prove the main results of the paper.

3.1. *General Setting*

First of all we set the notation. If $\mathcal{H}$ and $\mathcal{G}$ are Hilbert spaces, we denote by $\mathcal{L}(\mathcal{H},\mathcal{G})$ the Banach space of bounded linear operators from $\mathcal{H}$ into $\mathcal{G}$ endowed with the uniform norm $\|\cdot\|_{\mathcal{L}(\mathcal{H},\mathcal{G})}$. If $A\in\mathcal{L}(\mathcal{H},\mathcal{G})$ we denote by A^* the adjoint operator.

We consider the two linear problems

$$Af=g\qquad\qquad A:\mathcal{H}\to\mathcal{G} \tag{3.1}$$

and

$$B_\delta f=g_\delta\qquad\qquad B_\delta:\mathcal{H}\to\mathcal{Z} \tag{3.2}$$

where $\mathcal{H}$, $\mathcal{G}$, $\mathcal{Z}$ are Hilbert spaces, and A, B_δ are bounded operators. We will think of problem (3.2) as a discretization of problem (3.1), meaning that we regard B_δ and g_δ as approximations of A and g, respectively. Henceforth we assume that $M > 0$ and $\delta = (\delta_1, \delta_2) \in \mathbb{R}^2_+$ exist such that

$$\begin{aligned} \|g_\delta\|_{\mathcal{Z}} &\leq M \\ \|B_\delta^* g_\delta - A^* g\|_{\mathcal{H}} &\leq \delta_1 \\ \|B_\delta^* B_\delta - A^* A\|_{\mathcal{L}(\mathcal{H})} &\leq \delta_2. \end{aligned} \tag{3.3}$$

The constant M is regarded as a fixed a priori normalization of the norm in $\mathcal{Z}$, whereas the parameter $\delta = (\delta_1, \delta_2)$ gives a quantitative measure of the perturbation introduced by replacing the problem (3.1) with (3.2) and we are interested to the convergence of the regularized solutions of the discrete problem (3.2) when δ goes to zero.

In the following we focus on Tikhonov regularization, so that we consider the regularized solution

$$f_\delta^\lambda = (B_\delta^* B_\delta + \lambda I)^{-1} B_\delta^* g_\delta \tag{3.4}$$

where $\lambda > 0$. The regularization parameter $\lambda = \lambda_\delta$ is a suitable function of δ such that $f_\delta^{\lambda_\delta}$ converges to the least-squares solution of problem (3.1) as δ goes to 0. More precisely, to study the convergence of $f_\delta^{\lambda_\delta}$, we consider both the reconstruction error $\left\| f_\delta^{\lambda_\delta} - f^\dagger \right\|_{\mathcal{H}}$ and the residual $\left\| A f_\delta^{\lambda_\delta} - Pg \right\|_{\mathcal{G}}$ where $f^\dagger$ is the generalised solution of problem (3.1) and P is the projection on the closure of $Im(A)$ in $\mathcal{G}$. It is well known that $f^\dagger$ exists if and only if $Pg \in Im(A)$ and that $Af^\dagger = Pg$; however, as we study the residual we don't need to assume the existence of $f^\dagger$.

3.2. *Error Estimates*

In this section we give error estimates for both the residual and the reconstruction error in the setting presented in the previous section. To state our bounds, we need some kind of a priori information on the exact datum. Given $r > 0$ and $R > 0$, we let

$$\Omega_{r,R} = \{ g \in \mathcal{G} \,|\, Pg = (AA^*)^r \phi, \|\phi\|_{\mathcal{G}} \leq R \}, \tag{3.5}$$

whose role will be clarified by Prop. 3.2. Our main result is the following theorem.

Theorem 3.1. *Assume (3.3). If $g \in \Omega_{r,R}$ with $0 < r \leq 1$, then*

$$\left\| A f_\delta^\lambda - Pg \right\|_{\mathcal{G}} \leq \frac{M}{4\lambda} \delta_2 + \frac{1}{\sqrt{\lambda}} \delta_1 + R\lambda^r. \tag{3.6}$$

If $g \in \Omega_{r,R}$ with $\frac{1}{2} < r \leq \frac{3}{2}$, $f^\dagger$ exists and

$$\left\| f_\delta^\lambda - f^\dagger \right\|_{\mathcal{H}} \leq \frac{M}{2\lambda^{\frac{3}{2}}} \delta_2 + \frac{1}{\lambda} \delta_1 + R\lambda^{r - \frac{1}{2}}. \tag{3.7}$$

Before proving the theorem, notice that both in (3.6) and in (3.7) the last term in the bound depends only on λ, but not on the noise δ. On the other hand in the first two terms the dependence on δ and on λ is factorized. Moreover condition (3.5) on g affects only the third term in the bound.

Up to our knowledge, the first result similar to the above theorem was obtained in [7, Th. 2] in the framework of integral equations with white noise. In a deterministic setting, [8, Th. 1] gives a convergence analysis for integral equations assuming that $B_\delta^* B_\delta$ is a degenerate kernel and $g_\delta = g$. [24, Th. 3.1] and [25, Th. 2] consider a wider class of regularization methods, but B_δ has the form $Q_n A P_n$, where Q_n and P_n are orthogonal projections. Our bound is of the same kind of the estimates obtained in [26, Th. 2.2], [27, Th. 2.1] and [28, Th. 2.5]. Anyway in the above papers only the reconstruction error is studied and different estimates of the perturbation are considered.

To prove the theorem we let

$$f^\lambda = (A^*A + \lambda I)^{-1} A^* g \tag{3.8}$$

be the regularized solution of problem (3.1), so that the following decompositions can be considered

$$\begin{aligned} Af_\delta^\lambda - Pg &= A(f_\delta^\lambda - f^\lambda) + (Af^\lambda - Pg) \\ f_\delta^\lambda - f^\dagger &= (f_\delta^\lambda - f^\lambda) + (f^\lambda - f^\dagger). \end{aligned} \tag{3.9}$$

The following proposition gives a bound of the first term in the above decompositions, whereas Prop. 3.2 estimates the second term.

