

Discretization Error Analysis for Tikhonov Regularization in Learning Theory

E De Vito[†], A Caponnetto[‡] and L Rosasco[‡]

[†] 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

[‡] D.I.S.I., Università di Genova, Via Dodecaneso 35, 16146 Genova, Italy

E-mail: devito@unimo.it, caponnetto@disi.unige.it, rosasco@disi.unige.it

Submitted to: *Inverse Problems*

Abstract. We study the connections between learning from examples and inverse problems. We show that learning from examples can be seen as the discretization of a stochastic inverse problem defined by a Carleman operator. In particular we develop a discretization strategy for this class of inverse problems and we give a convergence analysis. Our approach can be applied to other classes of problems such as integral equations.

1. Introduction

The aim of the present paper is to clarify the relation between the theory of learning from examples, developed in the last two decades as theoretical framework in brain and cognitive science (Vapnik 1998, Cucker and Smale 2002b, Poggio and Smale 2003), and the theory of inverse problems.

We show that learning from examples can be regarded as the discretization of a stochastic inverse problem (Vapnik 1998). In particular we prove that the regularized least squares algorithm (Evgeniou et al. 2000) corresponds to the Tikhonov regularization of linear inverse problem defined by a Carleman operator (Halmos and Sunder 1978). As a byproduct, we obtain a convergence analysis for the discretization of a linear operator. Our result is based on a stability property of Tikhonov regularization and could be of interest also in the framework of discretization of integral equations (Kress 1999).

More precisely, let A be a bounded linear operator from a Hilbert space \mathcal{H} into a Hilbert space \mathcal{K} and consider the inverse problem associated to

$$Af = g, \tag{1}$$

where g is the datum belonging to \mathcal{K} . Usually the above problem is ill-posed and a regularization procedure is considered. For example, in the framework of Tikhonov regularization, the following minimization problem

$$\min_{f \in \mathcal{H}} (\|Af - h\|_{\mathcal{K}}^2 + \lambda \|f\|_{\mathcal{H}}^2)$$

replaces Problem (1).

Discretization is a procedure that replaces the exact problem with an approximated one

$$Bf = h, \tag{2}$$

where B is a bounded operator from \mathcal{H} into a finite dimensional Hilbert space \mathcal{Z} and h is an element of \mathcal{Z} that are *approximations* of the model A and the datum g , respectively.

Examples of such procedures are degenerate kernel methods, quadrature methods and projection methods. A common feature of these methods is the fact that they directly regularize the exact problem and that the reconstruction error of the discrete solution depends on the noise of the datum h and the dimension of the space \mathcal{Z} , where usually h is the projection of g with an additive noise and \mathcal{Z} is a finite dimensional subspace of \mathcal{K} (for a review see Groetsch (1984), Bertero et al. (1985, 1988), Engl et al. (1996), Kress (1999) and references therein).

Following Wahba (1977) and Groetsch (1990), we regard the datum h as a noisy approximation of the exact datum g and the operator B as a noisy approximation of the exact model A . The critical point is to give a measure of the discrepancy between h and g , and between B and A . For Tikhonov regularization this can be done by observing that the minimizer of Tikhonov functional is given by

$$f^\lambda = (B^*B + \lambda)^{-1}B^*h.$$

The above equation shows that f^λ depends on B^*B , which is an operator from \mathcal{H} to \mathcal{H} , and on B^*h , which is an element of \mathcal{H} , so that the output space \mathcal{Z} disappears. This observation suggests that the noise measures could be $\|B^*B - A^*A\|_{\mathcal{L}(\mathcal{H})}$ and $\|B^*h - A^*g\|_{\mathcal{H}}$.

The paper is organized as follows. Section 2 is devoted to the formalization of the above idea in an abstract setting. In Section 4 we apply the results to the problem of discretization of Carleman operators, whose properties are briefly recalled in Section 3. Carleman operators give an unifying framework both for solving integral equations and for approximation problems in reproducing kernel Hilbert spaces.

We provide a convergence analysis in two different settings. In Section 4.1 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 4.2 the discrete data are considered as random variables and a probabilistic bound on the discrete regularized solution is obtained. Our estimate holds for a very general type of noise and factorizes in a deterministic term due to

the regularization procedure and a probabilistic term due to the probabilistic nature of the data.

Finally, in Section 5 we show that learning from examples can be seen as the discretization of a Carleman operator with random discrete data. In particular, using the probabilistic estimate of Section 4.2 we prove the consistency of the regularized least squares algorithm.

2. Main results

In this section we prove that the regularized (*à la* Tikhonov) solution of the inverse problem $Af = g$ is a Lipschitz function of A^*A and A^*g , where the Lipschitz constants depend on a power of the regularization parameter λ .

First of all we set the notation. By Hilbert space we mean a separable Hilbert space over \mathbb{R} (our results also hold if \mathbb{R} is replaced by \mathbb{C}). We denote by $\|\cdot\|$ and $\langle \cdot, \cdot \rangle$ the corresponding norm and scalar product, respectively. If \mathcal{H} and \mathcal{K} are such spaces, we denote by $\mathcal{L}(\mathcal{H}, \mathcal{K})$ the Banach space of bounded linear operators from \mathcal{H} into \mathcal{K} endowed with the uniform norm $\|\cdot\|_{\mathcal{L}(\mathcal{H}, \mathcal{K})}$. If $A \in \mathcal{L}(\mathcal{H}, \mathcal{K})$ we denote by A^* the adjoint operator and by A^\dagger the Moore-Penrose generalized inverse (Groetsch 1984, Engl et al. 1996).

We consider a Hilbert space \mathcal{H} and we denote by \mathcal{T} the set of all possible triples (\mathcal{K}, g, A) where \mathcal{K} is an Hilbert space, $g \in \mathcal{K}$ and $A \in \mathcal{L}(\mathcal{H}, \mathcal{K})$. Given $(\mathcal{K}, A, g) \in \mathcal{T}$ and $\lambda > 0$, we recall that the Tikhonov functional (defined on \mathcal{H}) is

$$\|Af - g\|_{\mathcal{K}}^2 + \lambda \|f\|_{\mathcal{H}}^2 \tag{3}$$

and its unique minimizer is given by $(A^*A + \lambda)^{-1}A^*g$.

The following proposition studies the dependence of the minimizer of Tikhonov functional on the operator A and the datum g . To treat both the reconstruction error and the residue of the solution, we introduce a parameter $a \in [0, 1]$ and we let

$$C_a = \begin{cases} 1 & a = 0, a = 1 \\ a^a(1-a)^{(1-a)} & 0 < a < 1 \end{cases} . \tag{4}$$

Proposition 1 *Given $(\mathcal{K}, A, g) \in \mathcal{T}$ and $(\mathcal{Z}, B, h) \in \mathcal{T}$, let $\lambda > 0$ and*

$$\begin{aligned} f_0^\lambda &= (A^*A + \lambda)^{-1}A^*g, \\ f^\lambda &= (B^*B + \lambda)^{-1}B^*h. \end{aligned}$$

Then

$$\|(A^*A)^a(f^\lambda - f_0^\lambda)\|_{\mathcal{H}} \leq \frac{C_a}{\lambda^{1-a}} \left(\frac{\|h\|_{\mathcal{Z}}}{2\sqrt{\lambda}} \|B^*B - A^*A\|_{\mathcal{L}(\mathcal{H})} + \|B^*h - A^*g\|_{\mathcal{H}} \right) \tag{5}$$

for any $a \in [0, 1]$.

Proof. We let $T_0 = A^*A$, $T = B^*B$, $\phi_0 = A^*g$ and $\phi = B^*h$. We prove some preliminary facts.

The following equality is a known algebraic identity,

$$(T + \lambda)^{-1} - (T_0 + \lambda)^{-1} = (T_0 + \lambda)^{-1}(T_0 - T)(T + \lambda)^{-1}. \tag{6}$$

Consider now $a \in [0, 1]$, we claim that

$$\|T^a(T + \lambda)^{-1}\|_{\mathcal{L}(\mathcal{H})} \leq \frac{C_a}{\lambda^{1-a}}. \quad (7)$$

Indeed, since T is a self-adjoint positive operator, the spectral theorem gives

$$\|T^a(T + \lambda)^{-1}\|_{\mathcal{L}(\mathcal{H})} = \sup_{0 \leq t \leq \|T\|_{\mathcal{L}(\mathcal{H})}} \frac{t^a}{t + \lambda}.$$

Computing a derivative tells us that the maximizer of the function $\frac{t^a}{t+\lambda}$ on the interval $[0, \|T\|_{\mathcal{L}(\mathcal{H})}]$ is given by $\tau = \frac{\lambda a}{1-a}$ if $a < 1$, and $\tau = \|T\|_{\mathcal{L}(\mathcal{H})}$ if $a = 1$. Equation (7) follows.

