Optimal Rates for Regularized Least Squares Regression

Ingo Steinwart, Don Hush, and Clint Scovel
Modeling, Algorithms and Informatics Group, CCS-3
Los Alamos National Laboratory
{ingo,dhush,jcs}@lanl.gov

Abstract

We establish a new oracle inequality for kernel-based, regularized least squares regression methods, which uses the eigenvalues of the associated integral operator as a complexity measure. We then use this oracle inequality to derive learning rates for these methods. Here, it turns out that these rates are *independent* of the exponent of the regularization term. Finally, we show that our learning rates are asymptotically optimal whenever, e.g., the kernel is continuous and the input space is a compact metric space.

1 Introduction

Given a training set $((x_1, y_1), \ldots, (x_n, y_n))$ sampled from some unknown distribution P on $X \times [-M, M]$, the goal of non-parametric least squares regression is to find a function $f : X \to \mathbb{R}$, whose risk

$$\mathcal{R}_{L,P}(f) := \int_{X\times Y} L\big(y, f(x)\big)\, dP(x, y)\,,$$

where L is the least squares loss, i.e. $L(y, t) = (y - t)^2$, is close to the optimal risk

$$\mathcal{R}^*_{L,P} := \inf\big\{\mathcal{R}_{L,P}(f) \,\big|\, f : X \to \mathbb{R}\big\}\,.$$

It is well known that the regression function f^*_P, that is, the conditional expectation $x \mapsto \mathbb{E}_P(Y|x)$, is the P_X-almost surely minimizer of $\mathcal{R}_{L,P}$, where P_X denotes the marginal distribution of P. In other words, we have $\mathcal{R}_{L,P}(f^*_P) = \mathcal{R}^*_{L,P}$, and another well known result further shows

$$\mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P} = \int_X |f - f^*_P|^2\, dP_X = \|f - f^*_P\|^2_{L_2(P_X)}$$

for all $f : X \to \mathbb{R}$.

There exists a huge number of methods for solving the non-parametric least squares regression problem, many of which are thoroughly treated in [12]. In this work, we focus on kernel-based methods, i.e., on methods that find an

$$f_{D,\lambda} \in \arg\min_{f\in H}\Big(\lambda\|f\|^q_H + \mathcal{R}_{L,D}(f)\Big)\,. \qquad (1)$$

Here, H is a reproducing kernel Hilbert space (RKHS) of a bounded kernel k, $q \geq 1$ is some constant, $\lambda > 0$ is a regularization parameter, and $\mathcal{R}_{L,D}(f)$ is the empirical risk of f, that is

$$\mathcal{R}_{L,D}(f) = \frac{1}{n}\sum_{i=1}^n L\big(y_i, f(x_i)\big)\,.$$

Note that by a straightforward modification of [19, Lemma 5.1 & Theorem 5.2], see Lemma 12 for details, we find that there always exists exactly one $f_{D,\lambda}$ satisfying (1). Furthermore, by a general form of the representer theorem, see [16], it is known that this solution is of the form

$$f_{D,\lambda} = \sum_{i=1}^n \alpha_i k(x_i, \cdot\,)\,,$$

where $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ are suitable coefficients. Recall that, for $q = 2$, we recover the usual least squares support vector machine (SVM) without bias, which is also known as regularization network, see [15]. To the best of our knowledge, this value of q is also the only value for which efficient algorithms for finding $f_{D,\lambda}$ have been developed. On the other hand, the recent work [13] suggests that $q = 2$ may not be the optimal choice from a statistical point of view, that is, that least squares support vector machines may have a design flaw. Fortunately, we will see in this work that the exponent q has no influence on the learning rates and thus its value may be chosen on algorithmic considerations, only. In particular, there is no reason not to use $q = 2$.

The learning method (1), in particular in the SVM case, has recently attracted a lot of theoretical considerations, see, e.g., [8, 10, 18, 6, 13] and the references therein. In particular, optimal rates have been established in [6], if $f^*_P \in H$ and the eigenvalue behavior of the integral operator associated to H, see (5), is known. However, for distributions, for which the former assumption is *not* satisfied, i.e. $f^*_P \notin H$, the situation is substantially less understood. Indeed, the sharpest known result for this case was recently established in [13] for a learning method that, modulo some logarithmic terms, equals (1) in the case $q = \frac{2p}{1+p}$, where $p \in (0, 1)$ is an exponent describing the eigenvalue behavior of the integral operator.

As discussed in [13], the reason for this lack of understanding may be the fact that, for $f^*_P \notin H$, one has to deal with localized function classes whose $\|\cdot\|_\infty$-bounds are no longer bounded for $\lambda \to 0$. To address this difficulty, [13] assumes that the eigenfunctions of the integral operator are

uniformly bounded. This is then used to show that

$$\|f\|_\infty \le C \, \|f\|_H^p \, \|f\|_{L_2(P_X)}^{1-p} \tag{2}$$

for all $f \in H$, where $C > 0$ is some constant. Compared to the classical estimate $\|f\|_\infty \le \|k\|_\infty \|f\|_H$, Inequality (2) obviously gives a sharper $\|\cdot\|_\infty$-bound on those elements of H, for which we have a non-trivial bound on their $L_2(P_X)$-norm. [13] then shows that localization gives such non-trivial $L_2(P_X)$-bounds, which in turn results in the best known learning rates. Unfortunately, however, even without the extra logarithmic terms, it is unclear how to solve (1) algorithmically if $q \neq 2$. In addition, knowing the value of p that best describes the eigenvalue behavior seems to be rather unrealistic in many cases.

The goal of this work is to address these shortcomings by establishing an oracle inequality that holds for *all* $q \in [1, \infty)$. From this oracle inequality we then derive learning rates which *a)* equal those of [13], and *b)* turn out to be independent of q. Using an intimate relationship between eigenvalues and entropy numbers we further show that these learning rates are optimal in a minmax sense. From a statistical point of view, the SVM case $q = 2$ has therefore no advantages or disadvantages compared to other values of q. From an algorithmic point of view however, $q = 2$ is currently the only feasible case, which in turn makes SVMs the method of choice.

To achieve this new oracle inequality, we merge the idea of using (2) with another idea to improve certain $\|\cdot\|_\infty$-bounds. To explain the latter, recall that for some loss functions it was observed in [2, 22, 20, 19] that $\|\cdot\|_\infty$-bounds can be made smaller by *clipping* the decision function. Let us describe this clipping idea for the least squares loss. To this end, we first observe that for all $t \in \mathbb{R}$ and $y \in [-M, M]$ we have

$$L(y, \widehat{t}) \;\le\; L(y, t)\,, \tag{3}$$

where $\widehat{t}$ denotes the clipped value of t at $\pm M$, that is

$$\widehat{t} := \begin{cases} -M & \text{if } t < -M \\ t & \text{if } t \in [-M, M] \\ M & \text{if } t > M\,. \end{cases} \tag{4}$$

In other words, we never obtain a worse loss if we restrict our predictions to the interval from which the labels are drawn. Now note that (3) obviously implies that, for every function $f : X \to \mathbb{R}$, we have

$$\mathcal{R}_{L,P}(\widehat{f}\,) \le \mathcal{R}_{L,P}(f)\,,$$

where the clipping $\widehat{f}$ of f is meant to be a pointwise clipping at M. In other words, the clipping operation potentially reduces the risk. Now, given a decision function f_D of some learning method $\mathcal{A}$, the key idea of the clipping technique is to bound the risk $\mathcal{R}_{L,P}(\widehat{f_D})$ of the clipped decision function rather than the risk $\mathcal{R}_{L,P}(f_D)$ of the unclipped function. From a practical point of view, this approach means that the *training* algorithm for $\mathcal{A}$ remains unchanged and only the *evaluation* of the resulting decision functions needs to be slightly changed.

In the following section, we will present our main results, discuss the assumption (2) in more detail, and consider Sobolev spaces as a special case of our general results. Section 3 contains the proofs of the oracle inequality and the resulting learning rates, while in Section 4 we establish the corresponding lower bounds.

2 Results

In the following, L always denotes the least squares loss, and X denotes some arbitrary measurable space. Moreover, P is a probability measure on $X \times [-M, M]$, where $M > 0$ is some constant.

