Dipartimento di Informatica e
Scienze dell'Informazione

Discretization Error Analysis for Tikhonov Regularization

A. Caponnetto, C. Carmeli, E. De Vito,

L. Rosasco, A. Toigo

Technical Report DISI-TR-04-04

DISI, Università di Genova
v. Dodecaneso 35, 16146 Genova, Italy http://www.disi.unige.it/

Abstract

We discuss the problem of discretization for a linear inverse problem exploiting a stability property of Tikhonov regularization.

1. Introduction

In this report we discuss the problem of discretization for a linear inverse problem exploiting a stability property of Tikhonov regularization.

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

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

where g is the datum belonging to $\mathcal{K}$. Discretization is a procedure that replaces the exact problem (1) 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 $\mathcal{Z}$ such that they are *approximations* of the model A and the datum g.

Usually, the space $\mathcal{Z}$ is a finite dimensional subspace of $\mathcal{K}$, B is the restriction of A to a finite dimensional subspace of $\mathcal{H}$, and h is a noisy projection of g on $\mathcal{Z}$.

Moreover, since Problem (1) is usually ill-posed and Problem (2) gives rise to unstable numerical solutions, a regularized version of Problem (2) is usually solved. For example, in the framework of Tikhonov regularization, the following minimization problem is considered

$$\min_{f\in\mathcal{H}} \left(\|Bf - h\|_{\mathcal{Z}}^2 + \lambda \|f\|_{\mathcal{H}}^2 \right),$$

where λ is a positive parameter whose chioce depends on the datum, on the noise level on the datum and on the model (for a review, see, for example, Bertero et al. (1985, 1988), Engl et al. (1996), Groetsch (1984) and references therein).

We adopt a different point of view. Following Groetsch (1990), Nair and Schock (1998), Rajan (2003), Mathé and Pereverzev (2003) and reference therein, 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 observing that the minimizer of the Tikhonov functional is given by

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

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 $\|B^*h - A^*g\|_{\mathcal{H}}$.

Section 2 is devoted to the formalization of the above idea in an abstract setting. In Section 3 we apply the results of Section 2 to the problem of discretization. In particular, we discuss two examples: the problem of recovering the derivative of a function g when a finite set of points $y_i = g(x_i)$ is given and the consistency of the regularized least square algorithm in the context of statistical learning theory. In Appendix we review the main properties of reproducing kernel Hilbert spaces and of integral operators using the unifying notion of Carleman operator.

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 the regularization parameter λ through a power law.

First of all we fix the notation. By Hilbert space we mean a separable Hilbert space over $\mathbb{K}$, where $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$. We denote by $\|\cdot\|$ and $\langle \cdot, \cdot \rangle$ the corresponding norm and scalar product (if $\mathbb{K} = \mathbb{C}$, we assume that $\langle \cdot, \cdot \rangle$ is linear in the first argument). 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.

We fix 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 a 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 $(A^*A + \lambda I)^{-1} A^*g \in \mathcal{H}$.

The following proposition studies the dependence of the minimizer of Tikhonov functional on the operator A and the datum g. In order to treat both the reconstruction error and the residuum 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, for all $\lambda > 0$*

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

then, for any $a \in [0,1]$, it holds

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

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,

$$\begin{aligned} (T+\lambda)^{-1} - (T_0+\lambda)^{-1} &= (T_0+\lambda)^{-1}[(T_0+\lambda) - (T+\lambda)](T+\lambda)^{-1} \\ &= (T_0+\lambda)^{-1}(T_0 - T)(T+\lambda)^{-1} \end{aligned} \tag{6}$$

The second one is a consequence of spectral theorem. For any $a \in [0,1]$

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

Indeed, since T is a positive operator,

$$\begin{aligned} \left\| T^a (T+\lambda)^{-1} \right\|_{\mathcal{L}(\mathcal{H})} &= \sup_{0 \leq t \leq \|T\|_{\mathcal{L}(\mathcal{H})}} \frac{t^a}{t+\lambda} \\ &= \frac{\tau^a}{\tau+\lambda} \end{aligned}$$

where τ is a maximizer of the function $\frac{t^a}{t+\lambda}$ on the interval $[0, \|T\|_{\mathcal{L}(\mathcal{H})}]$. Computing the derivative, one has that $\tau = \frac{\lambda a}{1-a}$, if $a < 1$, and $\tau = \|T\|_{\mathcal{L}(\mathcal{H})}$, if $a = 1$. Replacing τ in the above formula, Equation (7) follows.

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

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

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

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

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

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

The thesis is now clear. ■

Up to our knowledge, the first result in this direction was obtained in Theorem 2 of Wahba (1977) in the framework of integral equations with white noise. Our bound is of the same kind of the one obtained in Groetsch (1990) and Theorem 2.5 of Rajan (2003) (where only the reconstruction error with respect to the the uniform norm is considered). In Theorem 2.1 of Nair and Schock (1998) a similar framework is considered, but with a different choice of noisy levels. In Mathé and Pereverzev (2003), a larger class of regularization is considered (see remark below), however the noisy data is assumed to be in the *exact data* space $\mathcal{K}$, compare with Theorem 2.

Remark 2 *Let now consider regularizations of the problem $Af = g$ different from the Tikhonov regularization. It is known (see for example Engl et al. (1996), Groetsch (1984)) that a large class of regularized solutions is of the form*

$$f_0^\lambda = r_\lambda(A^*A)A^*g, \tag{9}$$

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

$$\lim_{\lambda\to 0} r_\lambda(t) = \frac{1}{t} \qquad |r_\lambda(t)t| \leq M \ \forall t, \lambda.$$

A simple check shows the above proof still holds 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 this point can be found in Mathé and Pereverzev (2003).

The above result suggests the following definition of *parameter choice rule*.

Definition 3 *Given $(\mathcal{K}, g, A) \in \mathcal{T}$, for any $M > 0$, $\delta = (\delta_1, \delta_2) \in \mathbb{R}_+^2$ we let*

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

be 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*.*

Remark 4 *In the above definition the constant M plays the role of an a priori bound on the norm of the noisy data h. If this information is not available, one can still give a bound of the form of Equation (5), but with a worst dependence on the regularization parameter λ by replacing estimate given by Equation (8) with*

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

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

Corollary 5 *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 $\delta \in \mathbb{R}^2_+$, with the notations of Proposition 1, then*

1. *for any $\lambda > 0$,*

$$\sup_{(\mathcal{Z},h,B)\in\mathcal{U}_\delta} \left| \left\|Af^\lambda - Pg\right\|_{\mathcal{K}} - \left\|Af_0^\lambda - Pg\right\|_{\mathcal{K}} \right| \leq \frac{\delta_1}{2\sqrt{\lambda}} + \frac{M\delta_2}{4\lambda}; \tag{10}$$

2. *if $g \in \operatorname{dom} A^\dagger$, for any $\lambda > 0$,*

$$\sup_{(\mathcal{Z},h,B)\in\mathcal{U}_\delta} \left| \left\|f^\lambda - A^\dagger g\right\|_{\mathcal{H}} - \left\|f_0^\lambda - A^\dagger g\right\|_{\mathcal{H}} \right| \leq \frac{\delta_1}{\lambda} + \frac{M\delta_2}{2\lambda^{\frac{3}{2}}}; \tag{11}$$

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

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

   then

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

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

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

Equation (13) follows from inequality (11) and the definition of parameter choice rule. ■

The fact that $\left\|f_0^\lambda - A^\dagger g\right\|_{\mathcal{H}}$ goes to zero as λ goes to zero is a known fact. However to estimate the rate of convergence of f^λ to $A^\dagger g$ a bound on $\left\|f_0^\lambda - A^\dagger g\right\|_{\mathcal{H}}$ has to be given so that some a priori assumptions on the exact solution are required. A standard result, Groetsch (1984), Engl et al. (1996), shows that, if $A^\dagger g = \mathrm{Im}\ (A^*A)^\alpha A^*$ then $\left\|f_0^\lambda - A^\dagger g\right\| = 0(\lambda^\alpha)$. 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 ϕ convex function.

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

Corollary 6 *Let $(\mathcal{K}, g, A) \in \mathcal{T}$ with $g \notin \mathrm{dom}\, A^\dagger$. For any $\delta \in \mathbb{R}^2_+$, let $(\mathcal{Z}, h, B) \in \mathcal{U}_\delta$ and*

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

where $\lambda_\delta = \lambda(\delta; h, B)$ is a parameter choice rule satisfying Equation (12), then

$$\lim_{\delta \to 0} \left\|f^{\lambda_\delta}\right\|_{\mathcal{H}} = +\infty.$$

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

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

By definition of parameter choice rule, the second term goes to zero, whereas the first one goes to infinity since, by a well known result of Tikhonov regularization, if $g \notin \mathrm{dom}\, A^\dagger$, $\left\|f_0^\lambda\right\|_{\mathcal{H}}$ goes to $+\infty$ when λ goes to zero. ■

Remark 7 *We do not know if the bound in Equation (11) is tight, that is, if it holds that*

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

In particular, for the problem of discretization, one could 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 modelling error. To this aim, we consider only noisy problems $Bf = h$, where B is an operator from $\mathcal{H}$ to $\mathcal{K}$ and $h \in \mathcal{K}$ such that

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

With this assumptions, it is known, Tikhonov et al. (1995), that if

$$\lim_{\eta_1,\eta_2\to 0} \frac{(\eta_1+\eta_2)^2}{\lambda(\eta_1,\eta_2)} = 0 \tag{14}$$

the regularized solution f^λ goes to $A^\dagger g$. 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 (14) is weaker than Condition (12).

