Adaptation for Regularization Operators in Learning Theory
Andrea Caponnetto and Yuan Yao
1. REPORT DATE  
10 SEP 2006  

2. REPORT TYPE  

3. DATES COVERED  
00-09-2006 to 00-09-2006  

4. TITLE AND SUBTITLE  
Adaptation for Regularization Operators in Learning Theory  

5a. CONTRACT NUMBER  

5b. GRANT NUMBER  

5c. PROGRAM ELEMENT NUMBER  

5d. PROJECT NUMBER  

5e. TASK NUMBER  

5f. WORK UNIT NUMBER  

6. AUTHOR(S)  

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)  
Massachusetts Institute of Technology, Computer Science and Artificial Intelligence Laboratory (CSAIL), Cambridge, MA, 02139  

8. PERFORMING ORGANIZATION REPORT NUMBER  

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)  

10. SPONSOR/MONITOR’S ACRONYM(S)  

11. SPONSOR/MONITOR’S REPORT NUMBER(S)  

12. DISTRIBUTION/AVAILABILITY STATEMENT  
Approved for public release; distribution unlimited  

13. SUPPLEMENTARY NOTES  
The original document contains color images.  

14. ABSTRACT  

15. SUBJECT TERMS  

16. SECURITY CLASSIFICATION OF:  
   a. REPORT unclassified  
   b. ABSTRACT unclassified  
   c. THIS PAGE unclassified  

17. LIMITATION OF ABSTRACT  

18. NUMBER OF PAGES  
21  

19a. NAME OF RESPONSIBLE PERSON  

Standard Form 298 (Rev. 8-98)  
Prescribed by ANSI Std Z39-18
Adaptation for regularization operators
in learning theory

Andrea Caponnetto \textsuperscript{a b c} Yuan Yao \textsuperscript{d}

\textsuperscript{a} C.B.C.L., McGovern Institute, Massachusetts Institute of Technology, Bldg. 46-5155, 77 Massachusetts Avenue, Cambridge, MA 02139

\textsuperscript{b} D.I.S.I., Università di Genova, Via Dodecaneso 35, 16146 Genova, Italy

\textsuperscript{c} Department of Computer Science, University of Chicago, 1100 East 58th Street, Chicago, IL 60637

\textsuperscript{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 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 \( \nu \)-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 \( \rho \). 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 \( \tilde{f}_{z,\lambda} \) with a-priori choices of \( \lambda \), can be transferred to the estimators \( f_{z_{\text{tot}},\lambda} \) which use the validation examples \( z' \) in \( z_{\text{tot}} = (\tilde{z}, z') \) to determine \( \lambda \). The subsequent corollaries are applications of Theorem 3 to the choices of \( \lambda \) 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].


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

\[ z = (z_1, \ldots, z_m), \]

with \( z_i = (x_i, y_i) \) drawn i.i.d. according to the probability measure \( \rho \) over \( Z = X \times Y \). Moreover we assume that the unsupervised part of the training set is \( (x_{m+1}^u, \ldots, x_{\tilde{m}}^u) \), with \( x_i^u \) drawn i.i.d. according to the marginal probability measure over \( X, \rho_X \). For sake of brevity we also introduce the complete training set

\[ \tilde{z} = (\tilde{z}_1, \ldots, \tilde{z}_{\tilde{m}}), \]

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

\[
\tilde{x}_i = \begin{cases} 
  x_i & \text{if } 1 \leq i \leq m, \\
  x_i^u & \text{if } m < i \leq \tilde{m}, 
\end{cases}
\]

and

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

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

In the following we will study the generalization properties of a class of estimators \( f_{\tilde{z},\lambda} \) 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

\[
f_{x,\lambda} = G_{\lambda}(T_{x}) g_x, \tag{1}
\]

where \( T_x \in \mathcal{L}(\mathcal{H}) \) is given by,

\[
T_x f = \frac{1}{\tilde{m}} \sum_{i=1}^{\tilde{m}} K_{\tilde{x}_i} \langle K_{\tilde{x}_i}, f \rangle_{\mathcal{H}}.
\]
$g_x \in \mathcal{H}$ is given by,

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

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

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

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

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

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 in terms of the integral operator $L_K : L^2(X, \rho_X) \to L^2(X, \rho_X)$ 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 $\rho_{x|\lambda}$, through the conditional distributions $G_\lambda$. The function $G_\lambda$ is the “best” function which can be arbitrarily well approximated by elements of our hypothesis space $\mathcal{H}$. More formally, $g_H$ is the projection of the regression function $g(x) = \int_x y d\rho_x(y)$ onto the closure of $\mathcal{H}$ in $L^2(X, \rho_X)$.

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 $\rho_x$ through the effective dimension $N(\lambda) = \text{Tr}[T(T + \lambda)^{-1}]$, they are defined as follows

$$(4) \quad D_s = \begin{cases} \kappa^s \|L_K^{-s} f_H\| \rho & \text{if } f_H \in \text{Im} L_K^{-s}, \\ +\infty & \text{if } f_H \not\in \text{Im} L_K^{-s}, \end{cases} \quad r > 0.$$

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

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

$$(5) \quad \lim_{m \to \infty} \lambda_m = 0,$$

$$(6) \quad \lim_{m \to \infty} \sqrt{\lambda_m} = \infty.$$
Then, if $B_\bar{r} < +\infty$ for some $\bar{r} > 0$, it holds

$$
\lim_{m \to \infty} \|f_{\tilde{z}m, \lambda_m} - f_\mathcal{H}\|_\rho \overset{p} = 0.
$$

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 $\lambda$ satisfy the constraints $\lambda \leq \|T\|$ and

$$
\dot{\lambda} = q \left( \frac{4D_s \log \frac{6}{2r + s + t_{11}}}{\sqrt{m}} \right)^{2r - t_2},
$$

for some $q \geq 1$, $\delta \in (0, 1/3)$ and $t_1$ defined in eq. (10). Finally, assume $\tilde{m} \geq 4 \vee m \dot{\lambda}^{-1}$.