Proposition 3.1. *Assume (3.3). For any $\lambda > 0$, the following inequalities hold*

$$\left\| A(f_\delta^\lambda - f^\lambda) \right\|_{\mathcal{G}} \leq \frac{M}{4\lambda}\delta_2 + \frac{1}{2\sqrt{\lambda}}\delta_1 \tag{3.10}$$

$$\left\| f_\delta^\lambda - f^\lambda \right\|_{\mathcal{H}} \leq \frac{M}{2\lambda^{\frac{3}{2}}}\delta_2 + \frac{1}{\lambda}\delta_1. \tag{3.11}$$

Proof. To treat both the reconstruction error and the residual of the solution, we introduce a parameter $a \in [0,1]$ and we let

$$C_a = \begin{cases} 1 & a = 0,\ a = 1 \\ a^a (1-a)^{(1-a)} & 0 < a < 1 \end{cases}. \tag{3.12}$$

Moreover, we let $T = A^*A$, $T_\delta = B_\delta^* B_\delta$, $\phi = A^* g$ and $\phi_\delta = B_\delta^* g_\delta$, (3.3) ensures that

$$\|T - T_\delta\|_{\mathcal{L}(\mathcal{H})} \leq \delta_2 \qquad \|\phi - \phi_\delta\|_{\mathcal{H}} \leq \delta_1. \tag{3.13}$$

For all $a \in [0,1]$ the spectral theorem gives

$$\left\| T_\delta^a (T_\delta + \lambda)^{-1} \right\|_{\mathcal{L}(\mathcal{H})} \leq \frac{C_a}{\lambda^{1-a}}. \tag{3.14}$$

Moreover, the polar decomposition of B_δ yields $B_\delta = UT_\delta^{\frac{1}{2}}$, where U is a partial isometry from $\mathcal{H}$ to $\mathcal{Z}$, so that $\|U\|_{\mathcal{L}(\mathcal{H},\mathcal{Z})} = 1$. Since $T_\delta^{\frac{1}{2}}$ commutes with $(T_\delta + \lambda)^{-1}$

$$\left\|(T_\delta + \lambda)^{-1} B_\delta^*\right\|_{\mathcal{L}(\mathcal{Z},\mathcal{H})} = \left\|T_\delta^{\frac{1}{2}} (T_\delta + \lambda)^{-1} U^*\right\|_{\mathcal{L}(\mathcal{Z},\mathcal{H})},$$

and Equation (3.14) with $a = \frac{1}{2}$ implies

$$\left\|(T_\delta + \lambda)^{-1} B_\delta^*\right\|_{\mathcal{L}(\mathcal{Z},\mathcal{H})} \leq \frac{1}{2\sqrt{\lambda}} . \tag{3.15}$$

Finally, the definitions of f_δ^λ and f^λ give

$$\begin{aligned} f_\delta^\lambda - f^\lambda &= (T_\delta + \lambda)^{-1}\phi_\delta - (T + \lambda)^{-1}\phi \\ &= [(T_\delta + \lambda)^{-1} - (T + \lambda)^{-1}]\phi_\delta + (T + \lambda)^{-1}(\phi_\delta - \phi) . \end{aligned}$$

The known algebraic identity

$$(T_\delta + \lambda)^{-1} - (T + \lambda)^{-1} = (T + \lambda)^{-1}(T - T_\delta)(T_\delta + \lambda)^{-1}$$

and triangle inequality ensure

$$\begin{aligned} \left\|T^a(f^\lambda - f_0^\lambda)\right\|_{\mathcal{H}} &\leq \left\|T^a(T + \lambda)^{-1}(T - T_\delta)(T_\delta + \lambda)^{-1} B_\delta^* g_\delta\right\|_{\mathcal{H}} \\ &\quad + \left\|T^a(T + \lambda)^{-1}(\phi_\delta - \phi)\right\|_{\mathcal{H}} \\ &\leq \frac{C_a}{\lambda^{1-a}} \|T_\delta - T\|_{\mathcal{L}(\mathcal{H})} \frac{\|g_\delta\|_{\mathcal{Z}}}{2\sqrt{\lambda}} + \frac{C_a}{\lambda^{1-a}} \|\phi_\delta - \phi\|_{\mathcal{H}} \\ &\leq \frac{C_a}{\lambda^{1-a}} \delta_2 \frac{M}{2\sqrt{\lambda}} + \frac{C_a}{\lambda^{1-a}} \delta_2 \end{aligned}$$

by (3.13). Bound (3.11) is clear choosing $a = 0$, whereas the bound (3.10) follows choosing $a = \frac{1}{2}$ and using the polar decomposition of $A = WT^{\frac{1}{2}}$, which gives

$$\|Af\|_{\mathcal{G}} = \left\|T^{\frac{1}{2}} f\right\|_{\mathcal{H}} . \qquad \square$$

We now study the convergence of the second term in (3.9). The definition of regularization scheme ensures that both Af^λ and f^λ converge to Pg and $f^\dagger$, respectively, if λ goes to zero [2]. However, to have an explicit estimate of the error, we need some suitable a priori assumptions on the exact datum g or on $f^\dagger$. Such assumptions are usually referred to as source conditions in the inverse problem literature. For example a standard result [3,2] shows that, if $f^\dagger \in \operatorname{Im}(A^*A)^r A^*$, then $\left\|f^\lambda - f^\dagger\right\| = O(\lambda^r)$. The definition (3.5) is a slightly modification of the above source condition, which provides the desired error estimates.

Proposition 3.2. *If $g \in \Omega_{r,R}$ with $0 < r \leq 1$ then*

$$\left\|Af^\lambda - Pg\right\|_{\mathcal{G}} \leq R\lambda^r$$

for $0 < r \leq 1$.

