Found. Comput. Math. 331–368 (2007)

DOI: 10.1007/s10208-006-0196-8

FOUNDATIONS OF COMPUTATIONAL MATHEMATICS
The Journal of the Society for the Foundations of Computational Mathematics

Optimal Rates for the Regularized Least-Squares Algorithm

A. Caponnetto[1] and E. De Vito[2]

[1]Department of Computer Science
University of Chicago
1100 East 58th Street
Chicago, IL 60637
and
D.I.S.I.
Università di Genova
Via Dodecaneso 35
16146 Genova, Italy
caponnet@uchicago.edu

[2]Dipartimento di Matematica
Università di Modena
Via Campi 213/B
41100 Modena, Italy
and
I.N.F.N.
Sezione di Genova
Via Dodecaneso 33
16146 Genova, Italy
devito@unimo.it

Abstract. We develop a theoretical analysis of the performance of the regularized least-square algorithm on a reproducing kernel Hilbert space in the supervised learning setting. The presented results hold in the general framework of vector-valued functions; therefore they can be applied to multitask problems. In particular, we observe that the concept of effective dimension plays a central role in the definition of a criterion for the choice of the regularization parameter as a function of the number of samples. Moreover, a complete minimax analysis of the problem is described,

Date received: August 4, 2005. Final version received: May 1, 2006. Date accepted: May 16, 2006.
Communicated by Felipe Cucker. Online publication: August 8, 2006.
AMS classification: Primary 68T05, 68P30.
Key words and phrases: Learning theory, Model selection, Optimal rates, Least squares.

showing that the convergence rates obtained by regularized least-squares estimators are indeed optimal over a suitable class of priors defined by the considered kernel. Finally, we give an improved lower rate result describing worst asymptotic behavior on individual probability measures rather than over classes of priors.

1. Introduction

In this paper we investigate the estimation properties of the regularized least-squares (RLS) algorithm on a reproducing kernel Hilbert space (RKHS) in the regression setting. Following the general scheme of supervised statistical learning theory, the available input–output samples are assumed to be drawn (independently and identically distributed) i.i.d. according to an unknown probability distribution. The aim of a regression algorithm is estimating a particular invariant of the unknown distribution: the *regression function*, using only the available empirical samples. Hence the asymptotic performances of the algorithm are usually evaluated by the rate of convergence of its estimates to the regression function. The main result of this paper shows a choice for the regularization parameter of RLS, such that the resulting algorithm is optimal in a minimax sense for a suitable class of priors.

The RLS algorithm on an RKHS of real-valued functions (i.e., when the output space is equal to $\mathbb{R}$) has been extensively studied in the literature, for an account see [**33**], [**30**], [**17**], and references therein. For the case $X = \mathbb{R}^d$ and the RKHS a Sobolev space, optimal rates were established assuming a suitable smoothness condition on the regression function (see [**18**], and references therein). For an arbitrary RKHS and compact input space, in [**5**] a covering number technique was used to obtain nonasymptotic upper bounds expressed in terms of suitable complexity measures of the regression function (see also [**33**] and [**37**]). In [**8**], [**26**], [**9**], [**27**] the covering techniques were replaced by estimates of integral operators through concentration inequalities of vector-valued random variables. Although expressed in terms of easily computable quantities the last bounds do not exploit much information about the fine structure of the kernel. Here we show that such information can be used to obtain tighter bounds. The approach we consider is a refinement of the functional analytical techniques presented in [**9**]. The central concept in this development is the *effective dimension* of the problem. This idea was recently used in [**36**] and [**19**] in the analysis of the performances of kernel methods for learning. Indeed, in this paper we show that the effective dimension plays a central role in the definition of an optimal rule for the choice of the regularization parameter as a function of the number of samples.

Although the previous investigations in [**8**], [**26**], [**9**], [**27**] showed that operator and spectral methods are valuable tools for the performance analysis of kernel-based algorithms such as RLS, all these results failed to compare with similar results recently obtained using entropy methods (see [**10**], [**29**]). These results (e.g., [**29**, Theorem 1.3]) showed that the optimal rate of convergence is essentially determined by the entropy characteristic of the considered class of priors with respect to a suitable topology induced by ρ_X, the marginal probability measure over the input space. Clearly, entropy numbers, and therefore rates of convergence, depend dramatically on ρ_X. However, ρ_X seems not to be crucial in the rates found in [**8**], [**26**], [**9**], [**27**]. This observation was our original motivation for taking into account the effective dimension: a spectral theoretical parameter which quantifies some capacity properties of ρ_X by means of the kernel. In fact, the effective dimension turned out to be the right parameter, in our operator analytical framework, to get rates comparable to the ones defined in terms of entropy numbers.

Recently, various papers, [**34**], [**1**], [**20**], [**14**], have addressed the multitask learning problem using kernel techniques. For instance, [**34**] employs two kernels, one on the input space and the other on the output space, in order to represent similarity measures on the respective domains. The underlying similarity measures are supposed to capture some inherent regularity of the phenomenon under investigation and should be chosen according to the available prior knowledge. On the contrary, in [**1**] the prior knowledge is encoded by a single kernel on the space of input–output couples, and a generalization of standard support vector machines is proposed. It was in [**20**] and [**14**] that, for the first time in the learning theory literature, it was pointed out that particular scalar kernels defined on input–output couples can be profitably mapped onto operator-valued kernels defined on the input space. However, to our knowledge, a thorough error analysis for the RLS algorithm when the output space is a general Hilbert space had never been given before. Our result fills this gap and is based on the well-known fact (see, e.g., [**25**] and [**3**]) that the machinery of scalar positive defined kernels can be elegantly extended to cope with vector-valued functions using operator-valued positive kernels. An advantage of our treatment is the extreme generality of the mathematical setting, which in fact subsumes most of the frameworks of its type available in the literature. Here, the input space is an arbitrary Polish space and the output space any separable Hilbert space. We only assume that the output y has finite variance and the random variable $(y - \mathbb{E}[\, y \mid x\,])$, conditionally to the input x, satisfies a momentum condition à la Bennett (see *Hypothesis* 2 in Section 3).

The other characterizing feature of this paper is the minimax analysis. The first lower rate result (*Theorem* 2 in Section 4) shows that the error rate attained by the RLS algorithm with our choice for the regularization parameter is optimal on a suitable class of probability measures. The class of priors we consider depends on two parameters: the first is a measure of the complexity of the regression function, as in [**26**], the other one is related to the *effective dimension* of the marginal probability measure over the input space relative to the chosen kernel; roughly speaking, it counts the number of degrees of freedom associated to the kernel and the marginal measure, available at a given conditioning. This kind of minimax analysis is standard in the statistical literature but it has received less attention in the context of learning theory (see, for instance, [**17**], [**10**], [**29**], and references therein). The main issue with this kind of approach to minimax problems in statistical learning is that the *bad* distribution in the prior could depend on the number of available samples. In fact, we are mainly interested in a worst-case analysis for an increasing number of samples and a *fixed* probability measure. This type of problem is well known in approximation theory. The idea of comparing minimax rates of approximation over classes of functions with rates of approximation of individual functions goes back to Bernstein's problem (see [**28**, Sect. 2, for an historical account and some examples]). In the context of learning theory this problem was recently considered in [**17**, Sect. 3], where the notion of an *individual lower rate* was introduced. Theorem 3 in Section 4 gives a new lower rate of this type greatly generalizing analogous previous results.