We also have to introduce some notions related to kernels, see [19, Chapter 4] for more details. To this end, let $k : X \times X \to \mathbb{R}$ be a measurable kernel and H its associated reproducing kernel Hilbert space (RKHS). In the following, we always assume that k is bounded, that is, $\|k\|_\infty := \sup_{x \in X} k(x, x) < \infty$. To avoid superfluous notations, we further assume $\|k\|_\infty = 1$, where we note that this can always be achieved by properly scaling the kernel. Recall that in this case the RKHS contains bounded measurable functions, and we have $\|f\|_\infty \le \|f\|_H$ for all $f \in H$. Given a measurable kernel and a distribution ν on X we further define the integral operator $T_k : L_2(\nu) \to L_2(\nu)$ by

$$T_k f(x) := \int_X k(x', x) f(x') \, d\nu(x')\,, \tag{5}$$

where the definition is meant to be for ν-almost all $x \in X$. In the following, we usually assume $\nu = P_X$. It is well-known, see e.g. [19, Theorem 4.27], that this operator is compact, positive, and self-adjoint. In particular, it has at most countably many non-zero eigenvalues, and all of these eigenvalues are non-negative. Let us order these eigenvalues (with geometric multiplicities), and extend the corresponding sequence by zeros if there are only finitely many non-zero eigenvalues. Then the resulting sequence $(\mu_i(T_k))_{i \ge 1}$, which we call the *extended sequence of eigenvalues*, is summable, that is, $\sum_{i=1}^\infty \mu_i(T_k) < \infty$. Again, we refer to [19, Theorem 4.27] for a precise statement. Obviously, this summability implies $\mu_i(T_k) \le a \, i^{-1}$ for some constant a and all $i \ge 1$. In our analysis, we will assume that this sequence converges even faster to zero, which, as we will briefly discuss later, is known for many kernels.

Moreover, we need the q-approximation error function $A_q : [0, \infty) \to [0, \infty)$, which is defined by

$$A_q(\lambda) := \inf_{f \in H} \Big(\lambda \|f\|_H^q + \mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P} \Big)\,.$$

Note that for $q = 2$ this approximation error function was extensively studied in [19, Chapter 5.4]. In particular, its relation to the standard approximation error was described. Furthermore, [17] and [9, Chapter 4] relate the behavior of A_2 to both interpolation spaces and certain powers of the integral operator T_k. We will come back to these results after presenting the following main theorem of this work, which establishes an oracle inequality of the learning methods described by (1).

Theorem 1 *Let k be a bounded measurable kernel on X with $\|k\|_\infty = 1$ and separable RKHS H. Moreover, let P be a distribution on $X \times [-M, M]$, where $M > 0$ is some*

constant. For $\nu = P_X$ assume that the extended sequence of eigenvalues of the integral operator (5) satisfies

$$\mu_i(T_k) \le a\, i^{-\frac{1}{p}}\,, \qquad i \ge 1, \tag{6}$$

where $a \ge 16M^4$ and $p \in (0,1)$ are some constants. Moreover, assume that there exist constants $C \ge 1$ and $s \in (0,1]$ such that

$$\|f\|_\infty \le C\,\|f\|_H^s \cdot \|f\|_{L_2(P_X)}^{1-s} \tag{7}$$

for all $f \in H$. Then, for all $q \ge 1$, there exists a constant $c_{p,q}$ only depending on p and q such that, for all $\lambda \in (0,1]$, $\tau > 0$, and $n \ge 1$, the learning method described by (1) satisfies

$$\begin{aligned}
&\mathcal{R}_{L,P}(\widehat{f}_{D,\lambda}) - \mathcal{R}^*_{L,P} \\
&\le 9A_q(\lambda) + c_{p,q}\biggl(\frac{a^{pq}M^{2q}}{\lambda^{2p}n^q}\biggr)^{\frac{1}{q-2p+pq}} \\
&\quad + \frac{120C^2M^{2-2s}\tau}{n}\Bigl(\frac{A_q(\lambda)}{\lambda}\Bigr)^{\frac{2s}{q}} + \frac{3516M^2\tau}{n}
\end{aligned}$$

with probability P^n not less than $1 - 3e^{-\tau}$.

Recall that the eigenvalue assumption (6) was first used in [5] to establish an oracle inequality for SVMs using the hinge loss, while [6] considers (6) for SVMs using the least squares loss and $f^*_P \in H$. Moreover, [13] recently established an oracle inequality for a learning method that, besides some logarithmic terms, equals (1) in the case $q = \frac{2p}{1+p}$. In general, the eigenvalue assumption (6) is a tighter measure for the complexity of the RKHS than more classical covering, or entropy number assumptions. To recall the latter, let E be Banach space and $A \subset E$ be a bounded subset. Then, for $i \ge 1$, the ith entropy number $e_i(A,E)$ of A is the infimum over all $\varepsilon > 0$ for which there exist $x_1, \ldots, x_{2^{i-1}} \in A$ with

$$A \subset \bigcup_{j=1}^{2^{i-1}} (x_j + \varepsilon B_E)\,,$$

where B_E denotes the closed unit ball of E. Moreover, the ith entropy number of a bounded linear operator $T : E \to F$ is $e_i(T) := e_i(TB_E, F)$. Now one can show, see Theorem 15, that (6) is equivalent to

$$e_i(\mathrm{id} : H \to L_2(P_X)) \le \sqrt{a}\, i^{-\frac{1}{2p}}\,, \tag{8}$$

modulo a constant only depending on p. If $\ell_\infty(X)$ denotes the space of all bounded functions on X, the latter is clearly satisfied if the more classical, *distribution-free* entropy number assumption

$$e_i(\mathrm{id} : H \to \ell_\infty(X)) \le \sqrt{a}\, i^{-\frac{1}{2p}} \tag{9}$$

is satisfied. However, the converse is, in general, not true. For example, [19, Theorem 7.34] establishes a bound of the form (8) for Gaussian RBF kernels and certain distributions P_X having *unbounded* support, while a similar bound for the $\ell_\infty(X)$-entropy numbers is impossible since $\mathrm{id} : H \to \ell_\infty(X)$ is not even compact for unbounded X. Finally, for m-times differentiable kernels on Euclidean balls of $\mathbb{R}^d$, it is known that (9) holds for $p := \frac{d}{2m}$. We refer to e.g. [19, Theorem 6.26] for a precise statement.

Analogously, if X is a Euclidean ball in $\mathbb{R}^d$, $m > d/2$ is some integer, and P_X is the uniform distribution on X, then the Sobolev space $H := W^m(X)$ is an RKHS that satisfies (6) for $p := \frac{d}{2m}$, and this estimate is also asymptotically sharp. This can be checked by Theorem 15 and a well-known result by Birman and Solomyak [4] on the entropy numbers of the embedding $\mathrm{id} : W^m(X) \to L_2(P_X)$. We refer to [11] for a thorough treatment of such estimates and [19, Appendix A.5.6] for some explanation of the latter. Moreover, by this translation it is also easy to see that it suffices to assume that P_X has a density with respect to the uniform distribution that is bounded away from 0 and ∞.

Assumption (7) is always satisfied for $C = s = 1$, but obviously the more interesting case is $s < 1$, which was recently considered by [13]. In particular, they showed in their Lemma 5.1 essentially the following result:

Theorem 2 *Let X be a compact metric space, k be a continuous kernel on X, and P_X be a distribution on X whose support satisfies* $\operatorname{supp} P_X = X$. *If (6) holds for some $p \in (0,1)$ and the corresponding eigenfunctions $(e_i)_{i\ge1}$ are uniformly bounded, that is*

$$\sup_{i\ge1}\|e_i\|_\infty < \infty\,, \tag{10}$$

then (7) holds for $s = p$.

Theorem 2 shows that, in the case of uniformly bounded eigenfunctions, Condition (7) is automatically satisfied for $s = p$. However, recall from [23] that even for C^∞-kernels (10) is *not* always satisfied.

On the other hand, (7) has a clear meaning in terms of real interpolation of spaces. To be more precise, let us briefly recall the definition of these spaces. To this end, given two Banach spaces E and F such that $F \subset E$ and $\mathrm{id} : F \to E$ is continuous, we define the K-functional of $x \in E$ by

$$K(x,t) := \inf_{y\in F}\bigl(\|x-y\|_E + t\|y\|_F\bigr)\,, \qquad t > 0.$$

Then, following [3, Definition 1.7 on page 299], the real interpolation space $[E,F]_{\theta,r}$, where $0 < \theta < 1$ and $1 \le r \le \infty$, is the Banach space that consists of those $x \in E$ with finite norm

$$\|x\|_{\theta,r} := \begin{cases} \Bigl(\int_0^\infty (t^{-\theta}K(x,t))^r t^{-1}dt\Bigr)^{1/r} & \text{if } r < \infty \\ \sup_{t>0}(t^{-\theta}K(x,t) & \text{if } r = \infty\,. \end{cases}$$

Moreover, the limiting cases are defined by $[E,F]_{0,\infty} := E$ and $[E,F]_{1,\infty} := F$. It it well-known, see e.g. [3, Proposition 1.10 on page 301] that, for all $0 < \theta < 1$ and $1 \le r \le r' \le \infty$, the space $[E,F]_{\theta,r}$ is continuously embedded in $[E,F]_{\theta,r'}$, i.e.

$$\mathrm{id} : [E,F]_{\theta,r} \to [E,F]_{\theta,r'} \tag{11}$$

is well-defined and continuous. Moreover, the assumption that F is continuously embedded in E can be used to show by elementary means that

$$\mathrm{id} : [E,F]_{\theta',\infty} \to [E,F]_{\theta,1} \tag{12}$$

is well-defined and continuous for all $0 < \theta < \theta' \le 1$.