If $g \in \Omega_{r,R}$ with $\frac{1}{2} < r \leq \frac{3}{2}$, then $f^\dagger$ exists and

$$\left\|f^\lambda - f^\dagger\right\|_{\mathcal{H}} \leq R\lambda^{r-\frac{1}{2}}$$

for $1/2 < r \leq 3/2$.

Proof. The proof is standard [2]. Let $0 < r \leq 1$, since t^r is a concave function

$$\frac{\lambda}{\lambda+\sigma}\sigma^r \leq \lambda^r \qquad \lambda,\ \sigma > 0. \tag{3.16}$$

The assumption $g \in \Omega_{r,R}$ ensures

$$Pg - Af^\lambda = (I - A(A^*A + \lambda I)^{-1}A^*)Pg = (I - AA^*(AA^* + \lambda I)^{-1})(AA^*)^r\phi.$$

The spectral theorem with (3.16) gives $\left\|Af^\lambda - Pg\right\|_{\mathcal{G}} \leq R\lambda^r$.
To prove the second bound, let $A^* = U(AA^*)^{\frac{1}{2}}$ be the polar decomposition of A^*, then

$$Pg = (AA^*)^r = (AA^*)^{1/2}(AA^*)^c\phi = (AA^*)^{1/2}U^*U(AA^*)^cU^*U\phi = A(A^*A)^cU\phi,$$

where $c = r - 1/2 \in]0, 1]$. It follows that $Pg \in \operatorname{Im} A$, so that $f^\dagger$ exists and $f^\dagger = (A^*A)^cU\phi$. Mimicking the proof of the first bound,

$$f^\dagger - f^\lambda = (I - (A^*A + \lambda I)^{-1}A^*A)f^\dagger = (I - (A^*A + \lambda I)^{-1}A^*A)(A^*A)^cU\phi$$

and replacing r by c in (3.16) $\left\|f^\lambda - f^\dagger\right\|_{\mathcal{H}} \leq R\lambda^c$. □

The bounds claimed in Th. 3.1 follow from the above two propositions and Cauchy-Schwarz inequality applied to (3.9).

3.2.1. *Comparison with Results on Inverse Problem with Noisy Operator*

We compare Th. 3.1 with the known results for Tikhonov regularization in the presence of modeling error [30]. To this aim, we consider noisy problems $B_\delta f = g_\delta$, where B_δ is an operator from $\mathcal{H}$ to $\mathcal{G}$ and $g_\delta \in \mathcal{G}$ such that

$$\|g_\delta - g\|_{\mathcal{G}} \leq \eta_1 \qquad \|B_\delta - A\|_{\mathcal{L}(\mathcal{H},\mathcal{G})} \leq \eta_2.$$

In this case it is known [30] that if

$$\lim_{\eta_1,\eta_2\to 0} \frac{(\eta_1+\eta_2)^2}{\lambda(\eta_1,\eta_2)} = 0 \tag{3.17}$$

the regularized solution f_δ^λ approaches $f^\dagger$. Since

$$\|B_\delta^*B_\delta - A^*A\|_{\mathcal{L}(\mathcal{H})} \leq (\|B_\delta\|_{\mathcal{L}(\mathcal{H},\mathcal{G})} + \|A\|_{\mathcal{L}(\mathcal{H},\mathcal{G})}\,\eta_2 \leq C_1\eta_2 = \delta_2$$
$$\|B_\delta^*g_\delta - A^*g\|_{\mathcal{H}} \leq \|g_\delta\|_{\mathcal{G}}\,\eta_2 + \|A\|_{\mathcal{L}(\mathcal{H},\mathcal{G})}\,\eta_1 \leq C_2(\eta_1+\eta_2) = \delta_1$$

it follows that (3.17) is weaker than (3.11).

This observation suggests that the noise can be evaluated by means of $\|U_\delta^*g_\delta - U^*g\|_{\mathcal{H}} \leq \eta_1$ and $\||B_\delta| - |A|\|_{\mathcal{L}(\mathcal{H},\mathcal{G})} \leq \eta_2$, where $A = U|A|$ and $B_\delta =$

$U_\delta|B_\delta|$ are the polar decompositions of A and B_δ, respectively. In fact repeating the standard proof [30] for Tikhonov regularization in the presence of modeling error, we have that, if Condition (3.17) holds, then the regularized solution f_δ^λ approaches $f^\dagger$. However, in the applications it is difficult to evaluate the polar decomposition and, hence, to ensure that the noisy model is an approximation of the exact model.

Finally, we observe that the content of Prop. 3.1 can be regarded as regularization in the presence of modeling error. Indeed, the least-squares solutions of the exact problem $Af = g$ are the solutions of the inverse problem

$$A^*Af = A^*g.$$

This suggests to replace the noisy problem $B_\delta f = g_\delta$ with the problem

$$B_\delta^* B_\delta f = B_\delta^* g_\delta,$$

so that $B_\delta^* g_\delta$ is a noisy approximation of the exact datum A^*g, $B_\delta^* B_\delta$ is the noisy model of the exact model A^*A and the noise is controlled by two quantities

$$\|B_\delta^* B_\delta - A^*A\|_{\mathcal{L}(\mathcal{H})} \leq \delta_1 \qquad \|B_\delta^* g_\delta - A^*g\|_{\mathcal{H}} \leq \delta_2.$$

However, the regularized solution $f_\delta^\lambda = (B_\delta^* B_\delta + \lambda)^{-1} B_\delta^* g_\delta$ is not the Tikhonov regularization of the problem $B_\delta^* B_\delta f = B_\delta^* g_\delta$. Indeed, if $T_\delta = T_\delta^* = B_\delta^* B_\delta$ and $\phi_\delta = B_\delta^* g_\delta$, we have that

$$f_\delta^\lambda = (T_\delta + \lambda)^{-1}\phi_\delta = (T_\delta^* T_\delta + \lambda T_\delta)^{-1} T_\delta^* \phi_\delta,$$

whereas the Tikhonov regularized solution of $T_\delta f = \phi_\delta$ is $(T_\delta^* T_\delta + \lambda)^{-1} T_\delta^* \phi_\delta$.

