Optimal rates for regularization operators
in learning theory

Andrea Caponnetto \textsuperscript{a b c}

\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
Abstract

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

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


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

We consider the setting of semi-supervised statistical learning. We assume \( Y \subset [-M, M] \) and the supervised part of the training set equal to \( z = (z_1, \ldots, z_m) \), with \( z_i = (x_i, y_i) \) drawn i.i.d. according to the probability measure \( \rho \) over \( Z = X \times Y \). Moreover consider the unsupervised part of the training set \( (x_{m+1}, \ldots, x_n) \), with \( x_i^u \) drawn i.i.d. according to the marginal probability measure over \( X, \rho_X \). For sake of brevity we will also introduce the complete training set \( \tilde{z} = (\tilde{z}_1, \ldots, \tilde{z}_m) \), 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

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

and

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

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

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

\[
f_{\lambda} = G_{\lambda}(T_{\lambda}) g_{s},
\]

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

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

\( g_s \in \mathcal{H} \) is given by,

\[
g_s = \frac{1}{m} \sum_{i=1}^{m} K_{\tilde{x}_i} \tilde{y}_i = \frac{1}{m} \sum_{i=1}^{m} K_{x_i} y_i,
\]

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

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

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

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

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 [12]). Regularization methods have been recently studied in the context of learning theory in \([13, 9, 8, 10, 1]\).

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

The convergence rates in Theorems 1 and 2, will be described in terms of the constants \( C_r \) and \( D_{s} \) in \([0, +\infty[\) characterizing the probability measure \( \rho \). These constants can be
described in terms of the integral operator $L_K : \mathcal{L}^2(X, \rho_X) \rightarrow \mathcal{L}^2(X, \rho_X)$ 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 $\rho_{\perp x}$ through $f_{\mathcal{H}}$, they are defined as follows

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

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

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

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

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

The paper is organized as follows. In Section 2 we focus on the RLS estimators $f_{\mathcal{H},\lambda}^\delta$, defined by the optimization problem

$$f_{\mathcal{H},\lambda}^\delta = \arg\min_{f \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^m (f(\bar{x}_i) - \hat{y}_i)^2 + \lambda \| f \|_K^2,$$

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

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

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

2. Risk bounds for RLS.

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

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

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

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

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

$$\| f_{\mathcal{H},\lambda}^\delta - f_{\mathcal{H}} \|_B \leq 4(M + C_r) \left( \frac{4D_s \log \frac{2}{\delta}}{\sqrt{m}} \right)^{\frac{2}{2+s}}.$$
Let the eigenvalues of order is optimal over the class of probability measures in the minimax sense of \([11, 4]\). Indeed, in Th. 2 of \([4]\), it was showed that this asymptotic rate of convergence can be achieved by the RLS estimator, for finiteness of the constants (see Prop. 3 in \([4]\)).

Theorem 2. We adopt the same notations and definitions introduced in the previous section.

We need regularization algorithms of type described by equation (1). In this general framework, the unlabelled examples (enforcing the assumption \(\tilde{m} \geq m\lambda^{-1-2r} \geq 0\)) can be obtained observing that \(D_1 = 1\), for every kernel \(K\) and marginal distribution \(\mu\) (see Prop. 2).

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

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

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

\[
\|f_{\tilde{m}, \lambda} - f_\mu\|_\rho \leq 4(M + C_\rho) \left( \frac{4\log \frac{6}{\delta}}{\sqrt{m}} \right)^{\frac{2}{2r+1}}.
\]

3. Risk bounds for general regularization methods.

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

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

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

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

\[
\|f_{\tilde{m}, \lambda} - f_\mu\|_\rho \leq E_r \left( \frac{4D_s \log \frac{6}{\delta}}{\sqrt{m}} \right)^{\frac{2}{2r+1}}.
\]
where

\[ E_r = C_r (30A + 2(3 + r)B_r + 1) + 9MA. \]  

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

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

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

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

**Corollary 2.** Let \( \bar{m} \geq 4 \vee m \lambda^{-1[1-2\nu]} \) hold with \( r > 0 \). If \( \lambda \) satisfies the constraints

\[ \lambda \leq \|T\| \text{ and } \lambda \geq \frac{\log m}{\sqrt{m}}, \]

for some \( \delta \in (0, \frac{1}{2}) \), then, with probability greater than \( 1 - 3 \delta \), it holds

\[ \left\| f_{\lambda,\lambda}^{ls} - f_H \right\|_p \leq E_r \left( \frac{4 \log m}{\sqrt{m}} \right)^{\frac{2r}{2r+1}}, \]

with \( E_r \) defined by eq. (8).