This observation suggests that one can evaluate the noisy level of the noisy data $(\mathcal{Z}, g, B)$ 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. Repeating the standard proof, Tikhonov et al. (1995), for Tikhonov regularization in the presence of modelling error, one has that, if condition (14) holds, then the regularized solution f^λ goes to $A^\dagger g$. 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 one can restates the content of Proposition 1 as regularization in the presence of modelling error. Indeed, the least square 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 data A^*g, B^*B is the noisy model of the exact model A^*A and the noise levels are given by

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

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. Paper Wahba (1977) discusses the method of cross-validation, see also Wahba (1990), Groetsch (1984), 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 Nair and Schock (1998), Pereverzev and Schock (2000) and references therein. In Mathé and Pereverzev (2003) *adaptive strategy* is proposed for the choice of the parameter that provides the optimal order accuracy.

3. Discretization

In the present section, we study the discretization of the inverse problem whose direct problem is described by a Carleman operator, Halmos and Sunder (1978), see Appendix A for a review and notations.

We assume that X is a compact separable metric space X with a finite measure ν such that $\mathrm{supp}\ \nu = X$ and we simply set by $\mathcal{K} = L^2(X, \nu)$.

We fix a real Hilbert space $\mathcal{H}$ and a map $\gamma : X \to \mathcal{H}$

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

such that

1. the real function $(x, t) \mapsto \Gamma(x, t) = \langle \gamma_t, \gamma_x \rangle_{\mathcal{H}}$ is bounded on $X \times X$;
2. for all $t \in X$, the real function $x \mapsto \Gamma(x, t)$ is continuous on X.

We let A be the operator from $\mathcal{H}$ to $\mathcal{K}$ defined by

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

According to the theory of Appendix A, γ is a continuous bounded Carleman map (see Proposition 17 and Remark 21 taking into account that X is compact), and A is the corresponding Carleman operator, that is, $A = A_\gamma$ (see Equation (36)).

The following corollary recalls the main properties of the operator A.

Corollary 8 *The operator A is an injective Hilbert-Schmidt operator. The range of A is the reproducing kernel Hilbert space with kernel $\Gamma(x, t) = \langle \gamma_t, \gamma_x \rangle_{\mathcal{H}}$ and, in particular, is contained in $\mathcal{C}(X)$. Moreover it holds that*

$$A^* g = \int_X g(x)\gamma(x)\, d\nu(x), \tag{15}$$

where $g \in \mathcal{K}$ and the integral converges in the norm of $\mathcal{H}$,

$$A^*A = \int_X \langle \cdot, \gamma(x) \rangle_{\mathcal{H}} \gamma(x) \, d\nu(x), \tag{16}$$

where the integral converges in the trace norm.

Finally, if $g \in \operatorname{Im} A$ and $f^\dagger$ is the unique solution of $Af = g$, then

$$g(x) = \left\langle f^\dagger, \gamma_x \right\rangle_{\mathcal{H}} \quad x \in X. \tag{17}$$

Proof Since X is compact and Γ is bounded, Condition (44) holds. The content of the corollary is a restatement of Proposition 19 and Proposition 13. ■

Given a datum $g \in \mathcal{K}$, we consider the inverse problem $Af = g$, that is, the problem of finding $f \in \mathcal{H}$ such that

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

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

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

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

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

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

the Hilbert space $\mathcal{K}$ 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 with 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 with the vector

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

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

Corollary 9 *The operator $A_{\mathbf{x}}$ is bounded and*

$$\mathrm{Ker}\ A_{\mathbf{x}} = \mathrm{span}\{\, \gamma(x_i) \,|\, i = 1, \ldots, \ell\}^{\perp}$$

Moreover, it holds that

$$A_{\mathbf{x}}{}^{*} A_{\mathbf{x}} = \sum_{i=1}^{\ell} a_i \langle \cdot, \gamma(x_i) \rangle_{\mathcal{H}} \gamma(x_i) \tag{18}$$

$$A_{\mathbf{x}}{}^{*} \mathbf{y} = \sum_{i=1}^{\ell} a_i y_i \gamma(x_i) \tag{19}$$

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

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

being δ the Dirac measure at the point $i \in I$, then $A_{\mathbf{x}}$ is the Carleman operator associated to the map

$$I \ni i \mapsto \gamma(x_i) \in \mathcal{H}.$$

In particular, we have that $\mathrm{supp}\ \nu_{\mathbf{x}} = I$ so that $\mathcal{H}_{\mathbf{x}} = \mathrm{span}\{\, \gamma(x_i) \,|\, i = 1, \ldots, \ell\}$.

The content of the corollary is a restatement of Proposition 19 and the fact that the integrals reduce to sums. ■

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

$$f_0^{\lambda} = (A^* A + \lambda)^{-1} A^* g$$

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

where we add the pedix $\mathbf{z}$ to emphasize the dependence of the solution on the data. Finding the explicit form of $f_{\mathbf{z}}{}^{\lambda}$ reduces to solve a linear problem. Indeed, let $\Gamma_{\mathbf{x}}$ be 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 Equation (42),

$$f_{\mathbf{z}}{}^{\lambda} = \sum_{i,j=1}^{\ell} a_j \gamma(x_j) \left((\Gamma_{\mathbf{x}} + \lambda)^{-1} \right)_{ji} y_i. \tag{20}$$

Applying the results of Section 2, we propose a bound for the reconstruction error $\left\| f_{\mathbf{z}}{}^{\lambda} - f^{\dagger} \right\|_{\mathcal{H}}$ or the residuum $\left\| A f_{\mathbf{z}}{}^{\lambda} - g \right\|_{\mathcal{K}}$. To this aim we have to assume some hypotheses on the relation between ν and $\sum_{i=1}^{\ell} a_i \delta_{x_i}$, g and $\mathbf{y}$. We discuss two cases.

3.1 A priori approximation

In this section, we consider a framework where the measure ν is known, the points x_i and the values y_i are a sample 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.

Moreover, we study the reconstruction error of the approximated solution. To this aim, we assume that $g \in \mathrm{Im}\ A \subset \mathcal{C}(X)$ so that, by Equation (17), we can restate the hypothesis that the noise is zero by the fact that

$$y_i = g(x_i) = \left\langle f^\dagger, \gamma(x_i)\right\rangle_{\mathcal{H}} \qquad \forall i \in I \tag{21}$$

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

1. $x_i \in X_i$ for all $i \in I$;
2. $\nu(X_i \cap X_j) = 0$ for all $i \neq j$;
3. $\bigcup_i X_i = X$.

We let $a_i = \nu(X_i)$ and we have that following result.

Corollary 10 *We define*

$$\begin{aligned} \kappa &= \sup_{x\in X} \|\gamma(x)\|_{\mathcal{H}} = \sup_{x\in X} \sqrt{\Gamma(x,x)} < +\infty & (22)\\ \alpha &= \nu(X) < +\infty & (23)\\ c(\ell) &= \max_{i\in I} \sup_{x\in X_i} \|\gamma(x) - \gamma(x_i)\|_{\mathcal{H}} & (24)\\ &= \max_{i\in I} \sup_{x\in X_i} \sqrt{\Gamma(x,x) - 2\Gamma(x,x_i) + \Gamma(x_i,x_i)} \end{aligned}$$

then,

$$|\left\|f_{\mathbf{z}}{}^\lambda - f^\dagger\right\|_{\mathcal{H}} - \left\|f_0^\lambda - f^\dagger\right\|_{\mathcal{H}}| = \left\|f^\dagger\right\|_{\mathcal{H}} \kappa\, c(\ell)\, \alpha \left(\frac{2}{\lambda} + \frac{\kappa\sqrt{\alpha}}{\lambda^{\frac{3}{2}}}\right). \tag{25}$$

Proof We claim that

$$\begin{aligned} \|A^* g - A_{\mathbf{x}}{}^* \mathbf{y}\|_{\mathcal{L}(\mathcal{H})} &\leq 2\left\|f^\dagger\right\|_{\mathcal{H}} \kappa c(\ell) = \delta_1 & (26)\\ \|A^* A - A_{\mathbf{x}}{}^* A_{\mathbf{x}}\|_{\mathcal{H}} &\leq 2\kappa c(\ell) = \delta_2 & (27) \end{aligned}$$

We first prove Equation (27). By definition of X_i and a_i and Equations (16) and (18), one has that

$$\|A^* A - A_{\mathbf{x}}{}^* A_{\mathbf{x}}\|_{\mathcal{L}(\mathcal{H})} = \left\|\sum_i \int_{X_i} \left(\langle \cdot, \gamma(x)\rangle_{\mathcal{H}} \gamma(x) - \langle \cdot, \gamma(x_i)\rangle_{\mathcal{H}} \gamma(x_i)\right)\, d\nu(x)\right\|_{\mathcal{L}(\mathcal{H})}$$

$$
\begin{aligned}
&\le \sum_i \nu(X_i) \sup_{x\in X_i} \|\langle\cdot,\gamma(x)\rangle_{\mathcal{H}}\gamma(x) - \langle\cdot,\gamma(x_i)\rangle_{\mathcal{H}}\gamma(x_i)\|_{\mathcal{L}(\mathcal{H})} \\
&\le \max_{i\in I} \sup_{x\in X_i} \|\langle\cdot,\gamma(x)\rangle_{\mathcal{H}}\gamma(x) - \langle\cdot,\gamma(x_i)\rangle_{\mathcal{H}}\gamma(x_i)\|_{\mathcal{L}(\mathcal{H})} \sum_i \nu(X_i) \\
&\le 2\max_{i\in I} \sup_{x\in X_i} \left(\|\gamma(x)\|_{\mathcal{H}}\|\gamma(x)-\gamma(x_i)\|_{\mathcal{H}}\right)\nu(X) \\
&\le 2\kappa c(\ell)\alpha,
\end{aligned}
$$

so that Equations (27) is proved. By Equations (17), (15), (18), it follows that

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

so that Equation (26) is clear.