The paper is organized as follows. In Section 2 we briefly recall the main concepts of the regression problem in the context of supervised learning theory, [**6**], [**15**], [**23**]; however, the formalism could be easily rephrased using the language of nonparametric regression as in [**17**]. In particular, we define the notions of upper, lower, and optimal uniform rates over priors of probability measures. These concepts will be the main topic of Sections 4 and 5. In Section 3 we introduce the formalism of operator-valued kernels and the corresponding RKHS. Moreover, we describe the mathematical assumptions required by the subsequent developments. The assumptions specify conditions on both the RKHS (see *Hypothesis* 1) and the probability measure on the samples (see *Hypothesis* 2). Finally, we introduce (see *Definition* 1) the class of priors that will be considered throughout the minimax analysis.

In Section 4 we state the three main results of the paper, establishing upper and lower uniform rates for the RLS algorithm. The focus of our exposition on asymptotic rates, rather than on confidence analysis for finite sample size, was motivated by the decision to stress the aspects relevant to the main topic of investigation of the paper: optimality. However, all the results could be, with a relatively small effort, reformulated in terms of a nonasymptotic confidence analysis, see bounds (34) and (50).

The proofs of the theorems stated in Section 4 are postponed to Section 5.

2. Learning from Examples

We now introduce some basic concepts of statistical learning theory in the regression setting for vector-valued outputs (for details, see [**32**], [**15**], [**24**], [**6**], [**20**], and references therein).

In the framework of learning from examples there are two sets of variables: the input space X and the output space Y. The relation between the input $x \in X$ and the output $y \in Y$ is described by a probability distribution $\rho(x, y) = \rho_X(x)\rho(y|x)$ on $X \times Y$, where ρ_X is the marginal distribution on X and $\rho(\cdot|x)$ is the conditional distribution of y given $x \in X$. The distribution ρ is known only through a *training set* $\mathbf{z} = (\mathbf{x}, \mathbf{y}) = ((x_1, y_1), \ldots, (x_\ell, y_\ell))$ of ℓ examples drawn i.i.d. according to ρ. Given the sample $\mathbf{z}$, the aim of learning theory is to find a function $f_{\mathbf{z}} : X \to Y$ such that $f_{\mathbf{z}}(x)$ is a good estimate of the output y when a new input x is given.

The function $f_{\mathbf{z}}$ is called the *estimator* and a *learning algorithm* is the rule that, for any $\ell \in \mathbb{N}$, gives to every training set $\mathbf{z} \in Z^\ell$ the corresponding estimator $f_{\mathbf{z}}$.

We also use the notation $f_\ell(\mathbf{z})$ or, equivalently, $f_{\mathbf{z}}^\ell$, for the estimator, every time we want to stress its dependence on the number of examples. For the same reason, a learning algorithm will often be represented as a sequence $\{f_\ell\}_{\ell\in\mathbb{N}}$ of mappings f_ℓ from Z^ℓ to the set of functions Y^X.

If the output space Y is a Hilbert space, given a function $f : X \to Y$, the ability of f to describe the distribution ρ is measured by its *expected risk*,

$$\mathcal{E}[f] = \int_{X\times Y} \|f(x) - y\|_Y^2 \, d\rho(x, y).$$

The minimizer of the expected risk over the space of all the measurable Y-valued functions on X is the regression function

$$f_\rho(x) = \int_Y y \, d\rho(y|x).$$

The final aim of learning theory is to find an algorithm such that $\mathcal{E}[f_{\mathbf{z}}]$ is close to $\mathcal{E}[f_\rho]$ with high probability. However, if the estimators $f_{\mathbf{z}}$ are picked up from a *hypothesis space* $\mathcal{H}$ which is not dense in $L^2(X, \rho_X)$, approaching $\mathcal{E}[f_\rho]$ is too ambitious, and one can only hope to attain the expected error $\inf_{f\in\mathcal{H}} \mathcal{E}[f]$.

A learning algorithm $f_{\mathbf{z}}$ which, for every distribution ρ such that $\int_Y \|y\|_Y^2 \, d\rho_Y < +\infty$, achieves this goal, that is,

$$\lim_{\ell\to+\infty} \mathbb{P}_{\mathbf{z}\sim\rho^\ell}\left[\mathcal{E}[f_{\mathbf{z}}] - \inf_{f\in\mathcal{H}} \mathcal{E}[f] > \varepsilon\right] = 0 \quad \text{for all } \varepsilon > 0,$$

is said to be *universally consistent*.[1]

Universal consistency is an important and well-known propriety of many learning algorithms, among which is the RLS algorithm that will be introduced later. However, if $\mathcal{H}$ is infinite dimensional, the rate of convergence in the limit above, cannot be uniform on the set of all the distributions, but only on some restricted class $\mathcal{P}$ defined in terms of prior assumptions on ρ. In this paper the *priors*[2] are suitable classes $\mathcal{P}$ of distribution probabilities ρ encoding our knowledge on the relation between $\mathbf{x}$ and y. In particular, we consider a family of priors $\mathcal{P}(b, c)$ (see Definition 1) depending on two parameters: the *effective dimension* of $\mathcal{H}$ (with

[1] Various different definitions of consistency can be found in the literature (see [**17**], [**11**]): "in probability," "a.s.," "weak," and "strong." In more restrictive settings than ours, for example, assuming that $\|f_{\mathbf{z}}(x) - y\|_Y$ is bounded, it is possible to prove equivalence results between some of these definitions (e.g., between "weak" and "strong" consistency). However, this is not true under our assumptions. Moreover, our definition of consistency is weaker than analogous ones in the literature because we replaced $\mathcal{E}[f_\rho]$ with $\inf_{f\in\mathcal{H}} \mathcal{E}[f]$ in order to deal with hypothesis spaces which are not dense in $L^2(X, \rho_X)$.

[2] The concept of "prior" considered here should not be confused with its homologue in Bayesian statistics.

respect to ρ_X) and a notion of complexity of the regression function f_ρ which generalize to arbitrary RKHs the degree of smoothness of f_ρ, usually defined for regression in Sobolev spaces (see [**11**], [**30**], [**17**], [**5**], [**9**], and references therein).

Assuming ρ in a suitably small prior $\mathcal{P}$, it is possible to study the uniform convergence properties of learning algorithms. A natural way to do that is considering the *confidence function* (see [**10**], [**29**])

$$\inf_{f_\ell} \sup_{\rho \in \mathcal{P}} \mathbb{P}_{\mathbf{z} \sim \rho^\ell} \left[\mathcal{E}[f_{\mathbf{z}}^\ell] - \inf_{f \in \mathcal{H}} \mathcal{E}[f] > \varepsilon \right], \qquad \ell \in \mathbb{N}, \quad \varepsilon > 0,$$

where the infimum is over all the mappings $f_\ell : Z^\ell \to \mathcal{H}$. The learning algorithms $\{f_\ell\}_{\ell \in \mathbb{N}}$ attaining the minimization are optimal over $\mathcal{P}$ in the minimax sense. The main purpose of this paper (accomplished by Theorems 1 and 2) is showing that, for any $\mathcal{P}$ in the considered family of priors, the RLS algorithm (with a suitable choice of the regularization parameter) shares the asymptotic convergence properties of the optimal algorithms.

