Computer Science and Artificial Intelligence Laboratory

Technical Report

MIT-CSAIL-TR-2006-063
CBCL-265

September 10, 2006

Adaptation for Regularization Operators in Learning Theory

Andrea Caponnetto and Yuan Yao

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

Adaptation for regularization operators in learning theory

Andrea Caponnetto [a b c] Yuan Yao [d]

[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*

[d] *Department of Mathematics, University of California, Berkeley, CA 94720*

Abstract

We consider learning algorithms induced by regularization methods in the regression setting. We show that previously obtained error bounds for these algorithms using a-priori choices of the regularization parameter, can be attained using a suitable a-posteriori choice based on validation. In particular, these results prove adaptation of the rate of convergence of the estimators to the minimax rate induced by the "effective dimension" of the problem. We also show universal consistency for this class methods.

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 supported by the NSF grant 0325113. The first author is also supported 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 show that previous results in [2] about rates of convergence for regularization methods using a-priori choices of the regularization parameter, can be attained using a suitable a-posteriori choice based on validation. We also show universal consistency for this class methods. The framework for semi-supervised statistical learning theory is the same one considered in [2]. The algorithms we consider are based on the formalism of *regularization methods* for linear ill-posed inverse problems in their classical setting (see for example [11] for general reference). Some popular algorithms from this class are: regularized least-squares, truncated SVD, Landweber method and ν-method.

The paper is organized as follows. In Section 2 we focus on a-priori choices of the regularization parameter for regularization methods. Theorem 1 shows universal consistency for a large class of choice rules, and Theorem 2 shows specific rates of convergence under suitable prior assumptions (parameterized by the constants r, s, C_r and D_s) on the unknown probability measure ρ. Unlabelled data are added to the training set in order to improve the rates for a certain range of the parameters r and s.

In Section 3 we consider a validation technique for the a-posteriori choice of the regularization parameter. Theorem 3 shows how error bounds for the estimators $f_{\tilde{\mathbf{z}},\lambda}$, with a-priori choices of λ, can be transferred to the estimators $f_{\mathbf{z}^{\mathrm{tot}}}$ which use the validation examples $\mathbf{z}^{\mathrm{v}}$ in $\mathbf{z}^{\mathrm{tot}} = (\tilde{\mathbf{z}}, \mathbf{z}^{\mathrm{v}})$ to determine λ. The subsequent corollaries are applications of Theorem 3 to the choices of λ described in Section 2.

In Sections 4 and 5 we give the proofs of the results stated in the previous Sections, using some lemmas from [2].

2. A-priori choice of the regularization parameter.

We consider the setting of semi-supervised statistical learning. We assume that $Y \subset [-M, M]$ and we let the supervised part of the training set be 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 we assume that the unsupervised part of the training set is $(x^u_{m+1}, \ldots, x^u_{\tilde{m}})$, with x^u_i drawn i.i.d. according to the marginal probability measure over X, ρ_X. For sake of brevity we 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

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

and

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

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 [11]). Regularization methods have been recently studied in the context of learning theory in [12, 8, 7, 9, 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 Theorem 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

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

Finiteness of C_r is a common *source condition* in the inverse problems literature (see [11] for reference). This type of condition has been introduced in the statistical learning literature in [6, 16, 3, 15, 4].

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

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

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

The next theorem shows (strong) universal consistency (in probability) for the estimators $f_{\tilde{\mathbf{z}},\lambda}$ under mild assumptions on the choice of λ. 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 $\{\tilde{\mathbf{z}}_m\}_{m=1}^\infty$ be a sequence of training sets composed of m labelled examples drawn i.i.d. from a probability measure ρ over Z, and $\tilde{m}_m - m \geq 0$ unlabelled examples drawn i.i.d. from the marginal measure of ρ over X. Let the regularization parameter choice, $\lambda_m : \mathbb{N} \to (0, \kappa]$, fulfill the conditions*

$$\lim_{m \to \infty} \lambda_m = 0, \tag{6}$$

$$\lim_{m \to \infty} \sqrt{m} \lambda_m = \infty. \tag{7}$$

Then, if $B_{\bar{r}} < +\infty$ *for some* $\bar{r} > 0$, *it holds* [1]

$$\lim_{m\to\infty} \|f_{\tilde{\mathbf{z}}_m,\lambda_m} - f_{\mathcal{H}}\|_\rho \stackrel{P}{=} 0.$$

Theorem 2 below is a restatement in a slightly modified form of Theorem 2 in [2]. In particular the introduction of the parameter $q > 1$ will be useful when we will merge this result with Theorem 3 in the proof of Corollary 2.

Theorem 2. *Let* $r > 0$, $s \in (0,1]$ *and* $\alpha \in [0, |2-2r-s|_+]$. *Furthermore, let* m *and* λ *satisfy the constraints* $\lambda \leq \|T\|$ *and*

$$\dot{\lambda} = q\left(\frac{4D_s \log\frac{6}{\delta}}{\sqrt{m}}\right)^{\frac{2}{2r+s+t_1}}, \tag{8}$$

for some $q \geq 1$, $\delta \in (0, 1/3)$ *and* t_1 *defined in eq. (10). Finally, assume* $\tilde{m} \geq 4 \vee m\dot{\lambda}^{-\alpha}$. *Then, with probability greater than* $1 - 3\delta$, *it holds*

$$\|f_{\tilde{\mathbf{z}},\lambda} - f_{\mathcal{H}}\|_\rho \leq q^r E_r \left(\frac{4D_s \log\frac{6}{\delta}}{\sqrt{m}}\right)^{\frac{2r-t_2}{2r+s+t_1}},$$

where

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

$$t_1 = |2-2r-s|_+ - \alpha, \tag{10}$$

$$t_2 = |1-2r-2s-t_1|_+ . \tag{11}$$

The proofs of the above Theorems is postponed to Section 4.

3. Adaptation.

In this section we show the adaptation properties of the estimators obtained by a suitable data-dependent choice of the regularization parameter. The main results of this section are obtained assuming that

$$f_{\mathcal{H}} = f_\rho, \tag{12}$$

this is true for every ρ when the underlying kernel K is *universal* (see [17]). In fact for this class of kernels the RKHS $\mathcal{H}$ is always dense in $\mathcal{L}^2(X, \rho_X)$. The Gaussian kernel is a popular instance of a kernel in this family.

Let the *validation set*

$$\mathbf{z}^{\mathrm{v}} = (z_1^{\mathrm{v}}, \ldots, z_{m^{\mathrm{v}}}^{\mathrm{v}}),$$

be composed of m^{v} labelled examples $z_i^{\mathrm{v}} = (x_i^{\mathrm{v}}, y_i^{\mathrm{v}})$ drawn i.i.d. from the probability measure ρ over $Z = X \times Y$. The validation set $\mathbf{z}^{\mathrm{v}}$ is, by assumption, independent of the training set $\tilde{\mathbf{z}}$, and these two sets define the *learning set*

$$\mathbf{z}^{\mathrm{tot}} = (\tilde{\mathbf{z}}, \mathbf{z}^{\mathrm{v}}),$$

which represents the total input of the adaptive learning algorithm. Following the notations of the previous Section, we let $\tilde{m}$ be the total number of examples in $\tilde{\mathbf{z}}$, and m the number of its labelled examples.

---

[1] We say that the sequence of random variables $\{X_m\}_{m\in\mathbb{N}}$ converges in probability to the random variable X (and we write $\lim_{m\to\infty} X_m =^P X$ or $X_m \to^P X$), if for every $\epsilon > 0$, $\lim_{m\to\infty} \mathbb{P}[|X_m - X| \geq \epsilon] = 0$. This is equivalent to say that, for every $\delta \in (0,1)$, $\mathbb{P}[|X_m - X| \geq \epsilon(m,\delta)] \leq \delta$, with $\lim_{m\to\infty} \epsilon(m,\delta) = 0$.