We observe that, by Equation (21),

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

so that, replacing in Equation (11) the above bounds on M, δ_1 and δ_2 the inequality (25) follows. ■

In Equation (25) one needs an a priori bound on $\left\|f^\dagger\right\|_{\mathcal{H}}$. However, one can obtain a (worst) estimate of $\left\|f_{\mathbf{z}}{}^\lambda - f^\dagger\right\|_{\mathcal{H}}$ depending on $\|g\|_{\mathcal{Z}} = \left\|Af^\dagger\right\|_{\mathcal{Z}}$.

Let $\omega(\lambda) = \left\|f_0^\lambda - f^\dagger\right\|_{\mathcal{H}}$ 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 $\lim_{\ell\to\infty}\omega(\lambda_\ell) = 0$ then, as a consequence of Equation (25),

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

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

3.1.1 The problem of derivative

As an example we study the problem of the recovering the derivative of a function on the unit interval. The setting is the following one.

1. We let $X = [0, 1]$ with the Lebesgue measure $\nu = dx$;

2. we define the kernel K as

$$K(x,t) = \theta(x-t) = \begin{cases} 1 & t \leq x \\ 0 & t > x \end{cases} \qquad t \in [0,1]$$

   In particular, for all $x \in X$, $K(x, \cdot) \in L^2([0,1])$ and, for all $\phi \in L^2([0,1])$,

$$(L_K\phi)(x) = \int_0^x \phi(t)\, dt,$$

   so that $(L_K\phi)(\cdot) \in L^2([0,1])$

3. we choose $\mathcal{H} = H^1([0,1])$ be the Sobolev space of continuous functions with weak derivative in $L^2([0,1])$ endowed with the scalar product

$$\langle v, w \rangle = v(0)w(0) + \int_0^1 v'(x)w'(x)\, dx,$$

   so that $\mathcal{H}$ is a reproducing kernel Hilbert space on $[0,1]$ with kernel

$$\Delta(x,t) = \min\{x,t\} + 1,$$

   and, given $x \in X$,

$$\begin{aligned} \delta_x(t) &= \begin{cases} t+1 & t \leq x \\ x+1 & t > x \end{cases} \quad t \in [0,1] \\ \delta'_x(t) &= \begin{cases} 1 & t < x \\ 0 & t > x \end{cases} \qquad t \in [0,1]; \end{aligned}$$

4. we let j be the canonical immersion of $\mathcal{H}$ in $L^2([0,1])$ so that $A = L_K j$;

5. given $g \in L^2([0,1])$ the exact problem $Af = g$ amounts to find $f \in H^1([0,1])$ such that

$$\int_0^x f(t)\, dt = g(x);$$

6. the condition that $g \in \mathrm{Im}\ A$ implies that g is a continuous function with weak derivative in $\mathcal{H}$, so that $g \in C^1([0,1])$ and $f^\dagger = g'$;

7. for all $x \in X$,

$$\gamma_x(t) = j^*(K(x,\cdot))(t) = \begin{cases} x + tx - \frac{t^2}{2} & t \leq x \\ x + \frac{x^2}{2} & t > x \end{cases} \qquad t \in [0,1]$$

   so that

$$\gamma'_x(t) = \begin{cases} x - t & t \leq x \\ 0 & t > x \end{cases} \qquad t \in [0,1] \; ;$$

8. for all $x, t \in X$

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

9. if $0 \leq t \leq x \leq 1$,

$$\|\gamma(x) - \gamma(t)\|_{\mathcal{H}} = \sqrt{(x-t)^2 \frac{3 + x + 2t}{3}} \leq \sqrt{2}|x - t|; \qquad (28)$$

10. we choose $x_k = \frac{k}{\ell}$ for all $k = 0, \ldots, \ell$ and

$$X_k = [x_{k-1}, x_k]$$

    so that, by Equation (28), $c(\ell) = \frac{\sqrt{2}}{\ell}$;

11. clearly $\alpha = \nu([0,1]) = 1$ and, by Equation (28) with $t = 0$, it follows that $\kappa = \sqrt{2}$.

Replacing in Equation (25) we obtain that

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

from the above equation and assuming to have a parameter choice rule such that $\lambda(\mathbf{z}, \ell) = 0(\frac{1}{\ell^b})$, the approximated solution $f_{\mathbf{z}}{}^{\lambda_\ell}$ goes to the exact solution $f^\dagger$ if $b < \frac{2}{3}$.

According to Equation (20), the solution $f_{\mathbf{z}}{}^{\lambda}$ is a linear combination of the functions $\gamma(x_i)$ that are quadratic splines (piecewise polynomials of degree two with continuous derivative), Wahba (1990). The norm in $\mathcal{H}$ ensures that $f_{\mathbf{z}}{}^{\lambda}$ converges to $f^\dagger$ uniformly.

From a numerical point of view, the computation of $f_{\mathbf{z}}$ is reduced to compute the inverse of the $\ell \times \ell$ symmetric matrix

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

3.2 Stochastic approximation

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:

1. the marginal distribution of ρ on X is ν (this implies in particular that ν is normalized to 1);

2. 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 with the set$\{x_1, \ldots, x_\ell\}$.

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

$$\mathrm{Prob}\left\{\mathbf{z} \in (X \times \mathbb{R})^\ell \mid \left\| A f_{\mathbf{z}}{}^\lambda - Pg \right\|_{\mathcal{K}} \leq \epsilon \right\} \geq 1 - \eta.$$

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

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

Proposition 11 *Let* $\epsilon_1, \epsilon_2 > 0$ *and*

$$\kappa = \sup_{x \in X} \|\gamma(x)\| = \sup_{x \in X} \sqrt{\Gamma(x,x)},$$

then

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

$$\geq 1 - e^{-\frac{\epsilon_1^2 \ell}{2\kappa^2 L^2}} - e^{-\frac{\epsilon_2^2 \ell}{2\kappa^4}}$$

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

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

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

$$\text{Prob}\left\{\mathbf{z} \in (X \times \mathbb{R})^\ell \,|\, F(\mathbf{z}) \geq \mathbb{E}(F) + \epsilon\right\} \leq e^{-\frac{2\epsilon^2}{\sum_{i=1}^\ell c_i^2}}, \tag{29}$$

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).

To this aim, we let $Y = [-L, L]$ and $\varphi : X \times Y \to \mathcal{H}$

$$\varphi(x, y) = y\gamma(x)$$

and $\Phi : (X \times Y)^\ell \to \mathcal{H}$

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

Since φ is weakly continuous, it is measurable and bounded, so it is integrable on $X \times Y$ and

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

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-Schwartz inequality,

$$\begin{aligned}
\mathbb{E}(F) &= \int_{(X\times Y)^\ell} \|\Phi(\mathbf{z}) - A^*g\| \; d\rho^\ell(\mathbf{z}) \qquad (30) \\
&\leq \sqrt{\int_{(X\times Y)^\ell} \|\Phi(\mathbf{z}) - A^*g\|^2 \; d\rho^\ell(\mathbf{z})} \\
&\leq \frac{1}{\sqrt{\ell}}\sqrt{\int_{X\times Y} \|\varphi(x, y)\|^2 \; d\rho(x, y)} \\
(\ \|\phi(x, y)\| \leq L\kappa\ ) \quad &\leq \frac{L\kappa}{\sqrt{\ell}}. \qquad (31)
\end{aligned}$$

Moreover, by triangular inequality,

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

so that we can choose

$$c_i = \frac{2L\kappa}{\ell} \tag{32}$$

Finally, we observe that, given $\epsilon_1 > 0$, by Equation (31),

$$\begin{aligned} \mathrm{Prob}\left\{\mathbf{z} \in (X \times \mathbb{R})^\ell \,|\, F(\mathbf{z}) \geq \frac{L\kappa}{\sqrt{\ell}} + \epsilon_1\right\} &\leq \mathrm{Prob}\left\{\mathbf{z} \in (X \times \mathbb{R})^\ell \,|\, F(\mathbf{z}) \geq \mathbb{E}(F) + \epsilon_1\right\} \\ (\ Eqs.\ (29),\ (32)\ ) &\leq e^{-\frac{\epsilon_1^2\ell}{2\kappa^2L^2}} \end{aligned}$$

To prove the second part, we can mimic the previous 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 12 *Given $0 < \eta < 1$, with probability greater that $1 - \eta$,*

$$|\left\|Af_{\mathbf{z}}{}^\lambda - Pg\right\|_{\mathcal{K}} - \left\|Af_0^\lambda - Pg\right\|_{\mathcal{K}}| \leq \frac{\kappa L}{2\sqrt{\ell}}\left(\frac{1}{\sqrt{\lambda}} + \frac{\kappa}{2\lambda}\right)\left(1 + \log\sqrt{\frac{4}{\eta}}\right) \tag{33}$$

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,

$$\|Af_{\mathbf{z},\ell} - Pg\|_{\mathcal{K}} = \|Af_{0,\ell} - Pg\|_{\mathcal{K}} + O(\frac{1}{\ell^{1-2b}}).$$

Proof We observe that

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

The thesis follows from Equation (10) 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 (33) holds depends on ℓ and η, but does not depend on the regularization parameter.