Let us now introduce the RLS algorithm [**33**], [**22**], [**6**], [**37**]. In this framework the hypothesis space $\mathcal{H}$ is a given Hilbert space of functions $f : X \to Y$ and, for any $\lambda > 0$ and $\mathbf{z} \in Z^\ell$, the RLS estimator $f_{\mathbf{z}}^\lambda$ is defined as the solution of the minimizing problem

$$\min_{f \in \mathcal{H}} \left\{ \frac{1}{\ell} \sum_{i=1}^{\ell} \| f(x_i) - y_i \|_Y^2 + \lambda \| f \|_{\mathcal{H}}^2 \right\}.$$

In the following the regularization parameter $\lambda = \lambda_\ell$ is some function of the number of examples ℓ.

The first result of the paper is a bound on the *upper rate of convergence* for the RLS algorithm with a suitable choice of λ_ℓ, under the assumption $\rho \in \mathcal{P}$. That is, we prove the existence of a sequence $(a_\ell)_{\ell \geq 1}$ such that

$$\lim_{\tau \to \infty} \limsup_{\ell \to \infty} \sup_{\rho \in \mathcal{P}} \mathbb{P}_{\mathbf{z} \sim \rho^\ell} \left[\mathcal{E}[f_{\mathbf{z}}^{\lambda_\ell}] - \inf_{f \in \mathcal{H}} \mathcal{E}[f] > \tau a_\ell \right] = 0. \tag{1}$$

More precisely, Theorem 1 shows that there is a choice $\lambda = \lambda_\ell$ such that the rate of convergence is $a_\ell = \ell^{-bc/(bc+1)}$, where $1 < c \leq 2$ is a parameter related to the *complexity* of f_ρ and $b > 1$ is a parameter related to the effective dimension of $\mathcal{H}$.

The second result shows that this rate is optimal if Y is finite dimensional. Following the analysis presented in [**17**], we formulate this problem in the framework of minimax lower rates. More precisely, a *minimax lower rate of convergence* for the class $\mathcal{P}$ is a sequence $(a_\ell)_{\ell \geq 1}$ of positive numbers such that

$$\lim_{\tau \to 0} \liminf_{\ell \to +\infty} \inf_{f_\ell} \sup_{\rho \in \mathcal{P}} \mathbb{P}_{\mathbf{z} \sim \rho^\ell} \left[\mathcal{E}[f_{\mathbf{z}}^\ell] - \inf_{f \in \mathcal{H}} \mathcal{E}[f] > \tau a_\ell \right] > 0, \tag{2}$$

where the infimum is over all the mappings $f_\ell : Z^\ell \to \mathcal{H}$. The definitions of lower and upper rates are given with respect to the convergence in probability as in [**30**], and coherently with the optimization problem inherent to the definition of confidence function. On the contrary, in [**17**] convergence in expectation was considered. Clearly, an upper rate in expectation induces an upper rate in probability and a lower rate in probability induces a lower rate in expectation.

The choice of the parameter $\lambda = \lambda_\ell$ is *optimal* over the prior $\mathcal{P}$ if it is possible to find a minimax lower rate $(a_\ell)_{\ell \geq 1}$ which is also an upper rate for the algorithm $f_{\mathbf{z}}^{\lambda_\ell}$. Theorem 2 shows the optimality for the choice of λ_ℓ given by Theorem 1.

The minimax lower rates are not completely satisfactory in the statistical learning setting. In fact, from the definitions above, it is clear that the *bad* distribution (the one maximizing $\mathbb{P}_{\mathbf{z}\sim\rho^\ell}[\mathcal{E}[f_{\mathbf{z}}^\ell] - \inf_{f\in\mathcal{H}} \mathcal{E}[f] > \tau a_\ell])$ could change for different values of the number of samples ℓ. Instead, one would usually like to know how the excess error, $\mathcal{E}[f_{\mathbf{z}}^\ell] - \inf_{f\in\mathcal{H}} \mathcal{E}[f]$, decreases as the number of samples grows for a fixed probability measure in $\mathcal{P}$. This type of issue is well known, and has been extensively analyzed in the context of approximation theory (see [**28**, Sect. 2]). To overcome the problem, one needs to consider a different type of lower rate: the *individual lower rate*. Precisely, an individual lower rate of convergence for the prior $\mathcal{P}$ is a sequence $(a_\ell)_{\ell\geq 1}$ of positive numbers such that

$$\inf_{\{f_\ell\}_{\ell\in\mathbb{N}}} \sup_{\rho\in\mathcal{P}} \limsup_{\ell\to+\infty} \frac{\mathbb{E}_{\mathbf{z}\sim\rho^\ell}(\mathcal{E}[f_{\mathbf{z}}^\ell] - \inf_{f\in\mathcal{H}} \mathcal{E}[f])}{a_\ell} > 0, \tag{3}$$

where the infimum is over the set of learning algorithms $\{f_\ell\}_{\ell\in\mathbb{N}}$.

Theorem 3 proves an individual lower bound in expectation. However, in order to show the optimality of the RLS algorithm in the sense of individual rates, it remains to prove either an upper rate in expectation or an individual lower rate in probability.

3. Notations and Assumptions

The aim of this section is to set the notations, to state and discuss the main assumptions we need to prove our results, and to describe precisely the class of priors on which the bounds hold uniformly.

We assume that the input space X is a Polish space[3] and the output space Y is a real separable Hilbert space. We let Z be the product space $X \times Y$, which is a Polish space too.

We let ρ be the probability measure describing the relation between $x \in X$ and $y \in Y$. By ρ_X we denote the marginal distribution on X and by $\rho(\cdot|x)$ the conditional distribution on Y given $x \in X$, both existing since Z is a Polish space, see Theorem 10.2.2 of [**12**].

We state the main assumptions on $\mathcal{H}$ and ρ.

[3] A Polish space is a separable metrizable topological space such that it is complete with respect to a metric compatible with the topology. Any locally compact second countable space is Polish.

Hypothesis 1. *The space $\mathcal{H}$ is a* separable *Hilbert space of functions $f : X \to Y$ such that*:

- *for all $x \in X$ there is a Hilbert–Schmidt*[4] *operator $K_x : Y \to \mathcal{H}$ satisfying*

$$f(x) = K_x^* f, \qquad f \in \mathcal{H}, \tag{4}$$

  where $K_x^* : \mathcal{H} \to Y$ is the adjoint of K_x;
- *the real function from $X \times X$ to $\mathbb{R}$*

$$(x, t) \mapsto \langle K_t v, K_x w \rangle_{\mathcal{H}} \text{ is measurable } \forall v, w \in Y; \tag{5}$$

- *there is $\kappa > 0$ such that*

$$\mathrm{Tr}(K_x^* K_x) \le \kappa, \qquad \forall x \in X. \tag{6}$$

Hypothesis 2. *The probability measure ρ on Z satisfies the following properties*:

$$\int_Z \|y\|_Y^2 \, d\rho(x, y) < +\infty, \tag{7}$$

- *there exists $f_{\mathcal{H}} \in \mathcal{H}$ such that*

$$\mathcal{E}[f_{\mathcal{H}}] = \inf_{f \in \mathcal{H}} \mathcal{E}[f], \tag{8}$$

  where $\mathcal{E}[f] = \int_Z \|f(x) - y\|_Y^2 \, d\rho(x, y)$;
- *there are two positive constants Σ, M such that*

$$\int_Y \left(e^{\|y - f_{\mathcal{H}}(x)\|_Y / M} - \frac{\|y - f_{\mathcal{H}}(x)\|_Y}{M} - 1 \right) d\rho(y|x) \le \frac{\Sigma^2}{2M^2} \tag{9}$$

  for ρ_X-almost all $x \in X$.