4. **Proof of Theorem 1**

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

**Proposition 1.** Assume \( \lambda \leq \|T\| \) and

\[ \lambda \bar{m} \geq 16\kappa N(\lambda) \log \frac{6}{\delta}, \]

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

\[ \left\| (T + \lambda)\frac{1}{2}(f_{\lambda,\lambda}^{ls} - f_{\lambda,\lambda}^{L}) \right\|_{\mathcal{H}} \leq M + \kappa \bar{m} \left\| f_{\lambda}^{ls} \right\|_{\mathcal{H}} \left( \frac{2 \kappa}{\lambda} + \sqrt{\frac{N(\lambda)}{m}} \right) \log \frac{6}{\delta}, \]

where

\[ f_{\lambda}^{ls} = (T + \lambda)^{-1} L_K f_{\mathcal{H}}. \]

**Proof.** Assuming

\[ S_1 := \left\| (T + \lambda)^{-\frac{1}{2}} (T - T_{\lambda})(T + \lambda)^{-\frac{1}{2}} \right\|_{\mathcal{H}} < 1, \]

by simple algebraic computations we obtain
\[ f_{k,\lambda}^\lambda - f_{\lambda}^\lambda = (T_k + \lambda)^{-1} g_k - (T + \lambda)^{-1} g \]
\[ = (T_k + \lambda)^{-1} \left\{ (g_k - g) + (T - T_k)(T + \lambda)^{-1} g \right\} \]
\[ = (T_k + \lambda)^{-1} \left\{ (T + \lambda)^{-1} \left\{ (g_k - g) + (T + \lambda)^{-1} g \right\} \right\} \]
\[ = (T + \lambda)^{-2} \left\{ 1 - (T + \lambda)^{-1} (T - T_k)(T + \lambda)^{-1} g \right\} \]
\[ = \left\{ (T + \lambda)^{-1} (g_k - g) + (T + \lambda)^{-1} (T - T_k)f \right\} . \]

Therefore we get
\[ \left\| (T + \lambda)^{1/2} (f_{k,\lambda}^\lambda - f_{\lambda}^\lambda) \right\|_\mathcal{H} \leq \frac{S_2 + S_3}{1 - S_1} , \]

where
\[ S_2 := \left\| (T + \lambda)^{-1/2} (g_k - g) \right\|_\mathcal{H} , \]
\[ S_3 := \left\| (T + \lambda)^{-1/2} (T - T_k)f \right\|_\mathcal{H} . \]

Now we want to estimate the quantities \( S_1, S_2 \) and \( S_3 \) using Prop. 4. In fact, choosing the correct vector-valued random variables \( \xi_1, \xi_2 \) and \( \xi_3 \), the following common representation holds,
\[ S_h = \left\| \frac{1}{m_h} \sum_{i=1}^{m_h} \xi_i(x_i) - \mathbb{E}[\xi_i] \right\| , \quad h = 1, 2, 3. \]

Indeed, in order to let the equality above hold, \( \xi_1 : X \rightarrow \mathcal{L}_{HS}(\mathcal{H}) \) is defined by
\[ \xi_1(x)[\cdot] = (T + \lambda)^{-1/2} K_x (K_{x,1})_\mathcal{H} (T + \lambda)^{-1/2} , \]
and \( m_1 = \bar{m} \).
Moreover, \( \xi_2 : Z \rightarrow \mathcal{H} \) is defined by
\[ \xi_2(x, y) = (T + \lambda)^{-1/2} K_{xy} , \]
with \( m_2 = m \).
And finally, \( \xi_3 : X \rightarrow \mathcal{H} \) is defined by
\[ \xi_3(x) = (T + \lambda)^{-1/2} K_{x,1}\lambda(x) , \]
with \( m_3 = \bar{m} \).
Hence, applying three times Prop. 4, we can write
\[ P \left[ S_h \leq 2 \left( \frac{H_h}{m_h} + \frac{\sigma_h}{\sqrt{m_h}} \right) \log \frac{6}{\delta} \right] \geq 1 - \frac{3}{\delta} , \quad h = 1, 2, 3, \]
where, as it can be straightforwardly verified, the constants \( H_h \) and \( \sigma_h \) are given by the expressions
\[ H_1 = 2\frac{\kappa}{\lambda} , \quad \sigma_1^2 = \kappa \mathcal{N}(\lambda) , \]
\[ H_2 = 2M\sqrt{\frac{\kappa}{\lambda}} , \quad \sigma_2^2 = M^2\mathcal{N}(\lambda) , \]
\[ H_3 = 2 \left\| f_{\lambda}^\lambda \right\|_\mathcal{H} , \quad \sigma_3^2 = \kappa \left\| f_{\lambda}^\lambda \right\|_\mathcal{H} \mathcal{N}(\lambda). \]
Now, recalling the assumptions on $\lambda$ and $\tilde{m}$, with probability greater than $1 - \delta/3$, we get

