Computer Science and Artificial Intelligence Laboratory
Technical Report

MIT-CSAIL-TR-2006-062
CBCL-264

September 10, 2006

Optimal Rates for Regularization Operators in Learning Theory

Andrea Caponnetto

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

CSAIL

Optimal rates for regularization operators in learning theory

Andrea Caponnetto [a] [b] [c]

[a] *C.B.C.L.,McGovern Institute, Massachusetts Institute of Technology, Bldg. 46-5155, , 77 Massachusetts Avenue, Cambridge, MA 02139*

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

[c] *Department of Computer Science, University of Chicago, 1100 East 58th Street, Chicago, IL 60637*

Abstract

We develop some new error bounds for learning algorithms induced by regularization methods in the regression setting. The "hardness" of the problem is characterized in terms of the parameters r and s, the first related to the "complexity" of the target function, the second connected to the *effective dimension* of the marginal probability measure over the input space. We show, extending previous results, that by a suitable choice of the regularization parameter as a function of the number of the available examples, it is possible attain the optimal minimax rates of convergence for the expected squared loss of the estimators, over the family of priors fulfilling the constraint $r + s \geq \frac{1}{2}$. The setting considers both labelled and unlabelled examples, the latter being crucial for the optimality results on the priors in the range $r < \frac{1}{2}$.

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

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

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

This work was also supported by the NSF grant 0325113, by the FIRB Project ASTAA and the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778.

1. Introduction

We consider the setting of semi-supervised statistical learning. We assume $Y \subset [-M, M]$ and the supervised part of the training set equal to

$$\mathbf{z} = (z_1, \ldots, z_m),$$

with $z_i = (x_i, y_i)$ drawn i.i.d. according to the probability measure ρ over $Z = X \times Y$. Moreover consider the unsupervised part of the training set $(x^u_{m+1}, \ldots, x^u_{\tilde{m}})$, with x^u_i drawn i.i.d. according to the marginal probability measure over X, ρ_X. For sake of brevity we will also introduce the complete training set

$$\tilde{\mathbf{z}} = (\tilde{z}_1, \ldots, \tilde{z}_{\tilde{m}}),$$

with $\tilde{z}_i = (\tilde{x}_i, \tilde{y}_i)$, where we introduced the compact notations $\tilde{x}_i$ and $\tilde{y}_i$, defined by

$$\tilde{x}_i = \begin{cases} x_i & \text{if} \quad 1 \leq i \leq m, \\ x^u_i & \text{if} \quad m < i \leq \tilde{m}, \end{cases}$$

and

$$\tilde{y}_i = \begin{cases} \frac{\tilde{m}}{m} y_i & \text{if} \quad 1 \leq i \leq m, \\ 0 & \text{if} \quad m < i \leq \tilde{m}. \end{cases}$$

It is clear that, in the supervised setting, the semi-supervised part of the training set is missing, whence $\tilde{m} = m$ and $\tilde{\mathbf{z}} = \mathbf{z}$.

In the following we will study the generalization properties of a class of estimators $f_{\tilde{\mathbf{z}},\lambda}$ belonging to the *hypothesis space* $\mathcal{H}$: the RKHS of functions on X induced by the bounded Mercer kernel K (in the following $\kappa = \sup_{x \in X} K(x, x)$). The learning algorithms that we consider, have the general form

$$f_{\tilde{\mathbf{z}},\lambda} = G_\lambda(T_{\tilde{\mathbf{x}}})\, g_{\mathbf{z}}, \tag{1}$$

where $T_{\tilde{\mathbf{x}}} \in \mathcal{L}(\mathcal{H})$ is given by,

$$T_{\tilde{\mathbf{x}}} f = \frac{1}{\tilde{m}} \sum_{i=1}^{\tilde{m}} K_{\tilde{x}_i} \langle K_{\tilde{x}_i}, f \rangle_{\mathcal{H}},$$

$g_{\mathbf{z}} \in \mathcal{H}$ is given by,

$$g_{\mathbf{z}} = \frac{1}{\tilde{m}} \sum_{i=1}^{\tilde{m}} K_{\tilde{x}_i} \tilde{y}_i = \frac{1}{m} \sum_{i=1}^{m} K_{x_i} y_i,$$

and the *regularization parameter* λ lays in the range $(0, \kappa]$. We will often used the shortcut notation $\dot{\lambda} = \frac{\lambda}{\kappa}$.

The functions $G_\lambda : [0, \kappa] \to \mathbb{R}$, which select the *regularization method*, will be characterized in terms of the constants A and B_r in $[0, +\infty]$, defined as follows

$$A = \sup_{\lambda \in (0,\kappa]} \sup_{\sigma \in [0,\kappa]} |(\sigma + \lambda) G_\lambda(\sigma)| \tag{2}$$

$$B_r = \sup_{t \in [0,r]} \sup_{\lambda \in (0,\kappa]} \sup_{\sigma \in [0,\kappa]} |1 - G_\lambda(\sigma)\sigma|\, \sigma^t \lambda^{-t}, \quad r > 0. \tag{3}$$

Finiteness of A and B_r (with r over a suitable range) are standard in the literature of ill-posed inverse problems (see for reference [12]). Regularization methods have been recently studied in the context of learning theory in [13, 9, 8, 10, 1].

The main results of the paper, Theorems 1 and 2, describe the convergence rates of $f_{\tilde{\mathbf{z}},\lambda}$ to the *target function* $f_{\mathcal{H}}$. Here, the target function is the “best” function which can be arbitrarily well approximated by elements of our hypothesis space $\mathcal{H}$. More formally, $f_{\mathcal{H}}$ is the projection of the regression function $f_\rho(x) = \int_Y y d\rho_{|x}(y)$ onto the closure of $\mathcal{H}$ in $\mathcal{L}^2(X, \rho_X)$.

The convergence rates in Theorems 1 and 2, will be described in terms of the constants C_r and D_s in $[0, +\infty]$ characterizing the probability measure ρ. These constants can be described in terms of the integral operator $L_K : \mathcal{L}^2(X, \rho_X) \to \mathcal{L}^2(X, \rho_X)$ of kernel K. Note that the same integral operator is denoted by T, when seen as a bounded operator from $\mathcal{H}$ to $\mathcal{H}$.

The constants C_r characterize the conditional distributions $\rho_{|x}$ through $f_{\mathcal{H}}$, they are defined as follows

$$C_r \quad = \quad \begin{cases} \kappa^r \left\| L_K^{-r} f_{\mathcal{H}} \right\|_\rho & \text{if } f_{\mathcal{H}} \in \text{Im } L_K^r \\ +\infty & \text{if } f_{\mathcal{H}} \notin \text{Im } L_K^r \end{cases}, \quad r > 0. \tag{4}$$

Finiteness of C_r is a common *source condition* in the inverse problems literature (see [12] for reference). This type of condition has been introduced in the statistical learning literature in [7, 18, 3, 17, 4].

The constants D_s characterize the marginal distribution ρ_X through the *effective dimension* $\mathcal{N}(\lambda) = \text{Tr}\left[T(T+\lambda)^{-1}\right]$, they are defined as follows

$$D_s \quad = \quad 1 \vee \sup_{\dot{\lambda} \in (0,1]} \sqrt{\mathcal{N}(\lambda)\dot{\lambda}^s}, \quad s \in (0,1]. \tag{5}$$

Finiteness of D_s was implicitly assumed in [3, 4].

The paper is organized as follows. In Section 2 we focus on the RLS estimators $f^{\text{ls}}_{\tilde{\mathbf{z}},\lambda}$, defined by the optimization problem

$$f^{\text{ls}}_{\tilde{\mathbf{z}},\lambda} = \operatorname*{argmin}_{f \in \mathcal{H}} \frac{1}{\tilde{m}} \sum_{i=1}^{\tilde{m}} \left(f(\tilde{x}_i) - \tilde{y}_i\right)^2 + \lambda \left\|f\right\|_K^2,$$

and corresponding to the choice $G_\lambda(\sigma) = (\sigma + \lambda)^{-1}$ (see for example [5, 7, 18]). The main result of this Section, Theorem 1, extends the convergence analysis performed in [3, 4] from the range $r \geq \frac{1}{2}$ to arbitrary $r > 0$ and $s \geq \frac{1}{2} - r$. Corollary 1 gives optimal s-independent rates for $r > 0$.