We now briefly discuss the consequences of the above assumptions. If $Y = \mathbb{R}$, the operator K_x can be identified with the vector $K_x 1 \in \mathcal{H}$ and (4) reduces to

$$f(x) = \langle f, K_x \rangle, \qquad f \in \mathcal{H}, \quad x \in X,$$

so that $\mathcal{H}$ is an RKHS [**2**] with kernel

$$K(x, t) = \langle K_t, K_x \rangle_{\mathcal{H}}. \tag{10}$$

In fact, the theory of RKHSs can naturally be extended to vector-valued functions [**25**]. In particular, the assumption that K_x is a Hilbert–Schmidt operator is useful

[4] An operator $A : Y \to \mathcal{H}$ is a Hilbert–Schmidt operator if, for some (any) basis $(v_j)_j$ of Y, it holds that $\mathrm{Tr}(A^* A) = \sum_j \langle A v_j, A v_j \rangle_{\mathcal{H}} < +\infty$.

in keeping the generalized theory similar to the scalar one. Indeed, let $\mathcal{L}(Y)$ be the space of bounded linear operators on Y with the uniform norm $\|\cdot\|_{\mathcal{L}(\mathcal{H})}$. In analogy with (10), let $K : X \times X \to \mathcal{L}(Y)$ be the (vector-valued) reproducing kernel

$$K(x, t) = K_x^* K_t, \qquad x, t \in X.$$

Since K_x is a Hilbert–Schmidt operator, there is a basis $(v_j(x))_j$ of Y and an orthogonal sequence $(k_j(x))_j$ of vector in $\mathcal{H}$ such that

$$K_x v = \sum_j \langle v, v_j(x)\rangle_Y k_j(x), \qquad v \in Y,$$

with the condition $\sum_j \left\|k_j(x)\right\|_{\mathcal{H}}^2 < +\infty$. The reproducing kernel becomes

$$K(x, t)v = \sum_{j,m} \langle k_j(t), k_m(x)\rangle_{\mathcal{H}} \langle v, v_j(t)\rangle_Y v_m(x), \qquad v \in Y,$$

and (6) is equivalent to

$$\sum_j \left\|k_j(x)\right\|_{\mathcal{H}}^2 \leq \kappa, \qquad x \in X.$$

Remark 1. If Y is finite dimensional, any linear operator is Hilbert–Schmidt and (4) is equivalent to the fact that the evaluation functional on $\mathcal{H}$,

$$f \mapsto f(x) \in Y$$

is continuous for all $x \in X$. Moreover, the reproducing kernel K takes values in the space of $(d \times d)$-matrices (where $d = \dim Y$). In this finite-dimensional setting the vector-valued RKHS formalism can be rephrased in terms of ordinary scalar-valued functions. Indeed, let $(v_j)_{j=1}^d$ be a basis of Y, let $\widehat{X} = X \times \{1, \ldots, d\}$, and let $\widehat{K} : \widehat{X} \times \widehat{X} \to \mathbb{R}$ be the kernel

$$\widehat{K}(x, j; t, i) = \left\langle K_t v_i, K_x v_j\right\rangle_{\mathcal{H}}.$$

Since $\widehat{K}$ is symmetric and positive definite, let $\widehat{\mathcal{H}}$ be the corresponding RKHS, whose elements are real functions on $\widehat{X}$ [**2**]. Any element $f \in \mathcal{H}$ can be identified with the function in $\widehat{\mathcal{H}}$ given by

$$f(x, j) = \left\langle f(x), v_j\right\rangle_Y, \qquad x \in X, \quad j = 1, \ldots, d.$$

Moreover, the expected risk becomes

$$\begin{aligned}
\mathcal{E}[f] &= \int_Z \sum_j \left\langle f(x) - y, v_j\right\rangle_Y^2 \, d\rho(x, y) \\
&= \sum_j \int_{X \times \mathbb{R}} (f(x, j) - y_j)^2 \, d\rho_j(x, y_j) \\
&= d \int_{\widehat{X} \times \mathbb{R}} (f(x, j) - \xi)^2 \, d\widehat{\rho}(x, j, \xi) = d\, \widehat{\mathcal{E}}[f],
\end{aligned}$$

where ρ_j are the marginal distributions with respect to the projections

$$(x, y) \mapsto (x, \langle y, v_j \rangle_Y), \qquad (x, y) \in X \times Y,$$

and $\widehat{\rho}$ is the probability distribution on $\widehat{X} \times \mathbb{R}$ given by $\widehat{\rho} = (1/d) \sum_j \rho_j$. In a similar way, the regularized empirical risk becomes

$$\frac{1}{\ell} \sum_{i=1}^{\ell} \sum_{j=1}^{d} (f(x_i, j) - y_{i,j})^2 + \lambda \|f\|^2_{\widehat{\mathcal{H}}}, \qquad y_{i,j} = \langle y_i, v_j \rangle_Y,$$

where the example (x_i, y_i) is replaced by d-examples $(x_i, y_{i,1}), \ldots, (x_i, y_{i,d})$.

However, in this scalar setting the examples are not i.i.d., so we decide to state the results in the framework of vector-valued functions. Moreover, this does not result in more complex proofs, the theorems are stated in a basis independent form, and also hold for infinite-dimensional Y.

Coming back to the discussion on the assumptions. The requirement that $\mathcal{H}$ is separable avoids problems with measurability and allows us to employ vector-valued concentration inequalities. If (5) is replaced by the stronger condition

$$(x, t) \mapsto \langle K_t v, K_x w \rangle_{\mathcal{H}} \text{ is continuous } \forall v, w \in Y,$$

the fact that X and Y are separable implies that $\mathcal{H}$ is separable, too [**4**]. Conditions (5) and the fact that $\mathcal{H}$ is separable ensure that the functions $f \in \mathcal{H}$ are measurable from X to Y, whereas (6) implies that f are bounded functions. Indeed, (6) implies that

$$\left\| K_x^* \right\|_{\mathcal{L}(\mathcal{H}, Y)} = \|K_x\|_{\mathcal{L}(Y, \mathcal{H})} \leq \sqrt{\operatorname{Tr}(K_x^* K_x)} \leq \sqrt{\kappa}, \tag{11}$$

and (4) gives

$$\|f(x)\|_Y = \left\| K_x^* f \right\|_Y \leq \sqrt{\kappa}\, \|f\|_{\mathcal{H}}, \qquad \forall x \in X.$$

Regarding the distribution ρ, it is clear that if (7) is not satisfied, then $\mathcal{E}[f] = +\infty$ for all $f \in \mathcal{H}$ and the learning problem does not make sense. If it holds, (5) and (6) are the minimal requirements to ensure that any $f \in \mathcal{H}$ has a finite expected risk (see item (i) of Proposition 1).