Then, with probability greater than $1 - 3\delta$, it holds

$$
\|f_{\tilde{z}, \lambda} - f_\mathcal{H}\|_\rho \leq q' E_\nu \left( \frac{4D_s \log \frac{6}{\sqrt{m}}}{\sqrt{m}} \right)^{2r - t_2},
$$

where

$$
\begin{align*}
E_\nu &= C_\nu (30A + 2(3 + r)B_\nu + 1) + 9MA, \\
t_1 &= |2 - 2r - s|_+ - \alpha, \\
t_2 &= |1 - 2r - 2s - t_1|_+.
\end{align*}
$$

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

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

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

Let the *validation set*

$$
z^\nu = (z_1^\nu, \ldots, z_m^\nu),
$$

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

$$
z^\text{tot} = (\tilde{z}, z^\nu),
$$

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 $\tilde{z}$, and $m$ the number of its labelled examples.

---

1We say that the sequence of random variables $\{X_m\}_{m \in \mathbb{N}}$ converges in probability to the random variable $X$ (and we write $\lim_{m \to \infty} X_m = \rho X$ or $X_m \overset{P}{\to} X$), if for every $\epsilon > 0, \lim_{m \to \infty} P[|X_m - X| \geq \epsilon] = 0$. This is equivalent to say that, for every $\delta \in (0, 1)$, $P[|X_m - X| \geq \epsilon(m, \delta)] \leq \delta$, with $\lim_{m \to \infty} \epsilon(m, \delta) = 0$. 
Now let us explain how $\mathbf{z}^\ast$ is used for the choice of $\lambda$. We consider the finite set of positive reals $\Lambda_m$ depending on $m$, the number of labelled examples in $\mathbf{z}$, and the data-dependent choice for the regularization parameter is

$$
\hat{\lambda}_m = \arg\min_{\lambda \in \Lambda_m} \frac{1}{m^\nu} \sum_{i=1}^{m^\nu} (T_M f_{\hat{\lambda}_m}(x_i^\ast) - y_i^\ast)^2,
$$

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

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

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

$$
f_{\hat{\lambda}^{\text{tot}}} = T_M f_{\hat{\lambda}_m}.
$$

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

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

$$
\|f_{\hat{\lambda}_m} - f_{\rho}\|_\rho \leq \epsilon.
$$

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

$$
\|f_{\hat{\lambda}^{\text{tot}}} - f_{\rho}\|_\rho \leq \tilde{\epsilon},
$$

with

$$
\tilde{\epsilon}^2 = 2\epsilon^2 + \frac{80M^2}{m^\nu} \log \frac{2|\Lambda_m|}{\delta}.
$$

The proof of Theorem 3 is postponed to Section 5.

The first corollary of Theorem 3 proves universal consistency for the estimators $f_{\hat{\lambda}^{\text{tot}}}$ under mild assumptions on the cardinalities of the grids $\Lambda_m$ and validation sets $\mathbf{z}^\ast$.

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

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

with

$$
|\Lambda_m| = \omega(1).
$$

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

$$
\lim_{m \to \infty} \|f_{\hat{\lambda}^{\text{tot}}} - f_{\rho}\|_\rho \overset{P}{=} 0.
$$

**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 $\lambda_m = \kappa Q^{-\lfloor \log \log m \rfloor / |\Lambda_m|}$, which belongs to $\Lambda_m$, fulfills the assumptions (6) and (7). Hence, using the assumption on $m_m^\ast$, we get that for every $\delta \in (0, 1)$, with probability greater than $1 - 2\delta$

$$
\|f_{\hat{\lambda}^{\text{tot}}} - f_{\rho}\|_\rho^2 \leq o(1) + \frac{80M^2}{\omega(\log |\Lambda_m|)} \log \frac{2|\Lambda_m|}{\delta} \to 0.
$$

□
The second corollary proves explicit rates for the convergence of \( f^\text{tot}_z \) to \( f_\rho \) over specific prior classes defined in term of finiteness of the constants \( C_\rho \) and \( D_\rho \). The main assumption is the requirement \( m^* \geq m/\log m \). Since this constrain can be fulfilled still being \( m^* \) 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 \( z^\text{tot} \) with \( m^* \geq m \log m \) and \( \tilde{m} \geq 4 \sqrt{m^{1+\eta}} \), for some constants \( \eta \geq 0, r > 0, \) and \( s \in (0,1] \). Define \( \Lambda_m \) as in eq. (15) with \( Q \) an arbitrary constant greater than 1 and

\[
\frac{\eta}{\alpha} \log_Q m + 1 \leq |\Lambda_m| \leq m,
\]

with \( \alpha \) defined by eq. (20).

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

\[
Q \left( 4D_\rho \log \frac{6}{\sqrt{m}} \right)^{-\frac{2}{r-t_2}} \leq \kappa^{-1} ||T||.
\]

Then, with probability greater than 1 − 6\( \delta \)

\[
\|f^\text{tot}_z - f_\rho\|_\rho \leq 4(Q^\prime E_r D_\rho + 3M) \log(6m/\delta) \, m^{-\frac{1}{2}} \frac{2r-t_2}{2r+s+t_1},
\]

where \( E_r, t_1 \) and \( t_2 \) are the constants defined in equations (9), (10) and (11) substituting \( \alpha = |2-2r-s|_+ \) and \( \eta = \frac{|2-2r-s|_+}{2r+s+|2-2r-s|_+} \).