The analysis of the RLS algorithm is a useful preliminary step for the study of general regularization methods, which is performed in Section 3. The aim of this Section is develop a s-dependent analysis in the case $r > 0$ for general regularization methods G_λ. In Theorem 2 we extend the results given in Theorem 1 to general regularization methods. In fact, in Theorem 2 we obtain optimal minimax rates of convergence (see [3, 4]) for the involved problems, under the assumption that $r + s \geq \frac{1}{2}$. Finally, Corollary 2 extends Corollary 1 to general G_λ.

In Sections 4 and 5 we give the proofs of the results stated in the previous Sections.

2. Risk bounds for RLS.

We state our main result concerning the convergence of $f^{\text{ls}}_{\tilde{\mathbf{z}},\lambda}$ to $f_{\mathcal{H}}$. The function $|x|_+$, appearing in the text of Theorem 1, is the "positive part" of x, that is $\frac{x+|x|}{2}$.

Theorem 1. *Let r and s be two reals in the interval $(0,1]$, fulfilling the constraint $r + s \geq \frac{1}{2}$.*

Furthermore, let m and λ satisfy the constraints $\lambda \leq \|T\|$ and

$$\dot{\lambda} = \left(\frac{4D_s \log \frac{6}{\delta}}{\sqrt{m}}\right)^{\frac{2}{2r+s}}, \tag{6}$$

for $\delta \in (0,1)$. Finally, assume $\tilde{m} \geq m\dot{\lambda}^{-|1-2r|_+}$. Then, with probability greater than $1-\delta$, it holds

$$\left\| f^{\text{ls}}_{\tilde{\mathbf{z}},\lambda} - f_{\mathcal{H}} \right\|_\rho \leq 4(M + C_r) \left(\frac{4D_s \log \frac{6}{\delta}}{\sqrt{m}}\right)^{\frac{2r}{2r+s}}.$$

Some comments are in order.

First, while eq. (6) expresses λ in terms of m and δ, it is straightforward verifying that the condition $\lambda \leq \|T\|$ is satisfied for

$$\sqrt{m} \geq 4D_s \left(\frac{\kappa}{\|T\|}\right)^{r+\frac{s}{2}} \log \frac{6}{\delta}.$$

Second, the asymptotic rate of convergence $O\left(m^{-\frac{r}{2r+s}}\right)$ of $\left\|f_{\tilde{\mathbf{z}},\lambda}^{\mathrm{ls}} - f_{\mathcal{H}}\right\|_{\rho}$, is optimal in the minimax sense of [11, 4]. Indeed, in Th.2 of [4], it was showed that this asymptotic order is optimal over the class of probability measures ρ, such that $f_{\mathcal{H}} \in \mathrm{Im}\ L_K^r$, and the eigenvalues of T, λ_i, have asymptotic order $O\left(i^{-\frac{1}{s}}\right)$. In fact, the condition on $f_{\mathcal{H}}$ implies the finiteness of C_r and the condition on the spectrum of T implies the finiteness of D_s (see Prop.3 in [4]).

Upper bounds of the type given in [17] or [3] (and stated in [6, 4], under a weaker noise condition, and in the more general framework of vector-valued functions) can be obtained as a corollary of Theorem 1, considering the case $r \geq \frac{1}{2}$.

However, the advantage of using extra unlabelled data, is evident when $r < \frac{1}{2}$. In this case, the unlabelled examples (enforcing the assumption $\tilde{m} \geq m\dot{\lambda}^{2r-1}$) allow (if $s \geq \frac{1}{2} - r$) again the rate of convergence $O\left(m^{-\frac{r}{2r+s}}\right)$, over classes of measures ρ defined in terms of finiteness of the constants C_r and D_s. It is not known to the author whether the same rate of convergence can be achieved by the RLS estimator, for $s < \frac{1}{2} - r$.

A simple corollary of Theorem 1, encompassing all the values of r in $(0, 1]$, can be obtained observing that $D_1 = 1$, for every kernel K and marginal distribution ρ_X (see Prop. 2).

Corollary 1. *Let $\tilde{m} \geq m\dot{\lambda}^{-|1-2r|_+}$ hold with r in the interval $(0, 1]$. If λ satisfies the constraints $\lambda \leq \|T\|$ and*

$$\dot{\lambda} = \left(\frac{4 \log \frac{6}{\delta}}{\sqrt{m}}\right)^{\frac{2}{2r+1}},$$

for $\delta \in (0, 1)$, then, with probability greater than $1 - \delta$, it holds

$$\left\|f_{\tilde{\mathbf{z}},\lambda}^{\mathrm{ls}} - f_{\mathcal{H}}\right\|_{\rho} \leq 4(M + C_r)\left(\frac{4 \log \frac{6}{\delta}}{\sqrt{m}}\right)^{\frac{2r}{2r+1}}.$$

3. Risk bounds for general regularization methods.

In this Section we state a result which generalizes Theorem 1 from RLS to general regularization algorithms of type described by equation (1). In this general framework we need $(\dot{\lambda}^{-|2-2r-s|_+} - 1)m$ unlabelled examples in order to get minimax optimal rates, slightly more than the $(\dot{\lambda}^{-|1-2r|_+} - 1)m$ required in Theorem 1 for the RLS estimator. We adopt the same notations and definitions introduced in the previous section.

Theorem 2. *Let $r > 0$ and $s \in (0, 1]$ fulfill the constraint $r + s \geq \frac{1}{2}$. Furthermore, let m and λ satisfy the constraints $\lambda \leq \|T\|$ and*

$$\dot{\lambda} = \left(\frac{4D_s \log \frac{6}{\delta}}{\sqrt{m}}\right)^{\frac{2}{2r+s}}, \tag{7}$$

for $\delta \in (0, \frac{1}{3})$. Finally, assume $\tilde{m} \geq 4 \vee m\dot{\lambda}^{-|2-2r-s|_+}$. Then, with probability greater than $1 - 3\delta$, it holds

$$\|f_{\tilde{\mathbf{z}},\lambda} - f_{\mathcal{H}}\|_{\rho} \leq E_r \left(\frac{4D_s \log \frac{6}{\delta}}{\sqrt{m}}\right)^{\frac{2r}{2r+s}},$$

where

$$E_r = C_r\,(30A + 2(3+r)B_r + 1) + 9MA. \tag{8}$$

The proof of the above Theorem is postponed to Section 5.

For the particular case $G_\lambda(\sigma) = (\sigma+\lambda)^{-1}$, $f_{\tilde{\mathbf{z}},\lambda} = f^{\mathrm{ls}}_{\tilde{\mathbf{z}},\lambda}$ and the result above can be compared with Theorem 1. In this case, it is easy to verify that $A = 1$, $B_r \leq 1$ for $r \in [0,1]$ and $C_r = +\infty$ for $r > 1$. The maximal value of r for which $C_r < +\infty$ is usually denoted as the *qualification* of the regularization method.

For a description of the properties of common regularization methods, in the inverse problems literature we refer to [12]. In the context of learning theory a review of these techniques can be found in [10] and [1]. In particular in [10] some convergence results of algorithms induced by Lipschitz continuous G_λ can be found.

A simple corollary of Theorem 2 which generalizes Corollary 1 to arbitrary regularization methods, can be obtained observing that $D_1 = 1$, for every kernel K and marginal distribution ρ_X (see Prop. 2).

Corollary 2. *Let $\tilde{m} \geq 4 \vee m\dot{\lambda}^{-|1-2r|_+}$ hold with $r > 0$. If λ satisfies the constraints $\lambda \leq \|T\|$ and*

$$\dot{\lambda} = \left(\frac{4\log\frac{6}{\delta}}{\sqrt{m}}\right)^{\frac{2}{2r+1}},$$

for some $\delta \in (0, \frac{1}{3})$, then, with probability greater than $1 - 3\delta$, it holds

$$\left\| f^{\mathrm{ls}}_{\tilde{\mathbf{z}},\lambda} - f_{\mathcal{H}} \right\|_\rho \leq E_r \left(\frac{4\log\frac{6}{\delta}}{\sqrt{m}}\right)^{\frac{2r}{2r+1}},$$

with E_r defined by eq. (8).

4. Proof of Theorem 1

In this section we give the proof of Theorem 1. First we need some preliminary propositions.