In general $f_{\mathcal{H}}$ is not unique as an element of $\mathcal{H}$, but we recover uniqueness by choosing the one with minimal norm in $\mathcal{H}$.

If the regression function

$$f_\rho = \int_Y y \, d\rho(y|x)$$

belongs to $\mathcal{H}$, clearly $f_{\mathcal{H}} = f_\rho$. However, in general, the existence of $f_{\mathcal{H}}$ is a weaker condition than $f_\rho \in \mathcal{H}$, for example, if $\mathcal{H}$ is finite dimensional, $f_{\mathcal{H}}$ always exists. Condition (8) is essential to define the class of priors for which both the upper and lower bounds hold uniformly. Finally, (9) is a model of the noise of the output y and it is satisfied, for example, if the noise is bounded, Gaussian, or sub-Gaussian [**31**].

We now introduce some more notations we need to state our bounds. Let $\mathcal{L}_2(\mathcal{H})$ be the separable Hilbert space of Hilbert–Schmidt operators on $\mathcal{H}$ with scalar product

$$\langle A, B\rangle_{\mathcal{L}_2(\mathcal{H})} = \mathrm{Tr}(B^* A)$$

and norm

$$\|A\|_{\mathcal{L}_2(Y)} = \sqrt{\mathrm{Tr}(A^* A)} \geq \|A\|_{\mathcal{L}(\mathcal{H})}.$$

Given $x \in X$, let

$$T_x = K_x K_x^* \in \mathcal{L}(\mathcal{H}), \tag{12}$$

which is a positive operator. A simple computation shows that

$$\mathrm{Tr}\, T_x = \mathrm{Tr}\, K_x^* K_x \leq \kappa$$

so that T_x is a trace class operator and, a fortiori, a Hilbert–Schmidt operator. Hence

$$\|T_x\|_{\mathcal{L}(\mathcal{H})} \leq \|T_x\|_{\mathcal{L}_2(\mathcal{H})} \leq \mathrm{Tr}(T_x) \leq \kappa. \tag{13}$$

We let $T : \mathcal{H} \to \mathcal{H}$ be

$$T = \int_X T_x \, d\rho_X(x), \tag{14}$$

where the integral converges in $\mathcal{L}_2(\mathcal{H})$ to a positive trace class operator with

$$\|T\|_{\mathcal{L}(\mathcal{H})} \leq \mathrm{Tr}\, T = \int_X \mathrm{Tr}\, T_x \, d\rho_X(x) \leq \kappa \tag{15}$$

(see item (ii) of Proposition 1). Moreover, the spectral theorem gives

$$T = \sum_{n=1}^{N} t_n \, \langle \cdot, e_n\rangle_{\mathcal{H}} \, e_n, \tag{16}$$

where $(e_n)_{n=1}^N$ is a basis of $\mathrm{Ker}\, T^\perp$ (possibly $N = +\infty$), $0 < t_{n+1} \leq t_n$, with $\sum_{n=1}^N t_n = \mathrm{Tr}\, T \leq \kappa$.

We now discuss the class of priors.

Definition 1. Let us fix the positive constants M, Σ, R, α, and β.

Then, given $1 < b \leq +\infty$ and $1 \leq c \leq 2$, we define $\mathcal{P} = \mathcal{P}(b, c)$ the set of probability distributions ρ on Z such that:

(i) Hypotheses 2 holds with the given choice for M and Σ in (9);

(ii) there is $g \in \mathcal{H}$ such that $f_{\mathcal{H}} = T^{(c-1)/2} g$ with $\|g\|_{\mathcal{H}}^2 \leq R$;
(iii) if $b < +\infty$, then $N = +\infty$ and the eigenvalues of T given by (16) satisfy

$$\alpha \leq n^b t_n \leq \beta \quad \forall n \geq 1, \tag{17}$$

whereas if $b = +\infty$, then $N \leq \beta < +\infty$.

The first condition ensures that the constants appearing in the bounds do not depend on ρ, but only on $\mathcal{P}$. The second condition is a measure of the *complexity* of $f_{\mathcal{H}}$ depending both on the conditional distribution $\rho(y|x)$ and the marginal distribution ρ_X. If $\mathcal{H}$ is a Sobolev space, this is related to the smoothness of $f_{\mathcal{H}}$. About the last condition, observe that T depends only on ρ_X and (17) is related to the *effective dimension* of the space $\mathcal{H}$ with respect to ρ_X. If $b = +\infty$, $\mathcal{H}$ is finite dimensional, $f_{\mathcal{H}}$ always exists, and condition (ii) holds for any $1 < c \leq 2$.

Remark 2. The above conditions can be expressed in a different way. Let $L^2(X)$ be the Hilbert space of functions from X to Y square-integrable with respect to ρ_X, and denote by $\|\cdot\|_{\rho_X}$ and $\langle \cdot, \cdot \rangle_{\rho_X}$ the corresponding norm and scalar product. Define $L_K : L^2(X) \to L^2(X)$ to be the integral operator of kernel K,

$$(L_K \varphi)(t) = \int_X K(t, x) \varphi(x) \, d\rho_X(x),$$

which is bounded by (6). Based on the polar decomposition of the inclusion map from $\mathcal{H}$ into $L^2(X)$, in [**7**] it is shown that

$$L_K = \sum_{n=1}^{N} t_n \langle \cdot, \varphi_n \rangle_{\rho_X} \varphi_n, \qquad e_n = L_K^{1/2} \varphi_n,$$

where $(\varphi_n)_{n=1}^N$ is a basis of $(\ker L_K)^\perp$ and $L_K^{1/2}$ is the square root of L_K (so that $L_K^{1/2} \varphi_n = t_n^{1/2} \varphi_n = e_n$). Moreover, $f_{\mathcal{H}} = T^{(c-1)/2} g$ with $\|g\|_{\mathcal{H}}^2 \leq R$ if and only if $f_{\mathcal{H}} = L_K^{c/2} \varphi$ with $\|\varphi\|_{\rho_X}^2 \leq R$.

4. Upper and Lower Rates

In this section we report the main results of the paper. We first prove an upper bound on the expected risk for the RLS estimators. More precisely, we give a choice for the regularization parameter λ, as a function of ℓ, providing us with a rate of decay of the expected risk which is uniform on the prior $\mathcal{P}(b, c)$. Moreover, we obtain a minimax lower rate for $\mathcal{P}(b, c)$ showing that the above choice of the parameter is optimal. Both the upper and lower rates hold in probability. Finally, we prove an individual lower rate in expectation. The proofs are given in the next section.

We recall that, given $\lambda > 0$, for any $\ell \in \mathbb{N}$ and any training set $\mathbf{z} = (\mathbf{x}, \mathbf{y}) = ((x_1, y_1), \ldots, (x_\ell, y_\ell)) \in Z^\ell$, the estimator $f_{\mathbf{z}}^\lambda$ is defined as the solution of the minimization problem

$$\min_{f \in \mathcal{H}} \left(\frac{1}{\ell} \sum_{i=1}^{\ell} \| f(x_i) - y_i \|_Y^2 + \lambda \| f \|_{\mathcal{H}}^2 \right), \tag{18}$$

whose existence and uniqueness is well known (see item (v) of Proposition 1).