Now let us explain how $\mathbf{z}^{\mathrm{v}}$ is used for the choice of λ. We consider the finite set of positive reals Λ_m depending on m, the number of labelled examples in $\tilde{\mathbf{z}}$, and the data-dependent choice for the regularization parameter is

$$\hat{\lambda}_{\mathbf{z}^{\mathrm{v}}} = \operatorname*{argmin}_{\lambda \in \Lambda_m} \frac{1}{m^{\mathrm{v}}} \sum_{i=1}^{m^{\mathrm{v}}} (T_M f_{\tilde{\mathbf{z}},\lambda}(x_i^{\mathrm{v}}) - y_i^{\mathrm{v}})^2, \tag{13}$$

where the truncation operator $T_M : \mathcal{L}^2(X, \rho_X) \to \mathcal{L}^2(X, \rho_X)$ is defined by

$$T_M f(x) = (|f(x)| \wedge M) \ \operatorname{sign} f(x).$$

The final learning estimator, whose adaptation properties are investigated in this Section, is defined as follows

$$f_{\mathbf{z}^{\mathrm{tot}}} = T_M f_{\tilde{\mathbf{z}},\hat{\lambda}_{\mathbf{z}^{\mathrm{v}}}}. \tag{14}$$

Theorem 3 below is the main result of this Section and shows an important property of the estimator $f_{\mathbf{z}^{\mathrm{tot}}}$. It will be used to extend to $f_{\mathbf{z}^{\mathrm{tot}}}$ convergence results similar to the ones obtained in the previous Section.

Theorem 3. *Let ρ, K, m, $\tilde{m}$, m^{v}, Λ_m, $\delta \in (0,1)$, $\epsilon > 0$ and $\lambda_m \in \Lambda_m$ be such that with probability greater than $1 - \delta$, it holds*

$$\|f_{\tilde{\mathbf{z}},\lambda_m} - f_\rho\|_\rho \leq \epsilon.$$

Then, with probability greater than $1 - 2\delta$, it holds

$$\|f_{\mathbf{z}^{\mathrm{tot}}} - f_\rho\|_\rho \leq \hat{\epsilon},$$

with

$$\hat{\epsilon}^2 = 2\epsilon^2 + \frac{80M^2}{m^{\mathrm{v}}} \log \frac{2\,|\Lambda_m|}{\delta}.$$

The proof of Theorem 3 is postponed to Section 5.

The first corollary of Theorem 3 proves universal consistency for the estimators $f_{\mathbf{z}}^{\mathrm{tot}}$ under mild assumptions on the cardinalities of the grids Λ_m and validation sets $\mathbf{z}^{\mathrm{v}}$.

Corollary 1. *Let K be a universal kernel, Q be a constant greater than 1, and define*

$$\Lambda_m = \{\kappa, \kappa Q^{-1}, \ldots, \kappa Q^{-|\Lambda_m|+1}\}, \tag{15}$$

with

$$|\Lambda_m| = \omega(1). \tag{16}$$

Moreover let $\{\mathbf{z}_m^{\mathrm{tot}}\}_{m=1}^{\infty}$ be a sequence of learning sets drawn according to a probability measure ρ over Z. Assume $\mathbf{z}_m^{\mathrm{tot}} = (\tilde{\mathbf{z}}_m, \mathbf{z}_m^{\mathrm{v}})$, with the training sets $\tilde{\mathbf{z}}_m$ composed of m labelled examples and $\tilde{m}_m - m \geq 0$ unlabelled examples, and $\mathbf{z}_m^{\mathrm{v}}$ the validation sets composed by $m_m^{\mathrm{v}} = \omega(\log |\Lambda_m|)$ examples. Then, if $B_{\bar{r}} < +\infty$ for some $\bar{r} > 0$, it holds

$$\lim_{m \to \infty} \left\| f_{\mathbf{z}_m^{\mathrm{tot}}} - f_\rho \right\|_\rho \overset{P}{=} 0.$$

Proof. The result is a corollary of theorems 1 and 3. The universality of K enforces the equality (12) (see [17]). Condition (16) implies that the regularization parameter $\lambda_m = \kappa Q^{-(\lfloor \log \log m \rfloor \wedge |\Lambda_m|)}$, which belongs to Λ_m, fulfills the assumptions (6) and (7). Hence, using the assumption on m_m^{v}, we get that for every $\delta \in (0,1)$, with probability greater than $1 - 2\delta$

$$\left\| f_{\mathbf{z}_m^{\mathrm{tot}}} - f_\rho \right\|_\rho^2 \leq o(1) + \frac{80M^2}{\omega(\log |\Lambda_m|)} \log \frac{2\,|\Lambda_m|}{\delta} \to 0.$$

□

The second corollary proves explicit rates for the convergence of $f_{\mathbf{z}}^{\mathrm{tot}}$ to f_ρ over specific prior classes defined in term of finiteness of the constants C_r and D_s. The main assumption is the requirement $m^{\mathrm{v}} \geq m/\log m$. Since this constrain can be fulfilled still being m^{v} asymptotically negligible with respect to m, the rates (expressed in terms of m) that are obtained in the second part of the corollary are minimax optimal over the corresponding priors (see [4]).

Corollary 2. *Let K be a universal kernel. Consider a learning set $\mathbf{z}^{\mathrm{tot}}$ with $m^{\mathrm{v}} \geq \frac{m}{\log m}$ and $\tilde{m} \geq 4 \vee m^{1+\eta}$, for some constants $\eta \geq 0$, $r > 0$, and $s \in (0,1]$. Define Λ_m as in eq. (15) with Q an arbitrary constant greater than 1 and*

$$\frac{\eta}{\alpha} \log_Q m + 1 \leq |\Lambda_m| \leq m, \tag{17}$$

with α defined by eq. (20).

Moreover assume that for some $\delta \in (0, 1/6)$, m is large enough that it holds

$$Q \left(\frac{4 D_s \log \frac{6}{\delta}}{\sqrt{m}} \right)^{\frac{2\eta}{\alpha}} \leq \kappa^{-1} \|T\| . \tag{18}$$

Then, with probability greater than $1 - 6\delta$

$$\|f_{\mathbf{z}^{\mathrm{tot}}} - f_\rho\|_\rho \leq 4(Q^r E_r D_s + 3M) \log(6m/\delta)\ m^{-\frac{1}{2}\frac{2r-t_2}{2r+s+t_1}}, \tag{19}$$

where E_r, t_1 and t_2 are the constants defined in equations (9), (10) and (11) substituting

$$\alpha \ = \ |2 - 2r - s|_+ \ \wedge \ \frac{\eta}{1+\eta}(2r + s + |2 - 2r - s|_+). \tag{20}$$

In particular, if $r + s \geq \frac{1}{2}$ and $\eta = \frac{|2-2r-s|_+}{2r+s}$, and assuming

$$2 \log_Q m + 1 \leq |\Lambda_m| \leq m, \tag{21}$$

and

$$Q \left(\frac{4 D_s \log \frac{6}{\delta}}{\sqrt{m}} \right)^{\frac{2}{2r+s}} \leq \kappa^{-1} \|T\| ,$$

with probability greater than $1 - 6\delta$, it holds

$$\|f_{\mathbf{z}^{\mathrm{tot}}} - f_\rho\|_\rho \leq 4(Q^r E_r D_s + 3M) \log(6m/\delta)\ m^{-\frac{1}{2}\frac{2r}{2r+s}}.$$

Proof. The result is a corollary of theorems 2 and 3. The universality of K enforces the equality (12) (see [17]).

First, from equations (20) and (10), by simple algebra we get

$$\frac{\eta}{\alpha} = \frac{1}{2r + s + t_1}.$$

Therefore condition (18) is equivalent to

$$\dot{\lambda}_q = q \left(\frac{4 D_s \log \frac{6}{\delta}}{\sqrt{m}} \right)^{\frac{2}{2r+s+t_1}} \leq \kappa^{-1} \|T\| \qquad \forall q \in [1, Q],$$

and condition $\dot{\lambda} \leq \kappa^{-1} \|T\|$ in the text of Theorem 2 is verified by $\dot{\lambda}_q$ for every $q \in [1, Q]$. Moreover, since $D_s \geq 1$ and $\delta \leq 1/6$, for every $q \in [1, Q]$ we can write

$$\tilde{m} \geq 4 \vee m^{1+\eta} = 4 \vee m \left(m^{-\frac{\eta}{\alpha}} \right)^{-\alpha} \geq 4 \vee m \dot{\lambda}_q^{-\alpha},$$

which shows that also the other assumption of Theorem 2 is verified.

Hence, by Theorem 2 we get that for every $q \in [1, Q]$, with probability greater than $1 - 3\delta$, it holds

$$\|f_{\tilde{\mathbf{z}},\lambda} - f_\rho\|_\rho \leq \epsilon = Q^r E_r \left(\frac{4 D_s \log \frac{6}{\delta}}{\sqrt{m}} \right)^{\frac{2r-t_2}{2r+s+t_1}}.$$

The next step is verifying that for some $\bar{q} \in [1, Q]$, $\lambda_{\bar{q}} = \kappa\dot{\lambda}_{\bar{q}} \in \Lambda_m$, and hence applying Theorem 3.