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

\[
2 \log_Q m + 1 \leq |\Lambda_m| \leq m,
\]

and

\[
Q \left( 4D_\rho \log \frac{6}{\sqrt{m}} \right)^{-\frac{2}{r-t_2}} \leq \kappa^{-1} ||T||,
\]

with probability greater than 1 − 6\( \delta \), it holds

\[
\|f^\text{tot}_z - f_\rho\|_\rho \leq 4(Q^\prime E_r D_\rho + 3M) \log(6m/\delta) \, m^{-\frac{1}{2}} \frac{2r-t_2}{2r+s+t_1}.
\]

**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

\[
\hat{\lambda}_q = q \left( 4D_\rho \log \frac{6}{\sqrt{m}} \right)^{-\frac{2}{r-t_2}} \leq \kappa^{-1} ||T|| \quad \forall q \in [1, Q],
\]

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

\[
\hat{m} \geq 4 \vee m^{1+\eta} = 4 \vee m \left( m^{-\frac{\eta}{2}} \right)^{-\alpha} \geq 4 \vee m\hat{\lambda}_q^{-\alpha},
\]

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

\[
\|f^\text{tot}_z - f_\rho\|_\rho \leq \epsilon = Q^\prime E_r \left( 4D_\rho \log \frac{6}{\sqrt{m}} \right)^{-\frac{2r-t_2}{2r+s+t_1}}.
\]
The next step is verifying that for some $\bar{q} \in [1, Q]$, $\lambda_{\bar{q}} = \kappa \hat{\lambda}_{\bar{q}} \in \Lambda_m$, and hence applying Theorem 3.

In fact, from definition (15), assumption (17) and Proposition 5, it is clear that
\[
\min \Lambda_m \leq \kappa m^{-\frac{2}{m}} \leq \lambda_1 \leq \lambda_{\bar{q}} \leq \lambda_Q \leq \|T\| \leq \kappa = \max \Lambda_m,
\]
for some $\bar{q}$.

Applying Theorem 3, we get that with probability greater than $1 - 6\delta$ it holds
\[
\|f^t_{x^*} - f_{x^*}\| \leq \hat{\epsilon},
\]
with, using again condition (17) and the assumption $m^v \geq m \log m$, the chain of inequalities
\[
\hat{\epsilon} = \left( \epsilon^2 + \frac{80M^2}{m^v} \log \frac{6|\Lambda_m|}{\delta} \right)^{\frac{1}{2}} \leq \left( \epsilon^2 + \frac{80M^2}{m} \log \frac{6m}{\delta} \right)^{\frac{1}{2}} \leq 4Q'E_{r}D_{s} \log(6/\delta) m^{-\frac{1}{2}} \frac{2r-s}{2r+s} + 12M \log(6m/\delta) m^{-\frac{1}{2}} \leq 4Q'E_{r}D_{s} + 3M) \log(6m/\delta) m^{-\frac{1}{2}} \frac{2r-s}{2r+s} + 12M \log(6m/\delta) m^{-\frac{1}{2}} \frac{2r-s}{2r+s} + 12M \sqrt{m \log 2m},
\]
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 $\alpha = \frac{|2-2r-s|}{2r+s}$ 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 $\frac{2r-s}{2r+s} \leq 2$, 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 4.** 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
\[
\lim_{i \to \infty} a_i = 0,
\lim_{i \to \infty} b_i = 0.
\]

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
\[
c_i \geq a_i \quad \forall i \in \mathbb{N},
\lim_{i \to \infty} c_i = 0.
\]

**Proof.** We consider the sequence $\{c_i\}_{i \in \mathbb{N}}$ of positive numbers constructed by the recursive rule
\[
c_1 = a_1,
\]
\[
c_{i+1} = a_{i+1} \lor (b_{i+1})^{\log c_i / \log b_i}.
\]

Let us prove point i) by induction.
Since by assumption $a_1$ and $b_1$ belong to $(0, 1)$, by construction $d_1 = \frac{\log c_1}{\log b_1} = \frac{\log a_1}{\log b_1} > 0$.

Now, for $i \geq 1$ assume $d_i > 0$, then by construction, either $c_{i+1} = (b_{i+1})^{d_i}$, and hence $d_{i+1} = \frac{\log c_{i+1}}{\log b_{i+1}} = \frac{\log a_{i+1}}{\log b_{i+1}} > 0$, 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

$$d_{i+1} = \frac{\log a_{i+1}}{\log b_{i+1}} > 0,$$

$$(b_{i+1})^{d_i+1} \geq (b_{i+1})^{d_i} \Rightarrow d_{i+1} \leq d_i.$$

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

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

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

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

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

$$c_i = (b_i)^{d_i}, \quad \forall i > I.$$

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

The next proposition introduces the functions $f_\lambda^a$ and shows some simple results related to them.

**Proposition 2.** For any $\lambda > 0$ let the truncated function $f_\lambda^a$ be defined by

\begin{equation}
(22)
\quad f_\lambda^a = P_\lambda f_H
\end{equation}

where $P_\lambda$ is the orthogonal projector in $L^2(X, \rho_X)$ defined by

\begin{equation}
(23)
\quad P_\lambda = \Theta_\lambda(LK),
\end{equation}

with

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

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

\begin{equation}
(25)
\quad a(\lambda) = \| f_\lambda^a - f_H \|_\mu,
\end{equation}

is non-decreasing and fulfills the following properties

\begin{equation}
(26)
\quad 0 \leq a(\lambda) \leq M \quad \forall \lambda \in (0, \kappa],
\end{equation}

\begin{equation}
(27)
\quad \lim_{\lambda \to 0} a(\lambda) = 0.
\end{equation}
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 $L_K^2$ is an isometry between $L^2(X, \rho_X)$ and $\mathcal{H}$ (again, see [5]). Therefore, since $f_\mathcal{H}$ is the projection of $f_\rho$ over the closure of $\mathcal{H}$ in $L^2(X, \rho_X)$, it holds

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

Hence, by the definition of $f_\mathcal{H}$, 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. \(\square\)

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

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

\[
(29) \quad \lim_{\lambda \to 0} \lambda^{R(\lambda)} = 0.
\]