Theorem 1. *Given* $1 < b \le +\infty$ *and* $1 \le c \le 2$, *let*

$$\lambda_\ell = \begin{cases} (1/\ell)^{b/(bc+1)}, & b < +\infty,\ c > 1, \\ (\log \ell/\ell)^{b/(b+1)}, & b < +\infty,\ c = 1, \\ (1/\ell)^{1/2}, & b = +\infty, \end{cases} \tag{19}$$

and

$$a_\ell = \begin{cases} (1/\ell)^{bc/(bc+1)}, & b < +\infty,\ c > 1, \\ (\log \ell/\ell)^{b/(b+1)}, & b < +\infty,\ c = 1, \\ 1/\ell, & b = +\infty, \end{cases} \tag{20}$$

then

$$\lim_{\tau \to \infty} \limsup_{\ell \to \infty} \sup_{\rho \in \mathcal{P}(b,c)} \mathbb{P}_{\mathbf{z} \sim \rho^\ell} [\mathcal{E}[f_{\mathbf{z}}^{\lambda_\ell}] - \mathcal{E}[f_{\mathcal{H}}] > \tau a_\ell] = 0. \tag{21}$$

The above result gives a family of upper rates of convergence for the RLS algorithm as defined in (1). The following theorem proves that the corresponding minimax lower rates (see (2)) hold.

Theorem 2. *Assume that* $\dim Y = d < +\infty$, $1 < b < +\infty$ *and* $1 \le c \le 2$, *then*

$$\lim_{\tau \to 0} \liminf_{\ell \to +\infty} \inf_{f_\ell} \sup_{\rho \in \mathcal{P}(b,c)} \mathbb{P}_{\mathbf{z} \sim \rho^\ell} [\mathcal{E}[f_{\mathbf{z}}^\ell] - \mathcal{E}[f_{\mathcal{H}}] > \tau \ell^{-bc/(bc+1)}] = 1.$$

The above result shows that the rate of convergence given by the RLS algorithm is optimal when Y is finite dimensional for any $1 < b < +\infty$ (i.e., $N = +\infty$) and $1 < c \le 2$ and that it is optimal up to a logarithmic factor for $c = 1$.

Finally, we give a result about the individual lower rates in expectation (see (3)).

Theorem 3. *Assume that* $\dim Y = d < +\infty$, $1 < b < +\infty$ *and* $1 \le c \le 2$. *Then, for every* $B > b$, *the following individual lower rate holds*:

$$\inf_{\{f_\ell\}_{\ell \in \mathbb{N}}} \sup_{\rho \in \mathcal{P}(b,c)} \limsup_{\ell \to +\infty} \frac{\mathbb{E}_{\mathbf{z} \sim \rho^\ell} (\mathcal{E}[f_{\mathbf{z}}^\ell] - \mathcal{E}[f_{\mathcal{H}}])}{\ell^{-cB/(cB+1)}} > 0,$$

where the infimum is over the set of all learning algorithms $\{f_\ell\}_{\ell \in \mathbb{N}}$.

The advantages of individual lower rates over minimax lower rates have already been discussed in Sections 1 and 2. Here we add that the proof of the theorem above can be straightforwardly modified in order to extend the range of the infimum to general *randomized* learning algorithms, that is, algorithms whose outputs are random variables depending on the training set. Such a generalization seems not an easy task to accomplish in the standard minimax setting. It should also be remarked that the condition $1 \leq c \leq 2$ in Theorem 3 has been introduced to keep the notations homogeneous throughout the paper, but it could be relaxed to $0 \leq c \leq 2$.

5. Proofs

In this section we give the proofs of the three theorems stated above.

5.1. *Preliminary Results*

We recall some known facts without reporting their proofs.

The first proposition summarizes some mathematical properties of the RLS algorithm. It is well known in the framework of linear inverse problems (see [**13**]), and a proof in the context of learning theory can be found in [**7**] and, for the scalar case, in [**6**].

Proposition 1. *Assume Hypotheses* 1 *and* 2. *The following facts hold*:

(i) *For all* $f \in \mathcal{H}$, f *is measurable and*

$$\mathcal{E}[f] = \int_Z \|f(x) - y\|_Y^2 \, d\rho(x, y) < +\infty.$$

(ii) *The minimizers* $f_{\mathcal{H}}$ *are the solution of the following equation*:

$$T f_{\mathcal{H}} = g, \tag{22}$$

where T *is the positive trace class operator defined by*

$$T = \int_X K_x K_x^* = \int_X T_x \, d\rho_X(x)$$

with the integral converging in $\mathcal{L}_2(\mathcal{H})$ *and*

$$g = \int_X K_x f_\rho(x) \, d\rho_X(x) \in \mathcal{H}, \tag{23}$$

with the integral converging in $\mathcal{H}$.

(iii) *For all* $f \in \mathcal{H}$,

$$\mathcal{E}[f] - \mathcal{E}[f_{\mathcal{H}}] = \|\sqrt{T}(f - f_{\mathcal{H}})\|_{\mathcal{H}}^2, \qquad f \in \mathcal{H}. \tag{24}$$

(iv) *For any* $\lambda > 0$, *a unique minimizer* f^λ *of the regularized expected risk*

$$\mathcal{E}[f] + \lambda \|f\|_{\mathcal{H}}^2$$

exists and is given by

$$f^\lambda = (T + \lambda)^{-1} g = (T + \lambda)^{-1} T f_{\mathcal{H}}. \tag{25}$$

(v) *Given a training set* $\mathbf{z} = (\mathbf{x}, \mathbf{y}) = ((x_1, y_1), \ldots, (x_\ell, y_\ell)) \in Z^\ell$, *for any* $\lambda > 0$ *a unique minimizer* $f_{\mathbf{z}}^\lambda$ *of the regularized empirical risk*

$$\frac{1}{\ell} \sum_{i=1}^{\ell} \|f(x_i) - y_i\|_Y^2 + \lambda \|f\|_{\mathcal{H}}^2$$

exists and is given by

$$f_{\mathbf{z}}^\lambda = (T_{\mathbf{x}} + \lambda)^{-1} g_{\mathbf{z}}. \tag{26}$$

where $T_{\mathbf{x}} : \mathcal{H} \to \mathcal{H}$ *is the positive finite rank operator*

$$T_{\mathbf{x}} = \frac{1}{\ell} \sum_{i=1}^{\ell} T_{x_i} \tag{27}$$

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

$$g_{\mathbf{z}} = \frac{1}{\ell} \sum_{i=1}^{\ell} K_{x_i} y_i. \tag{28}$$

By means of (4), T and g are explicitly given by

$$(Tf)(x) = K_x^* T f = \int_X K_x^* (K_t K_t^*) f \, d\rho_X(t) = \int_X K(x,t) f(t) \, d\rho_X(t), \tag{29}$$

so T acts as the integral operator of kernel K, and

$$g(x) = K_t^* g = \int_X K(x,t) f_\rho(t) \, d\rho_X(t). \tag{30}$$

We also need the following probabilistic inequality based on a result of [**21**], see also Theorem 3.3.4 of [**35**].