In fact, from definition (15), assumption (17) and Proposition 5, it is clear that

$$\min \Lambda_m \leq \kappa m^{-\frac{\eta}{\alpha}} \leq \lambda_1 \leq \lambda_{\bar{q}} \leq \lambda_Q \leq \|T\| \leq \kappa = \max \Lambda_m,$$

for some $\bar{q}$.

Applying Theorem 3, we get that with probability greater than $1 - 6\delta$ it holds

$$\left\| f_{\mathbf{z}}^{\mathrm{tot}} - f_\rho \right\|_\rho \leq \hat{\epsilon},$$

with, using again condition (17) and the assumption $m^{\mathrm{v}} \geq \frac{m}{\log m}$, the chain of inequalities

$$\begin{aligned}
\hat{\epsilon} &= \left(\epsilon^2 + \frac{80M^2}{m^{\mathrm{v}}} \log \frac{6\,|\Lambda_m|}{\delta}\right)^{\frac{1}{2}} \\
&\leq \left(\epsilon^2 + \frac{80M^2}{m} \log^2 \frac{6m}{\delta}\right)^{\frac{1}{2}} \leq \epsilon + \frac{12M}{\sqrt{m}} \log \frac{6m}{\delta} \\
&\leq 4Q^r E_r D_s \log(6/\delta)\; m^{-\frac{1}{2}\frac{2r-t_2}{2r+s+t_1}} + 12M \log(6m/\delta)\; m^{-\frac{1}{2}} \\
&\leq 4(Q^r E_r D_s + 3M) \log(6m/\delta)\; m^{-\frac{1}{2}\frac{2r-t_2}{2r+s+t_1}},
\end{aligned}$$

which concludes the proof of the first part of the Corollary.

The second part of the Corollary is an instantiation of the previous result. In fact by equations (20) and (10), the assumption $\eta = \frac{|2-2r-s|_+}{2r+s}$ implies $\alpha = |2-2r-s|_+$ and $t_1 = 0$. Moreover from the assumption $r + s \geq \frac{1}{2}$ and eq. (11) we get $t_2 = 0$, and noticing that $\frac{\eta}{\alpha} = \frac{1}{2r+s} \leq 2$, it is clear that condition (21) implies condition (17).

□

4. Proofs of Theorems 1 and 2

In this section we give the proof of Theorems 1 and 2. We use various propositions taken [2], which we state without proof.

4.1. Before proving Theorem 1, we begin showing some preliminary propositions. The first one is a technical result about sequences of real numbers.

Proposition 1. *Let $\{a_i\}_{i\in\mathbb{N}}$ and $\{b_i\}_{i\in\mathbb{N}}$ be two non-increasing sequences of reals in the interval $(0,1)$ with*

$$\begin{aligned}
\lim_{i\to\infty} a_i &= 0, \\
\lim_{i\to\infty} b_i &= 0.
\end{aligned}$$

Then there exists a sequence $\{c_i\}_{i\in\mathbb{N}}$ of reals in the interval $(0,1)$ such that, defining $d_i = \log c_i / \log b_i$, the following properties hold,

i) $\{d_i\}_{i\in\mathbb{N}}$ is a non-increasing sequence of positive reals.

ii) $\{c_i\}_{i\in\mathbb{N}}$ is a non-increasing sequence of positive reals, with

$$\begin{aligned}
&c_i \geq a_i \qquad \forall i \in \mathbb{N}, \\
&\lim_{i\to\infty} c_i = 0.
\end{aligned}$$

Proof. We consider the sequence $\{c_i\}_{i\in\mathbb{N}}$ of positive numbers constructed by the recursive rule

$$\begin{aligned}
&c_1 = a_1, \\
&c_{i+1} = a_{i+1} \;\vee\; (b_{i+1})^{\frac{\log c_i}{\log b_i}}.
\end{aligned}$$

Let us prove point *i)* by induction.

Since by assumption a_1 and b_1 belong to $(0,1)$, by construction $d_1 = \frac{\log c_1}{\log b_1} = \frac{\log a_i}{\log b_i} > 0$. Now, for $i \geq 1$ assume $d_i > 0$, then by construction, either $c_{i+1} = (b_{i+1})^{d_i}$, and hence $d_{i+1} = \frac{\log c_{i+1}}{\log b_{i+1}} = d_i > 0$, or $c_{i+1} = (b_{i+1})^{d_{i+1}} = a_{i+1} \geq (b_{i+1})^{d_i}$, and hence, since a_{i+1} and b_{i+1} belong to $(0,1)$, it holds

$$\begin{aligned} d_{i+1} &= \frac{\log a_{i+1}}{\log b_{i+1}} > 0, \\ (b_{i+1})^{d_{i+1}} \geq (b_{i+1})^{d_i} &\Rightarrow d_{i+1} \leq d_i. \end{aligned}$$

Let us now prove point *ii)*.

First, by construction $c_i \geq a_i > 0$. Moreover, again by construction, either $c_{i+1} = a_{i+1}$, and hence,

$$c_{i+1} = a_{i+1} \leq a_i \leq c_i,$$

or $c_{i+1} = (b_{i+1})^{d_i}$ and hence, since $d_i > 0$ by point *i)*, it holds

$$c_{i+1} = (b_{i+1})^{d_i} \leq (b_i)^{d_i} = c_i.$$

Therefore the sequence $\{c_i\}_{i\in\mathbb{N}}$ is non-increasing and $c_i \leq c_1 = a_1 < 1$.

Finally, we prove that $\lim_i c_i = 0$.

Let us assume the there exists an infinite increasing sequence of naturals $\{i(k)\}_{k\in\mathbb{N}}$, such that

$$c_{i(k)} = a_{i(k)} \qquad \forall k \in \mathbb{N}.$$

Since, by assumption, $\lim_i a_i = 0$, then $\lim_k c_{i(k)} = 0$. Therefore, since we already proved that $\{a_i\}_{i\in\mathbb{N}}$ is non-increasing, $\lim_i c_i = 0$. Which proves the Proposition, if $\{i(k)\}_{k\in\mathbb{N}}$ exists.

If $\{i(k)\}_{k\in\mathbb{N}}$ does not exist, by construction, there exists $I \in \mathbb{N}$ such that

$$c_{i+1} = (b_{i+1})^{d_i} \qquad \forall i \geq I.$$

Therefore, recalling the definition of d_i, by induction, it follows

$$c_i = (b_i)^{d_I} \qquad \forall i > I.$$

Recalling that $d_I > 0$ and $\lim_i b_i = 0$, the relation above proves that, also in this case, $\lim_i c_i = 0$.

□

The next proposition introduces the functions f_λ^{tr} and shows some simple results related to them.

Proposition 2. *For any $\lambda > 0$ let the truncated function f_λ^{tr} be defined by*

$$f_\lambda^{\mathrm{tr}} = P_\lambda f_{\mathcal{H}} \tag{22}$$

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

$$P_\lambda = \Theta_\lambda(L_K), \tag{23}$$

with

$$\Theta_\lambda(\sigma) = \begin{cases} 1 & \text{if } \sigma \geq \lambda, \\ 0 & \text{if } \sigma < \lambda. \end{cases} \tag{24}$$

Then the function $a : (0, \kappa] \to \mathbb{R}$, defined by

$$a(\lambda) = \left\| f_\lambda^{\mathrm{tr}} - f_{\mathcal{H}} \right\|_\rho, \tag{25}$$

is non-decreasing and fulfills the following properties

$$0 \leq a(\lambda) \leq M \qquad \forall \lambda \in (0, \kappa], \tag{26}$$

$$\lim_{\lambda \to 0} a(\lambda) = 0. \tag{27}$$

Proof. Recall that the self-adjoint integral operator L_K has a countable eigensystem $\{(\lambda_i, \phi_i)\}_{i=1}^{\infty}$ with positive eigenvalues decreasing to zero (see [5]). Moreover $L_K^{\frac{1}{2}}$ is an isometry between $\mathcal{L}^2(X, \rho_X)$ and $\mathcal{H}$ (again, see [5]). Therefore, since $f_{\mathcal{H}}$ is the projection of f_ρ over the closure of $\mathcal{H}$ in $\mathcal{L}^2(X, \rho_X)$, it holds

$$f_{\mathcal{H}} = \sum_{i=1}^{\infty} |\langle f_\rho, \phi_i \rangle_\rho|^2 \phi_i.$$

Hence, by the definition of f_λ^{tr}, and recalling that $Y \subset [-M, M]$, we get

$$0 \leq a(\lambda)^2 = \sum_{\lambda_i < \lambda} |\langle f_\rho, \phi_i \rangle_\rho|^2 \leq \sum_{i=1}^{\infty} |\langle f_\rho, \phi_i \rangle_\rho|^2 \leq \|f_\rho\|_\rho^2 \leq M^2.$$

Monotonicity and convergence to zero for $a(\lambda)$ follow from the relation above by standard arguments on convergent series of positive numbers. □

The next proposition is used in the proof of Theorem 1.