Proposition 1. *Assume $\lambda \leq \|T\|$ and*

$$\lambda\tilde{m} \geq 16\kappa\mathcal{N}(\lambda)\log^2\frac{6}{\delta}, \tag{9}$$

for some $\delta \in (0,1)$. Then, with probability greater than $1-\delta$, it holds

$$\left\|(T+\lambda)^{\frac{1}{2}}(f^{\mathrm{ls}}_{\tilde{\mathbf{z}},\lambda} - f^{\mathrm{ls}}_\lambda)\right\|_{\mathcal{H}} \leq 8\left(M + \sqrt{\kappa\frac{m}{\tilde{m}}}\left\|f^{\mathrm{ls}}_\lambda\right\|_{\mathcal{H}}\right)\left(\frac{2}{m}\sqrt{\frac{\kappa}{\lambda}} + \sqrt{\frac{\mathcal{N}(\lambda)}{m}}\right)\log\frac{6}{\delta},$$

where

$$f^{\mathrm{ls}}_\lambda = (T+\lambda)^{-1}L_K f_{\mathcal{H}}.$$

Proof. Assuming

$$S_1 := \left\|(T+\lambda)^{-\frac{1}{2}}(T - T_{\tilde{\mathbf{x}}})(T+\lambda)^{-\frac{1}{2}}\right\|_{\mathrm{HS}} < 1, \tag{10}$$

by simple algebraic computations we obtain

$$\begin{aligned}
f^{\mathrm{ls}}_{\tilde{\mathbf{z}},\lambda} - f^{\mathrm{ls}}_{\lambda} &= (T_{\tilde{\mathbf{x}}} + \lambda)^{-1} g_{\mathbf{z}} - (T+\lambda)^{-1} g \\
&= (T_{\tilde{\mathbf{x}}} + \lambda)^{-1} \left\{ (g_{\mathbf{z}} - g) + (T - T_{\tilde{\mathbf{x}}})(T+\lambda)^{-1} g \right\} \\
&= (T_{\tilde{\mathbf{x}}} + \lambda)^{-1} (T+\lambda)^{\frac{1}{2}} \left\{ (T+\lambda)^{-\frac{1}{2}} (g_{\mathbf{z}} - g) + (T+\lambda)^{-\frac{1}{2}} (T - T_{\tilde{\mathbf{x}}})(T+\lambda)^{-1} g \right\} \\
&= (T+\lambda)^{-\frac{1}{2}} \left\{ \mathrm{Id} - (T+\lambda)^{-\frac{1}{2}} (T - T_{\tilde{\mathbf{x}}})(T+\lambda)^{-\frac{1}{2}} \right\}^{-1} \\
&\quad \left\{ (T+\lambda)^{-\frac{1}{2}} (g_{\mathbf{z}} - g) + (T+\lambda)^{-\frac{1}{2}} (T - T_{\tilde{\mathbf{x}}}) f_{\lambda} \right\}.
\end{aligned}$$

Therefore we get

$$\left\| (T+\lambda)^{\frac{1}{2}} (f^{\mathrm{ls}}_{\tilde{\mathbf{z}},\lambda} - f^{\mathrm{ls}}_{\lambda}) \right\|_{\mathcal{H}} \quad \leq \quad \frac{S_2 + S_3}{1 - S_1},$$

where

$$\begin{aligned}
S_2 &:= \left\| (T+\lambda)^{-\frac{1}{2}} (g_{\mathbf{z}} - g) \right\|_{\mathcal{H}}, \\
S_3 &:= \left\| (T+\lambda)^{-\frac{1}{2}} (T - T_{\tilde{\mathbf{x}}}) f_{\lambda} \right\|_{\mathcal{H}}.
\end{aligned}$$

Now we want to estimate the quantities S_1, S_2 and S_3 using Prop. 4. In fact, choosing the correct vector-valued random variables ξ_1, ξ_2 and ξ_3, the following common representation holds,

$$S_h \quad = \quad \left\| \frac{1}{m_h} \sum_{i=1}^{m_h} \xi_h(\omega_i) - \mathbb{E}[\xi_h] \right\|, \qquad\qquad h = 1, 2, 3.$$

Indeed, in order to let the equality above hold, $\xi_1 : X \to \mathcal{L}_{\mathrm{HS}}(\mathcal{H})$ is defined by

$$\xi_1(x)[\cdot] \quad = \quad (T+\lambda)^{-\frac{1}{2}} K_x \langle K_x, \cdot \rangle_{\mathcal{H}} (T+\lambda)^{-\frac{1}{2}},$$

and $m_1 = \tilde{m}$.
Moreover, $\xi_2 : Z \to \mathcal{H}$ is defined by

$$\xi(x, y) \quad = \quad (T+\lambda)^{-\frac{1}{2}} K_x y,$$

with $m_2 = m$.
And finally, $\xi_3 : X \to \mathcal{H}$ is defined by

$$\xi(x) \quad = \quad (T+\lambda)^{-\frac{1}{2}} K_x f^{\mathrm{ls}}_{\lambda}(x),$$

with $m_3 = \tilde{m}$.
Hence, applying three times Prop. 4, we can write

$$\mathbb{P}\left[S_h \leq 2 \left(\frac{H_h}{m_h} + \frac{\sigma_h}{\sqrt{m_h}} \right) \log \frac{6}{\delta} \right] \geq 1 - \frac{\delta}{3}, \qquad\qquad h = 1, 2, 3,$$

where, as it can be straightforwardly verified, the constants H_h and σ_h are given by the expressions

$$\begin{aligned}
H_1 &= 2\tfrac{\kappa}{\lambda}, & \sigma_1^2 &= \frac{\kappa}{\lambda} \mathcal{N}(\lambda), \\
H_2 &= 2M\sqrt{\tfrac{\kappa}{\lambda}}, & \sigma_2^2 &= M^2 \mathcal{N}(\lambda), \\
H_3 &= 2 \left\| f^{\mathrm{ls}}_{\lambda} \right\|_{\mathcal{H}} \tfrac{\kappa}{\sqrt{\lambda}}, & \sigma_3^2 &= \kappa \left\| f^{\mathrm{ls}}_{\lambda} \right\|_{\mathcal{H}} \mathcal{N}(\lambda).
\end{aligned}$$

Now, recalling the assumptions on λ and $\tilde{m}$, with probability greater than $1-\delta/3$, we get

$$\begin{array}{rcl} & S_1 \leq & 2\left(\dfrac{2\kappa}{\tilde{m}\lambda}+\sqrt{\dfrac{\mathcal{N}(\lambda)\kappa}{\tilde{m}\lambda}}\right)\log\dfrac{6}{\delta} \\ \left(\mathcal{N}(\lambda)\geq\dfrac{\|T\|}{\|T\|+\lambda}\geq\dfrac{1}{2}\right) & \leq & 4\dfrac{\kappa\mathcal{N}(\lambda)\log^2\frac{6}{\delta}}{\lambda\tilde{m}}+\sqrt{\dfrac{\kappa\mathcal{N}(\lambda)\log^2\frac{6}{\delta}}{\lambda\tilde{m}}} \\ \text{(eq. (9))} & \leq & \dfrac{1}{4}+\dfrac{1}{2}=\dfrac{3}{4}. \end{array}$$

Hence, since $\tilde{m}\geq m$, with probability greater than $1-\delta$,

$$\begin{array}{rcl} \left\|(T+\lambda)^{\frac{1}{2}}(f^{\mathrm{ls}}_{\tilde{\mathbf{z}},\lambda}-f^{\mathrm{ls}}_{\lambda})\right\|_{\mathcal{H}} & \leq & 4(S_2+S_3) \\ & \leq & 8\left(M+\sqrt{\kappa\dfrac{m}{\tilde{m}}}\left\|f^{\mathrm{ls}}_{\lambda}\right\|_{\mathcal{H}}\right)\left(\dfrac{2}{m}\sqrt{\dfrac{\kappa}{\lambda}}+\sqrt{\dfrac{\mathcal{N}(\lambda)}{m}}\right)\log\dfrac{6}{\delta}. \end{array}$$

□

Proposition 2. *For every probability measure ρ_X and $\lambda>0$, it holds*

$$\|T\|\leq\kappa,$$

and

$$\lambda\mathcal{N}(\lambda)\leq\kappa.$$

Proof. First, observe that

$$\mathrm{Tr}[T]=\int_X \mathrm{Tr}[K_x\langle K_x,\cdot\rangle_{\mathcal{H}}]d\rho_X(x)=\int_X K(x,x)d\rho_X(x)\leq\sup_{x\in X}K(x,x)\leq\kappa.$$