Proposition 2. *Let* $(\Omega, \mathcal{F}, P)$ *be a probability space and let* ξ *be a random variable on* Ω *taking value in a real separable Hilbert space* $\mathcal{K}$. *Assume that there are two positive constants* L *and* σ *such that*

$$\mathbb{E}[\|\xi - \mathbb{E}[\xi]\|_{\mathcal{K}}^m] \le \tfrac{1}{2} m! \, \sigma^2 L^{m-2} \quad \forall m \ge 2, \tag{31}$$

then, for all $\ell \in \mathbb{N}$ *and* $0 < \eta < 1$, *then*

$$\mathbb{P}_{(\omega_1,\ldots,\omega_\ell)\sim P^\ell}\left[\left\|\frac{1}{\ell}\sum_{i=1}^{\ell}\xi(\omega_i) - \mathbb{E}[\xi]\right\|_{\mathcal{K}} \le 2\left(\frac{L}{\ell} + \frac{\sigma}{\sqrt{\ell}}\right)\log\frac{2}{\eta}\right] \ge 1-\eta. \quad (32)$$

In particular, (31) *holds if*

$$\begin{aligned} \|\xi(\omega)\|_{\mathcal{K}} &\le L/2 \quad \text{a.s.} \\ \mathbb{E}[\|\xi\|_{\mathcal{K}}^2] &\le \sigma^2. \end{aligned} \quad (33)$$

5.2. *Upper Rates*

The main steps in the proof of the upper rate of convergence given in Theorem 1 are the following.

First, given a probability distribution ρ satisfying Hypothesis 2, Theorem 4 gives an upper bound for $\mathcal{E}[f_{\mathbf{z}}^\lambda] - \mathcal{E}[f_{\mathcal{H}}]$ that holds in probability for any small enough λ and any large enough ℓ (see (35)). The bound is controlled by the following quantities parametrized by $\lambda > 0$:

(1) The *residual*

$$\mathcal{A}(\lambda) = \mathcal{E}[f^\lambda] - \mathcal{E}[f_{\mathcal{H}}] = \|\sqrt{T}(f^\lambda - f_{\mathcal{H}})\|_{\mathcal{H}}^2,$$

where $f^\lambda \in \mathcal{H}$ is the minimizer of the regularized expected risk (see item (iv) of Proposition 1) and the second equality is a consequence of (24).

(2) The *reconstruction error*

$$\mathcal{B}(\lambda) = \left\|f^\lambda - f_{\mathcal{H}}\right\|_{\mathcal{H}}^2.$$

(3) The *effective dimension*

$$\mathcal{N}(\lambda) = \mathrm{Tr}[(T+\lambda)^{-1}T],$$

which is finite due to the fact that T is trace class (see item (ii) of Proposition 1).

Roughly speaking, the effective dimension $\mathcal{N}(\lambda)$ controls the complexity of the hypothesis space $\mathcal{H}$ according to the marginal measure ρ_X, whereas $\mathcal{A}(\lambda)$ and $\mathcal{B}(\lambda)$, which depend on ρ, control the *complexity* of $f_{\mathcal{H}}$.

Remark 3. In the framework of learning theory $\mathcal{A}(\lambda) = \|f^\lambda - f_{\mathcal{H}}\|_\rho^2$ is called the approximation error, whereas in inverse problems theory the approximation error is usually $\sqrt{\mathcal{B}(\lambda)} = \|f^\lambda - f_{\mathcal{H}}\|_{\mathcal{H}}$. In order to avoid confusion we adopt the nomenclature of inverse problems [**13**].

Next, Proposition 3 studies the asymptotic behavior of the above quantities when λ goes to zero under the assumption that $\rho \in \mathcal{P}(b, c)$. Finally, from this result, it is easy to derive a *best choice* for the parameter $\lambda = \lambda_\ell$ giving rise to the claimed rate of convergence.

The following theorem gives a nonasymptotic upper bound which is of interest by itself.

Theorem 4. *Let ρ satisfy Hypothesis 2, $\ell \in \mathbb{N}$, $\lambda > 0$ and $0 < \eta < 1$. Then, with probability greater than $1 - \eta$,*

$$\mathcal{E}[f_{\mathbf{z}}^\lambda] - \mathcal{E}[f_{\mathcal{H}}] \leq C_\eta \left(\mathcal{A}(\lambda) + \frac{\kappa^2 \mathcal{B}(\lambda)}{\ell^2 \lambda} + \frac{\kappa \mathcal{A}(\lambda)}{\ell \lambda} + \frac{\kappa M^2}{\ell^2 \lambda} + \frac{\Sigma^2 \mathcal{N}(\lambda)}{\ell} \right) \quad (34)$$

provided that

$$\ell \geq \frac{2 C_\eta \kappa \mathcal{N}(\lambda)}{\lambda} \qquad \text{and} \qquad \lambda \leq \|T\|_{\mathcal{L}(\mathcal{H})}, \quad (35)$$

where $C_\eta = 32 \log^2(6/\eta)$.

Proof. We split the proof in several steps. Let λ, η, and ℓ be as in the statement of the theorem.

Step 1. Given a training set $\mathbf{z} = (\mathbf{x}, \mathbf{y}) \in Z^\ell$, (24) gives

$$\mathcal{E}[f_{\mathbf{z}}^\lambda] - \mathcal{E}[f_{\mathcal{H}}] = \|\sqrt{T}(f_{\mathbf{z}}^\lambda - f_{\mathcal{H}})\|_{\mathcal{H}}^2.$$

Recalling the definition of f^λ, see item (iv) of Proposition 1, we split

$$f_{\mathbf{z}}^\lambda - f_{\mathcal{H}} = (f_{\mathbf{z}}^\lambda - f^\lambda) + (f^\lambda - f_{\mathcal{H}}).$$

Now (25) and (26) give

$$\begin{aligned}
f_{\mathbf{z}}^\lambda - f^\lambda \ &= \ ((T_{\mathbf{x}} + \lambda)^{-1} g_{\mathbf{z}}) - ((T + \lambda)^{-1} g) \\
&= \ (T_{\mathbf{x}} + \lambda)^{-1} \{(g_{\mathbf{z}} - g) + (T - T_{\mathbf{x}})(T + \lambda)^{-1} g\} \\
\text{(Eq. (22))} \ &= \ (T_{\mathbf{x}} + \lambda)^{-1} \{(g_{\mathbf{z}} - T_{\mathbf{x}} f_{\mathcal{H}} + T_{\mathbf{x}} f_{\mathcal{H}} - T f_{\mathcal{H}}) + (T - T_{\mathbf{x}}) f^\lambda\} \\
&= \ (T_{\mathbf{x}} + \lambda)^{-1} (g_{\mathbf{z}} - T_{\mathbf{x}} f_{\mathcal{H}}) + (T_{\mathbf{x}} + \lambda)^{-1} (T - T_{\mathbf{x}})(f^\lambda - f_{\mathcal{H}}),
\end{aligned}$$

where $T_{\mathbf{x}} \in \mathcal{L}(\mathcal{H})$, $g_{\mathbf{z}} \in \mathcal{H}$, and $g \in \mathcal{H}$ are given by (27), (28) and (23), respectively.