Proposition 3. *Let $\bar{r}$ be a positive number. Then, there exists a function*

$$R : (0, 1] \to (0, \bar{r}]$$

such that

$$\kappa^{R(\dot{\lambda})} \left\| L_K^{R(\dot{\lambda})} P_\lambda f_{\mathcal{H}} \right\|_\rho \leq 4M, \qquad \forall \dot{\lambda} \in (0, 1], \tag{28}$$

$$\lim_{\dot{\lambda} \to 0} \dot{\lambda}^{R(\dot{\lambda})} = 0. \tag{29}$$

Proof. Let $\{\lambda_i, \phi_i\}$ be the eigensystem of the positive compact operator L_K (we also use the shortcut notation $\dot{\lambda}_i = \kappa^{-1}\lambda_i$). First, if the range of L_K is finite dimensional, the choice $R(\dot{\lambda}) = \bar{r}$ fulfills trivially the required conditions. Second, from definition (25), it is clear that if the sequence $\{a(\lambda_i)\}_i$ has only a finite number of positive elements, $f_{\mathcal{H}}$ belongs to the finite dimensional range of the projector $P_{\bar{\lambda}}$, for some positive $\bar{\lambda}$, and the choice $R(\dot{\lambda}) = \bar{r}$ is again a trivial solution.

Therefore in the following we assume $\lambda_i > 0$ and $a(\lambda_i) > 0$ for every $i \in \mathbb{N}$. Moreover, from Proposition 5, $\lambda_i \leq \kappa$, and by eq. (26), $a(\lambda) \leq M$. Hence we can apply Proposition 1 to the non-increasing sequences $\{a_i\}_i$ and $\{b_i\}_i$ defined by

$$a_i = \frac{a(\lambda_i)}{2M},$$
$$b_i = \frac{\lambda_i}{2\kappa}.$$

The function R is defined in terms of the sequence $\{d_i\}_i$ constructed in Proposition 1 as follows

$$R(\lambda) \quad = \quad \begin{cases} \bar{r} & \text{if} \quad \lambda_1 < \lambda \leq 1, \\ \bar{r} d_i / (\bar{r} \vee d_1) & \text{if} \quad \lambda_{i+1} < \lambda \leq \lambda_i, \quad i \geq 1. \end{cases}$$

Equality (29) can be proved, recalling that by Proposition 1 $c_i = b_i^{d_i} \leq \dot{\lambda}_i^{d_i}$ goes to zero as $i \to \infty$, and hence

$$\lim_{\lambda \to 0} \dot{\lambda}^{R(\dot{\lambda})} = \lim_{i \to \infty} \dot{\lambda}_i^{R(\dot{\lambda}_i)} = \lim_{i \to \infty} \left((2b_i)^{d_i} \right)^{\bar{r}/(\bar{r} \vee d_1)} \leq 2^{\bar{r}} \left(\lim_{i \to \infty} c_i \right)^{\bar{r}/(\bar{r} \vee d_1)} = 0.$$

Since by Proposition 1 $\{d_i\}_i$ is a sequence of non-increasing positives, then R is non-decreasing. Therefore, defining $f_i = \langle f_{\mathcal{H}}, \phi_i \rangle_\rho$, we can write

$$\begin{aligned}
\kappa^{2R(\dot{\lambda})} \left\| L_K^{-R(\dot{\lambda})} P_\lambda f_{\mathcal{H}} \right\|_\rho^2 &= \sum_{\dot{\lambda}_i \geq \dot{\lambda}} f_i^2 \dot{\lambda}_i^{-2R(\dot{\lambda})} \leq \sum_i f_i^2 \dot{\lambda}_i^{-2R(\dot{\lambda}_i)} \\
&= \sum_i f_i^2 \left((2b_i)^{d_i} \right)^{-\overline{r}/(\overline{r} \vee d_1)} \\
\left(b_i^{d_i} = c_i < 1 \right) &\leq \sum_i f_i^2 c_i^{-1} \\
(c_i > a_i) &\leq \sum_i f_i^2 a_i^{-1} = 2M \sum_i f_i^2 a(\lambda_i)^{-1} \\
&= 2M \sum_{k=0}^{\infty} \sum_{2^{-k-1} < a(\lambda_i)/M \leq 2^{-k}} f_i^2 a(\lambda_i)^{-1} \\
&\leq 2 \sum_{k=0}^{\infty} 2^{k+1} \sum_{2^{-k-1} < a(\lambda_i)/M \leq 2^{-k}} f_i^2 \\
\left(a(\lambda)^2 = \sum_{\lambda_i \leq \lambda} f_i^2 \right) &\leq 4M^2 \sum_{k=0}^{\infty} 2^{-k} = 8M^2,
\end{aligned}$$

which proves inequality (28) and concludes the proof. □

We now state four propositions from [2]. The first one introduces the empirical and ideal estimators least-squares $f^{\mathrm{ls}}_{\tilde{\mathbf{z}},\lambda}$ and $f^{\mathrm{ls}}_{\lambda}$.

Proposition 4. *Assume* $\lambda \leq \|T\|$ *and*

$$\lambda \tilde{m} \geq 16 \kappa \mathcal{N}(\lambda) \log^2 \frac{6}{\delta}, \tag{30}$$

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

$$f^{\mathrm{ls}}_{\tilde{\mathbf{z}},\lambda} = (T_{\tilde{\mathbf{x}}} + \lambda)^{-1} g_{\mathbf{z}}, \tag{31}$$

$$f^{\mathrm{ls}}_{\lambda} = (T+\lambda)^{-1} L_K f_{\mathcal{H}}. \tag{32}$$

Proof. See Proposition 1 in [2]. □

The second one gives two simple properties of the operator T and the effective dimension $\mathcal{N}(\lambda)$.

Proposition 5. *For every probability measure* ρ_X *and* $\lambda > 0$, *it holds*

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

and

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

Proof. See Proposition 2 in [2]. □

The other two propositions from [2] estimate two different terms which appear in the proofs of Theorems 1 and 1. The symbol $\lfloor x \rfloor$ in the text below represents the greater integer less or equal to x.

Proposition 6. *Let f belong to* Im L_K^r *for some $r > 0$. Then, if $\lambda \in (0, \kappa]$, it holds*

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

where P_λ is defined in eq. (23), and

$$\begin{aligned} \gamma &= \lambda^{-1}\,\|T - T_{\tilde{\mathbf{x}}}\|\,, \\ \eta &= |r - \frac{1}{2}| - \lfloor |r - \frac{1}{2}| \rfloor. \end{aligned} \tag{33}$$

Proof. See Proposition 6 in [2]. □

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

$$\Omega_\lambda = \sqrt{T} G_\lambda(T_{\tilde{\mathbf{x}}})\,(T_{\tilde{\mathbf{x}}} + \lambda)(T + \lambda)^{-\frac{1}{2}}. \tag{34}$$

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

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

with γ defined in eq. (33).

Proof. See Proposition 7 in [2]. □

We finally need the following probabilistic inequality based on a result of [14], see also Th. 3.3.4 of [18]. We report it without proof.

Proposition 8. *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*

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

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

$$\mathbb{P}_{(\omega_1,\ldots,\omega_m)\sim P^m}\left[\left\| \frac{1}{m}\sum_{i=1}^m \xi(\omega_i) - \mathbb{E}[\xi] \right\|_{\mathcal{K}} \leq 2\left(\frac{H}{m} + \frac{\sigma}{\sqrt{m}}\right)\log\frac{2}{\delta} \right] \geq 1 - \delta. \tag{35}$$

We are now ready to prove Theorem 1.