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

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

and, using Equation (7) with $a = \frac{1}{2}$, we have that

$$\|(T + \lambda)^{-1}B^*\|_{\mathcal{L}(\mathcal{Z}, \mathcal{H})} \leq \frac{1}{2\sqrt{\lambda}}. \quad (8)$$

Finally, by definition of f^λ and f_0^λ , it follows that

$$\begin{aligned} f^\lambda - f_0^\lambda &= (T + \lambda)^{-1}\phi - (T_0 + \lambda)^{-1}\phi_0 \\ &= [(T + \lambda)^{-1} - (T_0 + \lambda)^{-1}]\phi + (T_0 + \lambda)^{-1}(\phi - \phi_0). \end{aligned}$$

By Equation (6) and triangular inequality, we obtain that

$$\begin{aligned} \|T_0^a(f^\lambda - f_0^\lambda)\|_{\mathcal{H}} &\leq \|T_0^a(T_0 + \lambda)^{-1}(T_0 - T)(T + \lambda)^{-1}B^*h\|_{\mathcal{H}} \\ &\quad + \|T_0^a(T_0 + \lambda)^{-1}(\phi - \phi_0)\|_{\mathcal{H}} \\ &\leq \frac{C_a}{\lambda^{1-a}}\|T - T_0\|_{\mathcal{L}(\mathcal{H})} \frac{\|h\|_{\mathcal{Z}}}{2\sqrt{\lambda}} + \frac{C_a}{\lambda^{1-a}}\|\phi - \phi_0\|_{\mathcal{H}}. \end{aligned}$$

The thesis is now clear. ■

Up to our knowledge, the first result in this direction was obtained in Wahba (1977, Theorem 2) in the framework of integral equations with white noise. In a deterministic setting, Groetsch (1990, Theorem 1) gives a convergence analysis for integral equations assuming that B^*B is a degenerate kernel and $h = g$. Plato and Vainikko (1990, Theorem 3.1) and Mathé and Pereverzev (2003, Theorem 2) consider a wider class of regularization methods, but B has the form Q_nAP_n , where Q_n and P_n are orthogonal projections. Our bound is of the same kind of the estimates obtained in Nair (1994, Theorem 2.2), Nair and Schock (1998, Theorem 2.1) and Rajan (2003, Theorem 2.5). Anyway in the above papers only the reconstruction error is considered and different choices of noisy levels are considered.

We briefly comment on regularization procedures other than Tikhonov regularization. It is known (Engl et al. 1996, Groetsch 1984) that a large class of regularized solutions is of the form

$$f_0^\lambda = r_\lambda(A^*A)A^*g, \quad (9)$$

where r_λ is a continuous function on $[0, \|A^*A\|_{\mathcal{L}(\mathcal{H})}]$ such that for some $K > 0$

$$|r_\lambda(t)t| \leq K \quad \forall t, \lambda \quad \text{and} \quad \lim_{\lambda \rightarrow 0} r_\lambda(t) = \frac{1}{t}.$$

The results of Proposition 1 still hold provided that

$$\|r_\lambda(A^*A) - r_\lambda(B^*B)\|_{\mathcal{L}(\mathcal{H})} \leq C_\lambda \|A^*A - B^*B\|_{\mathcal{L}(\mathcal{H})}.$$

A complete discussion on the above condition can be found in Mathé and Pereverzev (2002, 2003).

The results of Proposition 1 suggest the following definition of *parameter choice rule*.

Definition 1 Given $(\mathcal{K}, g, A) \in \mathcal{T}$ and $M > 0$, for any $\delta = (\delta_1, \delta_2) \in \mathbb{R}_+^2$ we let

$$\mathcal{U}_\delta = \{(\mathcal{Z}, h, B) \in \mathcal{T} \mid \|h\|_{\mathcal{Z}} \leq M, \|B^*h - A^*g\|_{\mathcal{H}} \leq \delta_1, \|B^*B - A^*A\|_{\mathcal{L}(\mathcal{H})} \leq \delta_2\},$$

the set of noisy data with noise level $\delta = (\delta_1, \delta_2)$.

A function $\lambda = \lambda(\delta; \mathcal{Z}, h, B)$, where $\delta \in \mathbb{R}_+^2$ and $(\mathcal{Z}, h, B) \in \mathcal{T}$, is called a *parameter choice rule*.

In the above definition the constant M plays the role of an a priori information on the noisy datum h . If this information is not available, a bound of the form of Equation (5) can still be given, but with a worse dependence on the regularization parameter λ , by replacing estimate given by Equation (8) with

$$\|(T + \lambda)^{-1}B^*h\|_{\mathcal{H}} \leq \frac{\|B^*h\|_{\mathcal{H}}}{\lambda} \leq \frac{\|A^*g\|_{\mathcal{H}} + \delta_1}{\lambda}.$$

As a consequence of Proposition 1 we have the following result.

Corollary 1 Let $(\mathcal{K}, g, A) \in \mathcal{T}$ and P be the projection on the closure of the range of A . Given $M > 0$ and $\delta \in \mathbb{R}_+^2$, then

(i) if $g \in \text{dom } A^\dagger$ and $\lambda > 0$,

$$\sup_{(\mathcal{Z}, h, B) \in \mathcal{U}_\delta} \left| \|f^\lambda - A^\dagger g\|_{\mathcal{H}} - \|f_0^\lambda - A^\dagger g\|_{\mathcal{H}} \right| \leq \frac{\delta_1}{\lambda} + \frac{M\delta_2}{2\lambda^{\frac{3}{2}}}; \quad (10)$$

(ii) if $\lambda > 0$,

$$\sup_{(\mathcal{Z}, h, B) \in \mathcal{U}_\delta} \left| \|Af^\lambda - Pg\|_{\mathcal{K}} - \|Af_0^\lambda - Pg\|_{\mathcal{K}} \right| \leq \frac{\delta_1}{2\sqrt{\lambda}} + \frac{M\delta_2}{4\lambda}; \quad (11)$$

(iii) if $g \in \text{dom } A^\dagger$ and $\lambda = \lambda(\delta; \mathcal{Z}, h, B)$ is a parameter choice rule such that

$$\begin{cases} \lim_{\delta \rightarrow 0} \sup_{(\mathcal{Z}, h, B) \in \mathcal{U}_\delta} \lambda(\delta; \mathcal{Z}, h, B) = 0 \\ \lim_{\delta \rightarrow 0} \sup_{(\mathcal{Z}, h, B) \in \mathcal{U}_\delta} \frac{\delta_1}{\lambda(\delta; \mathcal{Z}, h, B)} = 0 \\ \lim_{\delta \rightarrow 0} \sup_{(\mathcal{Z}, h, B) \in \mathcal{U}_\delta} \frac{\delta_2^{\frac{2}{3}}}{\lambda(\delta; \mathcal{Z}, h, B)} = 0, \end{cases} \quad (12)$$

then

$$\lim_{\delta \rightarrow 0} \sup_{(\mathcal{Z}, h, B) \in \mathcal{U}_\delta} \|f^\lambda - A^\dagger g\|_{\mathcal{H}} = 0. \quad (13)$$