$$S_1 \leq 2 \left( \frac{2\kappa}{\tilde{m}\lambda} + \sqrt{\frac{N(\lambda)\kappa}{\tilde{m}\lambda}} \right) \log \frac{6}{\delta}$$

$$\left( N(\lambda) \geq \frac{\|T\| + \lambda}{\|T\|} \geq \frac{1}{2} \right)$$

(eq. (9))

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

$$\| (T + \lambda)^{\frac{1}{2}} (f_{H,\lambda} - f_{\lambda}) \|_H \leq 4(S_2 + S_3)$$

$$\leq 8 \left( M + \sqrt{\frac{m}{\tilde{m}}} \left\| f_\lambda \right\|_H \right) \left( \frac{2}{m} \sqrt{\frac{\lambda}{\tilde{m}}} + \sqrt{\frac{N(\lambda)}{m}} \right) \log \frac{6}{\delta}.$$

□

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

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

and

$$\lambda N(\lambda) \leq \kappa.$$

**Proof.** First, observe that

$$\text{Tr}[T] = \int_X \text{Tr}[K_x (K_{x^*})_H] d\rho_X(x) = \int_X K(x, x) d\rho_X(x) \leq \sup_{x \in X} K(x, x) \leq \kappa.$$

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

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

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

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

□

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

$$\left\| f_{H,\lambda}^r - f_{H}^r \right\|_p \leq \lambda^r \left\| L_K^r f_H \right\|_p, \quad \text{if } r \leq 1$$

$$\left\| f_{H}^r \right\|_H \leq \left\{ \begin{array}{ll}
\lambda^{-1 + r} \left\| L_K^{-r} f_H \right\|_p & \text{if } r \leq \frac{1}{2}, \\
\kappa^{-1 + r} \left\| L_K^{-r} f_H \right\|_p & \text{if } r > \frac{1}{2}.
\end{array} \right.$$
Regarding the second estimate, if \( r \leq \frac{1}{2} \), since \( T \) is positive, we can write,

\[
\begin{align*}
\| f^n_\lambda \|_{\mathcal{H}} & \leq \| (T + \lambda)^{-1} L_K f_\mathcal{H} \| \mathcal{H} \\
& \leq \| (T + \lambda)^{-\frac{1}{2} + r} (T(T + \lambda)^{-1})^{\frac{1}{2} + r} L_K^{1-r} f_\mathcal{H} \|_{\mathcal{H}} \\
& \leq \| (T + \lambda)^{-1} \|^{\frac{1}{2} - r} \| L_K^{1-r} f_\mathcal{H} \|_\rho \leq \lambda^{-\frac{1}{2} + r} \| L_K^{1-r} f_\mathcal{H} \|_\rho .
\end{align*}
\]

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

\[
\begin{align*}
\| f^n_\lambda \|_{\mathcal{H}} & \leq \| (T + \lambda)^{-1} L_K f_\mathcal{H} \|_{\mathcal{H}} \\
& \leq \| T^{-\frac{1}{2}} T(T + \lambda)^{-1} L_K^{1-r} f_\mathcal{H} \|_{\mathcal{H}} \\
& \leq \| T \|^{r - \frac{1}{2}} \| L_K^{1-r} f_\mathcal{H} \|_\rho \leq \kappa^{r - \frac{1}{2}} \| L_K^{1-r} f_\mathcal{H} \|_\rho .
\end{align*}
\]

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

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

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

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

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

We are finally ready to prove Theorem 1.