3.2.1 Regularized least square algorithm

As an application, we consider the problem of consistency of regularized least square algorithm in the context of statistical learning theory in the regression setting (for an account of this theory and its application see Vapnik (1998), Evgeniou et al. (2000), Cucker and Smale (2002), Poggio and Smale (2003) and references therein). Some results in the same spirit can be found in Rudin (2004), Kurkova (2004). In the framework of learning, there are two sets of variables, namely, the input space X and the output space $\mathbb{R}$. 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 $X \times \mathbb{R}$. The distribution ρ is known only through a sample $\mathbf{z} = ((x_1, y_\ell), \ldots, (x_\ell, y_\ell))$ drawn according to ρ. The goal of learning is, given a sample $\mathbf{z}$, to find a function $f_{\mathbf{z}} : X \to \mathbb{R}$ such that $f_{\mathbf{z}}(x)$ is an *estimate* of the output y when the new input x is given. The function $f_{\mathbf{z}}$ is called *estimator* and the rule that, given a sample $\mathbf{z}$, provides us with $f_{\mathbf{z}}$ is called *learning algorithm*.

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

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

The regression function

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

is the minimizer of the expected risk over the set of all measurable functions. However, it can not be reconstructed by the sample $\mathbf{z}$ since the data are finite and noisy.

To overcome this problem, in the regularized least square algorithm one fixes an space of hypotheses $\mathcal{H}$, which is a reproducing kernel Hilbert space on X, and defines the estimator ${f_{\mathbf{z}}}^{\lambda}$

as the minimizer on $\mathcal{H}$ of the regularized least square functional, Cucker and Smale (2002),

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

where the regularization parameter has to be chosen according to the number of the data, $\lambda=\lambda_\ell$, in such a way that

$$\lim_{\ell\to+\infty}\mathrm{Prob}\left\{\mathbf{z}\in(X\times\mathbb{R})^\ell\,|\,I[f_{\mathbf{z}}{}^{\lambda_\ell}]-\inf_{f\in\mathcal{H}}I[f]\geq\epsilon\right\}=0,$$

(in general $\inf_{f\in\mathcal{H}}I[f]$ is greater that $I[g]$ and represents a sort of irreducible error associated with the choice of the space $\mathcal{H}$). This property is called *consistency* of the algorithm, Vapnik (1998).

We rewrite the above problem as discretization of an inverse problem defined by a Carleman map. We assume that X is a compact separable metric space, ρ satisfies the assumptions of Section 3.2 and the hypotheses space $\mathcal{H}$ is a reproducing kernel real Hilbert space on X with a continuous kernel Γ (see Appendix for definitions and main properties). Letting $\gamma_x=\Gamma(\cdot,x)$, the Carleman operator A is

$$(Af)(x)=\langle f,\gamma(x)\rangle_{\mathcal{H}}=f(x),$$

that is, A is the canonical immersion of $\mathcal{H}$ in $L^2(X,\nu)$. In particular, for all $f\in\mathcal{H}$, the expected risk becomes, (Cucker and Smale, 2002),

$$I[f]=\|Af-g\|_{L^2(X,\nu)}^2+I[g]$$

where

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

(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 $L^2(X,\nu)$).

Moreover, if $A_{\mathbf{x}}$ is 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 *á 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 we observe that, recalling that P is the orthogonal projection on the closure on the range of A, that is, on the closure in $L^2(X,\nu)$ of $\mathcal{H}$, for all $f \in \mathcal{H}$

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

Let now choose the parameter $\lambda_\ell = \lambda(\mathbf{z},\ell)$ such that $\lambda(\mathbf{z},\ell) = O(\frac{1}{\ell^b})$, with $0 < b < \frac{1}{2}$, Proposition 12 ensures that, with probability greater that $1-\eta$,

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

that is, in probability,

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

showing that the regularized least square algorithm is consistent.

We notice that the set of samples such that Equation (34) holds, it depends on ℓ and η, but does not depend on the parameter λ. This last fact allows us to consider a posteriori parameter choice rules $\lambda_\ell = \lambda(\mathbf{z},\ell)$ which can be in general problematic given the probabilistic setting of learning (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, A. Verri for useful discussions and suggestions. L. Rosasco is supported by an INFM fellowship. A. Caponnetto is supported by a PRIN fellowship within the project "Inverse problems in medical imaging", n. 2002013422. This research has been partially funded by the INFM Project MAIA, the FIRB Project ASTA, and by the EU Project KerMIT.

Appendix A. Carleman maps

In this section, we review some properties of reproducing kernel Hilbert spaces and integral equations using the notion of Carleman maps, (Halmos and Sunder, 1978) (Saitoh, 1997).

Let X be a locally compact second countable topological space. We denote by $\mathbb{C}^X$ the vector space of complex functions on X and by $\mathcal{C}(X)$ the space of continuous complex functions on X (the following results clearly hold by replacing $\mathbb{C}$ with $\mathbb{R}$). Let ν be a Radon measure on X (that is, an abstract positive measure defined on the Borel σ-algebra $\mathcal{B}(X)$ and finite on compact subsets of X).We denote by $\mathrm{supp}\ \nu$ the support of the measure ν, by $L^2(X,\nu)$ the Hilbert space of (equivalence classes) of complex functions on X that are square-integrable with respect to ν. Let $\mathcal{H}$ be a complex Hilbert space and γ be a map from X to $\mathcal{H}$

$$X \ni x \mapsto \gamma(x) = \gamma_x \in \mathcal{H}.$$

We define Γ as the function from $X \times X$ to $\mathbb{C}$

$$\Gamma(x,t) = \langle \gamma(t), \gamma(x) \rangle_{\mathcal{H}} \qquad x, t \in X. \tag{35}$$

and A_γ as the linear operator from $\mathcal{H}$ to $\mathbb{C}^X$

$$(A_\gamma v)(x) = \langle v, \gamma(x) \rangle_{\mathcal{H}} \qquad x \in X. \tag{36}$$

for all $v \in \mathcal{H}$. Moreover, we denote by

$$\mathcal{H}_\gamma = \overline{\mathrm{span}}\{\gamma(x) \,|\, x \in X\} \tag{37}$$

$$\mathcal{H}_\nu = \overline{\mathrm{span}}\{\gamma(x) \,|\, x \in \mathrm{supp}\ \nu\}, \tag{38}$$

where $\overline{\mathrm{span}}$ denotes the closure of the linear span.

We recall the following basic result, see Aronszajn (1950), Schwartz (1964), Saitoh (1997) (see definition of reproducing kernel Hilbert space in the following subsection).

Proposition 13 *The range of A_γ endowed with the scalar product*

$$\langle f, g \rangle_{\mathrm{Im}\ A_\gamma} = \langle v, w \rangle_{\mathcal{H}}\,, \tag{39}$$

where $f = A_\gamma v$, $g = A_\gamma w$ and $v, w \in \mathcal{H}_\gamma$, is the reproducing kernel Hilbert space with kernel Γ and, hence, A_γ is a unitary operator from $\mathcal{H}_\gamma$ onto $\mathrm{Im}\ A_\gamma$.

Proof By definition of $\mathcal{H}_\gamma$, A_γ is injective on $\mathcal{H}_\gamma$ so that the scalar product defined by Equation (39) is well defined. We prove that $\mathrm{Im}\ A_\gamma$ is a reproducing kernel Hilbert space showing that, given $x \in X$, the linear map

$$\mathrm{Im}\ A_\gamma \ni f \mapsto f(x) \in \mathbb{C}$$

is continuous. Indeed, for all $f = A_\gamma v$, $v \in \mathcal{H}_\gamma$,

$$\begin{aligned} f(x) &= \langle v, \gamma_x \rangle_{\mathcal{H}} \\ &= \langle A_\gamma v, A_\gamma \gamma_x \rangle_{\mathrm{Im}\ A_\gamma} \\ &= \langle f, A_\gamma \gamma_x \rangle_{\mathrm{Im}\ A_\gamma}\,, \end{aligned}$$

where we use that $\gamma_x \in \mathcal{H}_\gamma$. Hence $\mathcal{H}$ is a reproducing kernel Hilbert space. Observing that, for all $t \in X$,

$$(A_\gamma \gamma_x)(t) = (A_\gamma \gamma_x)(t) = \langle \gamma(x), \gamma(t) \rangle_{\mathcal{H}} = \Gamma(t, x),$$

it follows that the reproducing kernel is Γ. ■

We now give the definition of Carleman map, Halmos and Sunder (1978).

Definition 14 *A map $\gamma : X \to \mathcal{H}$ is a bounded Carleman map if*

1. *for all $v \in \mathcal{H}$, the function $\langle \gamma(\cdot), v \rangle_{\mathcal{H}}$ is measurable;*
2. *for all $v \in \mathcal{H}$, $\int_X |\langle \gamma(x), v \rangle_{\mathcal{H}}|^2 \, d\nu(x) < +\infty$.*

A bounded Carleman map is called continuous if

3. *for all $v \in \mathcal{H}$, the function $\langle \gamma(\cdot), v \rangle_{\mathcal{H}}$ is continuous.*

Remark 15 *The* boundedness *in definition of bounded Carleman map refers to the fact that the corresponding operator A_γ is bounded from $\mathcal{H}$ into $L^2(X, \nu)$ (see Proposition 19 below and Halmos and Sunder (1978) for a discussion). Condition (3) is the requirement that γ is weakly continuous and, clearly, implies Condition (1).*

We now discuss some characterizations of Carleman maps. We start with the problem of measurability.