Therefore, since T is a positive self-adjoint operator, the first inequality follows observing that

$$\|T\|\leq\mathrm{Tr}[T]\leq\kappa.$$

The second inequality can be proved observing that, since $\psi_\lambda(\sigma^2):=\frac{\lambda\sigma^2}{\sigma^2+\lambda}\leq\sigma^2$, it holds

$$\lambda\mathcal{N}(\lambda)=\mathrm{Tr}[\psi_\lambda(T)]\leq\mathrm{Tr}[T]\leq\kappa.$$

□

Proposition 3. *Let $f_{\mathcal{H}}\in\mathrm{Im}\,L_K^r$ for some $r>0$. Then, the following estimates hold,*

$$\begin{array}{rcl} \left\|f^{\mathrm{ls}}_{\lambda}-f_{\mathcal{H}}\right\|_{\rho} & \leq & \lambda^r\left\|L_K^{-r}f_{\mathcal{H}}\right\|_{\rho}, \quad \textit{if } r\leq 1 \\ \left\|f^{\mathrm{ls}}_{\lambda}\right\|_{\mathcal{H}} & \leq & \begin{cases} \lambda^{-\frac{1}{2}+r}\left\|L_K^{-r}f_{\mathcal{H}}\right\|_{\rho} & \textit{if } r\leq\frac{1}{2}, \\ \kappa^{-\frac{1}{2}+r}\left\|L_K^{-r}f_{\mathcal{H}}\right\|_{\rho} & \textit{if } r>\frac{1}{2}. \end{cases} \end{array}$$

Proof. The first estimate is standard in the theory of inverse problems, see, for example, [14, 12] or [18].

Regarding the second estimate, if $r \leq \frac{1}{2}$, since T is positive, we can write,

$$\begin{aligned} \left\| f_\lambda^{\mathrm{ls}} \right\|_{\mathcal{H}} &\leq \left\| (T+\lambda)^{-1} L_K f_{\mathcal{H}} \right\|_{\mathcal{H}} \\ &\leq \left\| (T+\lambda)^{-\frac{1}{2}+r} \left(T(T+\lambda)^{-1}\right)^{\frac{1}{2}+r} L_K^{\frac{1}{2}-r} f_{\mathcal{H}} \right\|_{\mathcal{H}} \\ &\leq \left\| (T+\lambda)^{-1} \right\|^{\frac{1}{2}-r} \left\| L_K^{-r} f_{\mathcal{H}} \right\|_\rho \leq \lambda^{-\frac{1}{2}+r} \left\| L_K^{-r} f_{\mathcal{H}} \right\|_\rho . \end{aligned}$$

On the contrary, if $r > \frac{1}{2}$, since by Prop. 2 $\|T\| \leq \kappa$, we obtain,

$$\begin{aligned} \left\| f_\lambda^{\mathrm{ls}} \right\|_{\mathcal{H}} &\leq \left\| (T+\lambda)^{-1} L_K f_{\mathcal{H}} \right\|_{\mathcal{H}} \\ &\leq \left\| T^{r-\frac{1}{2}} T(T+\lambda)^{-1} L_K^{\frac{1}{2}-r} f_{\mathcal{H}} \right\|_{\mathcal{H}} \\ &\leq \|T\|^{r-\frac{1}{2}} \left\| L_K^{-r} f_{\mathcal{H}} \right\|_\rho \leq \kappa^{r-\frac{1}{2}} \left\| L_K^{-r} f_{\mathcal{H}} \right\|_\rho . \end{aligned}$$

□

We also need the following probabilistic inequality based on a result of [16], see also Th. 3.3.4 of [19]. We report it without proof.

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

$$\begin{aligned} \|\xi(\omega)\|_{\mathcal{K}} &\leq \frac{H}{2} \qquad \text{a.}s, \\ \mathbb{E}[\|\xi\|_{\mathcal{K}}^2] &\leq \sigma^2, \end{aligned}$$

then, for all $m \in \mathbb{N}$ and $0 < \delta < 1$,

$$\mathbb{P}_{(\omega_1,\ldots,\omega_m)\sim P^m} \left[\left\| \frac{1}{m} \sum_{i=1}^m \xi(\omega_i) - \mathbb{E}[\xi] \right\|_{\mathcal{K}} \leq 2\left(\frac{H}{m} + \frac{\sigma}{\sqrt{m}}\right) \log\frac{2}{\delta} \right] \geq 1-\delta. \tag{11}$$

We are finally ready to prove Theorem 1.

Proof of Theorem 1. The Theorem is a corollary of Prop. 1. We proceed by steps.

First. Observe that, by Prop. 2, it holds

$$\dot{\lambda} \leq \frac{\|T\|}{\kappa} \leq 1.$$

Second. Condition (9) holds. In fact, since $\dot{\lambda} \leq 1$ and by the assumption $\tilde{m} \geq m\dot{\lambda}^{-|1-2r|_+}$, we get,

$$\dot{\lambda}\tilde{m} \geq \dot{\lambda}^{-|1-2r|_+ + 1} m \geq \dot{\lambda}^{2r} m.$$

Moreover, by eq. (6) and definition (5), we find

$$\dot{\lambda}^{2r} m = 16 D_s^2 \dot{\lambda}^{-s} \log^2 \frac{6}{\delta} \geq 16 \mathcal{N}(\lambda) \log^2 \frac{6}{\delta}.$$

Third. Since $\dot{\lambda} \leq 1$, recalling definition (4) and Prop. 3, for every r in $(0, 1]$, we can write,

$$\left\| f_\lambda^{\mathrm{ls}} - f_{\mathcal{H}} \right\|_\rho \leq \dot{\lambda}^r C_r,$$

$$\kappa \left\| f_\lambda^{\mathrm{ls}} \right\|_{\mathcal{H}}^2 \leq \dot{\lambda}^{-|1-2r|_+} C_r^2.$$

Therefore we can apply Prop. 1, and using the two estimates above, the assumption $\tilde{m} \geq m\dot{\lambda}^{-|1-2r|_+}$ and the definition of D_s, to obtain the following bound,

$$\begin{array}{rcl} \left\| f^{\mathrm{ls}}_{\tilde{\mathbf{z}},\lambda} - f_{\mathcal{H}} \right\|_{\rho} & \leq & \left\| f^{\mathrm{ls}}_{\tilde{\mathbf{z}},\lambda} - f^{\mathrm{ls}}_{\lambda} \right\|_{\rho} + \| f_{\lambda} - f_{\mathcal{H}} \|_{\rho} \\ \left(\|f\|_{\rho} = \|\sqrt{T} f\|_{\mathcal{H}} \right) & \leq & \left\| (T+\lambda)^{\frac{1}{2}} (f^{\mathrm{ls}}_{\tilde{\mathbf{z}},\lambda} - f^{\mathrm{ls}}_{\lambda}) \right\|_{\mathcal{H}} + \| f_{\lambda} - f_{\mathcal{H}} \|_{\rho} \\ & \leq & 8(M + C_r) \dfrac{1}{\sqrt{m}} \left(\dfrac{2}{\sqrt{m\dot{\lambda}}} + \dfrac{D_s}{\sqrt{\dot{\lambda}^s}} \right) \log \dfrac{6}{\delta} + \dot{\lambda}^r C_r \\ (eq.\ (6)) & = & 2(M + C_r) \dot{\lambda}^r \left(1 + \dfrac{\dot{\lambda}^{r+s-\frac{1}{2}}}{2 D_s^2 \log \frac{6}{\delta}} \right) + \dot{\lambda}^r C_r \\ \left(r + s \geq \dfrac{1}{2} \right) & \leq & (3(M + C_r) + C_r)\, \dot{\lambda}^r \leq 4(M + C_r) \dot{\lambda}^r . \end{array}$$

Substituting the expression (6) for $\dot{\lambda}$ in the inequality above, concludes the proof. □

5. Proof of Theorem 2

In this section we give the proof of Theorem 2. It is based on Proposition 1 which establishes an upper bound on the sample error for the RLS algorithm in terms of the constants C_r and D_s. When need some preliminary results. Proposition 5 shows properties of the truncated functions $f^{\mathrm{tr}}_{\lambda}$, defined by equation (12), analogous to those given in Proposition 3 for the functions $f^{\mathrm{ls}}_{\lambda}$.