(iv) if $\lambda = \lambda(\delta; \mathcal{Z}, h, B)$ is a parameter choice rule such that

$$\begin{cases} \lim_{\delta \rightarrow 0} \sup_{(\mathcal{Z}, h, B) \in \mathcal{U}_\delta} \lambda(\delta; \mathcal{Z}, h, B) = 0 \\ \lim_{\delta \rightarrow 0} \sup_{(\mathcal{Z}, h, B) \in \mathcal{U}_\delta} \frac{\delta_1}{\sqrt{\lambda(\delta; \mathcal{Z}, h, B)}} = 0 \\ \lim_{\delta \rightarrow 0} \sup_{(\mathcal{Z}, h, B) \in \mathcal{U}_\delta} \frac{\delta_2}{\lambda(\delta; \mathcal{Z}, h, B)} = 0, \end{cases} \quad (14)$$

then

$$\lim_{\delta \rightarrow 0} \sup_{(\mathcal{Z}, h, B) \in \mathcal{U}_\delta} \|A f^\lambda - g\|_{\mathcal{K}} = 0. \quad (15)$$

Proof. Inequalities (10) and (11) are consequences of Equation (5) taking $a = \frac{1}{2}$ and $a = 0$, respectively, and the fact that

$$\left| \|f^\lambda - A^\dagger g\|_{\mathcal{H}} - \|f_0^\lambda - A^\dagger g\|_{\mathcal{H}} \right| \leq \|f^\lambda - f_0^\lambda\|_{\mathcal{H}}.$$

Equations (13) and (15) follow from inequalities (10) and (11), and Conditions (12) and (14). ■

The fact that $\|f_0^\lambda - A^\dagger g\|_{\mathcal{H}}$ goes to zero as λ goes to zero is known. Estimating the rate of convergence of f^λ to $A^\dagger g$ requires a bound on $\|f_0^\lambda - A^\dagger g\|_{\mathcal{H}}$ and some a priori assumption on the exact solution. A standard result (Groetsch 1984, Engl et al. 1996) shows that, if $A^\dagger g \in \text{Im}(A^* A)^\alpha A^*$, then $\|f_0^\lambda - A^\dagger g\|_{\mathcal{H}} = O(\lambda^\alpha)$. In Mathé and Pereverzev (2003) a generalization is given under the assumption that $A^\dagger g = \phi(A^* A)v$ for some $v \in \mathcal{H}$, $\|v\|_{\mathcal{H}} \leq 1$, and some convex function ϕ .

As usual, if $g \notin \text{dom } A^\dagger$, f^λ cannot converge. Indeed, we have the following result.

Corollary 2 *Let $(\mathcal{K}, g, A) \in \mathcal{T}$ with $g \notin \text{dom } A^\dagger$. For any $\delta \in \mathbb{R}_+^2$, let $(\mathcal{Z}_\delta, h_\delta, B_\delta) \in \mathcal{U}_\delta$ and*

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

where $\lambda_\delta = \lambda(\delta; h_\delta, B_\delta)$ is a parameter choice rule satisfying Equation (12). Then

$$\lim_{\delta \rightarrow 0} \|f_\delta^{\lambda_\delta}\|_{\mathcal{H}} = +\infty.$$

Proof. The proof is standard. As in the first part of the above proof we have that

$$\begin{aligned} \|f_\delta^{\lambda_\delta}\|_{\mathcal{H}} &\geq \|f_0^{\lambda_\delta}\|_{\mathcal{H}} - \|f_\delta^{\lambda_\delta} - f_0^{\lambda_\delta}\|_{\mathcal{H}} \\ &\geq \|f_0^{\lambda_\delta}\|_{\mathcal{H}} - \left(\frac{\delta_1}{\lambda_\delta} + \frac{M\delta_2}{2\lambda_\delta^{\frac{3}{2}}} \right). \end{aligned}$$

By definition of parameter choice rule, the second term goes to zero, whereas, if $g \notin \text{dom } A^\dagger$, $\|f_0^\lambda\|_{\mathcal{H}}$ goes to $+\infty$ when λ goes to zero (Groetsch 1984, Engl et al. 1996). ■

Remark 1 *We do not know if the bound in Equation (10) is sharp, that is, the following holds*

$$\sup_{(\mathcal{Z}, h, B) \in \mathcal{U}_\delta} \|f^\lambda - f_0^\lambda\|_{\mathcal{H}} = \frac{\delta_1}{\lambda} + \frac{M\delta_2}{2\lambda^{\frac{3}{2}}}.$$

In particular, for the problem of discretization, could one obtain better bounds by considering only noisy data (\mathcal{Z}, h, B) with B having finite range?

We compare Proposition 1 with the known results for Tikhonov regularization in the presence of modeling error. To this aim, we consider noisy problems $Bf = h$, where B is an operator from \mathcal{H} to \mathcal{K} and $h \in \mathcal{K}$ such that

$$\|h - g\|_{\mathcal{K}} \leq \eta_1 \quad \|B - A\|_{\mathcal{L}(\mathcal{H})} \leq \eta_2.$$

In this case it is known (Tikhonov et al. 1995) that if

$$\lim_{\eta_1, \eta_2 \rightarrow 0} \frac{(\eta_1 + \eta_2)^2}{\lambda(\eta_1, \eta_2)} = 0 \tag{16}$$

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

$$\begin{aligned} \|B^*B - A^*A\|_{\mathcal{L}(\mathcal{H})} &\leq (\|B\|_{\mathcal{L}(\mathcal{H})} + \|A\|_{\mathcal{L}(\mathcal{H})})\eta_2 \leq C_1\eta_2 =: \delta_2 \\ \|B^*h - A^*g\|_{\mathcal{H}} &\leq \|h\|_{\mathcal{K}}\eta_2 + \|A\|_{\mathcal{L}(\mathcal{H})}\eta_1 \leq C_2(\eta_1 + \eta_2) =: \delta_1 \end{aligned}$$

it follows that Condition (16) is weaker than Condition (12).

This observation suggests that the noise levels can be evaluated by means of $\|V^*h - U^*g\|_{\mathcal{H}} \leq \eta_1$ and $\| |B| - |A| \|_{\mathcal{L}(\mathcal{H})} \leq \eta_2$, where $A = U|A|$ and $B = V|B|$ are the polar decompositions of A and B , respectively. In fact repeating the standard proof (Tikhonov et al. 1995) for Tikhonov regularization in the presence of modeling error, we have that, if Condition (16) holds, then the regularized solution f^λ approaches $A^\dagger g$. 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 Proposition 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 $Bf = h$ with the problem

$$B^*Bf = B^*h,$$

so that B^*h is a noisy approximation of the exact datum A^*g , B^*B is the noisy model of the exact model A^*A and the noise levels are given by

$$\|B^*B - A^*A\|_{\mathcal{L}(\mathcal{H})} = \delta_1 \quad \|B^*h - A^*g\|_{\mathcal{H}} = \delta_2.$$

However, the regularized solution $f^\lambda = (B^*B + \lambda)^{-1}B^*h$ is not the Tikhonov regularization of the problem $B^*Bf = B^*h$. Indeed, if $T = T^* = B^*B$ and $\phi = B^*h$, we have that

$$f^\lambda = (T + \lambda)^{-1}\phi = (T^*T + \lambda T)^{-1}T^*\phi,$$

whereas the Tikhonov regularized solution of $Tf = \phi$ is $(T^*T + \lambda)^{-1}T^*\phi$.

In this paper we do not discuss the problem of the choice of the parameter λ . Wahba (1977) discusses the method of cross-validation, see also Wahba (1990), Groetsch (1984) and Engl et al. (1996) for an account on cross validation. A clear discussion about the discrepancy principle in the framework of discretization of Tikhonov functional can be found in Plato and Vainikko (1990), Nair and Schock (1998) and Pereverzev and Schock (2000) and references therein. Mathé and Pereverzev (2003) propose *adaptive strategies* for the choice of the parameter that provides the optimal order accuracy.

3. 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 Halmos and Sunder (1978), where a clear exposition of the relation between Carleman operators and integral equations is given. The book of Saitoh (1997) is a source for results and bibliography on this topic. In Bertero et al. (1985), Wahba (1990) and references therein, there is an account of the theory of reproducing kernel Hilbert spaces in the context of inverse problems.

Let X be a compact separable metric space endowed with a finite measure ν . We denote by $L^2(X, \nu)$ the Hilbert space of (equivalence classes) of real functions on X that are square-integrable with respect to ν .

We consider a Hilbert space \mathcal{H} and a map γ from X to \mathcal{H} such that the kernel Γ

$$\Gamma(x, t) := \langle \gamma_t, \gamma_x \rangle_{\mathcal{H}} \quad x, t \in X$$

is bounded and measurable as a function on $X \times X$. If $f \in \mathcal{H}$, let Af be the function on X given by

$$(Af)(x) = \langle f, \gamma_x \rangle_{\mathcal{H}} \quad \forall x \in X.$$

The following result holds.