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

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

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

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

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

\[
\lim_{\lambda \to 0} \lambda^{R(\lambda)} = \lim_{i \to \infty} \lambda_i^{R(\lambda_i)} = \lim_{i \to \infty} \left( \frac{2h_i}{d_i} \right)^{x/(\bar{r} \vee d_i)} \leq 2^x \left( \lim_{i \to \infty} c_i \right)^{x/(\bar{r} \vee d_i)} = 0.
\]
Since by Proposition 1 \( \{d_i\}_i \) is a sequence of non-increasing positives, then \( R \) is non-decreasing. Therefore, defining \( f_i = (f_{H, \phi_i})_\rho \), we can write
\[
\kappa^{2R(\lambda)} \| L_K^{R(\lambda)} P_\lambda f_H \|_\rho^2 = \sum_{\lambda_i \geq \lambda} f_i^2 \lambda_i^{-2R(\lambda)} \leq \sum_{i} f_i^2 \lambda_i^{-2R(\lambda)} = \sum_{i} f_i^2 \left( \frac{2b_i^i}{c_i} \right)^{\frac{1}{\rho (\rho + d_i)}} \leq \sum_{i} f_i^2 c_i^{-1} (c_i > a_i) \leq \sum_{i} f_i^2 a_i^{-1} = 2M \sum_{i} f_i^2 a(\lambda_i)^{-1} \leq 2M \sum_{k=0}^{\infty} \sum_{2^{-k-1} < a(\lambda_i)/M \leq 2^{-k}} f_i^2 a(\lambda_i)^{-1} \leq 4M^2 \sum_{k=0}^{\infty} 2^{-k} = 8M^2,
\]
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 \( f_{\text{ls}} \tilde{z}, \lambda \) and \( f_{\text{ls}} \lambda \).

**Proposition 4.** Assume \( \lambda \leq \|T\| \) and
\[
\lambda \tilde{n} \geq 16\kappa N(\lambda) \log \frac{6}{\delta}, \tag{30}
\]
for some \( \delta \in (0, 1) \). Then, with probability greater than \( 1 - \delta \), it holds
\[
\| (T + \lambda)^{-1} (f_{\text{ls}} \tilde{z}, \lambda - f_{\text{ls}} \lambda) \|_H \leq 8 \left( M + \sqrt{\frac{m}{m}} \| f_{\text{ls}} \lambda \|_H \right) \left( \frac{2}{m} \sqrt{\frac{\kappa}{\lambda}} + \sqrt{\frac{N(\lambda)}{m}} \right) \log \frac{6}{\delta}
\]
where
\[
f_{\text{ls}} \tilde{z}, \lambda = (T_{\tilde{z}, \lambda})^{-1} \tilde{g}_{\tilde{z}}, \tag{31}
\]
\[
f_{\text{ls}} \lambda = (T + \lambda)^{-1} L_K f_H. \tag{32}
\]

**Proof.** See Proposition 1 in [2].

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

**Proposition 5.** For every probability measure \( \rho_X \) and \( \lambda > 0 \), it holds
\[
\| T \| \leq \kappa,
\]
and
\[
\lambda N(\lambda) \leq \kappa.
\]

**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 \( [x] \) in the text below represents the greater integer less or equal to \( x \).
Proposition 6. Let $f$ belong to $\text{Im} L^r_K$ for some $r > 0$. Then, if $\lambda \in (0, \kappa]$, it holds
\[ \| x_{d/\sqrt{T}}(G_{\lambda}(T_{z}) T_{z} - \ldots \text{following decomposition}, \| f_{z,\lambda} - f_{H} \|_{\rho} \leq D(\tilde{z}, \lambda) \]

where $P_{\lambda}$ is defined in eq. (23), and
\[ (33) \quad \gamma = \lambda^{-1} \| T - T_{k} \|, \]
\[ \eta = |r - \frac{1}{2}| - |r - \frac{1}{2}|. \]

Proof. See Proposition 6 in [2].

Proposition 7. Let the operator $\Omega_{\lambda}$ be defined by
\[ (34) \quad \Omega_{\lambda} = \sqrt{T} G_{\lambda}(T_{k})(T + \lambda)^{-\frac{1}{2}}. \]

Then, if $\lambda \in (0, \kappa]$, it holds
\[ \| \Omega_{\lambda} \| \leq (1 + 2\sqrt{T}) A, \]
with $\gamma$ 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 $\xi$ be a random variable on $\Omega$ taking value in a real separable Hilbert space $K$. Assume that there are two positive constants $H$ and $\sigma$ such that
\[ \| \xi(\omega) \|_{K} \leq \frac{H}{2} \quad \text{a.s.,} \]
\[ \mathbb{E}[\| \xi(\omega) \|_{K}^{2}] \leq \sigma^{2}, \]
then, for all $m \in \mathbb{N}$ and $0 < \delta < 1$,
\[ (35) \quad P(\omega_{1}, ..., \omega_{m}) = P \left[ \left\| \frac{1}{m} \sum_{i=1}^{m} \xi(\omega_{i}) - E[\xi] \right\|_{K} \leq 2 \left( \frac{H}{m} + \frac{\sigma}{\sqrt{m}} \right) \log \frac{2}{\delta} \geq 1 - \delta. \]

We are now ready to prove Theorem 1.

Proof of Theorem 1. Let us consider the expansion
\[ \sqrt{T}(f_{k,\lambda} - f_{H}) = \sqrt{T}(G_{\lambda}(T_{k}) g_{\omega} - f_{0}^{\omega}) + \sqrt{T}(f_{1}^{\omega} - f_{H}) = \Omega_{\lambda} \left( (T + \lambda)^{\frac{1}{2}}(f_{0,\lambda}^{\omega} - f_{0}^{\omega,\lambda}) + \sqrt{T}(G_{\lambda}(T_{k}) T_{k} - \text{Id}) f_{1}^{\omega} + \sqrt{T}(f_{1}^{\omega} - f_{H}) \right) = \Omega_{\lambda} \left( (T + \lambda)^{\frac{1}{2}}(j_{0,\lambda}^{\omega} - f_{0,\lambda}^{\omega}) + (T + \lambda)^{\frac{1}{2}}(j_{1}^{\omega} - f_{1}^{\omega,\lambda}) \right) + \sqrt{T}(G_{\lambda}(T_{k}) T_{k} - \text{Id}) f_{1}^{\omega} + \sqrt{T}(f_{1}^{\omega} - f_{H}) \]