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

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

\[
\hat{\lambda} \leq \frac{\| T \|}{\kappa} \leq 1.
\]

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

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

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

\[
\lambda^{2r} m = 16D_r^2 \lambda^{-\sigma} \log^2 \frac{6}{\delta} \geq 16N(\lambda) \log^2 \frac{6}{\delta}.
\]

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

\[
\| f^n_\lambda - f_\mathcal{H} \|_\rho \leq \hat{\lambda}^r C_r, \\
\kappa \| f^n_\lambda \|_{\mathcal{H}}^2 \leq \lambda^{-|1-2r|+} C_r.
\]

Therefore we can apply Prop. 1, and using the two estimates above, the assumption \( \bar{m} \geq m \lambda^{-|1-2r|+} \) and the definition of \( D_r \), to obtain the following bound,
The first estimate follows simply observing that

$$
\left\| f_{\lambda,\lambda} - f_\lambda \right\|_\rho \leq \left\| f_{\lambda,\lambda} - f_\lambda^\perp \right\|_\rho + \left\| f_\lambda - f_\lambda^\perp \right\|_\rho.
$$

Then, the following estimates hold,

$$
\left\| f_{\lambda,\lambda} \right\| = \left\| \sqrt{T} f \right\|_\rho 
\leq \left\| (T + \lambda)^{\frac{1}{2}} (f_{\lambda,\lambda} - f_\lambda^\perp) \right\|_H + \left\| f_\lambda - f_\lambda^\perp \right\|_\rho,
$$

$$
\leq 8(M + C_r) \frac{1}{\sqrt{m}} \left( \frac{2}{\sqrt{m\lambda}} + \frac{D_s}{\sqrt{\lambda'}} \right) \log \frac{6}{\lambda'} + \lambda' C_r.
$$

(eq. (6))

$$
= 2(M + C_r) \lambda' \left( 1 + \frac{\lambda'^{s+\frac{1}{2}}}{2D_s^2 \log \frac{6}{\lambda'}} \right) + \lambda' C_r,
$$

$$
\left( r + s \geq \frac{1}{2} \right) \leq (3(M + C_r) + C_r) \lambda' \leq 4(M + C_r) \lambda'.
$$

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

5. Proof of Theorem 2

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

Proposition 5. Let $f_\lambda \in \text{Im} L_K$ for some $r > 0$. For any $\lambda > 0$ let the truncated function $f^{\perp}_\lambda$ be defined by

$$
f^{\perp}_\lambda = P_\lambda f_\lambda
$$

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

$$
P_\lambda = \Theta_\lambda(L_K),
$$

with

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

Then, the following estimates hold,

$$
\left\| f^{\perp}_\lambda - f_\lambda \right\|_\rho \leq \lambda' \left\| L_K^{-r} f_\lambda \right\|_\rho,
$$

$$
\left\| f^{\perp}_\lambda \right\|_H \leq \begin{cases} \lambda^{\frac{1}{r} + \frac{1}{2}} \left\| L_K^{-r} f_\lambda \right\|_\rho & \text{if } r \leq \frac{1}{2}, \\ \kappa^{\frac{1}{r} + \frac{1}{2}} \left\| L_K^{-r} f_\lambda \right\|_\rho & \text{if } r > \frac{1}{2}. \end{cases}
$$

Proof. The first estimate follows simply observing that

$$
\left\| f^{\perp}_\lambda - f_\lambda \right\|_\rho = \left\| P_\lambda f_\lambda \right\|_\rho = \left\| P_\lambda L_K^{-r} f_\lambda \right\|_\rho \leq \lambda' \left\| L_K^{-r} f_\lambda \right\|_\rho,
$$

where we introduced the orthogonal projector $P_\lambda = \text{Id} - P_\lambda$.

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

$$
\left\| f^{\perp}_\lambda \right\|_H = \left\| P_\lambda f_\lambda \right\|_H \leq \left\| L_K^{-r} P_\lambda f_\lambda \right\|_\rho 
\leq \left\| L_K^{-\frac{1}{2} + r} \Theta_\lambda(L_K) \right\|_\rho \left\| L_K^{-r} f_\lambda \right\|_\rho.
$$
The proof is concluded observing that by Prop. 2, \( \|L_\kappa\| = \|T\| \leq \kappa \), and that, for every \( \sigma \in [0, \kappa] \), it holds
\[
\sigma^{-\frac{1}{2}+r}\Theta_\lambda(\sigma) \leq \begin{cases} 
\lambda^{-\frac{1}{2}+r} & \text{if } r \leq \frac{1}{2}, \\
\kappa^{-\frac{1}{2}+r} & \text{if } r > \frac{1}{2},
\end{cases}
\]
\( \Box \)