Now [3, Proposition 2.10 on page 316] shows that (7) is satisfied if and only if $[L_2(P_X), H]_{s,1}$, is continuously embedded in $\ell_\infty(X)$. If $H = W^m(X)$ and P_X is the uniform distribution on X, we further know by the discussion on page 230 of [1] that

$$\big[L_2(P_X), W^m(X)\big]_{s,1} = B^{sm}_{2,1}(X)\,,$$

where $B^{sm}_{2,1}(X)$ denotes a Besov space, see [1]. Moreover, for $s = \frac{d}{2m}$, this Besov space is continuously embedded in $\ell_\infty(X)$ by [1, Theorem 7.34]. In this case, (7) thus holds for $s = p = \frac{d}{2m}$, and it is obvious that this remains true, if we only assume that the marginal P_X has a density with respect to the uniform distribution that is bounded away from 0 and ∞. Finally, recall that the RKHSs of the Gaussian kernels are continuously embedded in all Sobolev spaces, and therefore they satisfy (7) for all $s \in (0,1]$, though in this case the appearing known constant C depends on both the kernel width and the used m.

As seen above, the case $s = p$ seems to be somewhat common, and in practice, the choice $q = 2$ is natural, since algorithmic solutions exist. Let us therefore restate Theorem 1 for these choices in a slightly simplified form:

Corollary 3 *Let k be a bounded measurable kernel on X with $\|k\|_\infty = 1$ and separable RKHS H. Moreover, let P be a distribution on $X \times [-M, M]$, where $M > 0$ is some constant. For $\nu = P_X$ assume that the extended sequence of eigenvalues of the integral operator (5) satisfies*

$$\mu_i(T_k) \le a\, i^{-\frac{1}{p}}\,, \qquad\qquad i \ge 1, \tag{13}$$

where $a \ge 16M^4$ and $p \in (0,1)$ are some constants. Moreover, assume that there exists a constants $C \ge 1$ such that

$$\|f\|_\infty \le C\,\|f\|_H^p \cdot \|f\|^{1-p}_{L_2(P_X)} \tag{14}$$

for all $f \in H$. Then, there exists a constant c only depending on p and C such that, for all $\lambda \in (0,1]$, $\tau > 0$, and $n \ge 1$, the least squares SVM described by (1) for $q = 2$ satisfies

$$\mathcal{R}_{L,P}(\widehat{f}_{D,\lambda}) - \mathcal{R}^*_{L,P} \le 9A_2(\lambda) + c\,\frac{a^p M^2 \tau}{\lambda^p n}$$

with probability P^n not less than $1 - 3e^{-\tau}$.

Our next goal is to investigate the influence of the regularization exponent q on the learning rates resulting from Theorem 1. To this end, we need the following result that translates the behavior of A_p into a behavior of A_q.

Lemma 4 *Let H be a separable RKHS over X that has a bounded measurable kernel, P be a distribution on $X \times [-M, M]$, and $p, q \ge 1$. Then for all $\lambda > 0$ and $\gamma \ge A_p(\lambda)$ we have*

$$A_q\big(\lambda^{q/p}\gamma^{1-q/p}\big) \le 2\gamma\,.$$

In particular, if there exist constants $c > 0$ and $\alpha > 0$ such that $A_p(\lambda) \le c\lambda^\alpha$, then for all $\lambda > 0$ we have

$$A_q(\lambda) \le 2\,c^{\frac{q}{q+\alpha(p-q)}}\,\lambda^{\frac{\alpha p}{q+\alpha(p-q)}}\,.$$

Let us illustrate the lemma above by assuming that the 2-approximation error function satisfies

$$A_2(\lambda) \le c\,\lambda^\beta\,, \qquad\qquad \lambda > 0, \tag{15}$$

where $c > 0$ and $\beta > 0$ are some constants.[1] Then Lemma 4 implies

$$A_q(\lambda) \le 2\,c^{\frac{q}{2\beta+q(1-\beta)}}\,\lambda^{\frac{2\beta}{2\beta+q(1-\beta)}} \tag{16}$$

for all $\lambda > 0$. Conversely, if (16) holds without the factor 2, then Lemma 4 yields $A_2(\lambda) \le 2c\lambda^\beta$ for all $\lambda > 0$. In other words, if A_2 has a polynomial behavior in λ, then this behavior completely determines the behavior of *all* A_q. In the following, it thus suffices to assume that the standard 2-approximation error function satisfies (15). Now, (15) also has a tight connection to interpolation spaces. Namely, [17], together with [19, Corollary 5.18] for the case $\beta = 1$, essentially showed the following result:

Theorem 5 *Let H be a separable RKHS over X that has a bounded measurable kernel, P be a distribution on $X \times [-M, M]$. Then (15) holds for some $\beta \in (0,1]$ if and only if*

$$f^*_P \in \big[L_2(P_X), H\big]_{\beta,\infty}\,.$$

With these preparations we can now establish learning rates for the learning method (1).

Corollary 6 *Consider Theorem 1 in the case $s = p$, and additionally assume that (15) holds. Define a sequence of regularization parameters (λ_n) by*

$$\lambda_n := n^{-\frac{2\beta+q(1-\beta)}{2\beta+2p}}\,. \tag{17}$$

Then there exists a constant $K \ge 1$ only depending on a, M, c, p, and q such that for all $\tau \ge 1$ and $n \ge 1$ the learning method described by (1) satisfies

$$\mathcal{R}_{L,P}(\widehat{f}_{D,\lambda_n}) - \mathcal{R}^*_{L,P} \;\le\; K\tau n^{-\frac{\beta}{\beta+p}}$$

with probability P^n not less than $1 - 3e^{-\tau n^{\frac{\beta p}{\beta+p}}}$.

An interesting observation from Corollary 6 is that the obtained learning rate is *not* affected by the choice of q. To understand the latter, recall that the *regularization path*, that is image of the function $\lambda \mapsto f_{D,\lambda}$, is also independent of q, see [19, Chapter 5.4] for a related discussion. In other words, for a given training set D, all learning methods considered in (1) produce the same decision function if we adjust $\lambda = \lambda(q)$ in a suitable way. From this perspective it is not surprising that the resulting learning rates are independent of q. However, note that the equality of the regularization path only suggests *equal* learning rates but it does not *prove* that the *optimal* learning rates are equal, since, in general, $\lambda(q)$ does depend on D.

The choice of λ_n in Corollary 6 ensures that the right hand side of the oracle inequality in Theorem 1 is asymptotically minimized. It thus yields the fastest rate we can expect from

[1] As shown in [19, Lemma 5.15] we may additionally assume $\beta \le 1$ since the case $\beta > 1$ implies $\mathcal{R}_{L,P}(0) = \mathcal{R}^*_{L,P}$, which is a rather uninteresting case for learning.

Theorem 1. Unfortunately, this choice of λ_n requires knowing p, and in most cases also β, which is unrealistic in almost all situations. However, [19, Chapter 7.4] shows that the best learning rates coming from oracle inequalities of our type can also be achieved *without* knowing p and β, if one splits D into a training and a validation set, and uses the validation set to identify a good value for λ from a suitable grid of candidates. To be more precise, let us recall the following definition from [19, Chapter 6.5] for the case $q = 2$:

Definition 7 *Let H be an RKHS over X and $\Lambda := (\Lambda_n)$ be a sequence of finite subsets $\Lambda_n \subset (0,1]$. Given a $D := ((x_1, y_1), \ldots, (x_n, y_n)) \in (X \times Y)^n$, we define*

$$\begin{aligned} D_1 &:= ((x_1, y_1), \ldots, (x_m, y_m)), \\ D_2 &:= ((x_{m+1}, y_{m+1}), \ldots, (x_n, y_n)), \end{aligned}$$

where $m := \lfloor n/2 \rfloor + 1$ and $n \geq 3$. Then use D_1 as a training set by computing the SVM decision functions

$$f_{D_1,\lambda} := \arg\min_{f \in H} \lambda \|f\|_H^2 + \mathcal{R}_{L,D_1}(f), \qquad \lambda \in \Lambda_n,$$

and use D_2 to determine λ by choosing a $\lambda_{D_2} \in \Lambda_n$ such that

$$\mathcal{R}_{L,D_2}(\widehat{f}_{D_1,\lambda_{D_2}}) = \min_{\lambda \in \Lambda_n} \mathcal{R}_{L,D_2}(\widehat{f}_{D_1,\lambda}).$$

Every learning method that produces the resulting decision functions $\widehat{f}_{D_1,\lambda_{D_2}}$ is called a training validation SVM with respect to Λ.