4. Discretization of Carleman Operators

In this section we discuss how to evaluate conditions (3.3) for inverse problems induced by a Carleman operator [31]. This setting is general enough to cope with learning theory as well as integral equations. After briefly recalling the definition and main properties of Carleman operator we discuss two different discretization settings. The first one is deterministic and is illustrated with the classical problem of differentiating a function. The second setting is stochastic and requires probabilistic estimates of vector valued random variables.

4.1. *Carleman Operators*

In the present section we briefly review the notion of Carleman operator that allows an unifying approach to the theories of reproducing kernel Hilbert spaces and integral equations. Our presentation follows the book of [31], where a clear exposition of the relation between Carleman operators and integral equations is given. The book of [32] is a source for results and bibliography on this topic. In [4,13] and references therein, there is an account of the theory of reproducing kernel Hilbert spaces in the context of inverse problems. Recent results on the Tikhonov regularization in the

framework of reproducing kernel Hilbert space can be found in [33] and references therein.

Let X be a closed subset of $\mathbb{R}^d$ and ρ_X be a finite measure on X, we define $\mathcal{G} = L^2(X, \rho_X)$ as the Hilbert space of functions $f : X \to \mathbb{R}$ square integrable with respect to ρ_X.
Given a (separable) Hilbert space $\mathcal{H}$, let $\gamma : X \to \mathcal{H}$ be a map such that γ is measurable and bounded. We define the Carleman operator $A : \mathcal{H} \to \mathcal{G}$ associated with the map γ [31] by

$$(Af)(x) = \langle f, \gamma_x \rangle_{\mathcal{H}} \qquad \rho_X-\text{almost all } x \in X$$

for all $f \in \mathcal{H}$. A natural way of discretizing the Carleman operator A is the following. Given $n \in \mathbb{N}$, we consider n points $\mathbf{x} = (x_1, \ldots, x_n)$ of X and we define the operator $B_{\mathbf{x},n}$ from $\mathcal{H}$ to $\mathcal{Z} = \mathbb{R}^n$ by

$$(B_{\mathbf{x},n} f)_i = \langle f, \gamma_{x_i} \rangle_{\mathcal{H}} \qquad i = 1, \ldots, n \qquad f \in \mathcal{H},$$

where $\mathcal{Z}$ is endowed with the scalar product

$$\langle \mathbf{y}, \mathbf{y}' \rangle_{\mathcal{Z}} = \sum_{i=1}^{n} a_i y_i y'_i, \tag{4.1}$$

with $a_i \in \mathbb{R}_+$ suitable functions of the sample $\mathbf{x}$.

The following proposition gives the main properties of A and $B_{\mathbf{x},n}$ (for the proof see, for example, [34,35]).

Proposition 4.1. *The operator A is a Hilbert-Schmidt operator from $\mathcal{H}$ into $\mathcal{G}$,*

$$A^*\phi = \int_X \phi(x)\gamma_x \, d\rho_X(x), \tag{4.2}$$

$$A^*A = \int_X \langle \cdot, \gamma_x \rangle_{\mathcal{H}} \, \gamma_x \, d\rho_X(x), \tag{4.3}$$

$$(AA^*\phi)(t) = \int_X \langle \gamma_t, \gamma_x \rangle_{\mathcal{H}} \, \phi(x) \, d\rho_X(x) = (L_\Gamma \phi)(t) \tag{4.4}$$

where $\phi \in \mathcal{G}$, the first integral converges in norm, the second integral in Hilbert-Schmidt norm and L_Γ is the integral operator with kernel $\Gamma(t, x) = \langle \gamma_t, \gamma_x \rangle_{\mathcal{H}}$.

The operator $B_{\mathbf{x},n}$ is a finite rank operator from $\mathcal{H}$ into $\mathcal{Z}$,

$$B^*_{\mathbf{x},n}\mathbf{y} = \sum_{i=1}^{n} a_i y_i \gamma_{x_i} \qquad \forall \ \mathbf{y} \in \mathcal{Z} \tag{4.5}$$

$$B^*_{\mathbf{x},n} B_{\mathbf{x},n} = \sum_{i=1}^{n} a_i \, \langle \cdot, \gamma_{x_i} \rangle_{\mathcal{H}} \, \gamma_{x_i}. \tag{4.6}$$

Let now $g \in L^2(X, \rho_X)$ be the exact datum, then problem (3.1) amounts to find $f \in \mathcal{H}$ such that

$$\langle f, \gamma_x \rangle_{\mathcal{H}} = g(x) \qquad \rho_X-\text{almost all } x \in X.$$

In particular, if the generalized solution $f^\dagger$ exists, it is the minimal norm solution of

$$(Pg)(x) = \left\langle f^\dagger, \gamma_x \right\rangle_{\mathcal{H}} \qquad \rho_X - \text{almost all } x \in X, \tag{4.7}$$

and the condition $g \in \Omega_{r,R}$ becomes $Pg \in \operatorname{Im} L_\Gamma^r$ with $\left\| L_\Gamma^{-r} Pg \right\|_\rho \leq R$. Notice that if γ is weakly continuous and the support of the measure ρ_X is X, then (4.7) holds for all $x \in X$ and $f^\dagger$ is the unique solution of (4.7) that belongs to the closure of the linear span of the set $\{\gamma_x \,|\, x \in X\}$.