Proposition 2 *For all $f \in \mathcal{H}$, $Af \in L^2(X, \nu)$ and the map A*

$$\mathcal{H} \ni f \rightarrow Af \in L^2(X, \nu)$$

is a Hilbert-Schmidt operator from \mathcal{H} into $L^2(X, \nu)$. Moreover,

$$A^* \phi = \int_X \phi(x) \gamma_x \, d\nu(x), \tag{17}$$

$$A^* A = \int_X \langle \cdot, \gamma_x \rangle_{\mathcal{H}} \gamma_x \, d\nu(x), \tag{18}$$

where $\phi \in L^2(X, \nu)$, the former integral converges in norm and the latter one in trace norm.

Proof. The proof is standard and we recall it for completeness. First of all, we show that, if $f \in \mathcal{H}$, Af is measurable. If $f = \gamma_t$ for some $t \in X$, the claim follows by the fact that Γ is measurable. Let now $\mathcal{H}_\gamma = \overline{\text{span}}\{\gamma_x \mid x \in X\}$, where $\overline{\text{span}}$ denotes the closure of the linear span. If $f \in \mathcal{H}_\gamma$, Af is the pointwise limit of a sequence of linear combinations of measurable functions of the form $A\gamma_{t_i}$, so that it is measurable. Finally, if $f \in \mathcal{H}_\gamma^\perp$, then $(Af)(x) = 0$ for all $x \in X$ and the claim is trivial.

We now prove that Af is in $L^2(X, \nu)$. The Cauchy-Schwarz inequality gives

$$\sup_{x \in X} |(Af)(x)| \leq \kappa \|f\|_{\mathcal{H}},$$

where $\kappa = \sup_{x \in X} \sqrt{\Gamma(x, x)}$ is finite since Γ is bounded. Then, recalling that ν is a finite measure, we have that $Af \in L^2(X, \nu)$ and

$$\|Af\|_{L^2(X, \nu)}^2 \leq \nu(X) \kappa^2 \|f\|_{\mathcal{H}}^2.$$

This last equation implies that A is a bounded operator from \mathcal{H} into $L^2(X, \nu)$. We will prove that A is a Hilbert-Schmidt operator as a consequence of Equation (18).

We now prove Equation (17). Since \mathcal{H} is separable and γ is weakly measurable, then γ is strong measurable and, if $\phi \in L^2(X, \nu)$, $\phi\gamma$ is strong measurable. Moreover, for all $x \in X$,

$$\|\phi(x)\gamma_x\|_{\mathcal{H}} = |\phi(x)|\sqrt{\Gamma(x, x)} \leq \kappa |\phi(x)|.$$

Since ν is finite, it follows that ϕ is in $L^1(X, \nu)$ and, hence, $\phi\gamma$ is integrable as an \mathcal{H} -valued function. Finally, for all $f \in \mathcal{H}$,

$$\int_X \phi(x) \langle \gamma_x, f \rangle_{\mathcal{H}} d\nu(x) = \langle \phi, Af \rangle_{L^2(X, \nu)} = \langle A^* \phi, f \rangle_{\mathcal{H}},$$

so, by uniqueness of the integral, Equation (17) holds.

Equation (18) is a consequence of Equation (17) and the fact that the integral commutes with the scalar product.

We now prove that A is a Hilbert-Schmidt operator. Let $(e_n)_{n \in \mathbb{N}}$ be a Hilbert basis of \mathcal{H} . Since A^*A is positive and $|\langle \gamma(\cdot), e_n \rangle_{\mathcal{H}}|^2$ is a positive function, by monotone convergence theorem we have that

$$\begin{aligned} \text{Tr}(A^*A) &= \sum_n \int_X |\langle e_n, \gamma_x \rangle_{\mathcal{H}}|^2 d\nu(x) \\ &= \int_X \sum_n |\langle e_n, \gamma_x \rangle_{\mathcal{H}}|^2 d\nu(x) \\ &= \int_X \langle \gamma_x, \gamma_x \rangle_{\mathcal{H}} d\nu(x) \leq \kappa \nu(X) < +\infty, \end{aligned}$$

and the thesis follows. ■

We call A the Carleman operator associated with the map γ (Halmos and Sunder (1978) give a weaker definition of Carleman operator).

4. Discretization of Carleman operators

In this section we study the discretization of the inverse problem $Af = g$ where A is the Carleman operator defined in the previous section. Given the exact datum $g \in L^2(X, \nu)$, the problem $Af = g$ amounts to find $f \in \mathcal{H}$ such that

$$\langle f, \gamma_x \rangle_{\mathcal{H}} = g(x) \quad x \in X.$$

In particular, if P denotes the orthogonal projection on $\overline{\text{Im } A}$ and $Pg \in \text{Im } A$, then

$$(Pg)(x) = \langle f^\dagger, \gamma_x \rangle_{\mathcal{H}} \quad x \in X, \tag{19}$$

where $f^\dagger = A^\dagger g$ is the generalized solution.

A natural way of discretizing the above problem is to replace the measure ν by a discrete measure so that integrals become weighted sums *à la* Cauchy-Riemann.

More precisely, given $\ell \in \mathbb{N}$, we consider a ℓ -sample \mathbf{z} of $X \times \mathbb{R}$, that is, a set of ℓ couples

$$((x_1, y_1), \dots, (x_\ell, y_\ell)) = (\mathbf{x}, \mathbf{y}) = \mathbf{z},$$

where $x_i \in X$ and $y_i \in \mathbb{R}$. We replace X by the finite set

$$I = \{1, \dots, \ell\},$$

and replace $L^2(X, \nu)$ by the finite dimensional Hilbert space $\mathcal{Z}_{\mathbf{z}} = \mathbb{R}^\ell$ endowed with the scalar product

$$\langle \mathbf{y}, \mathbf{y}' \rangle_{\mathcal{Z}_{\mathbf{z}}} = \sum_{i=1}^{\ell} a_i y_i y'_i,$$

where $a_i \in \mathbb{R}_+$ are chosen as suitable functions of the sample \mathbf{z} .

We approximate the operator A by the operator $A_{\mathbf{x}}$ from \mathcal{H} to $\mathcal{Z}_{\mathbf{z}}$ defined by

$$(A_{\mathbf{x}}f)_i = \langle f, \gamma_{x_i} \rangle_{\mathcal{H}} \quad i \in I,$$

and the exact datum g by the vector

$$\mathbf{y} = (y_1, \dots, y_\ell) \in \mathcal{Z}_{\mathbf{z}}.$$

The following proposition recalls the main properties of the operator $A_{\mathbf{x}}$.

Proposition 3 *The operator $A_{\mathbf{x}}$ is a bounded operator from \mathcal{H} into $\mathcal{Z}_{\mathbf{z}}$ and*

$$A_{\mathbf{x}}^* A_{\mathbf{x}} = \sum_{i=1}^{\ell} a_i \langle \cdot, \gamma_{x_i} \rangle_{\mathcal{H}} \gamma_{x_i}, \quad (20)$$

$$A_{\mathbf{x}}^* \mathbf{y} = \sum_{i=1}^{\ell} a_i y_i \gamma_{x_i} \quad \forall \mathbf{y} \in \mathcal{Z}_{\mathbf{z}}. \quad (21)$$

Proof. We have $\mathcal{Z}_{\mathbf{z}} = L^2(I, \nu_{\mathbf{x}})$, where $\nu_{\mathbf{x}}$ is the measure on I given by

$$\nu_{\mathbf{x}} = \sum_{i=1}^{\ell} a_i \delta_i,$$

and δ_i is the Dirac measure at the point $i \in I$. Therefore $A_{\mathbf{x}}$ is the Carleman operator associated to the map

$$I \ni i \mapsto \gamma_{x_i} \in \mathcal{H}.$$

Mimicking the proof of Proposition 2 and replacing integrals by sums, the thesis follows. \blacksquare

According to the notation of Section 2, if $\lambda > 0$, we denote by

$$\begin{aligned} f_0^\lambda &= (A^*A + \lambda)^{-1} A^*g \\ f_{\mathbf{z}}^\lambda &= (A_{\mathbf{x}}^*A_{\mathbf{x}} + \lambda)^{-1} A_{\mathbf{x}}^* \mathbf{y}, \end{aligned}$$

the regularized solutions of exact and discrete problems, respectively, where we add the subscript \mathbf{z} to emphasize the dependence of the solution on the data.