Lemma 16 *Let $\gamma : X \to \mathcal{H}$, the following four conditions are equivalent:*

1. *for all $v \in \mathcal{H}$, the function $\langle \gamma(\cdot), v \rangle_{\mathcal{H}}$ is measurable;*
2. *the map $x \mapsto \gamma(x)$ is measurable;*
3. *the map Γ is measurable;*
4. *for all $x \in X$, $\Gamma(x, \cdot)$ is measurable.*

Proof Since $\mathcal{H}$ is separable, the equivalence between the first two conditions is well known. Since the scalar product is continuous, condition (2) implies (3), and, clearly, (3) implies (4). Assume now that, given $x \in X$, $\Gamma(x, \cdot)$ is measurable. By Equation (35), $\langle \gamma(\cdot), \gamma_x \rangle_{\mathcal{H}} = \Gamma(x, \cdot)$, so that, by linearity and density, it follows that $\langle \gamma(\cdot), v \rangle_{\mathcal{H}}$ is measurable for all $v \in \mathcal{H}_\gamma$. However, if $v \in {\mathcal{H}_\gamma}^\perp$, $\langle \gamma(\cdot), v \rangle_{\mathcal{H}} = 0$, so Condition (1) is proved. ■

The following proposition faces up the problem of the continuity of the map γ and it is well known in the context of reproducing kernel Hilbert spaces, see Proposition 24 of Schwartz (1964).

Proposition 17 *The following facts are equivalent:*

1. *the map γ is weakly continuous;*
2. *the function Γ is locally bounded and separately continuous;*
3. *the operator A_γ is continuous from $\mathcal{H}$ into $\mathcal{C}(X)$, endowed with the topology of the uniform convergence on compact subset.*

Proof Assume Condition (2), we claim that the map γ is weakly continuous. Let $w = \sum_{i=1}^n a_i\gamma(x_i)$, by Equation (35), we have that

$$\langle \gamma(\cdot), w\rangle_{\mathcal{H}} = \sum_{i=1}^n a_i \Gamma(x_i, \cdot),$$

which is clearly continuous by the assumptions on Γ.

Let now $v \in \mathcal{H}_\gamma$ and $x_0 \in X$, we prove that $\langle \gamma(\cdot), v\rangle_{\mathcal{H}}$ is continuous at x_0. We fix a compact neighborhood K of x_0 and we let

$$M = \sup_{x\in K} \|\gamma_x\| = \sqrt{\Gamma(x,x)},$$

where M is finite due to locally boundedness of Γ. Fixed $\epsilon > 0$, there is a linear combinations $w = \sum a_i \gamma_{x_i}$ such that $\|v - w\| \le \epsilon$. By the above observation, $\langle \gamma(\cdot), w\rangle_{\mathcal{H}}$ is continuous, so, possibly replacing K with a small neighborhood, for all $x \in K$, $|\langle \gamma(x) - \gamma(x_0), w\rangle_{\mathcal{H}}| \le \epsilon$, and, hence,

$$\begin{aligned} |\langle \gamma(x), v\rangle_{\mathcal{H}} - \langle \gamma(x_0), v\rangle_{\mathcal{H}}| &\le |\langle \gamma(x) - \gamma(x_0), w\rangle_{\mathcal{H}}| + \|\gamma(x) - \gamma(x_0)\| \, \|w - v\| \\ &\le \epsilon(1 + 2M). \end{aligned}$$

Finally, if $v \in {\mathcal{H}_\gamma}^\perp$, $\langle \gamma(\cdot), v\rangle_{\mathcal{H}} = 0$, so that Condition (1) is proved.

Assume now that γ is weakly continuous, we prove that A_γ is a bounded operator. By definition of weak continuity, for all $v \in \mathcal{H}$, $A_\gamma v = \langle v, \gamma(\cdot)\rangle_{\mathcal{H}} \in \mathcal{C}(X)$. In particular, given a compact set K, for all $v \in \mathcal{H}$ $\langle v, \gamma(\cdot)\rangle_{\mathcal{H}}$ is bounded on K and, by Banach-Steinhaus theorem, there is a constant $M > 0$ such that

$$\|\gamma_x\| \le M \qquad \forall x \in K.$$

Moreover,

$$\begin{aligned} \sup_{x\in K} |(A_\gamma v)(x)| &= \sup_{x\in K} |\langle v, \gamma_x\rangle_{\mathcal{H}}| \\ &\le \|v\| \sup_{x\in K} \|\gamma_x\| \le M \|v\|, \end{aligned}$$

so that A_γ is a continuous linear map from $\mathcal{H}$ to $\mathcal{C}(X)$.

Assume now Condition (3), we prove Condition (2). For all $v \in \mathcal{H}$, $\langle v, \gamma(\cdot) \rangle_{\mathcal{H}} = A_\gamma v$ is continuous. It follows, by Equation (35), that Γ is separately continuous and, as above, that given a compact set K, there is $M > 0$ such that

$$\sqrt{\Gamma(x,x)} = \|\gamma_x\| \leq M \qquad \forall x \in K.$$

Observing that, if $t, x \in K$, by Cauchy-Schwartz inequality,

$$|\Gamma(x,t)| = |\langle \gamma_t, \gamma_x \rangle_{\mathcal{H}}| \leq \|\gamma_x\| \, \|\gamma_t\| \leq M^2,$$

that is, Γ is locally bounded. ■

The following corollary characterizes the compactness of A_γ in $\mathcal{C}(X)$, see Proposition 24 of Schwartz (1964).

Corollary 18 *The following facts are equivalent:*

1. *the function Γ is continuous;*
2. *the function Γ is continuous on the diagonal of $X \times X$;*
3. *the map γ is continuous;*
4. *the map γ is weakly continuous and the function $x \mapsto \Gamma(x,x)$ is continuous;*
5. *the operator $A_\gamma : \mathcal{H} \to \mathcal{C}(X)$ is compact.*

Proof Clearly, Condition (1) implies (2). Observing that, due to the definition of Γ, given $x, t \in X$

$$\|\gamma(x) - \gamma(t)\|^2 = \Gamma(x,x) + \Gamma(y,y) - \Gamma(x,y) - \Gamma(y,x)$$

Condition (2) implies Condition (3). Moreover, since the scalar product is continuous, by definition of Γ, Condition (3) implies (1). The equivalence between (3) and (4) is a restatement of the equivalence between strong convergence and weak convergence plus convergence in norm (so called $\mathcal{H}$-property).

We prove the equivalence between Condition (3) and Condition (5). Let B_1 the unit ball in $\mathcal{H}$, then A_γ is compact if and only if $A_\gamma(B_1)$ is compact in $\mathcal{C}(X)$. By Ascoli theorem (see page 233 of Kelley (1955)), the set of functions $A_\gamma(B_1)$ is compact if and only if $A_\gamma(B_1)$ is closed, for all $x \in X$, the set $(A_\gamma(B_1))(x)$ is bounded and $A_\gamma(B_1)$ is equicontinuous. First two conditions are always satisfied since A_γ is continuous from the Hilbert space $\mathcal{H}$ to $\mathcal{C}(X)$.

Moreover, given $x, t \in X$, we have that

$$\begin{aligned} \|\gamma(x)-\gamma(t)\| &= \sup_{v\in B_1} |\langle v, \gamma(x)-\gamma(t)\rangle_{\mathcal{H}}| \\ &= \sup_{v\in B_1} |(A_\gamma v)(x)-(A_\gamma v)(t)| \\ &= \sup_{f\in A_\gamma(B_1)} |f(x)-f(t)| \end{aligned} \tag{40}$$

Equation (40) shows that $A_\gamma(B_1)$ is equicontinuous if and only if map γ is continuous. ■

The following proposition collects the main properties of A_γ for γ being a bounded Carleman map.

Proposition 19 *Let γ a bounded Carleman map, then A_γ is a bounded linear operator from $\mathcal{H}$ into $L^2(X,\nu)$. For all $\phi \in L^2(X,\nu)$*

$$A_\gamma^*\phi = \int_X \phi(x)\gamma(x)\,d\nu(x), \tag{41}$$

where the integral converges in the weak topology, and

$$(A_\gamma A_\gamma^*\phi)(x) = \int_X \Gamma(x,t)\phi(t)\,d\nu(t) \qquad \nu\text{−a.e. } x \in X, \tag{42}$$

that is $A_\gamma A_\gamma^*$ is the integral operator of kernel Γ. Finally,

$$A_\gamma^* A_\gamma = \int_X \langle\cdot, \gamma(x)\rangle_{\mathcal{H}}\gamma(x)\,d\nu(x), \tag{43}$$

where the integral converges in the weak operator topology.

In particular A_γ is a Hilbert-Schmidt operator if and only if

$$\int_X \Gamma(x,x)\,d\nu(x) < +\infty. \tag{44}$$

If this last condition holds, the integral in Equation (41) converges in norm and the integral in Equation (43) converges in the trace norm.

Finally, if γ is a continuous bounded Carleman map, the kernel of A_γ is $\mathcal{H}_\nu{}^\perp$ and the range of A_γ is a subset of $\mathcal{C}(X)$.