The discretized version of g is a vector $\mathbf{y} = (y_1, \ldots, y_n) \in \mathbb{R}^n$, so that problem (3.2) becomes

$$\langle f, \gamma_{x_i} \rangle_{\mathcal{H}} = y_i \qquad i = 1, \ldots, n.$$

Letting $\mathbf{z} = (\mathbf{x}, \mathbf{y})$, according to the notation of Section 3, we denote by

$$f^\lambda = (A^*A + \lambda)^{-1} A^* g$$
$$f^\lambda_{\mathbf{z},n} = (B^*_{\mathbf{x},n} B_{\mathbf{x},n} + \lambda)^{-1} B^*_{\mathbf{x},n} \mathbf{y},$$

the regularized solutions of exact and discrete problems, respectively, where we add the subscript $\mathbf{z} = (\mathbf{x}, \mathbf{y})$ to emphasize the dependence of the solution on $\mathbf{x}$ and $\mathbf{y}$. Moreover, (4.6) and (4.5) gives

$$f^\lambda_{\mathbf{z},n} = \sum_{i,j=1}^n a_j \gamma_{x_j} \left((\Gamma_{\mathbf{x}} + \lambda)^{-1} \right)_{ji} y_i, \tag{4.8}$$

where $\Gamma_{\mathbf{x}}$ is the $n \times n$ matrix $(\Gamma_{\mathbf{x}})_{ij} = \left\langle \gamma_{x_j}, \gamma_{x_i} \right\rangle_{\mathcal{H}}$. In particular, $f_{\mathbf{z},n}$ belongs to the linear span of the set $\{\gamma_{x_i} \,|\, i = 1, \ldots, n\}$ [7,11].

In order to apply the results of Th 3.1, we discuss some reasonable hypotheses on the choice of the sample $\mathbf{x}$ and the noisy datum $\mathbf{y}$. We consider two different settings.

4.2. *Deterministic Discretization*

In this section, we consider a framework where the measure ρ_X is known, the points x_i are given and the values y_i are samples of the datum g without *noise*, that is, $y_i = g(x_i)$. Clearly, this is an ideal framework where the noise is due only to the finite dimensional approximation [23].

Moreover, we study the reconstruction error of the approximated solution. To this aim, we assume that $g \in \operatorname{Im} A$ so that, by (4.7), we can restate the hypothesis that the noise is zero by the fact that

$$y_i = g(x_i) = \left\langle f^\dagger, \gamma_{x_i} \right\rangle_{\mathcal{H}} \qquad \forall i \in I. \tag{4.9}$$

We choose the sample in the following way. We consider a family of measurable sets $X_1, \ldots, X_n \subset X$ such that

(1) $x_i \in X_i$ for all $i \in I$;

(2) $\rho_X(X_i \cap X_j) = 0$ for all $i \neq j$;
(3) $\cup_i X_i = X$.

Then we have the following result.

Proposition 4.2. *Let $a_i = \rho_X(X_i)$ in (4.1), then*

$$\|\mathbf{y}\|_{\mathcal{Z}_\mathbf{x}} \leq \sqrt{\alpha}\kappa \left\|f^\dagger\right\|_{\mathcal{H}} \tag{4.10}$$

$$\left\|A^*g - B^*_{\mathbf{x},n}\mathbf{y}\right\|_{\mathcal{H}} \leq 2\left\|f^\dagger\right\|_{\mathcal{H}} \alpha\kappa c(n) \tag{4.11}$$

$$\left\|A^*A - B^*_{\mathbf{x},n}B_{\mathbf{x},n}\right\|_{\mathcal{L}(\mathcal{H})} \leq 2\alpha\kappa c(n), \tag{4.12}$$

where $\alpha = \rho_X(X)$, $\kappa = \sup_{x\in X}\|\gamma_x\|_{\mathcal{H}}$ and

$$c(n) = \max_{i\in I}\left(\sup_{x\in X_i}\|\gamma_x - \gamma_{x_i}\|_{\mathcal{H}}\right) \tag{4.13}$$

Proof. We first prove (4.12). The definition of X_i and a_i with (4.3) and (4.6) gives

$$\begin{aligned}
\left\|A^*A - B^*_{\mathbf{x},n}B_{\mathbf{x},n}\right\|_{\mathcal{L}(\mathcal{H})} &= \left\|\sum_i \int_{X_i} \left(\langle\cdot,\gamma_x\rangle_{\mathcal{H}}\,\gamma_x - \langle\cdot,\gamma_{x_i}\rangle_{\mathcal{H}}\,\gamma_{x_i}\right) d\rho_X(x)\right\|_{\mathcal{L}(\mathcal{H})} \\
&\leq \sum_i \rho_X(X_i) \sup_{x\in X_i} \|\langle\cdot,\gamma_x\rangle_{\mathcal{H}}\,\gamma_x - \langle\cdot,\gamma_{x_i}\rangle_{\mathcal{H}}\,\gamma_{x_i}\|_{\mathcal{L}(\mathcal{H})} \\
&\leq \left(\max_{i\in I}\sup_{x\in X_i} \|\langle\cdot,\gamma_x\rangle_{\mathcal{H}}\,\gamma_x - \langle\cdot,\gamma_{x_i}\rangle_{\mathcal{H}}\,\gamma_{x_i}\|_{\mathcal{L}(\mathcal{H})}\right)\sum_i \rho_X(X_i) \\
&\leq 2\rho_X(X)\left(\max_{i\in I}\sup_{x\in X_i}\left(\|\gamma_x\|_{\mathcal{H}}\,\|\gamma_x - \gamma_{x_i}\|_{\mathcal{H}}\right)\right) \\
&\leq 2\kappa c(n)\alpha,
\end{aligned}$$

so that (4.12) is proved. Moreover (4.9) implies

$$\begin{aligned}
\|\mathbf{y}\|^2_{\mathcal{Z}} &= \sum_{i=1}^n a_i \left\langle f^\dagger, \gamma_{x_i}\right\rangle^2_{\mathcal{H}} \\
&\leq \sum_{i=1}^n \rho_X(X_i)\left\|f^\dagger\right\|^2_{\mathcal{H}} \|\gamma_{x_i}\|^2_{\mathcal{H}} \\
&\leq \alpha\left\|f^\dagger\right\|^2_{\mathcal{H}}\kappa^2.
\end{aligned}$$