Proposition 5. *Let $f_{\mathcal{H}} \in \mathrm{Im}\, L_K^r$ for some $r > 0$. For any $\lambda > 0$ let the truncated function $f^{\mathrm{tr}}_{\lambda}$ be defined by*

$$f^{\mathrm{tr}}_{\lambda} \quad = \quad P_{\lambda} f_{\mathcal{H}} \tag{12}$$

where P_{λ} is the orthogonal projector in $\mathcal{L}^2(X, \rho_X)$ defined by

$$P_{\lambda} = \Theta_{\lambda}(L_K), \tag{13}$$

with

$$\Theta_{\lambda}(\sigma) \quad = \quad \begin{cases} 1 & \text{if } \sigma \geq \lambda, \\ 0 & \text{if } \sigma < \lambda. \end{cases} \tag{14}$$

Then, the following estimates hold,

$$\begin{array}{rcl} \left\| f^{\mathrm{tr}}_{\lambda} - f_{\mathcal{H}} \right\|_{\rho} & \leq & \lambda^r \left\| L_K^{-r} f_{\mathcal{H}} \right\|_{\rho}, \\ \| f^{\mathrm{tr}}_{\lambda} \|_{\mathcal{H}} & \leq & \begin{cases} \lambda^{-\frac{1}{2}+r} \left\| L_K^{-r} f_{\mathcal{H}} \right\|_{\rho} & \text{if } r \leq \frac{1}{2}, \\ \kappa^{-\frac{1}{2}+r} \left\| L_K^{-r} f_{\mathcal{H}} \right\|_{\rho} & \text{if } r > \frac{1}{2}. \end{cases} \end{array}$$

Proof. The first estimate follows simply observing that

$$\| f^{\mathrm{tr}}_{\lambda} - f_{\mathcal{H}} \|_{\rho} = \left\| P^{\perp}_{\lambda} f_{\mathcal{H}} \right\|_{\rho} = \left\| P^{\perp}_{\lambda} L_K^r \right\| \left\| L_K^{-r} f_{\mathcal{H}} \right\|_{\rho} \leq \lambda^r \left\| L_K^{-r} f_{\mathcal{H}} \right\|_{\rho},$$

where we introduced the orthogonal projector $P^{\perp}_{\lambda} = \mathrm{Id} - P_{\lambda}$.

Now let us consider the second estimate. Firstly observe that, since the compact operators L_K and T have a common eigensystem of functions on X, then P_{λ} can also be seen as an orthogonal projector in $\mathcal{H}$, and $f^{\mathrm{tr}}_{\lambda} \in \mathcal{H}$. Hence we can write,

$$\begin{array}{rcl} \| f^{\mathrm{tr}}_{\lambda} \|_{\mathcal{H}} & = & \| P_{\lambda} f_{\mathcal{H}} \|_{\mathcal{H}} \leq \left\| L_K^{-\frac{1}{2}} P_{\lambda} f_{\mathcal{H}} \right\|_{\rho} \\ & \leq & \left\| L_K^{-\frac{1}{2}+r} \Theta_{\lambda}(L_K) \right\| \left\| L_K^{-r} f_{\mathcal{H}} \right\|_{\rho}. \end{array}$$

The proof is concluded observing that by Prop. 2, $\|L_K\| = \|T\| \leq \kappa$, and that, for every $\sigma \in [0, \kappa]$, it holds

$$\sigma^{-\frac{1}{2}+r}\Theta_\lambda(\sigma) \quad \leq \quad \begin{cases} \lambda^{-\frac{1}{2}+r} & \text{if } r \leq \frac{1}{2}, \\ \kappa^{-\frac{1}{2}+r} & \text{if } r > \frac{1}{2}. \end{cases}$$

□

Proposition 6 below estimates one of the terms appearing in the proof of Theorem 2 for any $r > 0$. The case $r \geq \frac{1}{2}$ had already been analyzed in the proof of Theorem 7 in [1].

Proposition 6. *Let $r > 0$ and define*

$$\gamma = \lambda^{-1} \|T - T_{\tilde{\mathbf{x}}}\| . \tag{15}$$

Then, if $\lambda \in (0, \kappa]$, it holds

$$\left\|\sqrt{T}\,(G_\lambda(T_{\tilde{\mathbf{x}}})\,T_{\tilde{\mathbf{x}}} - \mathrm{Id})\,f_\lambda^{\mathrm{tr}}\right\|_{\mathcal{H}} \leq B_r C_r (1 + \sqrt{\gamma})(2 + r\gamma\dot{\lambda}^{\frac{3}{2}-r} + \gamma^\eta)\dot{\lambda}^r,$$

where

$$\eta = |r - \frac{1}{2}| - \lfloor |r - \frac{1}{2}| \rfloor.$$

Proof. The two inequalities (16) and (17) will be useful in the proof. The first follows from Theorem 1 in [15],

$$\|T^\alpha - T_{\tilde{\mathbf{x}}}^\alpha\| \leq \|T - T_{\tilde{\mathbf{x}}}\|^\alpha, \quad \alpha \in [0, 1] \tag{16}$$

where we adopted the convection $0^0 = 1$. The second is a corollary of Theorem 8.1 in [2]

$$\|T^p - T_{\tilde{\mathbf{x}}}^p\| \leq p\kappa^{p-1} \|T - T_{\tilde{\mathbf{x}}}\|, \quad p \in \mathbb{N}. \tag{17}$$

We also need to introduce the orthogonal projector in $\mathcal{H}$, $P_{\tilde{\mathbf{x}},\lambda}$, defined by

$$P_{\tilde{\mathbf{x}},\lambda} = \Theta_\lambda(T_{\tilde{\mathbf{x}}}),$$

with Θ_λ defined in (14).

We analyze the cases $r \leq \frac{1}{2}$ and $r \geq \frac{1}{2}$ separately.

Case $r \leq \frac{1}{2}$: In the three steps below we subsequently estimate the norms of the three terms of the expansion

$$\begin{aligned} \sqrt{T}\,(G_\lambda(T_{\tilde{\mathbf{x}}})\,T_{\tilde{\mathbf{x}}} - \mathrm{Id})\,f_\lambda^{\mathrm{tr}} \quad = \quad & \sqrt{T}P_{\tilde{\mathbf{x}},\lambda}^\perp r_\lambda(T_{\tilde{\mathbf{x}}})\,f_\lambda^{\mathrm{tr}} \\ & + P_{\tilde{\mathbf{x}},\lambda} r_\lambda(T_{\tilde{\mathbf{x}}})\,T_{\tilde{\mathbf{x}}}^{\frac{1}{2}}\,f_\lambda^{\mathrm{tr}} \\ & + (\sqrt{T} - \sqrt{T_{\tilde{\mathbf{x}}}})P_{\tilde{\mathbf{x}},\lambda} r_\lambda(T_{\tilde{\mathbf{x}}})\,f_\lambda^{\mathrm{tr}}, \end{aligned} \tag{18}$$

where $P_{\tilde{\mathbf{x}},\lambda}^\perp = \mathrm{Id} - P_{\tilde{\mathbf{x}},\lambda}$ and $r_\lambda(\sigma) = \sigma G_\lambda(\sigma) - 1$.

Step 1: Observe that

$$\begin{aligned} \left\|\sqrt{T}P_{\tilde{\mathbf{x}},\lambda}^\perp\right\|^2 \quad & = \quad \left\|P_{\tilde{\mathbf{x}},\lambda}^\perp T P_{\tilde{\mathbf{x}},\lambda}^\perp\right\| \leq \sup_{\phi \in \mathrm{Im}\ P_{\tilde{\mathbf{x}},\lambda}^\perp} \frac{(\phi, T\phi)_{\mathcal{H}}}{\|\phi\|_{\mathcal{H}}^2} \\ & \leq \quad \sup_{\phi \in \mathrm{Im}\ P_{\tilde{\mathbf{x}},\lambda}^\perp} \frac{(\phi, T_{\tilde{\mathbf{x}}}\phi)_{\mathcal{H}}}{\|\phi\|_{\mathcal{H}}^2} + \sup_{\phi \in \mathcal{H}} \frac{(\phi, (T - T_{\tilde{\mathbf{x}}})\phi)_{\mathcal{H}}}{\|\phi\|_{\mathcal{H}}^2} \\ & \leq \quad \lambda + \|T_{\tilde{\mathbf{x}}} - T\| = \lambda(1 + \gamma). \end{aligned}$$