Proof The proof is standard. We first show that $A_\gamma \in \mathcal{B}(\mathcal{H}, L^2(X,\nu))$.

Since γ is a bounded Carleman map, A_γ is a linear operator defined on $\mathcal{H}$. We claim that it is closed. Indeed, let $(v_n)_{n\in\mathbb{N}}$ be a sequence such that it converges to $v \in \mathcal{H}$ and the sequence

$(A_\gamma v_n)_{n\in\mathbb{N}}$ converges to $\phi \in L^2(X,\nu)$. By construction, for all $x \in X$,

$$(A_\gamma v_n)(x) = \langle v_n, \gamma(x)\rangle_{\mathcal{H}} \to \langle v, \gamma(x)\rangle_{\mathcal{H}} = (A_\gamma v)(x).$$

By uniqueness of the limit, it follows that $A_\gamma v = \phi$, that is, A_γ is closed. By the closed graph theorem, one has that A_γ is bounded.

We show Equation (41). Let $\phi \in L^2(X,\nu)$, then for all $v \in \mathcal{H}$, the function $\phi(\cdot)\langle\gamma(\cdot), v\rangle_{\mathcal{H}}$ is clearly in $L^1(X,\nu)$ and

$$\int_X \phi(x)\,\langle\gamma(x), v\rangle_{\mathcal{H}}\,d\nu(x) = \langle \phi, A_\gamma v\rangle_{L^2(X,\nu)} = \left\langle A_\gamma^*\phi, v\right\rangle_{\mathcal{H}},$$

so that the function $\phi(\cdot)\,\gamma(\cdot)$ is weakly integrable and Equation (41) holds.

Equations (42), (43) are consequences of Equation (41) and the definition of Γ.

To prove Condition (44), let $(e_n)_{n\in\mathbb{N}}$ be a Hilbert basis of $\mathcal{H}$. Since A^*A is a positive operator and $|\langle\gamma(\cdot), e_n\rangle_{\mathcal{H}}|^2$ is a positive function, by monotone convergence theorem, one has that

$$\begin{aligned}
\operatorname{Tr}(A^*A) &= \sum_n \int_X |\langle e_n, \gamma(x)\rangle_{\mathcal{H}}|^2\,d\nu(x) \\
&= \int_X \sum_n |\langle e_n, \gamma(x)\rangle_{\mathcal{H}}|^2\,d\nu(x) \\
&= \int_X \langle\gamma(x), \gamma(x)\rangle_{\mathcal{H}}\,d\nu(x) \\
&= \int_X \Gamma(x,x)\,d\nu(x)
\end{aligned}$$

and the thesis follows.

Finally, we prove the statements about Equations (41) and (43). Indeed, by Lemma 16, the map γ is measurable, so that the maps $\phi(\cdot)\gamma(\cdot)$ and $x \mapsto \langle\cdot, \gamma(x)\rangle_{\mathcal{H}}\gamma(x)$ are measurable as maps taking value in $\mathcal{H}$ and in the space of trace class operators, respectively. Moreover,

$$\begin{aligned}
\|\phi(\cdot)\gamma(\cdot)\| &\leq |\phi(x)|\sqrt{\Gamma(x,x)} \in L^1(X,\nu) \\
\|\langle\cdot, \gamma(x)\rangle_{\mathcal{H}}\gamma(x)\|_1 &= \Gamma(x,x) \in L^1(X,\nu),
\end{aligned}$$

where $\|\cdot\|_1$ is the trace norm. So the integrals are well defined and the equalities follow by uniqueness of the integral.

Assume now that γ is weakly continuous, then the range of A_γ is in $\mathcal{C}(X)$. We characterize the kernel of A_γ. Clearly, it contains $\mathcal{H}_\nu$. Assume now that $A_\gamma v = 0$ in $L^2(X,\nu)$, then, for ν-almost all $x \in X$, $(A_\gamma v)(x) = 0$. However, $(A_\gamma v)(\cdot)$ is a continuous function, hence, by definition of support, it follows that, for all $x \in \mathrm{supp}\ \nu$, $(A_\gamma v)(x) = 0$, that is, $\langle v, \gamma(x)\rangle_{\mathcal{H}} = 0$. It follows that $v \in {\mathcal{H}_\nu}^\perp$. ■

We slight abuse of notation, we denote by A_γ both the operator in $\mathbb{C}^X$ and in $L^2(X,\nu)$, where the second one depends also on the measure ν. To overcome possible problems, the operator A_γ from $\mathcal{H}$ to $L^2(X,\nu)$ is called the Carleman operator associated with γ and ν (see the first remark below).

The following example shows that the hypothesis on the continuity of γ is essential to characterize the kernel of A_γ.

Example 1 *Let $X = [0,1]$ with the Lebesgue measure $\nu = dx$ and $\mathcal{H} = \mathbb{C}^2$. We define $\gamma_1 = (0,1)$ and, for all $x \neq 1$, $\gamma_x = (x,0)$. We have that $\mathcal{H}_\nu = \mathcal{H}$ and*

$$(A_\gamma(0,1))(x) = (0,1)\cdot(x,0) = 0 \qquad x \neq 1,$$

so that $A_\gamma(0,1) = 0$ in $L^2([0,1],dx)$ and $\mathrm{Ker}\ A_\gamma \neq 0$.

Remark 20 *In the statement of Proposition 19, there is a little abuse of notation that could generate some confusion in the following. Given $v \in \mathcal{H}$, according to Equation (36), $A_\gamma v$ is a function on X. However, in Proposition 19, $A_\gamma v$ is regarded as an element of $L^2(X,\nu)$, so that it is defined up to a set of null measure with respect to ν. However, for any $\phi \in \mathrm{Im}\ A_\gamma$, there is a unique $v \in \ker {A_\gamma}^\perp$, such that*

$$\phi(x) = \langle v, \gamma_x\rangle_{\mathcal{H}}$$

for ν-almost all $x \in X$. In particular, the function $x \mapsto \langle v, \gamma_x\rangle_{\mathcal{H}}$ is a preferential representative of the equivalence class of ϕ and we use it in the following. This preferential choice is not natural since it strongly depends on γ. In the above example, the equivalence class of the continuous function $\phi(x) = x$ is represented by the discontinuous function

$$f(x) = \left(A_\gamma(1,0)\right)(x) = \begin{cases} x & x \neq 1 \\ 0 & x = 1 \end{cases} .$$

On the contrary, if γ is weakly continuous and $\mathrm{supp}\ \nu = X$, f is easily characterized: it is the only continuous function in the equivalence class of ϕ and this choice of the preferential representative is the usual one in functional analysis.

Remark 21 *We observe that, if γ is weakly continuous and X is compact, γ is a bounded Carleman map and Condition (44) holds so that A_γ is a compact operator. To have compactness of A_γ as operator in $\mathcal{C}(X)$, we need that γ is continuous, see Proposition 18.*

The following proposition gives necessary and sufficient conditions to ensure that γ is a bounded Carleman map.

Proposition 22 *Assume that γ is measurable and locally bounded. The following conditions are equivalent:*

1. *γ is a bounded Carleman map;*
2. *The integral operator L_Γ with kernel Γ is everywhere defined.*

If one of the above conditions hold, then $A_\gamma^* A_\gamma = L_\Gamma$.

Proof Assume that γ is a bounded Carleman map, by Proposition (19), then A_γ is well defined bounded operator, the thesis follows from Equation (42).

Conversely, the second condition in the statement implies that, for all $\phi \in L^2(X, \nu)$, the function $\Gamma(x, \cdot)\phi(\cdot)$ is in $L^1(X, \nu)$ and that the function

$$(L_\Gamma \phi)(\cdot) := \int_X \Gamma(\cdot, t)\phi(t)\, \nu(t)$$

is in $L^2(X, \nu)$.

Let now $\mathcal{D}$ be the set of all $\phi \in L^2(X, \nu)$ such the map

$$X \ni x \mapsto \gamma(x)\phi(x) \in \mathcal{H}$$

is weakly integrable. We show that $\mathcal{D}$ is dense in $\mathcal{H}$. Indeed, let ϕ be a continuous function with compact support and $v \in \mathcal{H}$. Since $\langle \gamma(\cdot), v\rangle_{\mathcal{H}}$ is is bounded on the support of ϕ and ν is a Radon measure, the map $x \mapsto \phi(x)\langle \gamma(\cdot), v\rangle_{\mathcal{H}}$ is integrable. The claim follows by the density of continuous functions with compact support in $L^2(X, \nu)$.

We define $B : \mathcal{D} \to \mathcal{H}$

$$B\phi = \int_X \gamma(x)\phi(x)\, d\nu(x),$$

where the integral converges in the weak sense. We claim that, for all ϕ, $\psi \in \mathcal{D}$

$$\langle B\phi, B\psi\rangle_{\mathcal{H}} = \langle L_\Gamma \phi, \psi\rangle_{L^2(X,\nu)}\,. \tag{45}$$

Indeed, since the weak integral commutes with the scalar product, one has

$$\begin{aligned}
\langle B\phi, B\psi\rangle_{\mathcal{H}} &= \int_X \left(\int_X \left(\phi(t)\overline{\psi(x)}\, \langle \gamma(t), \gamma(x)\rangle_{\mathcal{H}} \right) d\nu(t) \right) d\nu(x). \\
&= \int_X \left(\int_X \Gamma(x,t)\phi(t)\, d\nu(t) \right) \overline{\psi(x)}\, d\nu(x). \\
&= \int_X \left(L_\Gamma \phi \right)(x)\overline{\psi(x)}\, d\nu(x)
\end{aligned}$$

$$= \langle L_\Gamma \phi, \psi \rangle_{L^2(X,\nu)}.$$

From Equation (45), it follows that L_Γ is a positive operator and that $\|B\phi\|_{\mathcal{H}} = \left\| L_\Gamma^{\frac{1}{2}} \phi \right\|_{L^2(X,\nu)}$. In particular, B extends to a bounded operator, denoted again by B, on $\mathcal{H}$.