Proof of Theorem 1. Let us consider the expansion

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

where the operator Ω_λ is defined by equation (34), the ideal RLS estimators are $f_\lambda^{\mathrm{ls}} = (T+\lambda)^{-1}T_M f_{\mathcal{H}}$ 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,

$$\|f_{\tilde{\mathbf{z}},\lambda} - f_{\mathcal{H}}\|_\rho \leq D(\tilde{\mathbf{z}},\lambda)\left(S^{\mathrm{ls}}(\tilde{\mathbf{z}},\lambda) + R(\lambda) + \bar{S}^{\mathrm{ls}}(\tilde{\mathbf{z}},\lambda)\right) + P(\tilde{\mathbf{z}},\lambda) + P^{\mathrm{tr}}(\lambda), \tag{36}$$

with

$$
\begin{aligned}
S^{\mathrm{ls}}(\tilde{\mathbf{z}},\lambda) &= \left\|(T+\lambda)^{\frac{1}{2}}(f^{\mathrm{ls}}_{\tilde{\mathbf{z}},\lambda}-f^{\mathrm{ls}}_{\lambda})\right\|_{\mathcal{H}}, \\
\bar{S}^{\mathrm{ls}}(\tilde{\mathbf{z}},\lambda) &= \left\|(T+\lambda)^{\frac{1}{2}}(f^{\mathrm{ls}}_{\tilde{\mathbf{z}}',\lambda}-\bar{f}^{\mathrm{ls}}_{\lambda})\right\|_{\mathcal{H}}, \\
D(\tilde{\mathbf{z}},\lambda) &= \|\Omega_\lambda\|, \\
P(\tilde{\mathbf{z}},\lambda) &= \left\|\sqrt{T}\,(G_\lambda(T_{\tilde{\mathbf{x}}})T_{\tilde{\mathbf{x}}}-\mathrm{Id})\,f^{\mathrm{tr}}_{\lambda}\right\|_{\mathcal{H}}, \\
P^{\mathrm{tr}}(\lambda) &= \left\|f^{\mathrm{tr}}_{\lambda}-f_{\mathcal{H}}\right\|_{\rho}, \\
R(\lambda) &= \left\|(T+\lambda)^{\frac{1}{2}}(\bar{f}^{\mathrm{ls}}_{\lambda}-f^{\mathrm{ls}}_{\lambda})\right\|_{\mathcal{H}}.
\end{aligned} \tag{37}
$$

Terms S^{ls} and $\bar{S}^{\mathrm{ls}}$ will be estimated by Proposition 4, terms P^{tr} and R by Proposition 2, term D by Proposition 7, and finally term P by Propositions 6, 3 and 2.

Step 1: Estimate of S^{ls}. Since $L_K^{\frac{1}{2}}$ is an isometry between $\mathcal{L}^2(X,\rho_X)$ and $\mathcal{H}$ (see [5]), we obtain

$$
\left\|f^{\mathrm{ls}}_{\lambda}\right\| \leq \left\|\sqrt{T}(T+\lambda)^{-1}\right\| \left\|L_K^{\frac{1}{2}} f_{\mathcal{H}}\right\|_{\mathcal{H}} \leq \frac{\|f_{\mathcal{H}}\|_\rho}{\sqrt{\lambda}} \leq \frac{M}{\sqrt{\lambda}}. \tag{38}
$$

Now, let δ be an arbitrary real in $(0,1)$. From the assumptions on λ_m, for large enough m, we have

$$
\begin{aligned}
\lambda_m\sqrt{m} &\geq 4\kappa\log\frac{6}{\delta}, \\
\lambda_m &\leq \|T\|.
\end{aligned}
$$

Hence, by Proposition 5, for large enough m, the assumptions of Proposition 4 are verified, and we get that with probability greater than $1-\delta$

$$
\begin{aligned}
S^{\mathrm{ls}}(\tilde{\mathbf{z}}_m,\lambda_m) &\leq 8\left(M+\sqrt{\kappa\frac{m}{\tilde{m}_m}}\left\|f^{\mathrm{ls}}_{\lambda_m}\right\|_{\mathcal{H}}\right)\left(\frac{2}{m}\sqrt{\frac{\kappa}{\lambda_m}}+\sqrt{\frac{\mathcal{N}(\lambda_m)}{m}}\right)\log\frac{6}{\delta} \\
(\textit{Prop.}5,\textit{eq. (38)}) &\leq 8M\left(1+\sqrt{\frac{\kappa}{\lambda_m}}\right)\frac{1}{\sqrt{m}}\left(2\sqrt{\frac{\kappa}{\lambda_m m}}+\sqrt{\frac{\kappa}{\lambda_m}}\right)\log\frac{6}{\delta} \\
(\lambda_m\leq\kappa, m\geq 4) &\leq \frac{32M\kappa}{\lambda_m\sqrt{m}}\log\frac{6}{\delta} \;\to\; 0.
\end{aligned}
$$

Hence it holds

$$
\lim_{m\to\infty} S^{\mathrm{ls}}(\tilde{\mathbf{z}}_m,\lambda_m) \overset{P}{=} 0. \tag{39}
$$

Step 2: Estimate of $\bar{S}^{\mathrm{ls}}$. This term 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 $X\times[-M',M']$, with $M'=\sup_{x\in X}f_{\rho'}(x)\leq\sqrt{\kappa}\left\|f^{\mathrm{tr}}_{\lambda}\right\|_{\mathcal{H}}$. Reasoning as in the analysis of S^{ls}, we obtain that, for every $\delta\in(0,1)$ and large enough m, with probability greater than $1-\delta$ it holds

$$
\begin{aligned}
\bar{S}^{\mathrm{ls}}(\tilde{\mathbf{z}}_m,\lambda_m) &\leq 8\left(M'+\sqrt{\kappa}\left\|\bar{f}^{\mathrm{ls}}_{\lambda_m}\right\|_{\mathcal{H}}\right)\left(\frac{2}{\tilde{m}_m}\sqrt{\frac{\kappa}{\lambda_m}}+\sqrt{\frac{\mathcal{N}(\lambda_m)}{\tilde{m}_m}}\right)\log\frac{6}{\delta} \\
(\textit{Prop.}5) &\leq 16\sqrt{\kappa}\left\|f^{\mathrm{tr}}_{\lambda_m}\right\|_{\mathcal{H}}\frac{1}{\sqrt{m}}\left(2\sqrt{\frac{\kappa}{\lambda_m m}}+\sqrt{\frac{\kappa}{\lambda_m}}\right)\log\frac{6}{\delta} \\
(m\geq 4) &\leq \frac{32\kappa}{\sqrt{\lambda_m m}}\left\|P_{\lambda_m}L_K^{-\frac{1}{2}}P_{\lambda_m}f_{\mathcal{H}}\right\|_\rho\log\frac{6}{\delta} \leq \frac{32\kappa M}{\lambda_m\sqrt{m}}\log\frac{6}{\delta} \;\to\; 0.
\end{aligned}
$$

Hence it holds

$$
\lim_{m\to\infty} \bar{S}^{\mathrm{ls}}(\tilde{\mathbf{z}}_m,\lambda_m) \overset{P}{=} 0. \tag{40}
$$

Step 3: Estimate of P^{tr}. By definition (25), $P^{\mathrm{tr}}(\lambda) = a(\lambda)$. Hence from eq. (27)

$$\lim_{m\to\infty} P^{\mathrm{tr}}(\lambda_m) = \lim_{m\to\infty} a(\lambda_m) = \lim_{\lambda\to 0} a(\lambda) = 0, \tag{41}$$

where we used the assumption (6).