The explicit form of $f_{\mathbf{z}}^\lambda$ amounts to solving a linear problem. Indeed, if $\Gamma_{\mathbf{x}}$ is the $\ell \times \ell$ matrix with entries

$$(\Gamma_{\mathbf{x}})_{ij} = \Gamma(x_i, x_j) = \langle \gamma_{x_j}, \gamma_{x_i} \rangle_{\mathcal{H}}$$

then, by a standard computation (Wahba 1977),

$$f_{\mathbf{z}}^\lambda = \sum_{i,j=1}^{\ell} a_j \gamma_{x_j} \left((\Gamma_{\mathbf{x}} + \lambda)^{-1} \right)_{ji} y_i. \quad (22)$$

Applying the results of Section 2, we propose a bound for the reconstruction error $\|f_{\mathbf{z}}^\lambda - f^\dagger\|_{\mathcal{H}}$ or the residue $\|Af_{\mathbf{z}}^\lambda - g\|_{L^2(X,\nu)}$. To this aim we have to assume some hypotheses on the relation between the measures ν and $\sum_{i=1}^{\ell} a_i \delta_{x_i}$ and between the data g and \mathbf{y} . We discuss two different settings.

4.1. A deterministic discretization

In this section, we consider a framework where the measure ν 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 (see also Smale and Zhou (2004a)).

Moreover, we study the reconstruction error of the approximated solution. To this aim, we assume that $g \in \text{Im } A$ so that, by Equation (19), we can restate the hypothesis that the noise is zero by the fact that

$$y_i = g(x_i) = \langle f^\dagger, \gamma_{x_i} \rangle_{\mathcal{H}} \quad \forall i \in I. \quad (23)$$

Moreover, we consider a family of measurable sets $X_1, \dots, X_\ell \subset X$ such that

- (i) $x_i \in X_i$ for all $i \in I$;
- (ii) $\nu(X_i \cap X_j) = 0$ for all $i \neq j$;
- (iii) $\cup_i X_i = X$.

We let $a_i = \nu(X_i)$ and recall that $\kappa = \sup_{x \in X} \sqrt{\Gamma(x, x)}$. Then we have the following result.

Corollary 3 *If $\lambda > 0$, then*

$$\left| \|f_{\mathbf{z}}^\lambda - f^\dagger\|_{\mathcal{H}} - \|f_0^\lambda - f^\dagger\|_{\mathcal{H}} \right| = \|f^\dagger\|_{\mathcal{H}} \kappa c(\ell) \alpha \left(\frac{2}{\lambda} + \frac{\kappa \sqrt{\alpha}}{\lambda^{\frac{3}{2}}} \right). \quad (24)$$

where $\alpha = \nu(X)$ and

$$\begin{aligned} c(\ell) &= \max_{i \in I} \left(\sup_{x \in X_i} \|\gamma_x - \gamma_{x_i}\|_{\mathcal{H}} \right) \\ &= \max_{i \in I} \left(\sup_{x \in X_i} \sqrt{\Gamma(x, x) - 2\Gamma(x, x_i) + \Gamma(x_i, x_i)} \right) \end{aligned} \quad (25)$$

Proof. We claim that

$$\|A^*g - A_{\mathbf{x}}^*\mathbf{y}\|_{\mathcal{H}} \leq 2\|f^\dagger\|_{\mathcal{H}} \alpha \kappa c(\ell) \quad (26)$$

$$\|A^*A - A_{\mathbf{x}}^*A_{\mathbf{x}}\|_{\mathcal{L}(\mathcal{H})} \leq 2\alpha \kappa c(\ell) \quad (27)$$

We first prove Equation (27). By definition of X_i and a_i and Equations (18) and (20), we have that

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

so that Equations (27) is proved. By Equations (19), (17), (20), it follows that

$$\|A^*Ag - A_{\mathbf{x}}^*A_{\mathbf{x}}\mathbf{y}\|_{\mathcal{H}} = \|(A^*A - A_{\mathbf{x}}^*A_{\mathbf{x}})f^\dagger\|_{\mathcal{H}},$$

so that Equation (26) is clear.

We observe that, by Equation (23),

$$\begin{aligned}
 \|\mathbf{y}\|_{\mathcal{Z}_{\mathbf{z}}}^2 &= \sum_{i=1}^{\ell} a_i \langle f^\dagger, \gamma_{x_i} \rangle_{\mathcal{H}}^2 \\
 &\leq \sum_{i=1}^{\ell} \nu(X_i) \|f^\dagger\|_{\mathcal{H}}^2 \|\gamma_{x_i}\|_{\mathcal{H}}^2 \\
 &\leq \alpha \|f^\dagger\|_{\mathcal{H}}^2 \kappa^2 = M^2,
 \end{aligned}$$

so that, replacing in Equation (10) the above bounds on M , δ_1 and δ_2 the inequality (24) follows. ■

In Equation (24) we need an a-priori bound on $\|f^\dagger\|_{\mathcal{H}}$. However, we can obtain a (worse) estimate of $\|f_{\mathbf{z}}^\lambda - f^\dagger\|_{\mathcal{H}}$ depending on $\|g\|_{\mathcal{Z}} = \|Af^\dagger\|_{\mathcal{Z}}$.

Let $\omega(\lambda) = \|f_0^\lambda - f^\dagger\|_{\mathcal{H}}$ be the reconstruction error of the regularized solution of the exact problem $Af = g$ and $\lambda_\ell = \lambda(\mathbf{z}, \ell)$ be a parameter choice rule so that $\lim_{\ell \rightarrow \infty} \omega(\lambda_\ell) = 0$ then, as a consequence of Equation (24),

$$\|f_{\mathbf{z}}^{\lambda_\ell} - f^\dagger\|_{\mathcal{H}} = \omega(\lambda_\ell) + O\left(\frac{c(\ell)}{\lambda_\ell^{\frac{3}{2}}}\right).$$

A sufficient condition ensuring that the approximated solution $f_{\mathbf{z}}^{\lambda_\ell}$ approaches to the exact solution f^\dagger is that $\lim_{\ell \rightarrow \infty} \frac{c(\ell)^{\frac{2}{3}}}{\lambda_\ell} = 0$.

4.1.1. 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 $g : [0, 1] \rightarrow \mathbb{R}$, 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

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

The space $H^1([0, 1])$ can be replaced by any Hilbert space \mathcal{H} (of functions on $[0, 1]$) having a continuous immersion into $L^2([0, 1], dx)$.

We define $A : \mathcal{H} \rightarrow L^2([0, 1], dx)$ 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 $f^\dagger = A^\dagger g = g'$ for all $g \in \text{Im } A$. 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

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

Hence A is the Carleman operator associated to the map γ

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

and we can apply the result of Section 4.1 with $X = [0, 1]$, $\nu = dx$, $\mathcal{H} = H^1([0, 1])$. To this aim, we notice that the kernel Γ is given by

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

which is bounded and measurable. As usual, we choose the points $x_i = \frac{i}{\ell}$ for all $i = 0, \dots, \ell$ and $X_i = [x_{i-1}, x_i]$.