One can show that training validation SVMs are adaptive to p and β if the sets Λ_n are chosen in a specific way. Namely, a simple combination of the techniques used in the proof of [19, Theorem 7.24] and the oracle inequality established in Theorem 1 yields:

Theorem 8 *Consider Theorem 1 in the case $s = p$ and $q = 2$ and assume that (15) is satisfied for some $\beta \in (0,1]$. Moreover, assume that $\Lambda_n \subset (0,1]$ is an n^{-2}-net of $(0,1]$ for all $n \geq 1$, and assume further that the cardinality $|\Lambda_n|$ grows polynomially in n. Then the training validation SVM with respect to $\Lambda := (\Lambda_n)$ learns with rate $n^{-\frac{\beta}{\beta+p}}$.*

Theorem 8 shows that the training/validation approach is *adaptive* in the sense that it yields the same rates as Corollary 6 does *without* knowing p or β. In this respect, it is interesting to note that, for $q := \frac{2p}{1+p}$, the definition (17) is actually *independent* of β, i.e.,the training validation approach for this particular regularization exponent is superfluous as long as p is known. Modulo some extra logarithmic factors in the regularization term, this has already been observed by [13].

Our last goal in this work is to show that the learning rates obtained in Corollary 6 are asymptotically optimal. Since we need fractional powers of the operator T_k for the formulation of a corresponding result, we now briefly recall the latter. To this end, let $\mu := (\mu_i)_{i \in I}$ be the ordered (with geometric multiplicities) sequence of non-zero eigenvalues of the integral operator T_k defined by (5) and $(e_i) \subset L_2(\nu)$ be an orthonormal system (ONS) of corresponding eigenfunctions. For $\beta \in [0,1]$, we then define $T_k^\beta : L_2(\nu) \to L_2(\nu)$ by

$$T_k^\beta f := \sum_{i \in I} \mu_i^\beta \langle f, e_i \rangle e_i, \tag{18}$$

where $\langle f, e_i \rangle := \langle f, e_i \rangle_{L_2(\nu)}$ is the inner product in $L_2(\nu)$. It is known that these fractional powers of T_k are closely related to the 2-approximation error function. Indeed, for continuous kernels k on compact metric spaces X and $\nu = P_X$, Smale and Zhou showed in [17] that

$$f_P^* \in T_k^{\beta/2}\big(L_2(P_X)\big) \tag{19}$$

implies (15). Moreover, they further showed that the converse implication is true up to arbitrary small $\varepsilon > 0$ if $\operatorname{supp} P_X = X$. In other words, for continuous kernels on compact metric spaces, the fractional powers of T_k provide an accurate mean to describe the behavior of the 2-approximation error function.

In the following theorem and its proof we write $a_i \sim b_i$ for two sequences (a_i) and (b_i) of positive real numbers, if there exist constants $c_1, c_2 > 0$ such that $c_1 a_i \leq b_i \leq c_2 a_i$ for all $i \geq 1$. With these preparations we can now formulate the lower bounds:

Theorem 9 *Let ν be a distribution on X, and k be a bounded measurable kernel on X with $\|k\|_\infty = 1$ and separable RKHS H. Assume that there exists a $p \in (0,1)$ such that the extended series of eigenvalues of the integral operator defined by (5) satisfies*

$$\mu_i(T_k) \sim i^{-\frac{1}{p}}. \tag{20}$$

Let $\beta \in (0,1]$ and assume that there exists a constant $c > 0$ such that

$$\|T_k^{\beta/2} f\|_\infty \leq c \|f\|_{L_2(\nu)}, \qquad f \in L_2(\nu). \tag{21}$$

Then, for all $M > 0$, there exist constants $\delta_0 > 0$, $c_1, c_2 > 0$, and $C > 0$ such that for all learning algorithms $\mathcal{A}$ there exists a distribution P on $X \times [-M, M]$ with $P_X = \nu$ and

$$f_P^* \in \frac{M}{4c} T_k^{\beta/2}\big(B_{L_2(\nu)}\big)$$

such that for all $\tau > 0$ and $n \geq 1$ we have

$$\begin{aligned} P^n\Big(D : \mathcal{R}_{L,P}(f_D) - \mathcal{R}_{L,P}^* \geq C\tau n^{-\frac{\beta}{\beta+p}}\Big) \\ \geq \begin{cases} \delta_0 & \text{if } \tau < 1 \\ c_1 e^{-c_2 \tau n^{\frac{p}{\beta+p}}} & \text{if } \tau \geq 1, \end{cases} \end{aligned}$$

where f_D is the decision function produced by $\mathcal{A}$ for a given training set D.

For $\alpha > 0$ and $\tau_n := \tau n^{-\alpha}$, Theorem 9 shows that the probability P^n of

$$\mathcal{R}_{L,P}(f_D) - \mathcal{R}_{L,P}^* \leq C\tau n^{-\frac{\beta}{\beta+p} - \alpha}$$

does not exceed $1 - \delta_0$ if $n > \tau^{1/\alpha}$. In this sense, Theorem 9 shows that the rates obtained in Corollary 6 are asymptotically optimal for continuous kernels on compact metric spaces.

To illustrate the assumption (21), we assume again that k is a continuous kernel on a compact metric space X. Then the proof of [9, Theorem 4.1] shows that the image of $T_k^{\beta/2}$ is continuously embedded into the real interpolation space $[L_2(\nu), H]_{\beta,\infty}$, and hence (21) is satisfied if $[L_2(\nu), H]_{\beta,\infty}$

is continuously embedded in $\ell_\infty(X)$. Note that in view of (11) and (11) this is slightly more than assuming (7) for $s = \beta$. To be more precise, let us assume that we have a $0 < p < 1$ such that $[L_2(\nu), H]_{p,1}$ is continuously embedded in $\ell_\infty(X)$, and that (20) is satisfied. As mentioned earlier, the first assumption then implies (7), while the second clearly implies (6), and hence Corollary 6 yields the learning rate

$$n^{-\frac{\beta}{\beta+p}}$$

whenever (19) is satisfied for some $\beta \in (0, 1]$. Moreover, (12) together with the assumption that $[L_2(\nu), H]_{p,1}$ is continuously embedded in $\ell_\infty(X)$ implies that $[L_2(\nu), H]_{\beta,\infty}$ is continuously embedded in $\ell_\infty(X)$ whenever $\beta \in (p, 1]$, and hence Theorem 9 shows that the learning rate obtained from Corollary 6 is asymptotically optimal for such β.

An interesting observation from this discussion is that the learning methods defined by (1) achieve the asymptotically optimal rates for *all* choices of regularization exponents q. In other words, the choice of q has no influence on the learning rates, which in turn means that q can be solely chosen on the basis of algorithmic considerations. Here we emphasize that these conclusions can, of course, only be made because we showed that the obtained learning rates are optimal.

To give a little more concrete example, let us now consider the case $H = W^m(X)$ for some $m > d/2$. In addition, we assume that $P_X = \nu$ has a Lebesgue density that is bounded away from 0 and ∞, and hence we have

$$B^{\beta m}_{2,\infty}(X) = \big[L_2(P_X), W^m(X)\big]_{\beta,\infty}\,.$$

Therefore, Corollary 6 yields the learning rate $n^{-\frac{2s}{2s+d}}$, whenever $f^*_P \in B^s_{2,\infty}(X)$, where $s := \beta m \in (0, m]$. Conversely, [1, Theorem 7.34] guarantees that $B^s_{2,\infty}(X)$ is continuously embedded into $\ell_\infty(X)$ for all $s > d/2$, and hence Theorem 9 shows that this rate is asymptotically optimal for such s. In other words, using $H = W^m(X)$, the learning methods defined by (1) can estimate *all* regression functions in the Besov scale $B^s_{2,\infty}(X)$, $s \in (d/2, m]$, with asymptotically optimal rate, which, in addition, is always faster than $n^{-1/2}$. Moreover, by using a validation set to determine the regularization parameter, these methods are adaptive, i.e., they do not need to know s. Finally, the covered scale $B^s_{2,\infty}(X)$, $s \in (d/2, m]$, can be arbitrarily large by choosing m large enough, that is, by choosing a sufficiently smooth kernel.