where the operator $\Omega_{\lambda}$ is defined by equation (34), the ideal RLS estimators are $f_{0}^{\omega} = (T + \lambda)^{-1} T_{k} f_{H}$ and $f_{0}^{\omega,\lambda} = (T + \lambda)^{-1} T f_{k,\lambda}^{\omega}$, and $f_{1}^{\omega,\lambda} = (T_{k} + \lambda)^{-1} T_{k} f_{1}^{\omega}$ is the RLS estimator constructed by the training set
\[ \tilde{z}' = ((\tilde{z}_{1}, f_{1}^{\omega}((\tilde{z}_{1})), \ldots, (\tilde{z}_{m}, f_{1}^{\omega}((\tilde{z}_{m})))), \]
Hence we get the following decomposition,
\[ (36) \quad \| f_{k,\lambda} - f_{H} \|_{\rho} \leq D(\tilde{z}, \lambda) \left( S^{\omega}(\tilde{z}, \lambda) + R(\lambda) + S^{\omega}(\tilde{z}, \lambda) \right) + P(\tilde{z}, \lambda) + P^{\omega}(\lambda), \]
with
\[
S^{ls}(\tilde{z}, \lambda) = \left\| (T + \lambda)^{1/2} (f^{ls}_{\lambda, \tilde{z}} - f^{ls}_\lambda) \right\|_{\mathcal{H}},
\]
\[
\bar{S}^{ls}(\tilde{z}, \lambda) = \left\| (T + \lambda)^{1/2} (f^{ls}_{\lambda, \tilde{z}} - f^{ls}_\lambda) \right\|_{\mathcal{H}},
\]
\[
D(\tilde{z}, \lambda) = \left\| \Omega \lambda \right\|,
\]
\[
P(\tilde{z}, \lambda) = \left\| \sqrt{T} (G_{\lambda} (T_{\lambda}) T_{\lambda} - \text{Id}) f^{ls}_\lambda \right\|_{\mathcal{H}},
\]
\[
P^{tr}(\lambda) = \left\| f^{ls}_\lambda - f^{ls}_\rho \right\|_{\rho},
\]
\[
R(\lambda) = \left\| (T + \lambda)^{1/2} (f^{ls}_\rho - f^{ls}_\lambda) \right\|_{\mathcal{H}}.
\]

Terms \( S^{ls} \) and \( \bar{S}^{ls} \) 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 \( L^2_{\Omega} \) is an isometry between \( L^2(X, \rho_X) \) and \( \mathcal{H} \) (see [5]), we obtain
\[
\left\| f^{ls}_\lambda \right\| \leq \left\| \sqrt{T} (T + \lambda)^{-1} \sqrt{T} f^{ls}_\lambda \right\|_{\mathcal{H}} \leq \frac{\left\| f^{ls}_\lambda \right\|_{\sqrt{X}}}{\sqrt{\lambda}} \leq \frac{M}{\sqrt{\lambda}}.
\]

Now, let \( \delta \) be an arbitrary real in \((0, 1)\). From the assumptions on \( \lambda_m \), for large enough \( m \), we have
\[
\lambda_m \sqrt{m} \geq 4\kappa \log \frac{6}{\delta},
\]
\[
\lambda_m \leq \left\| T \right\|.
\]

Hence, by Proposition 5, for large enough \( m \), the assumptions of Proposition 4 are verified, and we get that with probability greater than 1 − \( \delta \)
\[
S^{ls}(\tilde{z}_m, \lambda_m) \leq 8 \left( M + \sqrt{\frac{m}{m}} \frac{\left\| f^{ls}_{\lambda_m} \right\|_{\mathcal{H}}}{\left\| f^{ls}_\lambda \right\|_{\mathcal{H}}} \right) \left( \frac{2}{m} \sqrt{\frac{\kappa}{\lambda_m}} + \sqrt{\frac{N(\lambda_m)}{m}} \right) \log \frac{6}{\delta}
\]
\[
(\text{Prop. 5, eq. (38)}) \leq 8M \left( 1 + \sqrt{\frac{\kappa}{\lambda_m}} \right) \frac{1}{\sqrt{m}} \left( 2 \sqrt{\frac{\kappa}{\lambda_m m}} + \sqrt{\frac{\kappa}{\lambda_m}} \right) \log \frac{6}{\delta}
\]
\[
(\lambda_m \leq \kappa, m \geq 4) \leq \frac{32M\kappa}{\lambda_m \sqrt{m}} \log \frac{6}{\delta} \rightarrow 0.
\]

Hence it holds
\[
\lim_{m \to \infty} S^{ls}(\tilde{z}_m, \lambda_m) = 0.
\]

**Step 2:** Estimate of \( \bar{S}^{ls} \). This term can be estimated observing that \( \tilde{z} \) is a training set of \( m \) supervised samples drawn i.i.d. from the probability measure \( \rho \) with marginal \( \rho_X \) and conditional \( \rho_{\rho}(y) = \delta (y - f^{ls}(x)) \). Therefore the regression function induced by \( \rho \) is \( f_\rho = f^{ls}_\lambda \), and the support of \( \rho \) is included in \( X \times [-M', M'] \), with \( M' = \sup_{x \in X} f_\rho(x) \leq \sqrt{\pi} \left\| f^{ls}_\lambda \right\|_{\mathcal{H}} \). 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
\[
\bar{S}^{ls}(\tilde{z}_m, \lambda_m) \leq 8 \left( M' + \sqrt{\kappa} \left\| f^{ls}_{\lambda_m} \right\|_{\mathcal{H}} \right) \left( \frac{2}{m} \sqrt{\frac{\kappa}{\lambda_m}} + \sqrt{\frac{N(\lambda_m)}{m}} \right) \log \frac{6}{\delta}
\]
\[
(\text{Prop. 5}) \leq 16\sqrt{\kappa} \left\| f^{ls}_{\lambda_m} \right\|_{\mathcal{H}} \frac{1}{\sqrt{m}} \left( 2 \sqrt{\frac{\kappa}{\lambda_m m}} + \sqrt{\frac{\kappa}{\lambda_m}} \right) \log \frac{6}{\delta}
\]
\[
(m \geq 4) \leq \frac{32\kappa}{\lambda_m \sqrt{m}} \left\| P_{\lambda_m} L^{1/2} P_{\lambda_m} f^{ls}_\lambda \right\|_{\rho} \log \frac{6}{\delta} \leq \frac{32\kappa M}{\lambda_m \sqrt{m}} \log \frac{6}{\delta} \rightarrow 0.
\]

Hence it holds
\[
\lim_{m \to \infty} \bar{S}^{ls}(\tilde{z}_m, \lambda_m) = 0.
\]
**Step 3:** Estimate of $P^x$. By definition (25), $P^x(\lambda) = a(\lambda)$. Hence from eq. (27)
\[
\lim_{m \to \infty} P^x(\lambda_m) = \lim_{m \to \infty} a(\lambda_m) = \lim_{\lambda \to 0} a(\lambda) = 0,
\]
where we used the assumption (6).