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

**Proposition 6.** Let \( r > 0 \) and define
\[
\gamma = \lambda^{-1} \|T - T_k\|.
\]
Then, if \( \lambda \in (0, \kappa] \), it holds
\[
\left\| \sqrt{T} (G_\lambda(T_k) T_k - \text{Id}) f^\lambda \right\|_H \leq B_r C_r (1 + \sqrt{\gamma})(2 + \gamma \lambda^{3-r} + \gamma^2) \lambda^r,
\]
where
\[
\eta = |r - \frac{1}{2}| - \left| \frac{r - 1}{2} \right|.
\]

**Proof.** The two inequalities (16) and (17) will be useful in the proof. The first follows from Theorem 1 in [15],
\[
\|T^\alpha - T_k^\alpha\| \leq \|T - T_k\|, \quad \alpha \in [0, 1]
\]
where we adopted the convention \( 0^0 = 1 \). The second is a corollary of Theorem 8.1 in [2]
\[
\|T^p - T_k^p\| \leq p \eta^p \|T - T_k\|, \quad p \in \mathbb{N}.
\]
We also need to introduce the orthogonal projector in \( \mathcal{H} \), \( P_{\lambda,\kappa} \), defined by
\[
P_{\lambda,\kappa} = \Theta_\lambda(T_k),
\]
with \( \Theta_\lambda \) defined in (14).

We analyze the cases \( r \leq \frac{1}{2} \) and \( r > \frac{1}{2} \) separately.

**Case** \( r \leq \frac{1}{2} \): In the three steps below we subsequently estimate the norms of the three terms of the expansion
\[
\sqrt{T} (G_\lambda(T_k) T_k - \text{Id}) f^\lambda = \sqrt{T} P_{\lambda,\kappa} r_\lambda(T_k) f^\lambda + P_{\lambda,\kappa} r_\lambda(T_k) T_k f^\lambda + \sqrt{T} - \sqrt{T_k} P_{\lambda,\kappa} r_\lambda(T_k) f^\lambda,
\]
where \( P_{\lambda,\kappa} = \text{Id} - P_{\lambda,\kappa} \) and \( r_\lambda(\sigma) = \sigma G_\lambda(\sigma) - 1 \).

**Step 1:** Observe that
\[
\left\| \sqrt{T} P_{\lambda,\kappa} \right\|^2 = \left\| P_{\lambda,\kappa} T P_{\lambda,\kappa} \right\| \leq \sup_{\phi \in \mathcal{H}} \frac{(\phi, T \phi)_\mathcal{M}}{\|\phi\|_H^2} \leq 1 + \|T_k - T\| = \lambda(1 + \gamma).
\]
Therefore, from definitions (2) and (4) and Proposition 5, it follows
\[
\left\| \sqrt{T} P_{\lambda,\kappa} r_\lambda(T_k) f^\lambda \right\|_H \leq B_r C_r \sqrt{T + \gamma} \lambda^r.
\]
Step 2: Observe that, from inequality (16), definition (4) and Proposition 5
\[
\left\| T_k^{\frac{1}{2}} f\right\|_H \leq \|\Theta(T)\| \left\| T_k^{\frac{1}{2}} f_H\right\|_H + \left\| T_k^{\frac{1}{2}} f\right\|_H
\]
(19)
Therefore from definition (3), it follows
\[
\left\| P_{k,\lambda} r_\lambda(T_k) T_k^{\frac{1}{2}} f_H\right\|_H \leq \|P_{k,\lambda}\| \|r_\lambda(T_k)\| \left\| T_k^{\frac{1}{2}} f\right\|_H
\]
\[
\leq B_\varepsilon C_\varepsilon (1 + \frac{\sqrt{2}}{2} r) \lambda^r.
\]
Step 3: Recalling the definition of \( P_{k,\lambda} \), and applying again inequality (19) and Proposition 5, we get
\[
\left\| (\sqrt{T} - \sqrt{T_k}) P_{k,\lambda} r_\lambda(T_k) f_H\right\|_H \leq \left\| (\sqrt{T} - \sqrt{T_k}) T_k^{\frac{1}{2}} + P_{k,\lambda} (T_k^{\frac{1}{2}} f_H)\right\|_H
\]
\[
\leq \left\| \sqrt{T} - \sqrt{T_k}\right\| \left\| T_k^{\frac{1}{2}} P_{k,\lambda}\right\| \|r_\lambda(T_k)\| \left\| T_k^{\frac{1}{2}} f\right\|_H
\]
\[
\leq B_\varepsilon C_\varepsilon (1 + \frac{\sqrt{2}}{2} r) \lambda^r.
\]
Since we assumed \( 0 < r \leq \frac{1}{2} \), and therefore \( \eta = \frac{1}{2} - r \), the three estimates above prove the statement of the Theorem in this case.