In interesting observation from this discussion is that it is safe to over-specify the smoothness of the target function. Namely, if we are confident that the regression satisfies $f^*_P \in B^s_{2,\infty}(X)$ for some $s > d/2$, then we can learn this function with the optimal rate whenever we pick an $m > s$. In other words, from an asymptotic point of view we only need to know a crude upper bound m on the assumed smoothness s, and if we have such a bound, it it *not* necessary to fine tune the choice of m. In this sense, the "learning the kernel problem" is somewhat implicitly solved by the SVM.

Finally, we like to emphasize that, for $\beta < 1$, the obtained results implicitly assume that the used RKHS is infinite dimensional. To be more precise, let us assume that we have a finite dimensional RKHS H. Then we have either $f^*_P \in H$ or $f^*_P \notin H$. However, in the first case, we find $A_2(\lambda) \leq \|f^*_P\|^2_H \lambda$ for all $\lambda > 0$ by [19, Corollary 5.18], and hence $\beta < 1$ actually corresponds to the situation $f^*_P \notin H$. However, since H is finite dimensional, it is a closed subspace of $L_2(P_X)$, and hence we conclude

$$\inf_{f\in H} \mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P} = \inf_{f\in H} \|f - f^*_P\|^2_{L_2(P_X)} > 0\,.$$

Consequently, $A_2(\lambda)$ does not converge to 0 for $\lambda \to 0$, and therefore (15) cannot be satisfied for $\beta > 0$. In other words, for finite dimensional RKHSs we either have $\beta = 1$ or $\beta = 0$, and hence the interesting part of our results, namely the case $0 < \beta < 1$, never occurs. In particular, this is true if we use e.g. linear or polynomial kernels.

3 Proofs of the Upper Bounds

Since the proof of Theorem 1 is based [19, Theorem 7.20] we need to recall some notations related to the latter theorem. Let us begin by writing

$$r^* := \inf_{f\in H} \lambda\|f\|^q_H + \mathcal{R}_{L,P}(\widehat{f}\,) - \mathcal{R}^*_{L,P}\,. \tag{22}$$

Moreover, for $r > r^*$, we write $\mathcal{F}_{r,\lambda}$ for the set

$$\{f \in H \,:\, \lambda\|f\|^q_H + \mathcal{R}_{L,P}(\widehat{f}\,) - \mathcal{R}^*_{L,P} \leq r\}\,.$$

Finally, we need the set

$$\mathcal{H}_{r,\lambda} := \{L \circ \widehat{f} - L \circ f^*_P \,:\, f \in \mathcal{F}_{r,\lambda}\}\,,$$

where $L \circ g$ denotes the function $(x, y) \mapsto L(y, g(x))$. Furthermore, we obviously have

$$L(y, t) \leq 4M^2\,, \qquad y, t \in [-M, M]\,, \tag{23}$$

and a well known variance bound for the least squares loss, see e.g. [19, Example 7.3], shows

$$\mathbb{E}_P\big(L \circ \widehat{f} - L \circ f^*_P\big)^2 \leq 16M^2\mathbb{E}_P(L \circ \widehat{f} - L \circ f^*_P)$$

for all functions $f : X \to \mathbb{R}$. Last but not least, it is a simple exercise to show that the least squares loss restricted to $[-M, M]$ is Lipschitz continuous, that is

$$\big|L(y, t) - L(y, t')\big| \leq 4M|t - t'| \tag{24}$$

for all $y \in [-M, M]$ and $t, t' \in [-M, M]$.

Let us further recall the concept of empirical Rademacher averages. To this end, $(\Theta, \mathcal{C}, \nu)$ be a probability space, and $\varepsilon_1, \ldots, \varepsilon_n$ be a be a Rademacher sequence, that is, a sequence of independent random variables $\varepsilon_i : \Theta \to \{-1, 1\}$ with $\nu(\varepsilon_i = 1) = \nu(\varepsilon_i = -1) = 1/2$ for all $i = 1, \ldots, n$. For a non-empty set $\mathcal{H}$ consisting of functions that map from a measurable space Z to $\mathbb{R}$ and $D := (z_1, \ldots, z_n) \in Z^n$, the n-th empirical Rademacher average of $\mathcal{H}$ is then defined by

$$\mathrm{Rad}_D(\mathcal{H}, n) := \mathbb{E}_\nu \sup_{h\in\mathcal{H}} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i h(z_i)\right|.$$

Besides Rademacher averages we also need the following peeling argument:

Theorem 10 *Let $(Z, \mathcal{A}, P)$ be a probability space, (T, d) be a separable metric space, $h : T \to [0, \infty)$ be a continuous function, and $(g_t)_{t\in T}$ be a family of measurable functions $g_t : Z \to \mathbb{R}$ such that $t \mapsto g_t(z)$ is continuous for all $z \in Z$. We define $r^* := \inf\{h(t) : t \in T\}$. Moreover, let $\alpha \in (0, 1)$ be a constant and $\varphi : (r^*, \infty) \to [0, \infty)$ be a function such that $\varphi(2r) \le 2^\alpha \varphi(r)$ and*

$$\mathbb{E}_{z\sim P} \sup_{\substack{t\in T \\ h(t)\le r}} |g_t(z)| \;\le\; \varphi(r)$$

for all $r > r^*$. Then, for all $r > r^*$, we have

$$\mathbb{E}_{z\sim P} \sup_{t\in T} \frac{g_t(z)}{h(t) + r} \;\le\; \frac{2 + 2^\alpha}{2 - 2^\alpha} \cdot \frac{\varphi(r)}{r}\,.$$

Proof: This theorem can be proven by an almost literal repetition of [19, Theorem 7.7]. ∎

With these preparations, [19, Theorem 7.20] then becomes:

Theorem 11 *For fixed $q \in [1, \infty)$ we consider the learning method (1). Assume that for fixed $n \ge 1$ there exists a constant $\alpha \in (0, 1)$ and a function $\varphi_{n,\lambda} : [0, \infty) \to [0, \infty)$ such that $\varphi_{n,\lambda}(2r) \le 2^\alpha \varphi_{n,\lambda}(r)$ and*

$$\mathbb{E}_{D\sim P^n} \mathrm{Rad}_D(\mathcal{H}_{r,\lambda}, n) \;\le\; \varphi_{n,\lambda}(r) \tag{25}$$

for all $r > r^*$. Moreover, fix an $f_0 \in H$ and a $B_0 \ge 4M^2$ such that $\|L \circ f_0\|_\infty \le B_0$. Then, for all fixed $\tau > 0$ and $r > r^*$ satisfying

$$r > \max\left\{c(\alpha) \cdot \varphi_{n,\lambda}(r), \frac{1152M^2\tau}{n}, \frac{5B_0\tau}{n}\right\},$$

where $c(\alpha) := 8 \cdot \frac{2+2^\alpha}{2-2^\alpha}$, we have with probability P^n not less than $1 - 3e^{-\tau}$ that

$$\begin{aligned}&\lambda\|f_{D,\lambda}\|_H^q + \mathcal{R}_{L,P}(\widehat{f}_{D,\lambda}) - \mathcal{R}^*_{L,P}\\ \le\;& 6\big(\lambda\|f_0\|_H^q + \mathcal{R}_{L,P}(f_0) - \mathcal{R}^*_{L,P}\big) + 3r\,.\end{aligned}$$

Proof: As noted in (4), the least squares loss $L : [-M, M] \times \mathbb{R} \to [0, \infty)$ can be clipped at M in the sense of [19, Definition 2.22]. In addition, the learning method defined by (1) is a clipped regularized empirical risk minimizer in the sense of [19, Definition 7.18]. Moreover, the supremum bound (7.35) and the variance bound (7.36) in [19] are satisfied for $B := 4M^2$ and $\vartheta = 1$, $V = 16M^2$, respectively. Replacing the peeling step in the proof of [19, Theorem 7.20] in the middle of page 262 by the refined peeling result of Theorem 10, we then obtain the assertion by the otherwise unmodified proof of [19, Theorem 7.20]. ∎

Before we can prove Theorem 1 we finally need the following simple lemma, which ensures the existence of unique solutions of the "infinite-sample version" of (1).