Therefore, from definitions (2) and (4) and Proposition 5, it follows

$$\begin{aligned} \left\|\sqrt{T}P_{\tilde{\mathbf{x}},\lambda}^\perp r_\lambda(T_{\tilde{\mathbf{x}}})\,f_\lambda^{\mathrm{tr}}\right\|_{\mathcal{H}} \quad & \leq \quad \left\|\sqrt{T}P_{\tilde{\mathbf{x}},\lambda}^\perp\right\| \|r_\lambda(T_{\tilde{\mathbf{x}}})\| \left\|f_\lambda^{\mathrm{tr}}\right\|_{\mathcal{H}} \\ & \leq \quad B_r C_r \sqrt{1 + \gamma}\dot{\lambda}^r. \end{aligned}$$

Step 2: Observe that, from inequality (16), definition (4) and Proposition 5

$$\begin{aligned} \left\| T_{\tilde{\mathbf{x}}}^{\frac{1}{2}-r} f_\lambda^{\mathrm{tr}} \right\|_{\mathcal{H}} &\leq \|\Theta_\lambda(T)\| \left\| T^{\frac{1}{2}-r} f_{\mathcal{H}} \right\|_{\mathcal{H}} + \left\| T^{\frac{1}{2}-r} - T_{\tilde{\mathbf{x}}}^{\frac{1}{2}-r} \right\| \left\| f_\lambda^{\mathrm{tr}} \right\|_{\mathcal{H}} \\ &\leq \kappa^{-r} C_r (1 + \gamma^{\frac{1}{2}-r}). \end{aligned} \tag{19}$$

Therefore from definition (3), it follows

$$\begin{aligned} \left\| P_{\tilde{\mathbf{x}},\lambda} r_\lambda(T_{\tilde{\mathbf{x}}})\, T_{\tilde{\mathbf{x}}}^{\frac{1}{2}} f_\lambda^{\mathrm{tr}} \right\|_{\mathcal{H}} &\leq \|P_{\tilde{\mathbf{x}},\lambda}\| \, \|r_\lambda(T_{\tilde{\mathbf{x}}})\, T_{\tilde{\mathbf{x}}}^{r}\| \left\| T_{\tilde{\mathbf{x}}}^{\frac{1}{2}-r} f_\lambda^{\mathrm{tr}} \right\|_{\mathcal{H}} \\ &\leq B_r C_r (1 + \gamma^{\frac{1}{2}-r}) \dot{\lambda}^r. \end{aligned}$$

Step 3: Recalling the definition of $P_{\tilde{\mathbf{x}},\lambda}$, and applying again inequality (19) and Proposition 5, we get

$$\begin{aligned} &\left\| (\sqrt{T} - \sqrt{T_{\tilde{\mathbf{x}}}}) P_{\tilde{\mathbf{x}},\lambda} r_\lambda(T_{\tilde{\mathbf{x}}})\, f_\lambda^{\mathrm{tr}} \right\|_{\mathcal{H}} \\ &\leq \left\| (\sqrt{T} - \sqrt{T_{\tilde{\mathbf{x}}}}) T_{\tilde{\mathbf{x}}}^{-\frac{1}{2}+r} P_{\tilde{\mathbf{x}},\lambda} r_\lambda(T_{\tilde{\mathbf{x}}})\, T_{\tilde{\mathbf{x}}}^{\frac{1}{2}-r} f_\lambda^{\mathrm{tr}} \right\|_{\mathcal{H}} \\ &\leq \left\| \sqrt{T} - \sqrt{T_{\tilde{\mathbf{x}}}} \right\| \left\| T_{\tilde{\mathbf{x}}}^{-\frac{1}{2}+r} P_{\tilde{\mathbf{x}},\lambda} \right\| \|r_\lambda(T_{\tilde{\mathbf{x}}})\| \left\| T_{\tilde{\mathbf{x}}}^{\frac{1}{2}-r} f_\lambda^{\mathrm{tr}} \right\|_{\mathcal{H}} \\ &\leq B_r C_r \gamma^{\frac{1}{2}} (1 + \gamma^{\frac{1}{2}-r}) \dot{\lambda}^r. \end{aligned}$$

Since we assumed $0 < r \leq \frac{1}{2}$, and therefore $\eta = \frac{1}{2} - r$, the three estimates above prove the statement of the Theorem in this case.

Case $r \geq \frac{1}{2}$**:** Consider the expansion

$$\begin{aligned} (G_\lambda(T_{\tilde{\mathbf{x}}})\, T_{\tilde{\mathbf{x}}} - \mathrm{Id})\, f_\lambda^{\mathrm{tr}} &\leq r_\lambda(T_{\tilde{\mathbf{x}}})\, T^{r-\frac{1}{2}} v \\ &\leq r_\lambda(T_{\tilde{\mathbf{x}}})\, T_{\tilde{\mathbf{x}}}^{r-\frac{1}{2}} v + r_\lambda(T_{\tilde{\mathbf{x}}}) \left(T^{r-\frac{1}{2}} - T_{\tilde{\mathbf{x}}}^{r-\frac{1}{2}} \right) v \\ &\leq r_\lambda(T_{\tilde{\mathbf{x}}})\, T_{\tilde{\mathbf{x}}}^{r-\frac{1}{2}} v + r_\lambda(T_{\tilde{\mathbf{x}}})\, T_{\tilde{\mathbf{x}}}^{p} \left(T^{r-\frac{1}{2}-p} - T_{\tilde{\mathbf{x}}}^{r-\frac{1}{2}-p} \right) v \\ &\quad + r_\lambda(T_{\tilde{\mathbf{x}}})\, (T^p - T_{\tilde{\mathbf{x}}}^p)\, T^{r-\frac{1}{2}-p} v \end{aligned}$$

where $v = P_\lambda T^{\frac{1}{2}-r} f_{\mathcal{H}}$, $r_\lambda(\sigma) = \sigma G_\lambda(\sigma) - 1$ and $p = \lfloor r - \frac{1}{2} \rfloor$.

Now, for any $\beta \in [0, \frac{1}{2}]$, from the expansion above using inequalities (16) and (17), and definition (3), we get

$$\begin{aligned} &\left\| T_{\tilde{\mathbf{x}}}^{\beta} (G_\lambda(T_{\tilde{\mathbf{x}}})\, T_{\tilde{\mathbf{x}}} - \mathrm{Id})\, f_\lambda^{\mathrm{tr}} \right\|_{\mathcal{H}} \leq \left\| r_\lambda(T_{\tilde{\mathbf{x}}})\, T_{\tilde{\mathbf{x}}}^{r-\frac{1}{2}+\beta} \right\| \|v\|_{\mathcal{H}} \\ &+ \left\| r_\lambda(T_{\tilde{\mathbf{x}}})\, T_{\tilde{\mathbf{x}}}^{p+\beta} \right\| \left\| T^{r-\frac{1}{2}-p} - T_{\tilde{\mathbf{x}}}^{r-\frac{1}{2}-p} \right\| \|v\|_{\mathcal{H}} \\ &+ \left\| r_\lambda(T_{\tilde{\mathbf{x}}})\, T_{\tilde{\mathbf{x}}}^{\beta} \right\| \|T^p - T_{\tilde{\mathbf{x}}}^p\| \left\| T^{r-\frac{1}{2}-p} \right\| \|v\|_{\mathcal{H}} \\ \leq\;& B_r C_r \kappa^{-\frac{1}{2}+\beta} \left(\dot{\lambda}^{r-\frac{1}{2}+\beta} (1 + \gamma^{r-\frac{1}{2}-p}) + p \dot{\lambda}^{1+\beta} \gamma \right) \\ \leq\;& B_r C_r \kappa^{-\frac{1}{2}+\beta} \left(\dot{\lambda}^{-\frac{1}{2}+\beta} (1 + \gamma^{\eta}) + r\gamma \dot{\lambda}^{1+\beta-r} \right) \dot{\lambda}^r. \end{aligned} \tag{20}$$

Finally, from the expansion

$$\begin{aligned} \left\| \sqrt{T}\, (G_\lambda(T_{\tilde{\mathbf{x}}})\, T_{\tilde{\mathbf{x}}} - \mathrm{Id})\, f_\lambda^{\mathrm{tr}} \right\|_{\mathcal{H}} &\leq \left\| \sqrt{T} - \sqrt{T_{\tilde{\mathbf{x}}}} \right\| \left\| r_\lambda(T_{\tilde{\mathbf{x}}})\, f_\lambda^{\mathrm{tr}} \right\|_{\mathcal{H}} \\ &\quad + \left\| \sqrt{T_{\tilde{\mathbf{x}}}} r_\lambda(T_{\tilde{\mathbf{x}}})\, f_\lambda^{\mathrm{tr}} \right\|_{\mathcal{H}}, \end{aligned}$$

using (16) and inequality (20) with $\beta = 0$ and $\beta = \frac{1}{2}$, we get the claimed result also in this case. □

We need an additional preliminary result.