Finally (4.7), (4.2) and (4.6) give

$$\left\|A^*Ag - B^*_{\mathbf{x},n}B_{\mathbf{x},n}\mathbf{y}\right\|_{\mathcal{H}} = \left\|(A^*A - B^*_{\mathbf{x},n}B_{\mathbf{x},n})f^\dagger\right\|_{\mathcal{H}},$$

which implies (4.11) by means of (4.12). □

We are now in position to apply Th. 3.1. Assume that the exact datum $g \in \Omega_{r,R}$ with $\frac{1}{2} < r \leq \frac{3}{2}$, then (3.7) implies

$$\left\|f^\lambda_{\mathbf{z},n} - f^\dagger\right\|_{\mathcal{H}} \leq \left\|f^\dagger\right\|_{\mathcal{H}} \kappa\, c(n)\,\alpha\left(\frac{2}{\lambda} + \frac{\kappa\sqrt{\alpha}}{\lambda^{\frac{3}{2}}}\right) + R\lambda^{r-\frac{1}{2}}. \tag{4.14}$$

for any $\lambda > 0$, so that $f^{\lambda}_{\mathbf{z},n}$ converges to $f^{\dagger}$ if $\lambda = \lambda_{\mathbf{z},n}$ is chosen such that

$$\lim_{n\to+\infty} \lambda_{\mathbf{z},n} = 0 \quad \text{and} \quad \lim_{n\to+\infty} \frac{c(n)}{\lambda^{\frac{3}{2}}_{\mathbf{z},n}} = 0.$$

In the following example we show how to find $\lambda = \lambda_{\mathbf{z},n}$.

4.2.1. *The problem of differentiating a real function*

As a simple example of the above setting, we consider the problem of computing the derivative of a function $g : [0,1] \to \mathbb{R}$, when a finite set of samples $y_i = g(x_i)$ is given.

First of all, we rewrite the above problem by means of the formalism of Carleman operators. Let $H^1[0,1]$ be the Sobolev space of continuous real functions on $[0,1]$ whose weak derivative is in $L^2([0,1], dx)$, where dx is the Lebesgue measure on $[0,1]$. The scalar product in $H^1[0,1]$ is given by

$$\langle f, g\rangle_{H^1[0,1]} = f(0)g(0) + \int_0^1 f'(x)g'(x)\, dx.$$

We define $A : \mathcal{H} \to L^2([0,1], dx)$ as

$$(Af)(x) = \int_0^x f(t)\, dt \quad x \in [0,1],$$

for all $f \in \mathcal{H}$. Clearly, $Af = g$ if and only if $f = g'$, so that $f^{\dagger} = g'$ for all $g \in \operatorname{Im} A$. Moreover, a simple calculation shows that, if $x \in X$,

$$(Af)(x) = \langle f, \gamma_x\rangle_{H^1[0,1]}$$

where $\gamma_x \in H^1[0,1]$ is given by

$$\gamma_x(t) = \begin{cases} x + tx - \frac{t^2}{2} & t \le x \\ x + \frac{x^2}{2} & t > x \end{cases}.$$

Since the function

$$(x,t) \mapsto \langle \gamma_x, \gamma_t\rangle_{H^1[0,1]} = xt(1 + \frac{1}{2}\min\{x,t\}) - \frac{1}{6}(\min\{x,t\})^3$$

is continuous it follows that γ is measurable and, clearly, it is bounded. Hence A is the Carleman operator associated with the map γ

$$[0,1] \ni x \mapsto \gamma_x \in H^1[0,1]$$

and we can apply the result of Prop. 4.2 with $X = [0,1]$, $\rho_X = dx$, $\mathcal{H} = H^1[0,1]$.

For the discretization, we choose the points $x_i = \frac{i}{n}$ for all $i = 0, \dots, n$ and $X_i = [x_{i-1}, x_i]$. If $\lambda > 0$, $f^{\lambda}_{\mathbf{z},n}$ is the regularized solution of the discrete problem

$$\int_0^{x_i} f(t)\, dt = g(x_i) \quad i = 1, \dots, n,$$

where $f \in H^1[0,1]$. According to Equation (4.8), $f^\lambda_{\mathbf{z},n}$ is a linear combination of the functions γ_{x_i}, that are quadratic splines: piecewise polynomials of degree two with continuous derivative [13]. From a numerical point of view, the computation of $f^\lambda_{\mathbf{z},n}$ reduces to compute the inverse of the $n \times n$ symmetric matrix

$$\Gamma_{\mathbf{x}}(x_i, x_j) = x_i x_j (1 + \frac{1}{2}\min\{x_i, x_j\}) - \frac{1}{6}(\min\{x_i, x_j\})^3.$$

To apply Equation (4.14), we notice that $\alpha = \rho_X([0,1]) = 1$ and, if $0 \le t \le x \le 1$,

$$\|\gamma_x - \gamma_t\|_{\mathcal{H}} = \sqrt{(x-t)^2 \frac{3 + x + 2t}{3}} \le \sqrt{2}|x - t|.$$

It follows that $c(n) = \frac{\sqrt{2}}{n}$ and, letting $t = 0$, that $\kappa = \sqrt{2}$. Replacing these bounds in Equation (4.14) we obtain that

$$\left\| f^\lambda_{\mathbf{z},n} - f^\dagger \right\|_{H^1[0,1]} \le 2\sqrt{2}\, \|g'\|_{H^1[0,1]} \frac{1}{n} \left(\frac{\sqrt{2}}{\lambda} + \frac{1}{\lambda^{\frac{3}{2}}} \right) + R\lambda^{r-\frac{1}{2}}.$$

In this setting the optimal choice of the regularization parameter is $\lambda_n = n^{-\frac{1}{1+r}}$ and, with this choice,

$$\left\| f^\lambda_{\mathbf{z},n} - f^\dagger \right\|_{H^1[0,1]} = O\left(n^{-\frac{2r-1}{2+2r}} \right),$$

here the parameter r is related to the a priori assumption $g \in \operatorname{Im} L_\Gamma^r$, which is an assumption on the smoothness of g [13].