**Step 4:** Estimate of $R$. Since from the definitions of $f^x$ and $\tilde{f}^x$,\[
R(\lambda) = \left\| (T + \lambda)^{-\frac{1}{2}} T (\tilde{f}^x - f^x) \right\|_\mathcal{H} \leq \left\| \sqrt{T} (f^x - f^x) \right\|_\mathcal{H} \leq P^x(\lambda),
\]
from (41) we get
\[
\lim_{m \to \infty} R(\lambda_m) = 0.
\]

**Step 5:** Estimate of $D$. In order to estimate $D(\bar{z}, \lambda)$, we have first to estimate the quantity $\gamma = \gamma(\bar{z}, \lambda)$ (see definition (33)) appearing in the Proposition 7. Our estimate for $\gamma(\bar{z}, \lambda)$ follows from Proposition 8 applied to the random variable $\xi : X \to L_{HS}(\mathcal{H})$ defined by
\[
\gamma(\bar{z}, \lambda) = \lambda^{-1} K_{\xi, \gamma} (K_{\xi, \gamma})_\mathcal{H}.
\]
We can set $H = \frac{2\gamma}{\lambda}$ and $\sigma = \frac{2\gamma}{\lambda}$, and obtain that, for every $\delta \in (0, 1)$ and $m \geq 4$, with probability greater than $1 - \delta$
\[
\gamma(\bar{z}_m, \lambda) \leq \lambda^{-1} \| T - T_k \|_{HS} \leq 2 \left( \frac{2\gamma}{\lambda m} + \frac{\gamma}{\sqrt{m}} \right) \log 2 \frac{2\gamma}{\delta} \leq 4 \epsilon(\lambda, \delta).
\]
From the expression of $\epsilon(m, \lambda, \delta)$ we see that, by the assumption (7), for every $\delta \in (0, 1)$,
\[
\lim_{m \to \infty} \epsilon(m, \lambda, \delta) = 0,
\]
and hence,
\[
\lim_{m \to \infty} \gamma(\bar{z}_m, \lambda_m) = 0.
\]
Finally, from eq. (43) and Proposition 7 we find
\[
D(\bar{z}_m, \lambda_m) \leq \left( 1 + 2 \sqrt{\gamma(\bar{z}_m, \lambda_m)} \right) A \overset{P}{=} 3A.
\]

**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 $r_m = R(\lambda^{-1} \lambda_m) \leq \bar{r}$, getting
\[
P(\bar{z}_m, \lambda_m) \leq B(1 + \gamma(\bar{z}_m, \lambda_m)^{\frac{1}{2}})
\begin{align*}
(2 + r_m \gamma(\bar{z}_m, \lambda_m) + \gamma(\bar{z}_m, \lambda_m)^{\frac{1}{2}} m) & \kappa^{r_m} \left\| L_{\kappa}^{r_m} P_{\lambda_m} f_\mathcal{H} \right\|_\mathcal{H} \leq 4M \\
& \forall m,
\end{align*}
\]
This result together with eq. (43), and recalling that by Proposition 3 and assumption (6), the sequence $\{r_m\}_m$ verifies the two conditions
\[
\kappa^{r_m} \left\| L_{\kappa}^{r_m} P_{\lambda_m} f_\mathcal{H} \right\|_\mathcal{H} \leq 4M \quad \forall m,
\]
proves that
\[
\lim_{m \to \infty} P(\bar{z}_m, \lambda_m) = 0.
\]
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 $f^{\lambda}_N$ and $f^{\lambda}_H$ (defined in eq. (22) and eq. (32) respectively) when $f^N \in \text{Im } L^N_K$.

**Proposition 9.** Let $f^N_\lambda \in \text{Im } L^N_K$ for some $r > 0$. Then, the following estimates hold,

\[
\|f^N_\lambda - f^N\|_\rho \leq \lambda^r \|L^N_K f^N_\lambda\|_\rho,
\]

\[
\|f^N_\lambda\|_H \leq \begin{cases} \lambda^{-\frac{1}{2}+r} \|L^N_K f^N_\lambda\|_\rho & \text{if } r \leq \frac{1}{2}, \\ \kappa^{-\frac{1}{2}+r} \|L^N_K f^N_\lambda\|_\rho & \text{if } r > \frac{1}{2}. \end{cases}
\]

**Proof.** See Proposition 3 in [2]. \[ \square \]

**Proposition 10.** Let $f^N_\lambda \in \text{Im } L^N_K$ for some $r > 0$. Then, the following estimates hold,

\[
\|f^N_\lambda - f^N\|_\rho \leq \lambda^r \|L^N_K f^N\|_\rho, \quad \text{if } r \leq 1
\]

\[
\|f^N_\lambda\|_H \leq \begin{cases} \lambda^{-\frac{1}{2}+r} \|L^N_K f^N\|_\rho & \text{if } r \leq \frac{1}{2}, \\ \kappa^{-\frac{1}{2}+r} \|L^N_K f^N\|_\rho & \text{if } r > \frac{1}{2}. \end{cases}
\]

**Proof.** See Proposition 1 in [2]. \[ \square \]

We are now ready to prove Theorem 2.