Step 4: Estimate of R. Since from the definitions of f_λ^{ls} and $\bar{f}_\lambda^{\mathrm{ls}}$,

$$R(\lambda) = \left\| (T+\lambda)^{-\frac{1}{2}} T(\bar{f}_\lambda^{\mathrm{ls}} - f_\lambda^{\mathrm{ls}}) \right\|_{\mathcal{H}} \leq \left\| \sqrt{T}(f_\lambda^{\mathrm{tr}} - f_{\mathcal{H}}) \right\|_{\mathcal{H}} \leq P^{\mathrm{tr}}(\lambda),$$

from (41) we get

$$\lim_{m\to\infty} R(\lambda_m) = 0. \tag{42}$$

Step 5: Estimate of D. In order to estimate $D(\tilde{\mathbf{z}}, \lambda)$, we have first to estimate the quantity $\gamma = \gamma(\tilde{\mathbf{z}}, \lambda)$ (see definition (33)) appearing in the Proposition 7. Our estimate for $\gamma(\tilde{\mathbf{z}}, \lambda)$ follows from Proposition 8 applied to the random variable $\xi : X \to \mathcal{L}_{\mathrm{HS}}(\mathcal{H})$ defined by

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

We can set $H = \frac{2\kappa}{\lambda}$ and $\sigma = \frac{H}{2}$, and obtain that, for every $\delta \in (0,1)$ and $m \geq 4$, with probability greater than $1-\delta$

$$\gamma(\tilde{\mathbf{z}}_m, \lambda) \leq \lambda^{-1} \left\| T - T_{\tilde{\mathbf{x}}} \right\|_{\mathrm{HS}} \leq \frac{2}{\lambda} \left(\frac{2\kappa}{\tilde{m}_m} + \frac{\kappa}{\sqrt{\tilde{m}_m}} \right) \log \frac{2}{\delta} \leq 4 \frac{\kappa}{\lambda\sqrt{m}} \log \frac{2}{\delta} =: \epsilon(m, \lambda, \delta).$$

From the expression of $\epsilon(m, \lambda, \delta)$ we see that, by the assumption (7), for every $\delta \in (0,1)$,

$$\lim_{m\to\infty} \epsilon(m, \lambda_m, \delta) = 0,$$

and hence,

$$\lim_{m\to\infty} \gamma(\tilde{\mathbf{z}}_m, \lambda_m) \overset{P}{=} 0. \tag{43}$$

Finally, from eq. (43) and Proposition 7 we find

$$D(\tilde{\mathbf{z}}_m, \lambda_m) \leq \left(1 + 2\sqrt{\gamma(\tilde{\mathbf{z}}_m, \lambda_m)} \right) A \overset{P}{\to} 3A. \tag{44}$$

Step 6: Estimate of P. First, notice that by the definition (3), WLOG we can assume $\bar{r} < \frac{1}{2}$. Moreover by condition (6), we can assume m large enough that $\lambda_m \leq \kappa$. We consider the function R introduced by Proposition 3, and apply Proposition 6, with $f = P_{\lambda_m} f_{\mathcal{H}}$ and $r_m = R(\kappa^{-1}\lambda_m) \leq \bar{r}$, getting

$$P(\tilde{\mathbf{z}}_m, \lambda_m) \leq B_{\bar{r}} \left(1 + \gamma(\tilde{\mathbf{z}}_m, \lambda_m)^{\frac{1}{2}} \right) \left(2 + r_m \gamma(\tilde{\mathbf{z}}_m, \lambda_m) + \gamma(\tilde{\mathbf{z}}_m, \lambda_m)^{\frac{1}{2} - r_m} \right) \kappa^{r_m} \left\| L_K^{-r_m} P_{\lambda_m} f_{\mathcal{H}} \right\|_\rho (\kappa^{-1}\lambda_m)^{r_m}.$$

This result together with eq. (43), and recalling that by Proposition 3 and assumption (6), the sequence $\{r_m\}_m$ verifies the two conditions

$$\kappa^{r_m} \left\| L_K^{-r_m} P_{\lambda_m} f_{\mathcal{H}} \right\|_\rho \leq 4M \qquad \forall m,$$
$$\lim_{m\to\infty} (\kappa^{-1}\lambda_m)^{r_m} = 0,$$

proves that

$$\lim_{m\to\infty} P(\tilde{\mathbf{z}}_m, \lambda_m) \overset{P}{=} 0. \tag{45}$$

The proof of the Theorem is completed considering the limit $m \to \infty$ of estimate (36), and using equations (39), (40), (41), (42), (44) and (45).

□

4.2. Before showing the proof of Theorem 2, we state two propositions from [2] which describe properties of the functions f_λ^{tr} and f_λ^{ls} (defined in eq. (22) and eq. (32) respectively) when $f_{\mathcal{H}} \in \operatorname{Im} L_K^r$.

Proposition 9. *Let $f_{\mathcal{H}} \in \operatorname{Im} L_K^r$ for some $r > 0$. Then, the following estimates hold,*

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

Proof. See Proposition 3 in [2]. □

Proposition 10. *Let $f_{\mathcal{H}} \in \operatorname{Im} L_K^r$ for some $r > 0$. Then, the following estimates hold,*

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

Proof. See Proposition 1 in [2]. □

We are now ready to prove Theorem 2.

Proof of Theorem 2. We consider the same decomposition (see equations (36) and (37)) for $\|f_{\tilde{\mathbf{z}},\lambda} - f_{\mathcal{H}}\|_\rho$ that we used in the proof of Theorem 1.

Terms S^{ls} and $\bar{S}^{\mathrm{ls}}$ will be estimated by Proposition 4, term D by Proposition 7, term P by Proposition 6 and finally terms P^{tr} and R by Proposition 9.

Let us begin with the estimates of S^{ls} and $\bar{S}^{\mathrm{ls}}$. First observe that, by Proposition 5, it holds

$$\dot{\lambda} \leq \kappa^{-1} \|T\| \leq 1,$$

therefore, since by assumption $\tilde{m} \geq m\dot{\lambda}^{-|2-2r-s|_+ + t_1} \geq m\dot{\lambda}^{-|1-2r|_+ + t_1}$, we get,

$$\dot{\lambda}\tilde{m} \geq \dot{\lambda}^{-|1-2r|_+ + 1 + t_1} m \geq \dot{\lambda}^{2r+t_1} m.$$

Moreover, by eq. (8) and definition (5), we find

$$\dot{\lambda}^{2r+t_1} m = 16 q^{2r+s+t_1} D_s^2 \dot{\lambda}^{-s} \log^2 \frac{6}{\delta} \geq 16 \mathcal{N}(\lambda) \log^2 \frac{6}{\delta},$$

hence the hypothesis (30) in the text of Proposition 4 is verified.

Regarding the estimate of S^{ls}. Applying Proposition 4 and recalling that by assumption $\tilde{m} \geq m\dot{\lambda}^{-|2-2r-s|_+ + t_1} \geq m\dot{\lambda}^{-|1-2r|_+ + t_1}$ and from Proposition 10, $\sqrt{\kappa} \left\| f_\lambda^{\mathrm{ls}} \right\|_{\mathcal{H}} \leq C_r \dot{\lambda}^{-|\frac{1}{2}-r|_+}$, we get that with probability greater than $1-\delta$

$$\begin{aligned} S^{\mathrm{ls}}(\tilde{\mathbf{z}}, \lambda) &\leq 8 \left(M + \sqrt{\frac{m}{\tilde{m}}} C_r \dot{\lambda}^{-|\frac{1}{2}-r|_+} \right) \left(\frac{2}{m} \sqrt{\frac{\kappa}{\lambda}} + \sqrt{\frac{\mathcal{N}(\lambda)}{m}} \right) \log \frac{6}{\delta} \qquad (46) \\ &\leq 8(M + \dot{\lambda}^{-\frac{t_1}{2}} C_r) \frac{1}{\sqrt{m}} \left(\frac{2}{\sqrt{m\dot{\lambda}}} + \frac{D_s}{\sqrt{\dot{\lambda}^s}} \right) \log \frac{6}{\delta} \\ &= 2q^{-r-\frac{s}{2}-\frac{t_1}{2}} (M + C_r) \dot{\lambda}^r \left(1 + \frac{\dot{\lambda}^{-\frac{1}{2}(1-2r-2s-t_1)}}{2q^{r+\frac{s}{2}+\frac{t_1}{2}} D_s^2 \log \frac{6}{\delta}} \right) \qquad (\textit{eq. (8)}) \\ &\leq 3q^{-r-\frac{s}{2}-\frac{t_1}{2}} (M + C_r) \dot{\lambda}^{r-\frac{t_2}{2}} \qquad (t_1 \geq 0, q \geq 1) \\ &\leq 3(M + C_r) \dot{\lambda}^{r-\frac{t_2}{2}}. \qquad (q \geq 1) \end{aligned}$$