Proposition 7. *Let the operator* Ω_λ *be defined by*

$$\Omega_\lambda = \sqrt{T} G_\lambda(T_{\tilde{\mathbf{x}}})\,(T_{\tilde{\mathbf{x}}} + \lambda)(T + \lambda)^{-\frac{1}{2}}. \tag{21}$$

Then, if $\lambda \in (0, \kappa]$, *it holds*

$$\|\Omega_\lambda\| \leq (1 + 2\sqrt{\gamma})\,A,$$

with γ *defined in eq. (15).*

Proof. First consider the expansion

$$\Omega_\lambda = \left(\sqrt{T} - \sqrt{T_{\tilde{\mathbf{x}}}}\right) \Delta_\lambda\ (T + \lambda)^{-\frac{1}{2}} - \Delta_\lambda\ \left(\sqrt{T} - \sqrt{T_{\tilde{\mathbf{x}}}}\right) (T + \lambda)^{-\frac{1}{2}} + \Delta_\lambda\ \sqrt{T}(T + \lambda)^{-\frac{1}{2}},$$

where we introduced the operator

$$\Delta_\lambda = G_\lambda(T_{\tilde{\mathbf{x}}})\ \ (T_{\tilde{\mathbf{x}}} + \lambda).$$

By condition (2), it follows $\|\Delta_\lambda\| \leq A$. Moreover, from inequality (16)

$$\left\|\sqrt{T} - \sqrt{T_{\tilde{\mathbf{x}}}}\right\| \leq \sqrt{\|T - T_{\tilde{\mathbf{x}}}\|}. \tag{22}$$

From the previous observations we easily get

$$\begin{aligned} \|\Omega_\lambda\| &\leq 2\,\|\Delta_\lambda\| \left\|\sqrt{T} - \sqrt{T_{\tilde{\mathbf{x}}}}\right\| \left\|(T + \lambda)^{-\frac{1}{2}}\right\| + \|\Delta_\lambda\| \left\|\sqrt{T}(T + \lambda)^{-\frac{1}{2}}\right\| \\ &\leq A\,(1 + 2\sqrt{\gamma})\,, \end{aligned}$$

the claimed result.

□

We are now ready to show the proof of Theorem 2.

Proof of Theorem 2. We consider the expansion

$$\begin{aligned} \sqrt{T}(f_{\tilde{\mathbf{z}},\lambda} - f_{\mathcal{H}}) &= \sqrt{T}\left(G_\lambda(T_{\tilde{\mathbf{x}}})\, g_{\mathbf{z}} - f_\lambda^{\mathrm{tr}}\right) + \sqrt{T}(f_\lambda^{\mathrm{tr}} - f_{\mathcal{H}}) \\ &= \Omega_\lambda\ (T + \lambda)^{\frac{1}{2}}\,(f_{\tilde{\mathbf{z}},\lambda}^{\mathrm{ls}} - f_{\tilde{\mathbf{z}}',\lambda}^{\mathrm{ls}}) + \sqrt{T}\,(G_\lambda(T_{\tilde{\mathbf{x}}})\,T_{\tilde{\mathbf{x}}} - \mathrm{Id})\, f_\lambda^{\mathrm{tr}} + \sqrt{T}(f_\lambda^{\mathrm{tr}} - f_{\mathcal{H}}) \\ &= \Omega_\lambda\ \left((T + \lambda)^{\frac{1}{2}}\,(f_{\tilde{\mathbf{z}},\lambda}^{\mathrm{ls}} - f_\lambda^{\mathrm{ls}}) + (T + \lambda)^{\frac{1}{2}}\,(f_\lambda^{\mathrm{ls}} - \bar{f}_\lambda^{\mathrm{ls}}) + (T + \lambda)^{\frac{1}{2}}\,(\bar{f}_\lambda^{\mathrm{ls}} - f_{\tilde{\mathbf{z}}',\lambda}^{\mathrm{ls}})\right) \\ &\quad + \sqrt{T}\,(G_\lambda(T_{\tilde{\mathbf{x}}})\,T_{\tilde{\mathbf{x}}} - \mathrm{Id})\, f_\lambda^{\mathrm{tr}} + \sqrt{T}(f_\lambda^{\mathrm{tr}} - f_{\mathcal{H}}) \end{aligned}$$

where the operator Ω_λ is defined by equation (21), the ideal RLS estimators are $f_\lambda^{\mathrm{ls}} = (T+\lambda)^{-1} T f_{\mathcal{H}}$ and $\bar{f}_\lambda^{\mathrm{ls}} = (T+\lambda)^{-1} T f_\lambda^{\mathrm{tr}}$, and $f_{\tilde{\mathbf{z}}',\lambda}^{\mathrm{ls}} = (T_{\tilde{\mathbf{x}}}+\lambda)^{-1} T_{\tilde{\mathbf{x}}} f_\lambda^{\mathrm{tr}}$ is the RLS estimator constructed by the training set

$$\tilde{z}' = ((\tilde{x}_1, f_\lambda^{\mathrm{tr}}(\tilde{x}_1)) \ldots, (\tilde{x}_{\tilde{m}}, f_\lambda^{\mathrm{tr}}(\tilde{x}_{\tilde{m}}))).$$

Hence we get the following decomposition,

$$\|f_{\tilde{\mathbf{z}},\lambda} - f_{\mathcal{H}}\|_\rho \ \leq\ D\left(S^{\mathrm{ls}} + R + \bar{S}^{\mathrm{ls}}\right) + P + P^{\mathrm{tr}}, \tag{23}$$

with

$$\begin{aligned}
S^{\mathrm{ls}} &= \left\|(T+\lambda)^{\frac{1}{2}}(f^{\mathrm{ls}}_{\tilde{\mathbf{z}},\lambda} - f^{\mathrm{ls}}_{\lambda})\right\|_{\mathcal{H}}, \\
\bar{S}^{\mathrm{ls}} &= \left\|(T+\lambda)^{\frac{1}{2}}(f^{\mathrm{ls}}_{\tilde{\mathbf{z}}',\lambda} - \bar{f}^{\mathrm{ls}}_{\lambda})\right\|_{\mathcal{H}}, \\
D &= \|\Omega_\lambda\|, \\
P &= \left\|\sqrt{T}\,(G_\lambda(T_{\tilde{\mathbf{x}}})\,T_{\tilde{\mathbf{x}}} - \mathrm{Id})\,f^{\mathrm{tr}}_{\lambda}\right\|_{\mathcal{H}}, \\
P^{\mathrm{tr}} &= \left\|f^{\mathrm{tr}}_{\lambda} - f_{\mathcal{H}}\right\|_{\rho}, \\
R &= \left\|(T+\lambda)^{\frac{1}{2}}(\bar{f}^{\mathrm{ls}}_{\lambda} - f^{\mathrm{ls}}_{\lambda})\right\|_{\mathcal{H}}.
\end{aligned}$$

Terms S^{ls} and $\bar{S}^{\mathrm{ls}}$ will be estimated by Proposition 1, term D by Proposition 7, term P by Proposition 6 and finally terms P^{tr} and R by Proposition 5.

Let us begin with the estimates of S^{ls} and $\bar{S}^{\mathrm{ls}}$. First observe that, by the same reasoning in the proof of Theorem 1, the assumptions of the Theorem imply inequality (9) in the text of Proposition 1.