4.3. *Stochastic discretization: learning from examples*

In this section we consider the framework of learning theory as given in Section 2. In this setting, ρ_X is the marginal distribution of the unknown probability measure ρ and $\mathcal{H}$ is the (separable) reproducing kernel Hilbert space with kernel K. For all $x \in X$, we let $\gamma_x = K_x$, so that the map γ is bounded by (2.1) and measurable since K is measurable[a]. Eq. (2.2) implies that the corresponding Carleman operator A is the canonical inclusion I_K and the exact datum g is the regression function f_ρ. Moreover, with the choice $a_i = \frac{1}{n}$ in (4.1), for any training set $\mathbf{z} = (\mathbf{x}, \mathbf{y})$ $B_{\mathbf{x},n}$ is the sampling operator $S_{\mathbf{x},n}$ and $\mathbf{y}$ is the datum of the discretized problem.
Since both $S_{\mathbf{x},n}$ and $\mathbf{y}$ are random variables, we need a probabilistic estimate of the perturbation measure δ.

Proposition 4.3. *Let* $\kappa = \sup_{x \in X} \|K_x\|_{\mathcal{H}}$, $M = \sup_{y \in Y} |y|$ *and* $\delta = (\delta_1, \delta_2) \in \mathbb{R}^2_+$, *then*

$$\mathbb{P}_{\mathbf{z} \sim \rho^n} [\ \|\mathbf{y}\|_n > M] = 0 \tag{4.15}$$

$$\mathbb{P}_{\mathbf{z} \sim \rho^n} \left[\left\| I_K^* f_\rho - S^*_{\mathbf{x},n} \mathbf{y} \right\|_{\mathcal{H}} > \delta_1 \right] \le 2 \exp\left(-\frac{n {\delta_1}^2}{8\kappa^2 M^2} \right) \tag{4.16}$$

[a]Indeed, the measurability of K and the fact that $\{K_x,\ x \in X\}$ is total in $\mathcal{H}$ implies that $x \mapsto \langle K_x, f \rangle_{\mathcal{H}}$ is measurable for all $f \in \mathcal{H}$. Since $\mathcal{H}$ is separable, this ensures the measurability of γ.

$$\mathbb{P}_{\mathbf{z}\sim\rho^n}\left[\ \left\|I_K^* I_K - S_{\mathbf{x},n}^* S_{\mathbf{x},n}\right\|_{\mathcal{L}(\mathcal{H})} > \delta_2\right] \leq 2\exp\left(-\frac{n{\delta_2}^2}{8\kappa^4}\right). \tag{4.17}$$

Proof. The first equation is trivial. To prove (4.16) let $\xi_1 : Z \to \mathcal{H}$ be the random variable

$$\xi_1(x,y) = K_x y - I_K^* f_\rho.$$

Eq. (4.2) implies that $\mathbb{E}[\xi_1] = 0$. The definitions of κ and M ensure that $\|\xi(x,y)\|_{\mathcal{H}} \leq 2\kappa M$. Since $S_{\mathbf{x},n}^*\mathbf{y} - I_K^* f_\rho = \frac{1}{n}\sum_{i=1}^n \xi_1(x_i,y_i)$, Pinelis inequality (see Th.3.5 of [36]) gives

$$\mathbb{P}_{\mathbf{z}\sim\rho^n}\left[\ \left\|S_{\mathbf{x},n}^*\mathbf{y} - I_K^* f_\rho\right\|_{\mathcal{H}} > \delta_1\right] \leq 2\exp\left(-\frac{n{\delta_1}^2}{2(2\kappa M)^2}\right).$$

Let $\mathcal{L}_2(\mathcal{H})$ be the Hilbert space of the Hilbert-Schmidt operators on $\mathcal{H}$ and $\xi_2 : Z \to \mathcal{L}_2(\mathcal{H})$

$$\xi_2(x,y) = \langle\cdot, K_x\rangle_{\mathcal{H}} K_x - I_K^* I_K,$$

which is well defined since the first term is a rank one projection and the second one is a Hilbert-Schmidt operator, see Prop. 4.1. Eq. (4.3) implies that $\mathbb{E}[\xi_2] = 0$ and the definition of κ ensures

$$\|\xi(x,y)\|_{\mathcal{L}_2(\mathcal{H})} \leq 2\kappa^2.$$

Reasoning as above,

$$\mathbb{P}_{\mathbf{z}\sim\rho^n}\left[\ \left\|S_{\mathbf{x},n} S_{\mathbf{x},n}^* - I_K^* I_K\right\|_{\mathcal{L}_2(\mathcal{H})} > \delta_2\right] \leq 2\exp\left(-\frac{n{\delta_2}^2}{2(2\kappa^2)^2}\right).$$

The result follows observing that $\|\cdot\|_{\mathcal{L}(\mathcal{H})} \leq \|\cdot\|_{\mathcal{L}_2(\mathcal{H})}$. □

Th. 3.1 with the above probabilistic bounds gives the result claimed at the end of Section 2. Precisely,

Theorem 4.1. *Assume there is* $0 < r \leq 1$ *and* R *such that* $\left\|L_K^{-r} P f_\rho\right\|_\rho \leq R$ *and choose the regularization parameter according to the rule* $\lambda_n = n^{-\frac{1}{2r+1}}$. *For any* $\epsilon > 0$

$$\mathbb{P}_{\mathbf{z}\sim\rho^n}\left[\ \|f_{\mathbf{z},n} - P f_\rho\|_\rho^2 > (C\epsilon + R)^2 n^{-\frac{r}{r+1}}\right] \leq 2\left(e^{-C_1\epsilon^2} + e^{-C_1\epsilon^2}\right) \tag{4.18}$$

where C, C_1 *and* C_2 *are suitable constants depending only on* κ, M *and* r. *In particular*

$$\limsup_{n\to+\infty} \frac{\mathbb{E}[\|f_{\mathbf{z},n} - P f_\rho\|_\rho^2]}{n^{-\frac{r}{r+1}}} < +\infty. \tag{4.19}$$

The bound (4.19) means that the rate of convergence on average [15] is $n^{-\frac{r}{r+1}}$ and (4.19) follows from (4.18) by integrating the above tail inequality. The constants are explicitly given by

$$C_1 = \frac{1}{8\kappa^2 M^2} \qquad C_2 = \frac{1}{8\kappa^4} \qquad C = \frac{2M+1}{4},$$

which are independent from the distribution ρ, whereas the constant R could possibly strongly depend on ρ [37].
The error bounds obtained in this paper can be shown not to be optimal in the suitable minimax sense [14,37,38]. However, we believe, they deserved to be presented because of the simplicity of their derivation, and also because of their structure, which decouples the probabilistic analysis from the dependence on λ. This last characteristic allows a straightforward application of our error bounds to the case of data-dependent choice of the regularization parameter.