If $\lambda > 0$, $f_{\mathbf{z}}^\lambda$ is the regularized solution of the discrete problem

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

where $f \in H^1([0, 1])$. According to Equation (22), $f_{\mathbf{z}}^\lambda$ is a linear combination of the functions γ_{x_i} , that are quadratic splines: piecewise polynomials of degree two with continuous derivative (Wahba 1990). From a numerical point of view, the computation of $f_{\mathbf{z}}^\lambda$ reduces to compute the inverse of the $\ell \times \ell$ 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 (24), we notice that $\alpha = \nu([0, 1]) = 1$ and, if $0 \leq t \leq x \leq 1$,

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

It follows that $c(\ell) = \frac{\sqrt{2}}{\ell}$ and, letting $t = 0$, that $\kappa = \sqrt{2}$. Replacing these bounds in Equation (24) we obtain that

$$|\|f_{\mathbf{z}}^\lambda - f^\dagger\|_{H^1([0,1])} - \|f_0^\lambda - f^\dagger\|_{H^1([0,1])}| = 2\sqrt{2}\|g'\|_{H^1([0,1])} \frac{1}{\ell} \left(\frac{\sqrt{2}}{\lambda} + \frac{1}{\lambda^{\frac{3}{2}}} \right).$$

In particular, if $\lambda_\ell = \lambda(\mathbf{z}, \ell)$ is a parameter choice rule such that $\lambda_\ell = O(\frac{1}{\ell^b})$ with $b < \frac{2}{3}$, then the approximated solution $f_{\mathbf{z}}^{\lambda_\ell}$ approaches the exact solution $f^\dagger = g'$ in $H^1([0, 1])$. This implies that $f_{\mathbf{z}}^{\lambda_\ell}$ converges to g' uniformly on $[0, 1]$.

4.2. Stochastic discretization

In this section we consider a framework where the measure ν is unknown and the points (x_i, y_i) of the sample \mathbf{z} are drawn identically and independently distributed according to some probability distribution ρ . We assume the following facts:

- (i) the marginal distribution of ρ on X is ν (this implies in particular that ν is normalized to 1);
- (ii) if $\rho(y|x)$ denotes the conditional probability distribution of y given $x \in X$, then

$$g(x) = \int_{\mathbb{R}} y \, d\rho(y|x).$$

In this context we, let $a_i = \frac{1}{\ell}$ so that $\frac{1}{\ell} \sum_{i=1}^{\ell} \delta_{x_i}$ is the empirical measure of ν associated to the set $\{x_1, \dots, x_\ell\}$.

The aim is to give a probabilistic bound of the residue, that is, a bound on

$$\text{Prob} \left\{ \mathbf{z} \in (X \times \mathbb{R})^\ell \mid \|A f_{\mathbf{z}}^\lambda - P g\|_{L^2(X, \nu)} \leq \epsilon \right\},$$

where P is the orthogonal projection on the closure of $\text{Im } A$. A similar estimate can be obtained for the reconstruction error. We bound the residue since in Section 5 it will provide the consistency of learning algorithms.

To avoid technical problems, we assume that there is $L > 0$ such that $\text{supp } \rho(\cdot|x) \subset [-L, L] = Y$ for ν -almost all $x \in X$. In particular, with probability 1, any sample $\mathbf{z} = ((x_1, y_1), \dots, (x_\ell, y_\ell))$ is such that $|y_i| \leq L$.

The following proposition gives a probabilistic estimate of the noise levels (De Vito et al. 2004).

Proposition 4 *Let $\epsilon_1, \epsilon_2 > 0$ and $\kappa = \sup_{x \in X} \|\gamma_x\| = \sup_{x \in X} \sqrt{\Gamma(x, x)}$, then*

$$\begin{aligned} \text{Prob} \left\{ \mathbf{z} \in (X \times \mathbb{R})^\ell \mid \|A^* g - A_{\mathbf{x}}^* \mathbf{y}\|_{\mathcal{H}} \leq \frac{L\kappa}{\sqrt{\ell}} + \epsilon_1, \|A^* A - A_{\mathbf{x}}^* A_{\mathbf{x}}\|_{\mathcal{L}(\mathcal{H})} \leq \frac{\kappa^2}{\sqrt{\ell}} + \epsilon_2 \right\} \\ \geq 1 - e^{-\frac{\epsilon_1^2 \ell}{2\kappa^2 L^2}} - e^{-\frac{\epsilon_2^2 \ell}{2\kappa^4}} \end{aligned}$$

Proof. The idea of the proof is to apply McDiarmid inequality (McDiarmid 1989) to the random variables

$$F(\mathbf{z}) = \|A_{\mathbf{x}}^* \mathbf{y} - A^* g\|_{\mathcal{H}} \quad G(\mathbf{z}) = \|A_{\mathbf{x}}^* A_{\mathbf{x}} - A^* A\|_{\mathcal{L}(\mathcal{H})}.$$

This inequality ensures that, given $\epsilon > 0$,

$$\text{Prob} \left\{ \mathbf{z} \in (X \times \mathbb{R})^\ell \mid F(\mathbf{z}) \geq \mathbb{E}(F) + \epsilon \right\} \leq e^{-\frac{2\epsilon^2}{\sum_{i=1}^{\ell} c_i^2}}, \quad (28)$$

where $\mathbb{E}(F)$ is the expectation value of F , $c_i \geq \sup_{z, z^i} |F(z) - F(z^i)|$ and z^i is the training set with the i^{th} example being replaced by (\mathbf{x}', y') (a similar equation holds for G).

Let $Y = [-L, L]$ and $\varphi : X \times Y \rightarrow \mathcal{H}$

$$\varphi(x, y) = y\gamma_x.$$

Since Γ is a bounded measurable function and $|y| \leq L$, φ is measurable and

$$\|\varphi(x, y)\| \leq L\kappa$$

so that it is integrable and

$$\begin{aligned} \int_{X \times Y} \varphi(x, y) \, d\rho(x, y) &= \int_X \gamma_x \left(\int_Y y \, d\rho(y|x) \right) \, d\nu(x) \\ &= \int_X g(x)\gamma_x \, d\nu(x) \\ &= A^*g \end{aligned}$$

Let now $\Phi : (X \times Y)^\ell \rightarrow \mathcal{H}$

$$\Phi(\mathbf{z}) = \frac{1}{\ell} \sum_{i=1}^{\ell} \varphi(x_i, y_i) = A_{\mathbf{x}}^* \mathbf{y}.$$

Since the samples are drawn i.i.d., it follows that

$$\begin{aligned} \int_{(X \times Y)^\ell} \Phi(\mathbf{z}) \, d\rho^\ell(\mathbf{z}) &= A^*g \\ \frac{1}{\ell} \int_{X \times Y} \|\varphi(x, y)\|^2 \, d\rho(x, y) &= \int_{(X \times Y)^\ell} \|\Phi(\mathbf{z}) - A^*g\|^2 \, d\rho^\ell(\mathbf{z}) + \frac{1}{\ell} \|A^*g\|^2 \end{aligned}$$

so that, by Cauchy-Schwarz inequality,

$$\begin{aligned} \mathbb{E}(F) &= \int_{(X \times Y)^\ell} \|\Phi(\mathbf{z}) - A^*g\| \, d\rho^\ell(\mathbf{z}) \tag{29} \\ &\leq \sqrt{\int_{(X \times Y)^\ell} \|\Phi(\mathbf{z}) - A^*g\|^2 \, d\rho^\ell(\mathbf{z})} \\ &\leq \frac{1}{\sqrt{\ell}} \sqrt{\int_{X \times Y} \|\varphi(x, y)\|^2 \, d\rho(x, y)} \\ (\|\varphi(x, y)\| \leq L\kappa) &\leq \frac{L\kappa}{\sqrt{\ell}}. \tag{30} \end{aligned}$$

Moreover, using the triangular inequality,

$$\begin{aligned} |F(\mathbf{z}) - F(\mathbf{z}^i)| &= \left| \|A_{\mathbf{x}}^* \mathbf{y} - A^*g\| - \|A_{\mathbf{x}^i}^* h_{\mathbf{z}^i} - A^*g\| \right| \\ &\leq \|A_{\mathbf{x}}^* \mathbf{y} - A_{\mathbf{x}^i}^* h_{\mathbf{z}^i}\| \\ &= \frac{1}{\ell} \|y_i \gamma_{x_i} - y'_i \gamma_{x'_i}\| \\ &\leq \frac{2L\kappa}{\ell} \end{aligned}$$

so that we can choose

$$c_i = \frac{2L\kappa}{\ell}. \tag{31}$$

Finally, we observe that, given $\epsilon_1 > 0$, by Equation (30),

$$\begin{aligned} \text{Prob} \left\{ \mathbf{z} \in (X \times \mathbb{R})^\ell \mid F(\mathbf{z}) \geq \frac{L\kappa}{\sqrt{\ell}} + \epsilon_1 \right\} &\leq \text{Prob} \left\{ \mathbf{z} \in (X \times \mathbb{R})^\ell \mid F(\mathbf{z}) \geq \mathbb{E}(F) + \epsilon_1 \right\} \\ &\quad (\text{Eqs. (28), (31)}) \leq e^{-\frac{\epsilon_1^2 \ell}{2\kappa^2 L^2}} \end{aligned}$$