The adjoint of B is a bounded operator from $\mathcal{H}$ into $L^2(X,\nu)$ explicitly given by, if $v \in \mathcal{H}$ and $\phi \in \mathcal{D}$,

$$\begin{aligned} \langle B^* v, \phi \rangle_{L^2(X,\nu)} &= \langle v, B\phi \rangle_{\mathcal{H}} \\ &= \int_X \langle v, \gamma(x) \rangle_{\mathcal{H}} \overline{\phi(x)} \, d\nu(x) \end{aligned}$$

that is, $(B^* v)(x) = \langle v, \gamma(x) \rangle_{\mathcal{H}}$ for ν-almost all $x \in X$.

This fact implies, in particular, that $\langle v, \gamma(\cdot) \rangle_{\mathcal{H}}$ is in $L^2(X,\nu)$, that is, that γ is a bounded Carleman map.

■

We need the fact that γ is locally bounded only to ensure that B is densely defined. This assumption is clearly satisfied if γ is weakly continuous.

A.1 Reproducing kernel Hilbert spaces

We review the properties of reproducing kernel Hilbert spaces, Aronszajn (1950), Schwartz (1964), Saitoh (1988, 1997).

Let X be a locally compact second countable topological space and $\Gamma : X \times X \to \mathbb{C}$ be a Aronszajan kernel, Aronszajn (1950), Schwartz (1964), that is, a map such that

1. $\Gamma(x,t) = \overline{\Gamma(t,x)}$ for all $x, t \in X$;
2. for any $\ell \in \mathbb{N}$, $x_1, \ldots, x_\ell \in X$, the $\ell \times \ell$ matrix $(\Gamma(x_i, x_j))_{i,j}$ is positive.

For all $x \in X$, we let γ_x be the function $\Gamma(\cdot, x)$. We denote by $\mathcal{H}$ be the corresponding reproducing kernel Hilbert space, that is, the unique Hilbert space of complex functions on X such that

$$\mathcal{H} = \overline{\operatorname{span}}\{\gamma_x \,|\, x \in X\} \tag{46}$$

$$v(x) = \langle v, \gamma_x \rangle_{\mathcal{H}} \quad x \in X, v \in \mathcal{H}. \tag{47}$$

In the following we assume that $\mathcal{H}$ is separable. The following proposition gives a sufficient condition.

Proposition 23 *With the above notations, if the function Γ is locally bounded and separately continuous, then $\mathcal{H}$ is separable.*

Proof The proof of Proposition 17 does not depend on the fact that $\mathcal{H}$ is separable, so the assumption is equivalent to the fact that, for all $v \in \mathcal{H}$, the map $x \to \langle \gamma_x, v \rangle_{\mathcal{H}}$ is continuous. Since X is separable, there exists a dense denumerable subset X_0 of X and we let

$$\mathcal{H}_0 = \overline{\text{span}}\{\gamma_x \,|\, x \in X_0\}.$$

Clearly, $\mathcal{H}_0$ is separable, we claim that $\mathcal{H}_0 = \mathcal{H}$. Assume that $\mathcal{H}_0 \neq \mathcal{H}$ and choose $v \in \mathcal{H}^{\perp}$, $v \neq 0$, then, for all $x \in X_0$, $\langle \gamma_x, v \rangle_{\mathcal{H}} = 0$ and, by the fact that $x \to \langle \gamma_x, v \rangle_{\mathcal{H}}$ is continuous and X_0 is dense, it follows that, for all $x \in X$, $\langle \gamma_x, v \rangle_{\mathcal{H}} = 0$. By Equation (46), it follows that $v = 0$ and this is a contradiction. ■

We let γ from X to $\mathcal{H}$ defined by $\gamma(x) = \gamma_x$. By means of Equation (47), one has that

$$(A_\gamma v)(x) = v(x) \qquad \forall x \in X,$$

that is, A_γ is the canonical immersion j of $\mathcal{H}$ into $\mathbb{C}^X$. By Equation (46), $\mathcal{H}_\gamma = \mathcal{H}$ and

$$\Gamma(x,t) = \langle \gamma(t), \gamma(x) \rangle_{\mathcal{H}},$$

which agrees with Definition (35).

The following proposition characterizes the reproducing kernel Hilbert spaces being γ a continuous bounded Carleman map.

Proposition 24 *Let ν be a Radon measure on X such that $\operatorname{supp} \nu = X$ and γ be a continuous bounded Carleman map with respect to ν, then, for all $\phi \in L^2(X,\nu)$, there is a unique $v \in \mathcal{H}$ such that*

$$v = L_\Gamma^{\frac{1}{2}} \phi. \tag{48}$$

With this identification, $L_\Gamma^{\frac{1}{2}}$ is a unitary operator from $(\text{Ker } L_\Gamma)^\perp$ onto $\mathcal{H}$.

In particular, if L_Γ is compact, there is a basis $(\phi_n)_{n\in\mathbb{N}}$ of $L^2(X,\nu)$ and a sequence $(t_n)_{n\in\mathbb{N}}$ such that

$$\phi_n \in \mathcal{H} \quad \text{if } t_n > 0 \tag{49}$$

$$t_n \geq 0 \tag{50}$$

$$L_\Gamma = \sum_n t_n \langle \cdot, \phi_n \rangle_{L^2(X,\nu)} \phi_n \tag{51}$$

and

$$\begin{aligned}
\mathcal{H} &= \{\phi \in L^2(X,\nu) \mid \sum_n \frac{1}{t_n} |\left\langle \phi, \phi_n \right\rangle_{L^2(X,\nu)}|^2 < +\infty\} && (52) \\
\|\phi\|^2_{\mathcal{H}} &= \sum_n \frac{1}{t_n} |\left\langle \phi, \phi_n \right\rangle_{L^2(X,\nu)}|^2 && (53) \\
\Gamma(t,x) &= \sum_n t_n \phi_n(t) \overline{\phi_n(x)}, && (54)
\end{aligned}$$

where the series converges absolutely.

If γ is continuous, the series converges uniformly on the compact subsets of $X \times X$.

Proof Since γ is a continuous bounded Carleman map and $\mathrm{supp}\ \nu = X$, by Proposition (19), A_γ is a well defined bounded injective operator.

By polar decomposition of the adjoint A^*_γ, one has that

$$A^*_\gamma = W(A_\gamma A^*_\gamma)^{\frac{1}{2}} \tag{55}$$

where W is a partial isometry from $L^2(X,\nu)$ to $\mathcal{H}$ with kernel being equal to the kernel of A^*_γ and with range being equal to the closure of the range of A^*_γ.

By Proposition (19), one has that $A_\gamma A^*_\gamma = L_\Gamma$, so the kernel of W is the kernel of L_Γ and, since A_γ is injective, W is surjective, that is, WW^* is the identity.

By Equation (55), it follows that

$$A_\gamma W = L_\Gamma^{\frac{1}{2}}. \tag{56}$$

Let now $\phi \in L^2(X,\nu)$ and let $v = W\phi$. Since $\mathcal{H}$ is a reproducing kernel Hilbert space, then $A_\gamma v = v$ and, hence, $v = W\phi$. The fact that v is unique follows from the injectivity of A_γ and, with the identification given by Equation (48), $W = L_\Gamma^{\frac{1}{2}}$. The fact that W is surjective implies that $L_\Gamma^{\frac{1}{2}}$ is a unitary operator from $\mathrm{Ker}\ L_\gamma^\perp$ onto $\mathcal{H}$.

Assume now that L_γ is a compact positive operator, by spectral theorem,

$$L_\Gamma = \sum_n t_n \left\langle \cdot, \psi_n \right\rangle_{L^2(X,\nu)} \psi_n,$$

where $t_n \geq 0$ and $(\psi_n)_{n\in\mathbb{N}}$ is a basis of $L^2(X,\nu)$. If $t_n = 0$ let $\phi_n = \psi_n$ and, if $t_n > 0$, let $\phi_n = \frac{1}{\sqrt{t_n}} W\psi_n \in \mathcal{H}$. By the first part of the proposition, $\phi_n = \psi_n$, as elements of $L^2(X,\nu)$, so that Equations (49), (50) and (51) hold.

Equalities (52) and (53) are a restatement of the fact that W is a partial surjective isometry.

In particular, we have that $(\sqrt{t_n}\phi_n)_{t_n>0}$ is a basis of $\mathcal{H}$, so that, given $x \in X$,

$$\begin{aligned}\gamma(x) &= \sum_{t_n>0} \left\langle \gamma(x), \sqrt{t_n}\phi_n \right\rangle_{\mathcal{H}} \sqrt{t_n}\phi_n \\ &= \sum_n t_n \overline{\phi_n(x)} \phi_n\end{aligned}$$

Taking the scalar product with $\gamma(t)$ one has that

$$\Gamma(t,x) = \sum_n t_n \phi_n(t) \overline{\phi_n(x)},$$

where the convergence is pointwise. Observing that, by Schwartz inequality,

$$\sum_{n=M}^{N} t_n |e_n(t) \overline{e_n(x)}| \leq \sqrt{\sum_{n=M}^{N} t_n \Gamma(t,t)} \sqrt{\sum_{n=M}^{N} t_n \Gamma(x,x)},$$

the absolute converge follows.