Case \( r \geq \frac{1}{2} \): Consider the expansion
\[
(G_\lambda(T_k) T_k - \text{Id}) f_H^r \leq r_\lambda(T_k) T_k^{\lambda - \frac{1}{2}} v
\]
\[
\leq r_\lambda(T_k) T_k^{-\frac{1}{2}} v + r_\lambda(T_k) \left( T_k^{\frac{1}{2}} - T_k^{-\frac{1}{2}}\right) v
\]
\[
\leq r_\lambda(T_k) T_k^{-\frac{1}{2}} v + r_\lambda(T_k) T_k^{p} \left( T_k^{\frac{1}{2}} - T_k^{-\frac{1}{2}}\right) v
\]
\[
+ r_\lambda(T_k) (T_k^{p} - T_k^{p}) T_k^{\frac{1}{2}} v
\]
where \( v = P_{k,\lambda} T_k^{\frac{1}{2}} f_H, r_\lambda(\sigma) = \sigma G_\lambda(\sigma) - 1 \) and \( p = [r - \frac{1}{2}] \).

Now, for any \( \beta \in [0, \frac{1}{2}] \), from the expansion above using inequalities (16) and (17), and definition (3), we get
\[
\left\| T_k^{\beta} (G_\lambda(T_k) T_k - \text{Id}) f_H^r\right\|_H \leq \left\| r_\lambda(T_k) T_k^{\beta + 1} v\right\|_H
\]
(20)
\[
+ \left\| r_\lambda(T_k) T_k^{p + \beta} \left( T_k^{\frac{1}{2}} - T_k^{-\frac{1}{2}}\right) v\right\|_H
\]
\[
+ \left\| r_\lambda(T_k) T_k^{p} \left( T_k^{\frac{1}{2}} - T_k^{-\frac{1}{2}}\right) v\right\|_H
\]
\[
\leq B_\varepsilon C_\varepsilon (1 + \frac{\sqrt{2}}{2} r) \lambda^r.
\]
Finally, from the expansion
\[
\left\| \sqrt{T} (G_\lambda(T_k) T_k - \text{Id}) f_H^r\right\|_H \leq \left\| \sqrt{T} - \sqrt{T_k}\right\| \left\| r_\lambda(T_k) f_H^r\right\|_H
\]
\[
+ \left\| \sqrt{T_k} r_\lambda(T_k) f_H^r\right\|_H.
\]
using (16) and inequality (20) with \( \beta = 0 \) and \( \beta = \frac{1}{2} \), we get the claimed result also in this case.

We need an additional preliminary result.

**Proposition 7.** Let the operator \( \Omega_\lambda \) be defined by

\[
\Omega_\lambda = \sqrt{T} G_\lambda(T_k)(T_k + \lambda) (T + \lambda)^{-\frac{1}{2}}.
\]

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

\[
\| \Omega_\lambda \| \leq (1 + 2\sqrt{T}) A,
\]

with \( \gamma \) defined in eq. (15).

**Proof.** First consider the expansion

\[
\Omega_\lambda = \left( \sqrt{T} - \sqrt{T_k} \right) \Delta_\lambda (T + \lambda)^{-\frac{1}{2}} - \Delta_\lambda \left( \sqrt{T} - \sqrt{T_k} \right) (T + \lambda)^{-\frac{1}{2}} + \Delta_\lambda \sqrt{T}(T + \lambda)^{-\frac{1}{2}},
\]

where we introduced the operator

\[
\Delta_\lambda = G_\lambda(T_k)(T_k + \lambda).
\]

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

\[
\left\| \sqrt{T} - \sqrt{T_k} \right\| \leq \sqrt{T - T_k}.
\]

From the previous observations we easily get

\[
\| \Omega_\lambda \| \leq 2 \| \Delta_\lambda \| \left\| \sqrt{T} - \sqrt{T_k} \right\| (T + \lambda)^{-\frac{1}{2}} + \| \Delta_\lambda \| \left\| \sqrt{T}(T + \lambda)^{-\frac{1}{2}} \right\|
\]

\[
\leq A(1 + 2\sqrt{T}),
\]

the claimed result.