To prove the second part, we can mimic the same proof, observing that

$$\|A_{\mathbf{x}^*} A_{\mathbf{x}} - A^* A\|_{\mathcal{L}(\mathcal{H})} \leq \|A_{\mathbf{x}^*} A_{\mathbf{x}} - A^* A\|_2,$$

where $\|A_{\mathbf{x}^*} A_{\mathbf{x}} - A^* A\|_2$ is the Hilbert-Schmidt norm in the Hilbert space of the Hilbert-Schmidt operators (we use the Hilbert-Schmidt norm since we need to assume that G takes value in a Hilbert space). ■

From the above proposition, we can deduce that

Proposition 5 *Given $0 < \eta < 1$, with probability greater than $1 - \eta$,*

$$\left| \|Af_{\mathbf{z}}^\lambda - Pg\|_{L^2(X,\nu)} - \|Af_0^\lambda - Pg\|_{L^2(X,\nu)} \right| \leq \frac{\kappa L}{2\sqrt{\ell}} \left(\frac{1}{\sqrt{\lambda}} + \frac{\kappa}{2\lambda} \right) \left(1 + \log \sqrt{\frac{4}{\eta}} \right) \quad (32)$$

for all $\lambda > 0$.

If we choose, according to some parameter choice rule, $\lambda = \lambda(\mathbf{z}, \ell) = O(\frac{1}{\ell^b})$, with $0 < b < \frac{1}{2}$, and we let $f_{\mathbf{z},\ell} = f_{\mathbf{z}}^{\lambda(\ell)}$ and $f_{0,\ell} = f_0^{\lambda(\ell,\mathbf{z})}$, then, in probability,

$$\|Af_{\mathbf{z},\ell} - Pg\|_{L^2(X,\nu)} = \|Af_{0,\ell} - Pg\|_{L^2(X,\nu)} + O\left(\frac{1}{\sqrt{\ell^{1-2b}}}\right).$$

Proof. We observe that

$$\|\mathbf{y}\|_{\mathcal{Z}_{\mathbf{z}}}^2 = \frac{1}{\ell} \sum_{i=1}^{\ell} y_i^2 \leq L^2.$$

The thesis follows from Equation (11) with the choice $M = L$ and from the above proposition with the choice

$$\begin{aligned} e^{-\frac{\epsilon_1^2 \ell}{2\kappa^2 L^2}} &= \frac{\eta}{2} \\ e^{-\frac{\epsilon_2^2 \ell}{2\kappa^4}} &= \frac{\eta}{2} \end{aligned}$$

■

We stress that the set of samples \mathbf{z} such that Equation (32) holds depends on ℓ and η , but does not depend on λ . This is a consequence of the fact that the analytic dependence of the bound on λ is due to regularization, whereas its probabilistic dependence on the sample \mathbf{z} is due to the estimate of the noise levels δ_1 and δ_2 .

5. Learning from examples

In this section we show that the learning from examples can be set in the context of a stochastic discretization of a Carleman operator. To have a linear problem, we treat only the regression setting with quadratic loss function and we study the regularized least squares algorithm (for an account of learning theory and its application see Vapnik

(1998), Evgeniou et al. (2000), Cucker and Smale (2002b), Poggio and Smale (2003) and references therein). Some results in the same spirit can be found in Cucker and Smale (2002a), Mukherjee et al. (2002), Poggio et al. (2004), Smale and Zhou (2004b), Rudin (2004) and Kurkova (2004).

In the framework of learning theory, there are two sets, namely, the input space and the output space. The former is a compact separable metric space X , the latter is \mathbb{R} . The relation between the input $x \in X$ and the output $y \in \mathbb{R}$ is probabilistic and it is described by a probability distribution ρ on $X \times \mathbb{R}$. We denote by ν the marginal distribution of ρ on X and by $\rho(y|x)$ the conditional probability of y given $x \in X$. We assume that there is $L > 0$ such that $\text{supp } \rho(\cdot|x) \subset [-L, L] = Y$ for ν -almost all $x \in X$.

The distribution ρ is known only through a sample $\mathbf{z} = (\mathbf{x}, \mathbf{y}) = ((x_1, y_1), \dots, (x_\ell, y_\ell))$, called *training set*, drawn independently and identically distributed according to ρ . The goal of learning is, given a sample \mathbf{z} , to find a function $f_{\mathbf{z}} : X \rightarrow \mathbb{R}$ such that $f_{\mathbf{z}}(x)$ is an *estimate* of the output y when a new input x is given. The function $f_{\mathbf{z}}$ is called *estimator* and the rule provides $f_{\mathbf{z}}$ for any given sample \mathbf{z} is called *learning algorithm*.

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

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

which is finite if and only if $f \in L^2(X, \nu)$. The minimizer of the expected risk over the set of all measurable functions on X is the regression function

$$g(x) = \int_{\mathbb{R}} y d\rho(y|x) \quad x \in X,$$

(since the support of ρ is contained in the compact subset $X \times [-L, L]$ the regression function g is well defined and belongs to $L^2(X, \nu)$). However, such a regression function cannot be reconstructed starting from the sample \mathbf{z} since the data are finite and noisy.

To overcome this problem, in learning theory we choose a hypotheses space \mathcal{H} , which is a reproducing kernel Hilbert space on X , (Aronszajn 1950, Schwartz 1964). We recall that \mathcal{H} is a Hilbert space of real functions on X such that, if $x \in X$, there is $\gamma_x \in \mathcal{H}$ satisfying

$$f(x) = \langle f, \gamma_x \rangle_{\mathcal{H}} \quad \forall f \in \mathcal{H}. \tag{33}$$

We assume that the kernel $\Gamma(x, t) = \langle \gamma_t, \gamma_x \rangle_{\mathcal{H}}$ is a bounded measurable function on $X \times X$, so that the expected risk of any $f \in \mathcal{H}$ is finite.

Given $\lambda > 0$, in the regularized least squares algorithm, the estimator $f_{\mathbf{z}}^\lambda$ is defined as the minimizer on \mathcal{H} of the regularized least squares functional (Cucker and Smale 2002b)

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

In learning theory, the regularization parameter λ has to be chosen, as a function of ℓ , in such a way that the algorithm is consistent (Vapnik 1998), that is, for all $\epsilon > 0$

$$\lim_{\ell \rightarrow +\infty} \text{Prob} \left\{ \mathbf{z} \in (X \times \mathbb{R})^\ell \mid I[f_{\mathbf{z}}^{\lambda_\ell}] - \inf_{f \in \mathcal{H}} I[f] \geq \epsilon \right\}.$$

Notice that, in general, $\inf_{f \in \mathcal{H}} I[f]$ is greater than $I[g]$ and represents a sort of irreducible error associated to the choice of the space \mathcal{H} .

We now show that the above problem is the discretization of an inverse problem defined by a Carleman operator. Indeed, following Cucker and Smale (2002b), the expected risk of $f \in \mathcal{H}$ can be rewritten as

$$\begin{aligned} I[f] &= \int_X (f(x) - g(x))^2 d\nu(x) + \int_{X \times \mathbb{R}} (y - g(x))^2 d\rho(x, y) \\ &= \|Af - g\|_{L^2(X, \nu)}^2 + I[g], \end{aligned}$$

where $I[g] = \int_{X \times \mathbb{R}} (y - g(x))^2 d\rho(x, y)$ and A is the canonical immersion of \mathcal{H} into $L^2(X, \nu)$. The operator A is well defined since Γ is bounded and measurable, and Equation (33) implies that A is the Carleman operator associated with the map γ

$$x \ni X \mapsto \gamma_x \in \mathcal{H}.$$

We are now in position to apply the result of Section 4.2. Given a training set \mathbf{z} , let $A_{\mathbf{x}}$ be the discretized version of A , that is,

$$(A_{\mathbf{x}}f)_i = \langle f, \gamma_{x_i} \rangle_{\mathcal{H}} = f(x_i),$$

then

$$\frac{1}{\ell} \sum_{i=1}^{\ell} (f(x_i) - y_i)^2 = \|A_{\mathbf{x}}f - \mathbf{y}\|_{\mathcal{Z}_{\mathbf{z}}}^2,$$

so that the estimator $f_{\mathbf{z}}^{\lambda}$ given by the regularized least squares algorithm is the regularized solution *à la* Tikhonov of the discrete problem $A_{\mathbf{x}}f = \mathbf{y}$ (in the context of learning theory the functional $\frac{1}{\ell} \sum_{i=1}^{\ell} (f(x_i) - y_i)^2$ is called *empirical risk* of f).