Lemma 12 *Let P be a distribution on $X \times [-M, M]$ and k be a bounded measurable kernel on X with separable RKHS H. Then, for all $q \ge 1$ and $\lambda > 0$, there exists a unique function $f_{P,\lambda} \in H$ that minimizes*

$$f \mapsto \lambda\|f\|_H^q + \mathcal{R}_{L,P}(f)\,.$$

Proof: Since the function $f \mapsto \|f\|_H^q$ is convex and the function $f \mapsto \mathcal{R}_{L,P}(f)$ is strictly convex, the uniqueness follows. The existence can be shown by repeating the proof of [19, Theorem 5.2] for $q \ne 2$. ∎

Proof of Theorem 1: For $f \in \mathcal{F}_{r,\lambda}$, a simple calculation shows

$$\lambda\|f\|_H^q \;\le\; \lambda\|f\|_H^q + \mathcal{R}_{L,P}(\widehat{f}\,) - \mathcal{R}^*_{L,P} \;\le\; r\,,$$

and hence we conclude that $\|f\|_H \le (r/\lambda)^{1/q}$. In other words, we have $\mathcal{F}_{r,\lambda} \subset (r/\lambda)^{1/q} B_H$. By [19, Corollary 7.31], the second line on page 276 of [19], and Carl's inequality [7, Theorem 3.1.1], see also the proof of Theorem 15 for this argument, we further see that the eigenvalue assumption (6) implies

$$\mathbb{E}_{D_X\sim P_X^n} e_i(\mathrm{id} : H \to L_2(D_X)) \le c_p\sqrt{a}\, i^{-\frac{1}{2p}},$$

where $c_p \ge 1$ is a constant only depending on p. Here, the entropy numbers $e_i(\mathrm{id} : H \to L_2(D_X))$ are defined with respect to the space $L_2(D_X)$, where D_X denotes the empirical measure with respect to $D_X = (x_1, \ldots, x_n)$ sampled from P^n. The Lipschitz continuity (24) of the restricted least squares loss thus yields

$$\mathbb{E}_{D\sim P^n} e_i(\mathcal{H}_{r,\lambda}, L_2(D)) \le 8c_pM\Big(\frac{r}{\lambda}\Big)^{1/q}\sqrt{a}\, i^{-\frac{1}{2p}}\,.$$

Moreover, for $f \in \mathcal{F}_{r,\lambda}$, we have

$$\mathbb{E}_P(L \circ \widehat{f} - L \circ f^*_{L,P})^2 \le 16M^2 r\,,$$

and consequently [19, Theorem 7.16] applied to $\mathcal{H} := \mathcal{H}_{r,\lambda}$ shows that (25) is satisfied for

$$\begin{aligned}\varphi_{n,\lambda}(r) \;:=\; \max\Bigg\{&C_1(p)C\Big(\frac{r}{\lambda}\Big)^{\frac{p}{q}}(16M^2r)^{\frac{1-p}{2}}n^{-\frac{1}{2}},\\ &C_2(p)C^{\frac{2}{1+p}}\Big(\frac{r}{\lambda}\Big)^{\frac{2p}{(1+p)q}}(4M^2)^{\frac{1-p}{1+p}}n^{-\frac{1}{1+p}}\Bigg\},\end{aligned}$$

where $C_1(p)$ and $C_2(p)$ are the constants appearing in [19, Theorem 7.16] and $C := 8^p c_p^p M^p a^{p/2}$. Moreover, $0 < p < 1$ and $q \ge 1$ imply $q \ge \frac{2p}{1+p}$, and since $\frac{2p}{(1+p)q} - \frac{p}{q} = \frac{p(1-p)}{(1+p)q}$, we conclude that

$$\alpha := \frac{p}{q} + \frac{1-p}{2} \ge \frac{2p}{(1+p)q}\,.$$

In turn, this inequality can be used to show that $\varphi_{n,\lambda}(2r) \le 2^\alpha\varphi_{n,\lambda}(r)$ for all $r > r^*$, and since $0 < p < 1$ and $q \ge 1$ further imply $\alpha \in (0, 1)$, we see that $\varphi_{n,\lambda}$ satisfies the assumptions of Theorem 11. Furthermore, a simple yet tedious calculation shows that there exists a constant $c_{p,q}$ only depending on p and q such that $r \ge 8 \cdot \frac{2+2^\alpha}{2-2^\alpha}\varphi_{n,\lambda}(r)$ is satisfied if

$$r \ge c_{p,q}\left(\frac{a^{pq}M^{2q}}{\lambda^{2p}n^q}\right)^{\frac{1}{q-2p+pq}}\,.$$

Let us now fix $f_0 := f_{P,\lambda}$, where $f_{P,\lambda}$ is the function considered in Lemma 12. To find a B_0, we first observe that

$\|L \circ f_0\|_\infty = \|L \circ f_{P,\lambda}\|_\infty \leq 2M^2 + 2\|f_{P,\lambda}\|_\infty^2$. Moreover, we have

$$\begin{aligned} & \int |f_{P,\lambda}(x)|^2 \, dP_X(x) \\ \leq & \int 2|f_{P,\lambda}(x) - y|^2 + 2|y|^2 \, dP(x,y) \\ \leq & \; 2\mathcal{R}_{L,P}(f_{P,\lambda}) + 2M^2 \\ \leq & \; 4M^2 \,, \end{aligned}$$

where in the last step we used $\mathcal{R}_{L,P}(f_{P,\lambda}) \leq \mathcal{R}_{L,P}(0) \leq M^2$. Consequently, our assumption (7) yields

$$\|f_{P,\lambda}\|_\infty \leq 2^{1-s} M^{1-s} C \, \|f_{P,\lambda}\|_H^s \,,$$

and from this we conclude

$$\|L \circ f_{P,\lambda}\|_\infty \leq 4M^2 + 8C^2 M^{2-2s} \Big(\frac{A_q(\lambda)}{\lambda} \Big)^{\frac{2s}{q}},$$

where we also used the estimate $\lambda \|f_{P,\lambda}\|_H^q \leq A_q(\lambda)$. Using Theorem 11 now yields the assertion. ∎

Proof of Lemma 4: By Lemma 12 there exists a function $f_{P,\lambda} \in H$ that satisfies

$$\lambda \|f_{P,\lambda}\|^p + \mathcal{R}_{L,P}(f_{P,\lambda}) - \mathcal{R}^*_{L,P} = A_p(\lambda) \leq \gamma \,,$$

and since $\mathcal{R}_{L,P}(f_{P,\lambda}) - \mathcal{R}^*_{L,P} \geq 0$, we find $\lambda \|f_{P,\lambda}\|^p \leq \gamma$. For $\kappa := \lambda^{q/p} \gamma^{1-q/p}$ we hence obtain

$$\begin{aligned} A_q(\kappa) & \leq \kappa \|f_{P,\lambda}\|^q + \mathcal{R}_{L,P}(f_{P,\lambda}) - \mathcal{R}^*_{L,P} \\ & \leq \lambda^{q/p} \gamma^{1-q/p} \cdot \gamma^{q/p} \lambda^{-q/p} + \gamma \\ & = 2\gamma \,, \end{aligned}$$

i.e., we have shown the first assertion. The second assertion now follows from some simple algebraic transformations. ∎

Proof of Corollary 6: Since Assumption (15) implies (16) we find

$$A_q(\lambda_n) \leq 2 \, c^{\frac{q}{2\beta + q(1-\beta)}} \, n^{-\frac{\beta}{\beta+p}} \,.$$

Moreover, we have

$$\lambda_n^{2p} n^q = n^{\frac{\beta(q-2p+pq)}{\beta+p}} \,,$$

and from this it is easy to conclude that the second term in the oracle inequality of Theorem 1 reduces to

$$c_{p,q} \big(a^{pq} M^{2q} \big)^{\frac{1}{q-2p+pq}} n^{-\frac{\beta}{\beta+p}} \,.$$

In addition, our first estimate shows

$$\frac{A_q(\lambda_n)}{\lambda_n} \leq 2 \, c^{\frac{q}{2\beta+q(1-\beta)}} \, n^{\frac{q(1-\beta)}{2\beta+2p}} \,,$$

and hence the third term of Theorem 1 can be estimated by

$$\begin{aligned} & \frac{120 C^2 M^{2-2p} \tau}{n} \Big(\frac{A_q(\lambda)}{\lambda} \Big)^{\frac{2p}{q}} \\ \leq & \; 480 \, c^{\frac{2p}{2\beta+q(1-\beta)}} \, C^2 M^{2-2p} \tau n^{-\frac{\beta(1+p)}{\beta+p}} \,. \end{aligned}$$

By considering $\tau_n := \tau n^{\frac{\beta p}{\beta+p}}$ in Theorem 1, the assertion now follows. ∎