Regarding the estimate of S^{ls}. Applying Proposition 1 and reasoning as in the proof of Theorem 1 (recall that by assumption $\tilde{m} \geq m\dot{\lambda}^{-|2-2r-s|_+} \geq m\dot{\lambda}^{-|1-2r|_+}$ and from Proposition 3, $\sqrt{\kappa}\left\|f^{\mathrm{ls}}_{\lambda}\right\|_{\mathcal{H}} \leq C_r\dot{\lambda}^{-\left|\frac{1}{2}-r\right|_+}$), we get that with probability greater than $1-\delta$

$$\begin{aligned}
(24) \qquad S^{\mathrm{ls}} &\leq 8\left(M + \sqrt{\frac{m}{\tilde{m}}}C_r\dot{\lambda}^{-\left|\frac{1}{2}-r\right|_+}\right)\left(\frac{2}{m}\sqrt{\frac{\kappa}{\lambda}} + \sqrt{\frac{\mathcal{N}(\lambda)}{m}}\right)\log\frac{6}{\delta} \\
&\leq 8(M+C_r)\frac{1}{\sqrt{m}}\left(\frac{2}{\sqrt{m\dot{\lambda}}} + \frac{D_s}{\sqrt{\dot{\lambda}^s}}\right)\log\frac{6}{\delta} \\
(eq.\ (7)) \qquad &= 2(M+C_r)\dot{\lambda}^r\left(1 + \frac{\dot{\lambda}^{r+s-\frac{1}{2}}}{2D_s^2\log\frac{6}{\delta}}\right) \\
\left(r+s \geq \frac{1}{2}\right) \qquad &\leq 3(M+C_r)\dot{\lambda}^r
\end{aligned}$$

The term $\bar{S}^{\mathrm{ls}}$ can be estimated observing that $\tilde{\mathbf{z}}'$ is a training set of $\tilde{m}$ supervised samples drawn i.i.d. from the probability measure ρ' with marginal ρ_X and conditional $\rho'_{|x}(y) = \delta(y - f^{\mathrm{tr}}_{\lambda}(x))$. Therefore the regression function induced by ρ' is $f_{\rho'} = f^{\mathrm{tr}}_{\lambda}$, and the support of ρ' is included in $[-M', M'] \times X$, with $M' = \sup_{x\in X} f_{\rho'}(x) \leq \sqrt{\kappa}\left\|f^{\mathrm{tr}}_{\lambda}\right\|_{\mathcal{H}}$. Again applying Proposition 1 and reasoning as in the proof of Theorem 1, we obtain that with probability greater than $1-\delta$ it holds

$$\begin{aligned}
&(25) & \bar{S}^{\mathrm{ls}} &\leq 8\left(M' + \sqrt{\kappa}\left\|\bar{f}_\lambda^{\mathrm{ls}}\right\|_{\mathcal{H}}\right)\left(\frac{2}{\tilde{m}}\sqrt{\frac{\kappa}{\lambda}} + \sqrt{\frac{\mathcal{N}(\lambda)}{\tilde{m}}}\right)\log\frac{6}{\delta} \\
& & &\leq 16\sqrt{\kappa}\left\|f_\lambda^{\mathrm{tr}}\right\|_{\mathcal{H}}\left(\frac{2}{\tilde{m}}\sqrt{\frac{\kappa}{\lambda}} + \sqrt{\frac{\mathcal{N}(\lambda)}{\tilde{m}}}\right)\log\frac{6}{\delta} \\
&(Prop.5) & &\leq 16\sqrt{\frac{m}{\tilde{m}}}C_r\dot{\lambda}^{-\left|\frac{1}{2}-r\right|_+}\left(\frac{2}{m}\sqrt{\frac{\kappa}{\lambda}} + \sqrt{\frac{\mathcal{N}(\lambda)}{m}}\right)\log\frac{6}{\delta} \\
& & &\leq 16C_r\frac{1}{\sqrt{m}}\left(\frac{2}{\sqrt{m\dot{\lambda}}} + \frac{D_s}{\sqrt{\dot{\lambda}^s}}\right)\log\frac{6}{\delta} \\
&(eq.\ (7)) & &= 4C_r\dot{\lambda}^r\left(1 + \frac{\dot{\lambda}^{r+s-\frac{1}{2}}}{2D_s^2\log\frac{6}{\delta}}\right) \\
&\left(r+s\geq\frac{1}{2}\right) & &\leq 6C_r\dot{\lambda}^r
\end{aligned}$$

In order to get an upper bound for D and P, we have first to estimate the quantity γ (see definition (15)) appearing in the Propositions 6 and 7. Our estimate for γ follows from Proposition 4 applied to the random variable $\xi : X \to \mathcal{L}_{\mathrm{HS}}(\mathcal{H})$ defined by

$$\xi(x)[\cdot] \quad = \quad \lambda^{-1}K_x\left\langle K_x,\cdot\right\rangle_{\mathcal{H}}.$$

We can set $H = \frac{2\kappa}{\lambda}$ and $\sigma = \frac{H}{2}$, and obtain that with probability greater than $1-\delta$

$$\begin{aligned}
&\gamma \leq \lambda^{-1}\left\|T - T_{\tilde{\mathbf{x}}}\right\|_{\mathrm{HS}} \leq \frac{2}{\lambda}\left(\frac{2\kappa}{\tilde{m}} + \frac{\kappa}{\sqrt{\tilde{m}}}\right)\log\frac{2}{\delta} \leq 4\frac{1}{\dot{\lambda}\sqrt{\tilde{m}}}\log\frac{2}{\delta} \\
&\leq 4\frac{\dot{\lambda}^{\left|1-r-\frac{s}{2}\right|_+-1}}{\sqrt{m}}\log\frac{2}{\delta} \leq \dot{\lambda}^{\left|1-r-\frac{s}{2}\right|_+-(1-r-\frac{s}{2})} \leq \dot{\lambda}^{\left|r+\frac{s}{2}-1\right|_+} \leq 1,
\end{aligned}$$

where we used the assumption $\tilde{m} \geq 4 \vee m\dot{\lambda}^{-|2-2r-s|_+}$ and the expression for $\dot{\lambda}$ in the text of the Theorem.

Hence, since $\gamma \leq 1$, from Proposition 7 we get

$$(26) \qquad D \quad \leq \quad 3A,$$

and from Proposition 6

$$\begin{aligned}
&(27) & P &\leq 2B_rC_r(3 + r\gamma\dot{\lambda}^{\frac{3}{2}-r})\dot{\lambda}^r \\
& & &\leq 2B_rC_r(3 + r\dot{\lambda}^{\left|r+\frac{s}{2}-1\right|_++\frac{3}{2}-r})\dot{\lambda}^r \\
& & &\leq 2B_rC_r(3 + r\dot{\lambda}^{\frac{s+1}{2}})\dot{\lambda}^r \leq 2B_rC_r(3+r)\dot{\lambda}^r.
\end{aligned}$$

Regarding terms P^{tr} and R. From Proposition 5 we get

$$(28) \qquad P^{\mathrm{tr}} \quad \leq \quad C_r\dot{\lambda}^r,$$

and hence,

$$\begin{aligned}
&(29) & R &= \left\|(T+\lambda)^{-\frac{1}{2}}T(\bar{f}_\lambda^{\mathrm{ls}} - f_\lambda^{\mathrm{ls}})\right\|_{\mathcal{H}} \\
& & &\leq \left\|\sqrt{T}(\bar{f}_\lambda^{\mathrm{ls}} - f_\lambda^{\mathrm{ls}})\right\|_{\mathcal{H}} \leq P^{\mathrm{tr}} \leq C_r\dot{\lambda}^r.
\end{aligned}$$

The proof is completed by plugging inequalities (24), (25), (26), (27), (28) and (29) in (23) and recalling the expression for $\dot{\lambda}$. □

Acknowledgements

The author wish to thank E. De Vito, T. Poggio, L. Rosasco, S. Smale, A. Verri and Y. Yao for useful discussions and suggestions.

References

[1] F. Bauer, S. Pereverzev, and L. Rosasco. On regularization algorithms in learning theory. preprint, 2005.

[2] M. S. Birman and M. Solomyak. Double operators integrals in hilbert scales. *Integr. Equ. Oper. Theory*, pages 131–168, 2003.

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

[4] A. Caponnetto and E. De Vito. Optimal rates for regularized least-squares algorithm. 2005. *to appear in* Foundations of Computational Mathematics.

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

[6] E. De Vito and A. Caponnetto. Risk bounds for regularized least-squares algorithm with operator-valued kernels. Technical report, Massachusetts Institute of Technology, Cambridge, MA, May 2005. CBCL Paper #249/AI Memo #2005-015.

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

[8] E. De Vito, L. Rosasco, and A. Caponnetto. Discretization error analysis for tikhonov regularization. to appear in *Analisys and Applications*, 2005.

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

[10] E. De Vito, L. Rosasco, and A. Verri. Spectral methods for regularization in learning theory. preprint, 2005.

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

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

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

[14] 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.

[15] P. Mathé and S. V. Pereverzev. Moduli of continuity for operator valued functions. *Numer. Funct. Anal. Optim.*, 23(5-6):623–631, 2002.

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

[17] S. Smale and D. Zhou. Learning theory estimates via integral operators and their approximations. *preprint*, 2005.

[18] S. Smale and D.X. Zhou. Shannon sampling II : Connections to learning theory. *Appl. Comput. Harmonic Anal.* to appear.

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