Finally, since P is the orthogonal projection on the closure of \mathcal{H} in $L^2(X, \nu)$, by definition of projection,

$$I[f] - \inf_{f \in \mathcal{H}} I[f] = \|Af - Pg\|_{L^2(X, \nu)}^2,$$

for all $f \in \mathcal{H}$. The above equation clarifies that in learning theory we have to study the convergence of the residue instead of the reconstruction error, as in the theory of inverse problems.

As a consequence of Proposition 5, we have the following bound on the performance of the learning algorithm: with probability greater than $1 - \eta$,

$$I[f_{\mathbf{z}}^{\lambda}] \leq I[f_0^{\lambda}] + \frac{\kappa L}{2\sqrt{\ell}} \left(\frac{1}{\sqrt{\lambda}} + \frac{\kappa}{2\lambda} \right) \left(1 + \log \sqrt{\frac{4}{\eta}} \right) \quad (34)$$

for all $\lambda > 0$.

Let now choose the parameter $\lambda_\ell = \lambda(\mathbf{z}, \ell)$ such that $\lambda_\ell = O(\frac{1}{\ell^b})$, with $0 < b < \frac{1}{2}$, then, in probability,

$$\lim_{\ell \rightarrow \infty} I[f_{\mathbf{z}}^{\lambda(\mathbf{z}, \ell)}] = \inf_{f \in \mathcal{H}} I[f],$$

showing that the regularized least square algorithm is consistent.

We notice that the set of samples such that Equation (34) holds, depends on ℓ and η , but does not depend on λ . This last fact allows us to consider a posteriori parameter choice rule $\lambda_\ell = \lambda(\mathbf{z}, \ell)$ which is problematic in general due to the probabilistic setting of learning theory (see De Vito et al. (2004), Devroye et al. (1996)).

Acknowledgments

We would like to thank M. Bertero, U. De Giovannini, C. De Mol, M. Piana, T. Poggio, G. Talenti and A. Verri for useful discussions and suggestions. This research has been partially funded by the INFM Project MAIA, the FIRB Project ASTAA and the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778.

References

- Aronszajn N 1950 Theory of reproducing kernels *Trans. Amer. Math. Soc.* **68** 337–404
- Bertero M, De Mol C and Pike E R 1985 Linear inverse problems with discrete data. I. General formulation and singular system analysis *Inverse Problems* **1**(4) 301–330
- Bertero M, De Mol C and Pike E R 1988 Linear inverse problems with discrete data. II. Stability and regularisation *Inverse Problems* **4**(3) 573–594
- Cucker F and Smale S 2002a Best choices for regularization parameters in learning theory: on the bias-variance problem *Found. Comput. Math.* **2**(4) 413–428
- Cucker F and Smale S 2002b On the mathematical foundations of learning *Bull. Amer. Math. Soc. (N.S.)* **39**(1) 1–49 (electronic)
- De Vito E, Caponnetto A and Rosasco L 2004 Model selection for regularized least-squares algorithm in learning theory to appear in *Foundations Computational Mathematics*
- Devroye L, Györfi L and Lugosi G 1996 *A Probabilistic Theory of Pattern Recognition* (New York: Springer)
- Engl H W, Hanke M and Neubauer A 1996 *Regularization of inverse problems* Vol. 375 of *Mathematics and its Applications* (Dordrecht: Kluwer Academic Publishers Group)
- Evgeniou T, Pontil M and Poggio T 2000 Regularization networks and support vector machines *Adv. Comput. Math.* **13**(1) 1–50
- Groetsch C W 1984 *The theory of Tikhonov regularization for Fredholm equations of the first kind* Vol. 105 of *Research Notes in Mathematics* (Boston, MA: Pitman, Advanced Publishing Program)
- Groetsch C W 1990 Convergence analysis of a regularized degenerate kernel method for Fredholm integral equations of the first kind *Integral Equations Operator Theory* **13**(1) 67–75
- Halmos P R and Sunder V S 1978 *Bounded integral operators on L^2 spaces* Vol. 96 of *Ergebnisse der Mathematik und ihrer Grenzgebiete* (Berlin: Springer-Verlag)
- Kress R 1999 *Linear integral equations* Vol. 82 of *Applied Mathematical Sciences* second edn (New York: Springer-Verlag)
- Kurkova V 2004 Supervised learning as an inverse problem Technical Report 960 Institute of Computer Science, Academy of Sciences of the Czech Republic

- Mathé P and Pereverzev S V 2002 Moduli of continuity for operator valued functions *Numer. Funct. Anal. Optim.* **23**(5-6) 623–631
- Mathé P and Pereverzev S V 2003 Discretization strategy for linear ill-posed problems in variable Hilbert scales *Inverse Problems* **19**(6) 1263–1277
- McDiarmid C 1989 On the method of bounded differences in ‘Surveys in combinatorics, 1989 (Norwich, 1989)’ Vol. 141 of *London Math. Soc. Lecture Note Ser.* Cambridge Univ. Press Cambridge pp. 148–188
- Mukherjee S, Rifkin R and Poggio T 2002 Regression and classification with regularization in D Denison, M Hansen, C Holmes, B Mallick and B Yu, eds, ‘Lectures Notes in Statistics: Nonlinear Estimation and Classification’ Springer-Verlag pp. 107–124. Proceedings from MSRI Workshop
- Nair M T 1994 A unified approach for regularized approximation methods for Fredholm integral equations of the first kind *Numer. Funct. Anal. Optim.* **15**(3-4) 381–389
- Nair M T and Schock E 1998 A discrepancy principle for Tikhonov regularization with approximately specified data *Ann. Polon. Math.* **69**(3) 197–205
- Pereverzev S and Schock E 2000 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
- Plato R and Vainikko G 1990 On the regularization of projection methods for solving ill-posed problems *Numer. Math.* **57**(1) 63–79
- Poggio T, Rifkin R, Mukherjee S and Niyogi P 2004 General conditions for predictivity in learning theory *Nature* **428** 419–422
- Poggio T and Smale S 2003 The mathematics of learning: dealing with data *Notices Amer. Math. Soc.* **50**(5) 537–544
- Rajan M P 2003 Convergence analysis of a regularized approximation for solving Fredholm integral equations of the first kind *J. Math. Anal. Appl.* **279**(2) 522–530
- Rudin C 2004 A different type of convergence for statistical learning algorithms Technical report Program in Applied and Computational Mathematics Princeton University
- Saitoh S 1997 *Integral transforms, reproducing kernels and their applications* Vol. 369 of *Pitman Research Notes in Mathematics Series* (Harlow: Longman)
- Schwartz L 1964 Sous-espaces hilbertiens d’espaces vectoriels topologiques et noyaux associés (noyaux reproduisants) *J. Analyse Math.* **13** 115–256
- Smale S and Zhou D 2004a Shannon sampling and function reconstruction from point values *Bull. Amer. Math. Soc. (N.S.)* **39**(1) 1–49 (electronic)
- Smale S and Zhou D 2004b Shannon sampling II. Connections to learning theory *to appear*
- Tikhonov A N, Goncharsky A V, Stepanov V V and Yagola A G 1995 *Numerical methods for the solution of ill-posed problems* Vol. 328 of *Mathematics and its Applications* (Dordrecht: Kluwer Academic Publishers Group)
- Vapnik V N 1998 *Statistical learning theory* (New York: John Wiley & Sons Inc.)
- Wahba G 1977 Practical approximate solutions to linear operator equations when the data are noisy *SIAM J. Numer. Anal.* **14**(4) 651–667
- Wahba G 1990 *Spline models for observational data* Vol. 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics* (Philadelphia, PA: Society for Industrial and Applied Mathematics)