The term $\bar{S}^{\rm ls}$ 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 $\rho'_{|x}(y) = \delta(y - f_\lambda^{\rm tr}(x))$. Therefore the regression function induced by ρ' is $f_{\rho'} = f_\lambda^{\rm tr}$, and the support of ρ' is included in $X \times [-M', M']$, with $M' = \sup_{x\in X} f_{\rho'}(x) \leq \sqrt{\kappa}\left\|f_\lambda^{\rm tr}\right\|_{\mathcal{H}}$. Again applying Proposition 4, we obtain that with probability greater than $1-\delta$ it holds

$$
\begin{aligned}
(47) \qquad \bar{S}^{\rm ls}(\tilde{\mathbf{z}}, \lambda) &\leq 8\left(M' + \sqrt{\kappa}\left\|\bar{f}_\lambda^{\rm ls}\right\|_{\mathcal{H}}\right)\left(\frac{2}{\tilde{m}}\sqrt{\frac{\kappa}{\lambda}} + \sqrt{\frac{\mathcal{N}(\lambda)}{\tilde{m}}}\right)\log\frac{6}{\delta} \\
&\leq 16\sqrt{\kappa}\left\|f_\lambda^{\rm tr}\right\|_{\mathcal{H}}\left(\frac{2}{\tilde{m}}\sqrt{\frac{\kappa}{\lambda}} + \sqrt{\frac{\mathcal{N}(\lambda)}{\tilde{m}}}\right)\log\frac{6}{\delta} \\
(Prop.9) \qquad &\leq 16\sqrt{\frac{m}{\tilde{m}}}C_r\dot{\lambda}^{-\left|\frac{1}{2}-r\right|_+}\left(\frac{2}{m}\sqrt{\frac{\kappa}{\lambda}} + \sqrt{\frac{\mathcal{N}(\lambda)}{m}}\right)\log\frac{6}{\delta} \\
&\leq 16C_r\frac{\dot{\lambda}^{-\frac{t_1}{2}}}{\sqrt{m}}\left(\frac{2}{\sqrt{m\dot{\lambda}}} + \frac{D_s}{\sqrt{\dot{\lambda}^s}}\right)\log\frac{6}{\delta} \\
(eq.\ (8)) \qquad &= 4q^{-r-\frac{s}{2}-\frac{t_1}{2}}C_r\dot{\lambda}^r\left(1 + \frac{\dot{\lambda}^{-\frac{1}{2}(1-2r-2s-t_1)}}{2q^{r+\frac{s}{2}+\frac{t_1}{2}}D_s^2\log\frac{6}{\delta}}\right) \\
(t_1 \geq 0, q \geq 1) \qquad &\leq 6q^{-r-\frac{s}{2}-\frac{t_1}{2}}C_r\dot{\lambda}^{r-\frac{t_2}{2}} \\
(q \geq 1) \qquad &\leq 6C_r\dot{\lambda}^{r-\frac{t_2}{2}}.
\end{aligned}
$$

In order to get an upper bound for D and P, we have first to estimate the quantity $\gamma = \gamma(\tilde{\mathbf{z}}, \lambda)$ (see definition (33)) appearing in the Propositions 6 and 7. Our estimate for $\gamma(\tilde{\mathbf{z}}, \lambda)$ follows from Proposition 8 applied to the random variable $\xi : X \to \mathcal{L}_{\rm HS}(\mathcal{H})$ 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 $1-\delta$

$$
\gamma(\tilde{\mathbf{z}}, \lambda) \leq \lambda^{-1}\left\|T - T_{\tilde{\mathbf{x}}}\right\|_{\rm HS} \leq \frac{2}{\lambda}\left(\frac{2\kappa}{\tilde{m}} + \frac{\kappa}{\sqrt{\tilde{m}}}\right)\log\frac{2}{\delta} \leq 4\frac{1}{\dot{\lambda}\sqrt{\tilde{m}}}\log\frac{2}{\delta}
$$

$$
\leq 4\frac{\dot{\lambda}^{\left|1-r-\frac{s}{2}\right|_+ - 1 - \frac{t_1}{2}}}{\sqrt{m}}\log\frac{2}{\delta} \leq \dot{\lambda}^{\left|1-r-\frac{s}{2}\right|_+ - (1-r-\frac{s}{2})} \leq \dot{\lambda}^{\left|r+\frac{s}{2}-1\right|_+} \leq 1,
$$

where we used the assumption $\tilde{m} \geq 4 \vee m\dot{\lambda}^{-|2-2r-s|_+ + t_1}$ and the expression for $\dot{\lambda}$ in the text of the Theorem.

Hence, since $\gamma(\tilde{\mathbf{z}}, \lambda) \leq \dot{\lambda}^{\left|r+\frac{s}{2}-1\right|_+}$, from Proposition 7 we get

$$
(48) \qquad D(\tilde{\mathbf{z}}, \lambda) \quad \leq \quad (1 + 2\sqrt{\gamma})A \leq 3A,
$$

and from Proposition 6

$$
\begin{aligned}
(49) \qquad P(\tilde{\mathbf{z}}, \lambda) &\leq B_rC_r(1+\sqrt{\gamma})(2 + r\gamma\dot{\lambda}^{\frac{3}{2}-r} + \gamma^\eta)\dot{\lambda}^r \\
&\leq 2B_rC_r(3 + r\gamma\dot{\lambda}^{\frac{3}{2}-r})\dot{\lambda}^r \\
&\leq 2B_rC_r(3 + r\dot{\lambda}^{\left|r+\frac{s}{2}-1\right|_+ + \frac{3}{2}-r})\dot{\lambda}^r \\
&\leq 2B_rC_r(3 + r\dot{\lambda}^{\frac{s+1}{2}})\dot{\lambda}^r \leq 2B_rC_r(3+r)\dot{\lambda}^r.
\end{aligned}
$$

Regarding terms $P^{\rm tr}$ and R. From Proposition 9 we get

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

and hence,

$$R(\lambda) = \left\|(T+\lambda)^{-\frac{1}{2}}T(\vec{f}_\lambda^{\,\mathrm{ls}} - f_\lambda^{\mathrm{ls}})\right\|_{\mathcal{H}} \leq \left\|\sqrt{T}(\vec{f}_\lambda^{\,\mathrm{ls}} - f_\lambda^{\mathrm{ls}})\right\|_{\mathcal{H}} \leq P^{\mathrm{tr}} \leq C_r \dot{\lambda}^r. \tag{51}$$

The proof is completed by plugging inequalities (46), (47), (48), (49), (50) and (51) in (36), recalling the expression for $\dot{\lambda}$. □

5. Proof of Theorem 3

The following result is due to [13], adapted to a suitable form used in this paper.

Proposition 11. *Let $\{X_i\}_{i\in 1}^n$ be a set of real valued i.i.d. random variables with mean μ, $|X_i| \leq B$ and $\mathbb{E}[(X_i-\mu)^2] \leq \sigma^2$, for all $i \in \{1,\ldots,n\}$. Then for arbitrary $\alpha > 0$, $\epsilon > 0$,*

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n X_i - \mu \geq \alpha\sigma^2 + \epsilon\right] \leq e^{-\frac{6n\alpha\epsilon}{3+4\alpha B}}, \tag{52}$$

and

$$\mathbb{P}\left[\mu - \frac{1}{n}\sum_{i=1}^n X_i \geq \alpha\sigma^2 + \epsilon\right] \leq e^{-\frac{6n\alpha\epsilon}{3+4\alpha B}}. \tag{53}$$

Proof. It suffices to prove the one side inequality (52). For any $s > 0$,

$$\begin{aligned}
&\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n X_i - \mu \geq \alpha\sigma^2 + \epsilon\right] = \mathbb{P}\left[e^{\frac{s}{n}\sum_{i=1}^n (X_i-\mu)} \geq e^{s(\alpha\sigma^2+\epsilon)}\right] \\
\leq\ & e^{-s\epsilon - s\alpha\sigma^2}\mathbb{E}e^{\frac{s}{n}\sum_{i=1}^n (X_i-\mu)}, \quad \text{by Markov inequality} \\
=\ & e^{-s\epsilon - s\alpha\sigma^2}\prod_{i=1}^n \mathbb{E}e^{\frac{s}{n}(X_i-\mu)}, \quad \text{by independence of } X_i
\end{aligned}$$

Denote $Z_i = X_i - \mu$, $t = s/n$ and $B_1 = 2B$. Thus for those s such that $sB < 3n/2$ (or equivalently $B_1 t/3 < 1$),

$$\begin{aligned}
\mathbb{E}e^{tZ_i} &= 1 + \sum_{k=1}^\infty \frac{t^k}{k!}\mathbb{E}[Z_i^k] \leq 1 + 0 + \sum_{k=2}^\infty \frac{t^k}{k!}B_1^{k-2}\sigma^2 \leq 1 + \frac{t^2\sigma^2}{2}\sum_{k=0}^\infty \left(\frac{B_1 t}{3}\right)^k \\
&= 1 + \frac{3t^2\sigma^2}{6-4Bt} \leq \exp\left(\frac{3t^2\sigma^2}{6-4Bt}\right) = \exp\left(\frac{3s^2\sigma^2}{n^2(6-4sB/n)}\right)
\end{aligned}$$

whence

$$e^{-s\epsilon - s\alpha\sigma^2}\prod_{i=1}^n \mathbb{E}e^{\frac{s}{n}(X_i-\mu)} \leq e^{-s\epsilon}\exp\left\{\frac{s\sigma^2}{n}\left(\frac{3s}{6-4sB/n} - n\alpha\right)\right\}.$$

Setting $s = s_0 = \dfrac{6\alpha n}{3+4\alpha B}$ (one can check that $s_0 B = \frac{6n\alpha B}{3+4\alpha B} < 3n/2$), we have $\frac{3s_0}{6-4s_0 B/n} - n\alpha = 0$ and thus $r.h.s. \leq e^{-s_0\epsilon} = \exp\left(-\dfrac{6n\alpha\epsilon}{3+4\alpha B}\right)$, which gives estimate (52). □

We are now ready to prove Theorem 3.