Acknowledgements

We would like to thank M.Bertero, C. De Mol, M. Piana, T. Poggio, S. Smale, G. Talenti, A. Verri and Y. Yao for useful discussions and suggestions. This research has been partially funded by the FIRB Project ASTAA and the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. 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), as well as in the Dipartimento di Informatica e Scienze dell'Informazione (DISI) at University of Genoa, Italy. 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.

References

[1] A.N. Tikhonov and V.Y. Arsenin. *Solutions of Ill Posed Problems*. W. H. Winston, Washington, D.C., 1977.
[2] 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.
[3] 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.
[4] M. Bertero, C. De Mol, and E. R. Pike. Linear inverse problems with discrete data.

I. General formulation and singular system analysis. *Inverse Problems*, 1(4):301–330, 1985.

[5] M. Bertero, C. De Mol, and E. R. Pike. Linear inverse problems with discrete data. II. Stability and regularisation. *Inverse Problems*, 4(3):573–594, 1988.

[6] R. Kress. *Linear Integral Equations*. Number 82 in Applied Mathematical Sciences. Springer, Heidelberg, 1989.

[7] G. Wahba. Practical approximate solutions to linear operator equations when the data are noisy. *SIAM J. Numer. Anal.*, 14(4):651–667, 1977.

[8] C. W. Groetsch. Convergence analysis of a regularized degenerate kernel method for Fredholm integral equations of the first kind. *Integral Equations Operator Theory*, 13(1):67–75, 1990.

[9] 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.

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

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

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

[13] 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.

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

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

[16] S. Mukherjee, R. Rifkin, and T. Poggio. Regression and classification with regularization. *Lectures Notes in Statistics: Nonlinear Estimation and Classification, Proceedings from MSRI Workshop*, 171:107–124, 2002.

[17] S. Smale and D. Zhou. Shannon sampling II : Connections to learning theory. *preprint*, 2004.

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

[19] 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.

[20] F. Cucker and S. Smale. Best choices for regularization parameters in learning theory: on the bias-variance problem. *Foundations of Computationals Mathematics*, 2:413–428, 2002.

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

[22] E. De Vito, L. Rosasco, A. Caponnetto, M. Piana and A. Verri. Some Properties of Regularized Kernel Methods. *Journal of Machine Learning Research*, 5(Oct):1363–1390, 2004.

[23] 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.

[24] R. Plato and G. Vainikko. On the regularization of projection methods for solving

ill-posed problems. *Numer. Math.*, 57(1):63–79, 1990.

[25] P. Mathé and S. V. Pereverzev. Discretization strategy for linear ill-posed problems in variable Hilbert scales. *Inverse Problems*, 19(6):1263–1277, 2003.

[26] M.T. Nair. A unified approach for regularized approximation methods for Fredholm integral equations of the first kind. *Numer. Funct. Anal. Optim.*, 15(3-4):381–389, 1994.

[27] M. T. Nair and E. Schock. A discrepancy principle for Tikhonov regularization with approximately specified data. *Ann. Polon. Math.*, 69(3):197–205, 1998.

[28] M. P. Rajan. Convergence analysis of a regularized approximation for solving Fredholm integral equations of the first kind. *J. Math. Anal. Appl.*, 279(2):522–530, 2003.

[29] S. Pereverzev and E. Schock. Morozov's discrepancy principle for Tikhonov regularization of severely ill-posed problems in finite-dimensional subspaces. *Numer. Funct. Anal. Optim.*, 21(7-8):901–916, 2000.

[30] A. N. Tikhonov, A. V. Goncharsky, V. V. Stepanov, and A. G. Yagola. *Numerical methods for the solution of ill-posed problems*, volume 328 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1995. Translated from the 1990 Russian original by R. A. M. Hoksbergen and revised by the authors.

[31] P.R. Halmos and V.S. Sunder. *Bounded integral operators on L^2 spaces*, volume 96 of *Results in Mathematics and Related Areas*. Springer-Verlag, Berlin, 1978.

[32] S. Saitoh. *Integral transforms, reproducing kernels and their applications*, volume 369 of *Pitman Research Notes in Mathematics Series*. Longman, Harlow, 1997.

[33] S. Saitoh Best approximation, Tikhonov regularization and reproducing kernels. *Kodai Math. J.*, 28(2):359–367, 2005.

[34] E. De Vito, L. Rosasco, A. Caponnetto, U. De Giovannini, and F. Odone. Learning from examples as an inverse problem. *Journal of Machine Learning Research*, 6:883–904, 2005.

[35] C. Carmeli and E. De Vito, E.and Toigo. Reproducing kernel hilbert spaces and mercer theorem. Technical report, arvXiv:math.FA/0504071, 2005.

[36] I. Pinelis. Optimum bounds for the distributions of martingales in Banach spaces. *Ann. Probab.*, 22(4):1679–1706, 1994.

[37] S. Smale and D. Zhou. Learning Theory Estimates via Integral Operators and Their Approximations. Technical report, Toyota Technological Insititute, April 2005 (to appear). Avaible at http://www.tti-c.org/smale_papers/sampIII5412.pdf.

[38] A. Caponnetto and E. De Vito. Optimal rates for regularized least-squares algorithm. Technical report, Massachusetts Institute of Technology, Cambridge, MA, CBCL Memo 248/AI Memo 2005-013, April 2005.