4 Proof of the Lower Bound

The core of the proof of Theorem 9 is based on the following reformulation of [21, Theorem 2.2]:

Theorem 13 *Let ν be a distribution on X and $\Theta \subset L_2(\nu)$ be a subset such that $\|f\|_\infty \leq M/4$ for all $f \in \Theta$ and some $M > 0$. In addition, assume that there exists an $r \in (0,1)$ such that*

$$e_i(\Theta, L_2(P_X)) \sim i^{-1/r} \,.$$

Then there exist constants $\delta_0 > 0$, $c_1, c_2 > 0$ and a sequence (ε_n) with

$$\varepsilon_n \sim n^{-\frac{2}{2+r}}$$

such that for all learning methods $\mathcal{A}$ there exists a distribution P on $X \times [-M, M]$ satisfying $P_X = \nu$ and $f^*_P \in \Theta$ such that for all $\varepsilon > 0$ and $n \geq 1$ we have

$$\begin{aligned} & P^n \big(D : \mathcal{R}_{L,P}(f_D) - \mathcal{R}^*_{L,P} \geq \varepsilon \big) \\ & \qquad \geq \begin{cases} \delta_0 & \text{if } \varepsilon < \varepsilon_n \\ c_1 e^{-c_2 \varepsilon n} & \text{if } \varepsilon \geq \varepsilon_n \,, \end{cases} \end{aligned}$$

where f_D is the decision function produced by $\mathcal{A}$ for a given training set D.

Proof: The proof of [21, Theorem 2.2] identifies the "bad" distribution P with the help of [21, Theorem 2.1]. Since the latter result provides a distribution P for which the lower bound on $P^n(D : \mathcal{R}_{L,P}(f_D) - \mathcal{R}^*_{L,P} \geq \varepsilon)$ holds for *all* $\varepsilon > 0$ and $n \geq 1$, the same is true in [21, Theorem 2.2]. Analogously, we check that the compactness assumption on Θ made in [21, Theorem 2.2] is superfluous. ∎

Our next goal is to apply Theorem 13. To this end, we need to translate the eigenvalue assumption (20) into an assumption on the behavior of entropy numbers of a suitable set Θ. The key step in this direction is Lemma 14 below, which relates the fractional powers of the integral operator T_k to some spaces of functions. For its formulation, we need to introduce a weighted sequence space. More precisely, given a set of integers $I \subset \mathbb{N}$ and a decreasing sequence $\mu := (\mu_i)_{i \in I}$ of strictly positive numbers, we define

$$\|b\|^2_{\ell_2(\mu^{-1})} := \sum_{i \in I} \frac{b_i^2}{\mu_i}$$

for all $b := (b_i)_{i \in I} \subset \mathbb{R}$. Then it is easy to see that $\|\cdot\|_{\ell_2(\mu^{-1})}$ is a Hilbert space norm on the set

$$\ell_2(\mu^{-1}) := \big\{ (b_i)_{i \in I} : \|(b_i)\|_{\ell_2(\mu^{-1})} < \infty \big\} \,.$$

Moreover, $\mu_i \leq \mu_1$ for all $i \in I$ implies

$$\|b\|_{\ell_2(I)} \leq \mu_1 \|b\|_{\ell_2(\mu^{-1})}$$

for all $b \in \ell_2(\mu^{-1})$.

With these preparations we can now investigate the fractional powers T_k^β.

Lemma 14 *Let X be a measurable space, ν be a distribution on X, and k be a bounded measurable kernel on X that has a separable RKHS H. Let $\mu := (\mu_i)_{i \in I}$ be the ordered (with geometric multiplicities) sequence of non-zero eigenvalues of the integral operator T_k defined by (5) and*

$(e_i) \subset L_2(\nu)$ be an ONS of corresponding eigenfunctions. For $\beta \in [0,1]$ we define $T_k^\beta : L_2(\nu) \to L_2(\nu)$ by

$$T_k^\beta f := \sum_{i\in I} \mu_i^\beta \langle f, e_i\rangle e_i\,, \qquad f \in L_2(\nu),$$

where $\langle f, e_i\rangle := \langle f, e_i\rangle_{L_2(\nu)}$ is the inner product in $L_2(\nu)$. Moreover, we define

$$H_\beta := \left\{ \sum_{i\in I} b_i e_i : (b_i) \in \ell_2(\mu^{-\beta}) \right\}$$

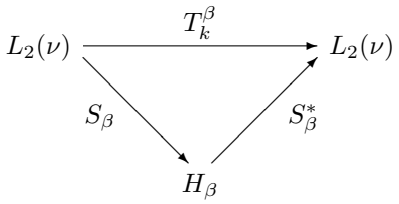
and equip this space with the Hilbert space norm

$$\| \sum_{i\in I} b_i e_i\|_{H_\beta} := \|(b_i)\|_{\ell_2(\mu^{-\beta})}\,.$$

Then we have $H_\beta \subset L_2(\nu)$, and $T_k^\beta f \in H_\beta$ with

$$\|T_k^\beta f\|_{\ell_2(\mu^{-\beta})} \leq \mu_1^{\beta/2} \|f\|_{L_2(\nu)}$$

for all $f \in L_2(\nu)$. Moreover, the bounded linear operator $S_\beta : L_2(\nu) \to H_\beta$ defined by $S_\beta f := T_k^\beta f$ satisfies $S_\beta^* = \mathrm{id} : H_\beta \to L_2(\nu)$, and hence we have



Furthermore, $f \mapsto T_k^{\beta/2} f$ defines an isometric isomorphism between the closed supspace $H_0 = \overline{\mathrm{span}\{e_i : i \in I\}}$ of $L_2(\nu)$ and H_β. Finally, if H is dense in $L_2(\nu)$ and $\mathrm{id} : H \to L_2(\nu)$ is injective, then $H_0 = L_2(\nu)$.

Proof: The definition of T_k^β is the standard way to define fractional powers of operators and is known to be independent of the choice of the ONS of eigenfunctions. Moreover, H_β is clearly a Hilbert space since the definition ensures that it is isometrically isomorphic to $\ell_2(\mu^{-\beta})$. From this it is easy to see that the inclusion $\ell_2(\mu^{-1}) \subset \ell_2$ implies $H_\beta \subset L_2(\nu)$. Let us now fix an $f \in L_2(\nu)$. Then we have $\|(\langle f, e_i\rangle)\|_{\ell_2(I)}^2 \leq \|f\|_{L_2(\nu)}^2$ by Bessel's inequality, and hence we obtain

$$\|T_k^\beta f\|_{H_\beta}^2 = \sum_{i\in I} \frac{\mu_i^{2\beta} |\langle f, e_i\rangle|^2}{\mu_i^\beta} \leq \mu_1^\beta \|f\|_{L_2(\nu)}^2.$$

Let us now show that $S_\beta^* = \mathrm{id} : H_\beta \to L_2(\nu)$. To this end, we fix an $f \in L_2(\nu)$ and an $h \in H_\beta$ with $h = \sum_{i\in I} b_i e_i$. Then the definition of the norm $\|\cdot\|_{H_\beta}$ and the corresponding inner product yields

$$\begin{aligned} \langle h, S_\beta f\rangle_{H_\beta} &= \sum_{i\in I} \frac{b_i \mu_i^\beta \langle f, e_i\rangle_{L_2(\nu)}}{\mu_i^\beta} \\ &= \sum_{i\in I} b_i \langle f, e_i\rangle_{L_2(\nu)} \\ &= \langle h, f\rangle_{L_2(\nu)}\,, \end{aligned}$$

i.e., we have shown $S_\beta^* = \mathrm{id} : H_\beta \to L_2(\nu)$. From this the diagram easily follows. Let us show the claimed isometric isomorphism. To this end, we observe that (e_i) is an orthonormal basis (ONB) of H_0, and hence Parseval's identity yields

$$\|T_k^{\beta/2} f\|_{H_\beta}^2 = \sum_{i\in I} \frac{\mu_i^\beta |\langle f, e_i\rangle|^2}{\mu_i^\beta} = \|f\|_{H_0}^2\,,$$

i.e., $f \mapsto T_k^{\beta/2} f$ is an isometric injection from H_0 to H_β. To show the surjectivity, we fix an $h \in H_\beta$ with $h = \sum_{i\in I} b_i e_i$. For $a_i := \mu_i^{-\beta/2} b_i$ we then have

$$\sum_{i\in I} a_i^2 = \sum_{i\in I} \mu_i^{-\beta} b_i^2 < \infty\,,$$

and hence we can define $f := \sum_{i\in I} a_i e_i \in H_0$. An easy calculation then shows

$$T_k^{\beta/2} f = \sum_{i\in I} \mu_i^{\beta/2} a_i e_i = \sum_{i\in I} b_i e_i = h\,,$$

i.e., we have shown the surjectivity. By [19, Theorem 4.26] and the denseness of H in $L_2(\nu)$ we finally see that the operator $S : L_2(\nu) \to H$ defined by $Sf := T_k f$ is injective, and hence so is T_k by the assumed injectivity of $\mathrm{id} : H \to L_2(\nu)$. Consequently, (e_i) is an ONB of $L_2(\nu)$. ∎

With the help of the lemma above we can now describe the relationsship between the eigenvalues of T_k and the entropy numbers of $\mathrm{id} : H_\beta \to L_2(\nu)$.