Proof of Theorem 3. The strategy of the proof is the following. Define

$$\lambda_m^* = \operatorname*{argmin}_{\lambda\in\Lambda_m}\int_Z (T_M f_{\mathbf{z},\lambda}(x) - y)^2 d\rho. \tag{54}$$

Notice that, since for every $f \in \mathcal{L}^2(X,\rho_X)$,

$$\int_Z (f(x)-y)^2 d\rho = \|f - f_\rho\|_\rho^2 + \int_Z (f_\rho(x)-y)^2 d\rho,$$

definition (54) of λ_m^*, is equivalent to

$$\lambda_m^* = \operatorname*{argmin}_{\lambda \in \Lambda_m} \|T_M f_{\tilde{\mathbf{z}},\lambda} - f_\rho\|_\rho .$$

Now, from the equality above, the assumption of the Theorem, and recalling that $f_\rho(x) \in Y \subset [-M, M]$, we get that with probability greater than $1-\delta$ it holds

$$\left\|T_M f_{\tilde{\mathbf{z}},\lambda_m^*} - f_\rho\right\|_\rho \leq \|T_M f_{\tilde{\mathbf{z}},\lambda_m} - f_\rho\|_\rho \leq \|f_{\tilde{\mathbf{z}},\lambda_m} - f_\rho\|_\rho \leq \epsilon. \tag{55}$$

We claim that for every $\tilde{\mathbf{z}}$, $\lambda > 0$ and $\delta \in (0,1)$, with probability greater than $1-\delta$ over the probability measure $\rho^{m^{\mathrm{v}}}$, it holds

$$\|f_{\mathbf{z}^{\mathrm{tot}}} - f_\rho\|_\rho^2 \leq 2\left\|T_M f_{\tilde{\mathbf{z}},\lambda_m^*} - f_\rho\right\|_\rho^2 + \frac{80M^2}{m^{\mathrm{v}}} \log \frac{2\,|\Lambda_m|}{\delta}. \tag{56}$$

Estimates (55) and (56) together will complete the proof of the Theorem.

We now proceed to proving eq. (56). For $i = 1, \ldots, m^{\mathrm{v}}$, let us define the random variables

$$\xi_i^\lambda = (T_M f_{\tilde{\mathbf{z}},\lambda}(x_i^{\mathrm{v}}) - y_i^{\mathrm{v}})^2 - (f_\rho(x_i^{\mathrm{v}}) - y_i^{\mathrm{v}})^2 .$$

Clearly

$$\begin{aligned}
|\xi_i^\lambda| &\leq 4M^2, \\
\mathbb{E}[\xi_i^\lambda] &= \int_Z (T_M f_{\tilde{\mathbf{z}},\lambda}(x) - y)^2 \, d\rho - \int_Z (f_\rho(x) - y)^2 \, d\rho = \|T_M f_{\tilde{\mathbf{z}},\lambda} - f_\rho\|_\rho^2, \\
\mathbb{E}[(\xi_i^\lambda)^2] &= \int_Z (T_M f_{\tilde{\mathbf{z}},\lambda}(x) - f_\rho(x))^2 \, (T_M f_{\tilde{\mathbf{z}},\lambda}(x) + f_\rho(x) - 2y)^2 \, d\rho \\
&\leq 16M^2 \, \|T_M f_{\tilde{\mathbf{z}},\lambda} - f_\rho\|_\rho^2 = 16M^2 \mathbb{E}[\xi_i^\lambda].
\end{aligned}$$

Hence, using Proposition 11 with $X_i = \xi_i^\lambda$, $\mu = \mathbb{E}[\xi_i^\lambda]$, $B = 4M^2$ and $\sigma^2 = \mathbb{E}[(\xi_i^\lambda)^2] \leq 16M^2\mu$, we obtain that for all $\lambda \in \Lambda_m$ with probability greater than $1-\delta$,

$$\frac{1}{m^{\mathrm{v}}} \sum_{i=1}^{m^{\mathrm{v}}} \xi_i^\lambda \leq (1+\alpha')\mathbb{E}[\xi_i^\lambda] + \epsilon,$$

and

$$\mathbb{E}[\xi_i^\lambda] \leq \frac{1}{1-\alpha'} \left(\frac{1}{m^{\mathrm{v}}} \sum_{i=1}^{m^{\mathrm{v}}} \xi_i^\lambda \right) + \frac{\epsilon}{1-\alpha'},$$

where $\alpha' = 16\alpha M^2$ and $\epsilon = \dfrac{3+\alpha'}{6\alpha m^{\mathrm{v}}} \log \dfrac{2\,|\Lambda_m|}{\delta}$. Therefore

$$\begin{aligned}
\|f_{\mathbf{z}^{\mathrm{tot}}} - f_\rho\|_\rho^2 &= \mathbb{E}[\xi_i^{\hat{\lambda}_{\mathbf{z}^{\mathrm{v}}}}] \leq \frac{1}{1-\alpha'} \left(\frac{1}{m^{\mathrm{v}}} \sum_{i=1}^{m^{\mathrm{v}}} \xi_i^{\hat{\lambda}_{\mathbf{z}^{\mathrm{v}}}} \right) + \frac{\epsilon}{1-\alpha'} \\
&\leq \frac{1}{1-\alpha'} \left(\frac{1}{m^{\mathrm{v}}} \sum_{i=1}^{m^{\mathrm{v}}} \xi_i^{\lambda_m^*} \right) + \frac{\epsilon}{1-\alpha'} \\
&\leq \frac{1+\alpha'}{1-\alpha'} \mathbb{E}[\xi_i^{\lambda_m^*}] + \frac{2\epsilon}{1-\alpha'} \\
&= \frac{1+\alpha'}{1-\alpha'} \left\|T_M f_{\tilde{\mathbf{z}},\lambda_m^*} - f_\rho\right\|_\rho^2 + \frac{2\epsilon}{1-\alpha'}.
\end{aligned}$$

Setting $\alpha = 1/(48M^2)$, this gives $\alpha' = 1/3$ and

$$\|f_{\mathbf{z}^{\mathrm{tot}}} - f_\rho\|_\rho^2 \leq 2\left\|T_M f_{\tilde{\mathbf{z}},\lambda_m^*} - f_\rho\right\|_\rho^2 + \frac{80M^2}{m^{\mathrm{v}}} \log \frac{2\,|\Lambda_m|}{\delta},$$

which proves eq. (56), as desired.

□

Acknowledgements

The authors wish to thank Peter Bickel for an inspiring discussion leading to this work, Bo Li for pointing out reference [10] and E. De Vito, T. Poggio, L. Rosasco, S. Smale and A. Verri for useful discussions and suggestions.

References

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

[2] A. Caponnetto. Optimal rates for regularization operators in learning theory. Preprint, 2006.

[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, 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.

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

[8] 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.

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

[10] S. Dudoit and M.J. van der Laan. Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Statistical Methodology*, 2(2):131–154, 2001.

[11] 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.

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

[13] M. Hamers and M. Kohler. A bound on the expected maximal deviations of sample averages from their means. Preprint 2001-9, Mathematical Institute A, University of Stuttgart.

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

[15] S. Smale and D. Zhou. Learning theory estimates via integral operators and their approximations. Preprint, Toyota Technological Institute, Chicago, 2005.

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

[17] I. Steinwart. On the influence of the kernel on hte consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2002.

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