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

**Proof of Theorem 2.** We consider the expansion

\[
\sqrt{T}(f_{\lambda} - f_H) = \sqrt{T}(G_\lambda(T_k)g_k - f_{\lambda}^R) + \sqrt{T}(f_{\lambda}^R - f_H)
\]

\[
= \Omega_\lambda (T + \lambda)^{\frac{1}{2}} (f_{\lambda}^R - f_{\lambda}^R, \lambda) + \sqrt{T}(G_\lambda(T_k)T_k - \text{Id}) f_{\lambda}^R + \sqrt{T}(f_{\lambda}^R - f_H)
\]

\[
= \Omega_\lambda \left( (T + \lambda)^{\frac{1}{2}} (f_{\lambda}^R - f_{\lambda}^R, \lambda) + (T + \lambda)^{\frac{1}{2}} (f_{\lambda}^R - f_{\lambda}^R, \lambda) + (T + \lambda)^{\frac{1}{2}} (f_{\lambda}^R - f_{\lambda}^R, \lambda) \right)
\]

\[
+ \sqrt{T}(G_\lambda(T_k)T_k - \text{Id}) f_{\lambda}^R + \sqrt{T}(f_{\lambda}^R - f_H)
\]

where the operator \( \Omega_\lambda \) is defined by equation (21), the ideal RLS estimators are \( f_{\lambda}^R = (T + \lambda)^{-1}T f_H \) and \( f_{\lambda}^\ast = (T + \lambda)^{-1}T f_{\lambda}^\ast \), and \( f_{\lambda}^{\ast, \lambda} = (T_k + \lambda)^{-1}T_k f_{\lambda}^R \) is the RLS estimator constructed by the training set

\[
\tilde{z}' = ((\tilde{x}_1, f_{\lambda}^R(\tilde{x}_1)), \ldots, (\tilde{x}_n, f_{\lambda}^R(\tilde{x}_n))).
\]

Hence we get the following decomposition,

\[
\| f_{\lambda} - f_H \| \leq D \left( S^R + R + S^A \right) + P + P^S,
\]
with

\[ S_{ls}^o = \left\| (T + \lambda) \frac{1}{2} (f_{\lambda, \lambda}^o - f_{\lambda}^o) \right\|_H, \]
\[ \bar{S}_{ls} = \left\| (T + \lambda) \frac{1}{2} (f_{\lambda, \lambda}^o - \tilde{f}_{\lambda}^o) \right\|_H, \]
\[ D = \| \Omega \lambda \|, \]
\[ P = \sqrt{T} \left( G_\lambda (T \tilde{x}) \right. \left. T \tilde{x} - \text{Id} \right) f_{\lambda, \lambda}^o, \]
\[ P_{tr} = \| f_{\lambda, \lambda}^o - f_{\lambda}^o \|_\rho, \]
\[ R = \left\| (T + \lambda) \frac{1}{2} (\tilde{f}_{\lambda}^o - f_{\lambda}^o) \right\|_H. \]

Terms \( S_{ls}^o \) and \( \bar{S}_{ls} \) will be estimated by Proposition 1, term \( D \) by Proposition 7, term \( P \) by Proposition 6 and finally terms \( P_{tr} \) and \( R \) by Proposition 5.

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

Regarding the estimate of \( S_{ls} \). Applying Proposition 1 and reasoning as in the proof of Theorem 1 (recall that by assumption \( \tilde{m} \geq m \hat{\lambda} - |2r - s| - |2r - s| \)), we get that with probability greater than \( 1 - \delta \)

\[
S_{ls} \leq 8 \left( M + \sqrt{\frac{m}{m}} C_v \hat{\lambda}^{-\frac{1}{2} - r} \right) \left( \frac{2}{m} \sqrt{\frac{\lambda}{m}} + \sqrt{\frac{N(\lambda)}{m}} \right) \log \frac{6}{\delta} \]
\[
\leq 8 (M + C_v) \frac{1}{\sqrt{m}} \left( \frac{2}{\sqrt{m}} + \frac{D_s}{\sqrt{\lambda}} \right) \log \frac{6}{\delta} \]
\[
(\text{eq. (7)}) \]
\[
= 2 (M + C_v) \hat{\lambda}^r \left( 1 + \frac{\hat{\lambda}^{s+\frac{1}{2}}}{2D_s^2 \log \frac{6}{\delta}} \right) \]
\[
\left( r + s \geq \frac{1}{2} \right) \]
\[
\leq 3 (M + C_v) \hat{\lambda}^r.
\]

The term \( S_{ls}^o \) can be estimated observing that \( \tilde{x}' \) is a training set of \( \tilde{m} \) supervised samples drawn i.i.d. from the probability measure \( \rho' \) with marginal \( \rho_X \) and conditional \( \rho'_y(y) = \delta(y - f^o_\lambda(x)) \). Therefore the regression function induced by \( \rho' \) is \( f_\rho = f_X^o \), and the support of \( \rho' \) is included in \([-M', M'] \times X \), with \( M' = \sup_{x \in X} f_\rho(x) \leq \sqrt{\pi} \| f_X^o \|_H \).