**Proof of Theorem 2.** We consider the same decomposition (see equations (36) and (37)) for $\|f^N_\lambda - f^N\|_\rho$ that we used in the proof of Theorem 1.

Terms $S^{a}\lambda$ and $\tilde{S}^{a}\lambda$ will be estimated by Proposition 4, term $D$ by Proposition 7, term $P$ by Proposition 6 and finally terms $P^t$ and $R$ by Proposition 9.

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

\[
\hat{\lambda} \leq \kappa^{-1} \|T\| \leq 1,
\]

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

\[
\hat{\lambda}\hat{m} \geq \hat{\lambda}^{-|1-2r|_++1+t_1} m \geq \hat{\lambda}^{2r+t_1} m.
\]

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

\[
\hat{\lambda}^{2r+t_1} m = 16q^{2r+t_1} D^2_\hat{\lambda}^{-s} \log^2 \frac{6\hat{\lambda}^{-t}}{\delta} \geq 16N(\lambda) \log^2 \frac{6\hat{\lambda}^{-t}}{\delta},
\]

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

Regarding the estimate of $\tilde{S}^{a}\lambda$. Applying Proposition 4 and recalling that by assumption $\hat{m} \geq m\hat{\lambda}^{-|2-2r-s|_++t_1} \geq m\hat{\lambda}^{-|1-2r|_++t_1}$ and from Proposition 10, $\sqrt{\kappa} \|f^N_\lambda\|_H \leq C_\tau \hat{\lambda}^{-\frac{1}{2}+r} r$, we get that with probability greater than $1 - \delta$

\[
(46) \quad \tilde{S}^{a}\hat{\lambda}(\tilde{\alpha}, \lambda) \leq 8 \left(M + \frac{m}{\sqrt{m}} C_\tau \hat{\lambda}^{-\frac{1}{2}+r} r\right) \left(2 \frac{\kappa}{\sqrt{\lambda}} + \sqrt{\frac{N(\lambda)}{m}} \right) \log^2 \frac{6\hat{\lambda}^{-t}}{\delta}
\]

\[
\leq 8(M + \hat{\lambda}^{-\frac{1}{2}+r} C_\tau) \frac{1}{\sqrt{m}} \left(2 \frac{\kappa}{\sqrt{m\lambda}} + \frac{D_\tilde{s}}{\sqrt{\lambda}^t} \right) \log^2 \frac{6\hat{\lambda}^{-t}}{\delta}
\]

(eq. (8))

\[
= 2q^{-r-\frac{1}{2}} \left(M + C_\tau\right) \hat{\lambda}^{-\frac{1}{2}} \left(1 + \frac{\hat{\lambda}^{-\frac{1}{2}+r} (1-2r-2s-t_1)}{2q^{r-\frac{1}{2}+\frac{1}{2}+t_1} D^2_\lambda \log^2 \frac{6\hat{\lambda}^{-t}}{\delta}}\right)
\]

\[
(t_1 \geq 0, q \geq 1) \leq 3q^{-r-\frac{1}{2}} \left(M + C_\tau\right) \hat{\lambda}^{-\frac{1}{2}}
\]

\[
(q \geq 1) \leq 3(M + C_\tau) \hat{\lambda}^{-\frac{1}{2}+r}.
\]
The term \( Sls \) can be estimated observing that \( z' \) is a training set of \( \tilde{m} \) supervised samples drawn i.i.d. from the probability measure the support of \( \gamma \). Hence, since \( |z'| \leq \text{tr} \gamma f' \), we get

\[
\gamma(z, \lambda) \leq 2q^{-r} - \frac{r}{2} C_r \lambda^{-\frac{3}{2}} \left( 1 + \frac{\lambda^{-\frac{1}{2}} (1 - 2r - 2s - t_1)}{2q^{r/2} + \frac{r}{2} D_r \log \frac{q}{\lambda}} \right)
\]

(Prop. 9)

\[
(\xi(x)[]) = \lambda^{-1} K_{x, \cdot} (K_{x, \cdot})_{\mathcal{H}}.
\]

We can set \( H = \frac{2k}{m} \) and \( \sigma = \frac{r}{2} \), and obtain that with probability greater than \( 1 - \delta \)

\[
\gamma(z, \lambda) \leq \lambda^{-1} \|T - T_s\|_{\mathcal{H}} \leq \frac{2k}{m} + \frac{\kappa}{\sqrt{m}} \log \frac{2}{\delta} \leq 4 \lambda^{-1} \frac{\lambda^{-\frac{3}{2}}}{\lambda^{-1}} \leq 4 \lambda^{-1} - \frac{1}{2} \lambda^{-\frac{3}{2}} \leq 4 \lambda^{-1} - \frac{1}{2} \lambda^{-\frac{3}{2}} \leq 1,
\]

where we used the assumption \( m \geq 4 \sqrt{m} \lambda^{-2r - 2s - 1} + t_1 \) and the expression for \( \lambda \) in the text of the Theorem.

Hence, since \( \gamma(z, \lambda) \leq \lambda^{r + \frac{1}{2}} - 1 \), from Proposition 7 we get

\[
D(z, \lambda) = (1 + 2\sqrt{\mathcal{L}}) \lambda \leq 3A,
\]

and from Proposition 6

\[
P(z, \lambda) \leq B_r C_r (1 + \sqrt{\mathcal{L}}) (2 + r\gamma \lambda^{\frac{3}{2} - r} + \gamma^r) \lambda^r
\]

(49)

\[
P(z, \lambda) \leq 2B_r C_r (3 + r\gamma \lambda^{\frac{3}{2} - r}) \lambda^r
\]

\[
P(z, \lambda) \leq 2B_r C_r (3 + \gamma^r) \lambda^r \leq 2B_r C_r (3 + r) \lambda^r.
\]

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

\[
P^{tr} (\lambda) \leq C_r \lambda^r,
\]