If γ is continuous, the map $x \mapsto \Gamma(x,x)$ is continuous, so, letting $t = x$, one has a series of positive continuous functions converging to a continuous function. By Dini theorem, the convergence is uniform on compact subsets. ∎

The above proposition allows us to identify the elements of $\mathcal{H}$ with the only continuous function on X whose equivalence class belongs to the range of $L_\Gamma^{\frac{1}{2}}$, extending a result of Cucker and Smale (2002). The second half of the proposition is an extension of the Mercer theorem.

Remark 25 *If* $\operatorname{supp} \nu \neq X$*, the statements of the above proposition holds replacing* X *with* $\operatorname{supp} \nu$ *and* $\mathcal{H}$ *with* $\mathcal{H}_\nu$*. We use the assumption that* γ *is weakly continuous only to characterize the range of* $L_\Gamma^{\frac{1}{2}}$*. Without this hypothesis,* $L_\Gamma^{\frac{1}{2}}$ *is a unitary operator onto* $\operatorname{Ker} A_\gamma^\perp$ *(however see Remark 20 above for the choice of the preferential representative).*

Remark 26 *Given an abstract Hilbert space* $\mathcal{H}$ *and a map* $\gamma : X \to \mathcal{H}$ *such that Condition (46) holds, that is* $\mathcal{H}_\gamma = \mathcal{H}$*, by Proposition 13, we can identify* $\mathcal{H}$ *with the reproducing kernel Hilbert space with kernel* $\Gamma(x,t)$ *and, with this identification, the results of the present section hold (see, also, Remark below).*

A.2 Integral operator

We review some properties of integral operators defined either on $L^2(X,\nu)$ or on a reproducing kernel Hilbert space.

Let X and T be two locally compact second countable topological spaces, ν and μ two Radon measures on X and T, respectively. Let $K : X \times T \to \mathbb{C}$ be a measurable function such that

1. for ν-almost all $x \in X$, $K(x, \cdot) \in L^2(T, \mu)$;
2. for all $\phi \in L^2(X, \nu)$, the map $\int_T K(\cdot, t)\phi(t)\, d\mu(t)$ is in $L^2(X, \nu)$.

The above assumptions ensure that the integral operator L_K

$$(L_K\phi)(x) = \int_Y K(x, t)\phi(t)\, d\mu(t)$$

is a bounded operator from $L^2(T, \mu)$ into $L^2(X, \nu)$.

Let now $\mathcal{H}$ be an Hilbert space and $j : \mathcal{H} \to L^2(T, \mu)$ a bounded operator. We define γ from X to $\mathcal{H}$ as

$$\gamma(x) = j^*(\overline{K(x, \cdot)}).$$

The following proposition gives the relation with the formalism of the Carleman maps.

Proposition 27 *The map γ is a bounded Carleman map and*

$$A_\gamma = L_K j.$$

If $\mathcal{H}$ is $L^2(T, \mu)$ and j is the identity, then

$$\begin{aligned} \gamma(x) &= \overline{K(x, \cdot)} \\ \Gamma(x, x') &= \int_Y \overline{K(x', t)} K(x, t)\, d\mu(t) \end{aligned}$$

If $\mathcal{H}$ is a reproducing kernel Hilbert space on Y with kernel Δ and $j = A_\Delta$ is the canonical immersion, then

$$\begin{aligned} \gamma(x) &= \int_Y \overline{K(x, t)}\delta_t\, \mu(t) \\ \Gamma(x, x') &= \int_Y \left(\int_Y \left(\overline{K(x, t)} K(x', t')\Delta(t', t) \right) d\mu(t) \right) d\mu(t'), \end{aligned}$$

where the first integral converges in the weak topology and $\delta_t = \Delta(\cdot, t)$.

Proof Let $v \in \mathcal{H}$, then, for all $x \in X$

$$\begin{aligned} \langle v, \gamma(x) \rangle_{\mathcal{H}} &= \int_Y K(x, t)(jv)(t)\, d\mu(t) \\ &= (L_K jv)(x) \end{aligned}$$

Since the range of L_K is in $L^2(X,\nu)$, it follows that γ is a bounded Carleman map and that $A_\gamma = L_K j$.

If $\mathcal{H} = L^2(X,\nu)$, clearly $\gamma(x) = \overline{K(x,\cdot)}$, and

$$\begin{aligned}\Gamma(x,x') &= \langle \gamma(t), \gamma(x) \rangle_{\mathcal{H}} \\ &= \int_Y \overline{K(x',t)} K(x,t)\, d\mu(t)\end{aligned}$$

If $\mathcal{H}$ is a reproducing kernel Hilbert space, since $j^* = A_\delta^*$, by Equation (41),

$$\gamma(x) = \int_Y \overline{K(x,t)} \delta_t \, d\mu(t),$$

and, since the weak integral commutes with the scalar product we have the last equation in the statement of the proposition. ■

The class of integral operators we considered in this section is not the most general. There are integral operators with kernel K such that $K(x,\cdot)$ is not in $L^2(T,\mu)$, Halmos and Sunder (1978). The class of kernel we considered is precisely the class of Carleman kernel, according to the definition of Halmos and Sunder (1978).

Remark 28 *Let now γ be a bounded Carleman map from X into $\mathcal{H} = L^2(T,\mu)$, a result of Halmos and Sunder (1978) ensures that there is a measurable complex function K defined on $X \times Y$ such that, for ν-almost all $x \in X$, $\gamma(x) = K(x,\cdot)$ and A_γ is the integral operator of kernel K. However, if $\{K(x,\cdot) \,|\, x \in X\}$ is dense in $L^2(X,\nu)$, by Remark 28, we can identify $L^2(T,\mu)$ with the reproducing kernel Hilbert space on X and kernel $\Gamma(x,x')$ and, with this identification, A_γ is the canonical immersion of $\mathcal{H}$ into $L^2(X,\nu)$.*

References

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

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. ISSN 0266-5611.

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. ISSN 0266-5611.

Felipe Cucker and Steve Smale. On the mathematical foundations of learning. *Bull. Amer. Math. Soc. (N.S.)*, 39(1):1–49 (electronic), 2002. ISSN 0273-0979.

E. De Vito, A. Caponnetto, and L. Rosasco. Model selection for regularized least-squares algorithm in learning theory. *submitted*, 2004.

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

Heinz W. Engl, Martin Hanke, and Andreas Neubauer. *Regularization of inverse problems*, volume 375 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1996. ISBN 0-7923-4157-0.

Theodoros Evgeniou, Massimiliano Pontil, and Tomaso Poggio. Regularization networks and support vector machines. *Adv. Comput. Math.*, 13(1):1–50, 2000. ISSN 1019-7168.

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. ISBN 0-273-08642-1.

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. ISSN 0378-620X.

Paul Richard Halmos and Viakalathur Shankar Sunder. *Bounded integral operators on L^2 spaces*, volume 96 of *Ergebnisse der Mathematik und ihrer Grenzgebiete [Results in Mathematics and Related Areas]*. Springer-Verlag, Berlin, 1978. ISBN 3-540-08894-6.

John L. Kelley. *General topology*. D. Van Nostrand Company, Inc., Toronto-New York-London, 1955.

Vera Kurkova. Supervised learning as an inverse problem. Technical Report 960, Institute of Computer Science, Academy of Sciences of the Czech Republic, April 2004.

Peter Mathé and Sergei V. Pereverzev. Discretization strategy for linear ill-posed problems in variable Hilbert scales. *Inverse Problems*, 19(6):1263–1277, 2003. ISSN 0266-5611.

Colin McDiarmid. On the method of bounded differences. In *Surveys in combinatorics, 1989 (Norwich, 1989)*, volume 141 of *London Math. Soc. Lecture Note Ser.*, pages 148–188. Cambridge Univ. Press, Cambridge, 1989.

M. Thamban Nair and Eberhard Schock. A discrepancy principle for Tikhonov regularization with approximately specified data. *Ann. Polon. Math.*, 69(3):197–205, 1998. ISSN 0066-2216.

Sergei Pereverzev and Eberhard 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. ISSN 0163-0563.

Tomaso Poggio and Steve Smale. The mathematics of learning: dealing with data. *Notices Amer. Math. Soc.*, 50(5):537–544, 2003. ISSN 0002-9920.

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. ISSN 0022-247X.

Cynthia Rudin. A different type of convergence for statistical learning algorithms. Technical report, Program in Applied and Computational Mathematics Princeton University, 2004.

S. Saitoh. *Integral transforms, reproducing kernels and their applications*, volume 369 of *Pitman Research Notes in Mathematics Series*. Longman, Harlow, 1997. ISBN 0-582-31758-4.

Saburou Saitoh. *Theory of reproducing kernels and its applications*, volume 189 of *Pitman Research Notes in Mathematics Series*. Longman Scientific & Technical, Harlow, 1988. ISBN 0-582-03564-3.

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

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. ISBN 0-7923-3583-X. Translated from the 1990 Russian original by R. A. M. Hoksbergen and revised by the authors.

Vladimir N. Vapnik. *Statistical learning theory*. Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons Inc., New York, 1998. ISBN 0-471-03003-1. A Wiley-Interscience Publication.

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

Grace 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. ISBN 0-89871-244-0.