Again applying Proposition 1 and reasoning as in the proof of Theorem 1, we obtain that
with probability greater than $1 - \delta$ it holds

\begin{equation}
S_{\alpha}^{\tilde{\gamma}} \leq 8 \left( M' + \sqrt{\kappa} \left\| f_{\alpha}^{\tilde{\gamma}} \right\|_{m} \right) \left( \frac{2}{\tilde{m}} \sqrt{\frac{\kappa}{\lambda}} + \sqrt{\frac{\kappa}{\tilde{m}}} \right) \log \frac{6}{\delta}
\end{equation}

\begin{equation}
\leq 16 \sqrt{\kappa} \left\| f_{\alpha}^{\tilde{\gamma}} \right\|_{m} \left( \frac{2}{\tilde{m}} \sqrt{\frac{\kappa}{\lambda}} + \sqrt{\frac{\kappa}{\tilde{m}}} \right) \log \frac{6}{\delta}
\end{equation}

\text{(Prop.5)} \leq 16 \frac{\sqrt{m}}{\tilde{m}} C_\gamma \lambda^{-\frac{1}{2} - |r|} \left( \frac{2}{\tilde{m}} \sqrt{\frac{\kappa}{\lambda}} + \sqrt{\frac{\kappa}{\tilde{m}}} \right) \log \frac{6}{\delta}

\begin{equation}
\leq 16 C_\gamma \frac{1}{\sqrt{m}} \left( \frac{2}{\sqrt{m}} \lambda + \frac{D_s}{\lambda} \right) \log \frac{6}{\delta}
\end{equation}

\text{(eq. (7))} = 4 C_\gamma \lambda^r \left( 1 + \frac{\lambda^{r+s} - 1}{2 D_s^2 \log \frac{6}{\delta}} \right)

\begin{equation}
\left( r + s \geq \frac{1}{2} \right) \leq 6 C_\gamma \lambda^r
\end{equation}

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

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

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

\[ \gamma \leq \lambda^{-1} \left\| T - T_k \right\|_{HS} \leq \frac{2}{\lambda} \left( \frac{2\kappa}{\tilde{m}} + \frac{\kappa}{\sqrt{\tilde{m}}} \right) \log \frac{2}{\delta} \leq 4 \frac{1}{\lambda \sqrt{\tilde{m}}} \log \frac{2}{\delta}
\]

\[ \leq 4 \frac{\lambda^{1-r} - 1}{\sqrt{\tilde{m}}} \log \frac{2}{\delta} \leq \lambda^{1-r} \left( \lambda^{1-r} \right)^{-1} \leq \lambda^{r+s} \leq 1,
\]

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

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

\begin{equation}
D \leq 3 A,
\end{equation}

and from Proposition 6

\begin{equation}
P \leq 2B_C (3 + r \lambda^{1-r}) \lambda^r
\end{equation}

\[ \leq 2B_C (3 + r \lambda^{1-r} + 1 + 1) \lambda^r
\]

\[ \leq 2B_C (3 + r \lambda^{1-r} + 1) \lambda^r \leq 2B_C (3 + r) \lambda^r.
\]

Regarding terms $P^{\alpha}$ and $R$. From Proposition 5 we get

\begin{equation}
P^{\alpha} \leq C_\lambda \lambda^r
\end{equation}

and hence,

\begin{equation}
R = \left\| (T + \lambda)^{-\frac{1}{2}} T (f_{z}^{\alpha} - f_{z}^{\lambda}) \right\|_{\mathcal{H}}
\end{equation}

\[ \leq \left\| \sqrt{T} (f_{z}^{\alpha} - f_{z}^{\lambda}) \right\|_{\mathcal{H}} \leq P^{\alpha} \leq C_\lambda \lambda^r.
\]

The proof is completed by plugging inequalities (24), (25), (26), (27), (28) and (29) in (23) and recalling the expression for $\lambda$. \( \square \)
The author wish to thank E. De Vito, T. Poggio, L. Rosasco, S. Smale, A. Verri and Y. Yao for useful discussions and suggestions.

Acknowledgements

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

References