Theorem 15 *Let X be a measurable space, ν be a distribution on X, and k be a bounded measurable kernel on X that has a separable RKHS H. Then for all $q > 0$ there exists a constant $c_q > 0$ only depending on q such that for all $m \geq 1$ and $\beta \in (0,1]$ we have*

$$\begin{aligned} &\sup_{i\leq m} i^{1/q} e_i(\mathrm{id} : H_\beta \to L_2(\nu)) \\ \leq\ & c_q \sup_{i\leq m} i^{1/q} \mu_i^{\beta/2} (T_k : L_2(\nu) \to L_2(\nu)) \end{aligned}$$

and

$$\mu_m^{\beta/2}(T_k : L_2(\nu) \to L_2(\nu)) \leq 2 e_m(\mathrm{id} : H_\beta \to L_2(\nu))\,.$$

Moreover, given a fixed $p \in (0,1)$, we have

$$e_i(\mathrm{id} : H_\beta \to L_2(\nu)) \sim i^{-\frac{\beta}{2p}}$$

if and only if $\mu_i(T_k) \sim i^{-\frac{1}{2p}}$.

Proof: The complete proof would require introducing some heavy machinery from functional analysis, in particular from the theory of so-called s-numbers, see [14]. Since this is clearly out of the scope of this paper, we refer [19, Appendix 5.2] for a brief summary of these techniques. Let us now show the inequalities. To this end, we first observe that

$$\mu_i^\beta(T_k) = \mu_i(T_k^\beta)$$

by construction. Furthermore, from the last two pages of [19, Appendix 5.2] we conclude that

$$\mu_i(T_k^\beta) = s_i(T_k^\beta) = s_i^2(S_\beta^*) = a_i^2(S_\beta^*)$$

where $s_i(\cdot)$ and $a_i(\cdot)$ denote the ith singular and the ith approximation number defined in [19, (A.25)] and [19, (A.29)], respectively. From Carl's inequality, see [7, Theorem 3.1.1], we then obtain the first inequality. The second inequality follows from the relation

$$a_m(R : H_1 \to H_2) \leq 2e_m(R : H_1 \to H_2)$$

that holds for all bounded linear operators R between Hilbert spaces H_1 and H_2, see [7, p. 120].

Let us now assume that $\mu_i(T_k) \sim i^{-\frac{1}{p}}$. Then the upper bound on the entropy numbers easily follows from the first inequality for $q := p/\beta$, while the lower bound on the entropy numbers is a trivial consequence of the second inequality. Conversely, if we assume

$$e_i(\mathrm{id} : H_\beta \to L_2(\nu)) \sim i^{-\frac{\beta}{2p}},$$

then the second inequality immediately gives the desired upper bound on the eigenvalues. To establish the lower bound, let $j, m \geq 1$ be integers, where m will be specified later. For suitable constants $c_1, c_2, c_3 > 0$, we then have

$$\begin{aligned}
(m \cdot j)^{\frac{\beta}{2p}} &\leq c_1 (m \cdot j)^{\frac{\beta}{p}} e_{m \cdot j}(\mathrm{id} : H_\beta \to L_2(\nu)) \\
&\leq c_2 \sup_{i \leq m \cdot j} i^{\frac{\beta}{p}} \mu_i^{\beta/2}(T_k) \\
&\leq c_2 \sup_{i \leq j} i^{\frac{\beta}{p}} \mu_i^{\beta/2}(T_k) + c_2 \sup_{j \leq i \leq m \cdot j} i^{\frac{\beta}{p}} \mu_i^{\beta/2}(T_k) \\
&\leq c_3\, j^{\frac{\beta}{2p}} + c_2 (m \cdot j)^{\frac{\beta}{p}} \mu_j^{\beta/2}(T_k),
\end{aligned}$$

where in the last step we used the already established upper bound on the eigenvalues. If we now *fix* an m such that $m^{\frac{\beta}{2p}} \geq 1 + c_3$, we obtain the desired lower bound by some simple algebraic transformations. ∎

Proof of Theorem 9: Let us write

$$\Theta := \frac{M}{4c} S_\beta^*(B_{H_\beta}) = \frac{M}{4c} T_k^{\beta/2}\big(B_{L_2(\nu)}\big).$$

Then we have $\|f\|_\infty \leq M/4$ for all $f \in \Theta$ by our assumption on $T_k^{\beta/2}$. Moreover, the assumption on the eigenvalues together with Theorem 15 implies

$$e_i(\mathrm{id} : H_\beta \to L_2(P_X)) \sim i^{-\frac{\beta}{2p}},$$

and hence Θ satisfies the assumptions of Theorem 13 for $r := \frac{2p}{\beta}$. Consequently the assertion follows by considering $\varepsilon := \tau\varepsilon_n$ in Theorem 13. ∎

References

[1] R. A. Adams and J. J. F. Fournier. *Sobolev Spaces*. Academic Press, New York, 2nd edition, 2003.

[2] P. L. Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE Trans. Inf. Theory*, 44:525–536, 1998.

[3] C. Bennett and R. Sharpley. *Interpolation of Operators*. Academic Press, Boston, 1988.

[4] M. Sh. Birman and M. Z. Solomyak. Piecewise-polynomial approximations of functions of the classes W_p^α (Russian). *Mat. Sb.*, 73:331–355, 1967.

[5] G. Blanchard, O. Bousquet, and P. Massart. Statistical performance of support vector machines. *Ann. Statist.*, 36:489–531, 2008.

[6] A. Caponnetto and E. De Vito. Optimal rates for regularized least squares algorithm. *Found. Comput. Math.*, 7:331–368, 2007.

[7] B. Carl and I. Stephani. *Entropy, Compactness and the Approximation of Operators*. Cambridge University Press, Cambridge, 1990.

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

[9] F. Cucker and D. X. Zhou. *Learning Theory: An Approximation Theory Viewpoint*. Cambridge University Press, Cambridge, 2007.

[10] E. De Vito, A. Caponnetto, and L. Rosasco. Model selection for regularized least-squares algorithm in learning theory. *Found. Comput. Math.*, 5:59–85, 2005.

[11] D. E. Edmunds and H. Triebel. *Function Spaces, Entropy Numbers, Differential Operators*. Cambridge University Press, Cambridge, 1996.

[12] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, New York, 2002.

[13] S. Mendelson and J. Neeman. Regularization in kernel learning. Technical report, Australian National University, `http://wwwmaths.anu.edu.au/~mendelso/papers/MN29-02-08.pdf`, 2008.

[14] A. Pietsch. *Eigenvalues and s-Numbers*. Geest & Portig K.-G., Leipzig, 1987.

[15] T. Poggio and F. Girosi. A theory of networks for approximation and learning. *Proc. IEEE*, 78:1481–1497, 1990.

[16] B. Schölkopf, R. Herbrich, and A. J. Smola. A generalized representer theorem. In D. Helmbold and B. Williamson, editors, *Proceedings of the 14th Annual Conference on Computational Learning Theory*, pages 416–426. Springer, New York, 2001.

[17] S. Smale and D.-X. Zhou. Estimating the approximation error in learning theory. *Anal. Appl.*, 1:17–41, 2003.

[18] S. Smale and D.-X. Zhou. Learning theory estimates via integral operators and their approximations. *Constr. Approx.*, 26:153–172, 2007.

[19] I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, New York, 2008.

[20] I. Steinwart, D. Hush, and C. Scovel. An oracle inequality for clipped regularized risk minimizers. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 1321–1328. MIT Press, Cambridge, MA, 2007.

[21] V. Temlyakov. Optimal estimators in learning theory. *Banach Center Publications, Inst. Math. Polish Academy of Sciences*, 72:341–366, 2006.

[22] Q. Wu, Y. Ying, and D.-X. Zhou. Multi-kernel regularized classifiers. *J. Complexity*, 23:108–134, 2007.

[23] D. X. Zhou. The covering number in learning theory. *J. Complexity*, 18:739–767, 2002.