In order to get an upper bound for \( D \) and \( P \), we have first to estimate the quantity \( \gamma(z, \lambda) \) (see definition (33)) appearing in the Proposition 6 and 7. Our estimate for \( \gamma(z, \lambda) \) follows from Proposition 8 applied to the random variable \( \xi : X \rightarrow \mathcal{L}_{\mathcal{H}}(\mathcal{H}) \) defined by

\[
(47)
\]
and hence,
\[
R(\lambda) = \left\| \frac{1}{\sqrt{T}} (\hat{\lambda} - \overline{\lambda}) \right\|_H \leq \sqrt{T} (\overline{\lambda} - \lambda) \leq C_r \lambda^r.
\]

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

\[\square\]

5. Proof of Theorem 3

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

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

\[
P \left[ \frac{1}{n} \sum_{i=1}^n X_i - \mu \geq \alpha \sigma^2 + \epsilon \right] \leq e^{-\frac{\alpha n \sigma^2 + \epsilon}{\mathbb{E}[\sum_{i=1}^n (X_i - \mu)]}},
\]

and

\[
P \left[ \mu - \frac{1}{n} \sum_{i=1}^n X_i \geq \alpha \sigma^2 + \epsilon \right] \leq e^{-\frac{\alpha n \sigma^2 + \epsilon}{\mathbb{E}[\sum_{i=1}^n (X_i - \mu)]}}.
\]

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

\[
P \left[ \frac{1}{n} \sum_{i=1}^n X_i - \mu \geq \alpha \sigma^2 + \epsilon \right] = \mathbb{P} \left[ e^{\epsilon \sum_{i=1}^n (X_i - \mu)} \right] \leq e^{-s \alpha \sigma^2 + \epsilon} \mathbb{E} [ e^{\epsilon \sum_{i=1}^n (X_i - \mu)} ] = e^{-s \alpha \sigma^2} \mathbb{E} e^{\epsilon \sum_{i=1}^n (X_i - \mu)},
\]

by Markov inequality

\[
e^{-s \alpha \sigma^2} \prod_{i=1}^n \mathbb{E} e^{\epsilon (X_i - \mu)}, \quad \text{by independence of } X_i.
\]

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

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

whence

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

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

We are now ready to prove Theorem 3.

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

\[
\lambda_m^* = \arg \min_{\lambda \in \Lambda_m} \int_Z (T_M f_{\lambda, \lambda}(x) - y)^2 dp.
\]

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

\[
\int_Z (f(x) - y)^2 dp = \|f - f_0\|_p^2 + \int_Z (f_\lambda(x) - y)^2 dp.
\]
\( \lambda_m^* = \arg \min_{\lambda \in \Lambda_m} \|T_M f_{\lambda, \lambda} - f_{\rho}\|_\rho \).

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
\[
\|T_M f_{\lambda, \lambda} - f_{\rho}\|_\rho \leq \|T_M f_{\lambda, \lambda} - f_{\rho}\|_\rho \leq \|f_{\lambda, \lambda} - f_{\rho}\|_\rho \leq \epsilon.
\]

We claim that for every \( \bar{\lambda}, \lambda > 0 \) and \( \delta \in (0, 1) \), with probability greater than 1 − \( \delta \) over the probability measure \( \rho^m \), it holds
\[
\|f_{\lambda}^{\star \rho} - f_{\rho}\|_\rho^2 \leq 2 \|T_M f_{\lambda, \lambda} - f_{\rho}\|_\rho^2 + \frac{80M^2}{m^\alpha} \log \frac{2|\Lambda_m|}{\delta}.
\]

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

We now proceed to proving eq. (56). For \( i = 1, \ldots, m^\alpha \), let us define the random variables
\[
\xi_i^\lambda = (T_M f_{\lambda, \lambda}(x_i^\lambda) - y_i^\lambda)^2 - (f_{\rho}(x_i^\lambda) - y_i^\lambda)^2.
\]

Clearly
\[
\text{E}[\xi_i^\lambda]\leq 4M^2, \quad \text{E}[(\xi_i^\lambda)^2] \leq 16M^2 \text{E}[\xi_i^\lambda].
\]

Hence, using Proposition 11 with \( X_i = \xi_i^0, \mu = \text{E}[\xi_i^0], B = 4M^2 \) and \( \sigma^2 = \text{E}[(\xi_i^0)^2] \leq 16M^2 \mu \), we obtain that for all \( \lambda \in \Lambda_m \) with probability greater than 1 − \( \delta \),
\[
\frac{1}{m^\alpha} \sum_{i=1}^{m^\alpha} \xi_i^\lambda \leq (1 + \alpha') \text{E}[\xi_i^0] + \epsilon,
\]
and
\[
\text{E}[\xi_i^\lambda] \leq \frac{1}{1 - \alpha'} \left( \frac{1}{m^\alpha} \sum_{i=1}^{m^\alpha} \xi_i^\lambda \right) \leq \frac{1}{1 - \alpha'} \text{E}[\xi_i^0] + \epsilon,
\]
where \( \alpha' = 16\alpha M^2 \) and \( \epsilon = \frac{3 + \alpha'}{60m^\alpha} \log \frac{2|\Lambda_m|}{\delta} \). Therefore
\[
\|f_{\lambda}^{\star \rho} - f_{\rho}\|_\rho^2 = \text{E}[\xi_i^{\lambda \rho}] \leq \frac{1}{1 - \alpha'} \left( \frac{1}{m^\alpha} \sum_{i=1}^{m^\alpha} \xi_i^{\lambda \rho} \right) + \frac{\epsilon}{1 - \alpha'}
\]
\[
\leq \frac{1 + \alpha'}{1 - \alpha'} \text{E}[\xi_i^{\lambda \rho}] + \frac{2\epsilon}{1 - \alpha'}.
\]

Setting \( \alpha = 1/(48M^2) \), this gives \( \alpha' = 1/3 \) and
\[
\|f_{\lambda}^{\star \rho} - f_{\rho}\|_\rho^2 \leq 2 \|T_M f_{\lambda, \lambda} - f_{\rho}\|_\rho^2 + \frac{80M^2}{m^\alpha} \log \frac{2|\Lambda_m|}{\delta},
\]
which proves eq. (56), as desired.\[ \square \]
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