The inequality $\|f_1 + f_2 + f_3\|_{\mathcal{H}}^2 \leq 3(\|f_1\|_{\mathcal{H}}^2 + \|f_2\|_{\mathcal{H}}^2 + \|f_3\|_{\mathcal{H}}^2)$ implies

$$\mathcal{E}[f_{\mathbf{z}}^\lambda] - \mathcal{E}[f_{\mathcal{H}}] \leq 3(\mathcal{A}(\lambda) + \mathcal{S}_1(\lambda, \mathbf{z}) + \mathcal{S}_2(\lambda, \mathbf{z})), \quad (36)$$

where $\mathcal{A}(\lambda)$ is the residual and

$$\begin{aligned}
\mathcal{S}_1(\lambda, \mathbf{z}) \ &= \ \|\sqrt{T}(T_{\mathbf{x}} + \lambda)^{-1}(g_{\mathbf{z}} - T_{\mathbf{x}} f_{\mathcal{H}})\|_{\mathcal{H}}^2, \\
\mathcal{S}_2(\lambda, \mathbf{z}) \ &= \ \|\sqrt{T}(T_{\mathbf{x}} + \lambda)^{-1}(T - T_{\mathbf{x}})(f^\lambda - f_{\mathcal{H}})\|_{\mathcal{H}}^2.
\end{aligned}$$

Step 2. Probabilistic bound on $\mathcal{S}_2(\lambda, \mathbf{z})$. Clearly,

$$\mathcal{S}_2(\lambda, \mathbf{z}) \leq \|\sqrt{T}(T_{\mathbf{x}} + \lambda)^{-1}\|^2_{\mathcal{L}(\mathcal{H})} \|(T - T_{\mathbf{x}})(f^\lambda - f_{\mathcal{H}})\|^2_{\mathcal{H}}. \tag{37}$$

Step 2.1. Probabilistic bound on $\|\sqrt{T}(T_{\mathbf{x}} + \lambda)^{-1}\|_{\mathcal{L}(\mathcal{H})}$. Assume that

$$\Theta(\lambda, \mathbf{z}) = \|(T + \lambda)^{-1}(T - T_{\mathbf{x}})\|_{\mathcal{L}(\mathcal{H})} = \|(T - T_{\mathbf{x}})(T + \lambda)^{-1}\|_{\mathcal{L}(\mathcal{H})} \leq \tfrac{1}{2}, \tag{38}$$

(the second inequality holds since if A, B are self-adjoint operators in $\mathcal{L}(\mathcal{H})$, then $\|AB\|_{\mathcal{L}(\mathcal{H})} = \|(AB)^*\|_{\mathcal{L}(\mathcal{H})} = \|BA\|_{\mathcal{L}(\mathcal{H})}$). Then the Neumann series gives

$$\begin{aligned}\sqrt{T}(T_{\mathbf{x}} + \lambda)^{-1} &= \sqrt{T}(T + \lambda)^{-1}(I - (T - T_{\mathbf{x}})(T + \lambda)^{-1})^{-1} \\ &= \sqrt{T}(T + \lambda)^{-1} \sum_{n=0}^{+\infty} ((T - T_{\mathbf{x}})(T + \lambda)^{-1})^n,\end{aligned}$$

so that

$$\begin{aligned}\|\sqrt{T}(T_{\mathbf{x}} + \lambda)^{-1}\|_{\mathcal{L}(\mathcal{H})} &\leq \|\sqrt{T}(T + \lambda)^{-1}\|_{\mathcal{L}(\mathcal{H})} \sum_{n=0}^{+\infty} \|(T - T_{\mathbf{x}})(T + \lambda)^{-1}\|^n_{\mathcal{L}(\mathcal{H})} \\ &\leq \|\sqrt{T}(T + \lambda)^{-1}\|_{\mathcal{L}(\mathcal{H})} \frac{1}{1 - \Theta(\lambda, \mathbf{z})} \\ &\leq 2\|\sqrt{T}(T + \lambda)^{-1}\|_{\mathcal{L}(\mathcal{H})}.\end{aligned}$$

The spectral theorem ensures that $\|\sqrt{T}(T + \lambda)^{-1}\|_{\mathcal{L}(\mathcal{H})} \leq 1/2\sqrt{\lambda}$ so that

$$\|\sqrt{T}(T_{\mathbf{x}} + \lambda)^{-1}\|_{\mathcal{L}(\mathcal{H})} \leq \frac{1}{\sqrt{\lambda}}. \tag{39}$$

We claim that (35) implies (38) with probability greater than $1 - \eta/3$. To this aim we apply Proposition 2 to the random variable $\xi_1 : X \to \mathcal{L}_2(\mathcal{H})$,

$$\xi_1(x) = (T + \lambda)^{-1} T_x$$

so that

$$\mathbb{E}[\xi_1] = (T + \lambda)^{-1} T \qquad \text{and} \qquad \frac{1}{\ell} \sum_{i=1}^{\ell} \xi_1(x_i) = (T + \lambda)^{-1} T_{\mathbf{x}}.$$

Moreover, (13) and $\|(T + \lambda)^{-1}\|_{\mathcal{L}(\mathcal{H})} \leq 1/\lambda$ imply

$$\|\xi\|_{\mathcal{L}_2(\mathcal{H})} \leq \|(T + \lambda)^{-1}\|_{\mathcal{L}(\mathcal{H})} \|T_x\|_{\mathcal{L}_2(\mathcal{H})} \leq \frac{\kappa}{\lambda} = \frac{L_1}{2}.$$

Condition (13) ensures that T_x is of trace class and the inequality

$$\operatorname{Tr}(AB) \leq \|A\|_{\mathcal{L}(\mathcal{H})} \operatorname{Tr} B \tag{40}$$

(A positive bounded operator, B positive trace class operator) implies

$$
\begin{aligned}
\mathbb{E}[\|\xi_1\|^2_{\mathcal{L}_2(\mathcal{H})}] &= \int_X \operatorname{Tr}(T_x(T_x^{1/2}(T+\lambda)^{-2}T_x^{1/2}))\, d\rho_X(x) \\
&\le \int_X \|T_x\|_{\mathcal{L}(\mathcal{H})} \operatorname{Tr}((T+\lambda)^{-2}T_x)\, d\rho_X(x), \\
((13)) \quad &\le \kappa \operatorname{Tr}((T+\lambda)^{-2}T) \\
&= \kappa \operatorname{Tr}((T+\lambda)^{-1}((T+\lambda)^{-1/2}T(T+\lambda)^{-1/2})), \\
((40)) \quad &\le \kappa \|(T+\lambda)^{-1}\|_{\mathcal{L}(\mathcal{H})} \operatorname{Tr}((T+\lambda)^{-1}T) \\
&\le \frac{\kappa}{\lambda}\mathcal{N}(\lambda) = \sigma_1^2,
\end{aligned}
$$

by definition of effective dimension $\mathcal{N}(\lambda)$. Hence, (33) of Proposition 2 holds and (32) gives

$$
\|(T+\lambda)^{-1}(T_{\mathbf{x}}-T)\|_{\mathcal{L}_2(\mathcal{H})} \le 2\log(6/\eta)\left(\frac{2\kappa}{\lambda\ell} + \sqrt{\frac{\kappa\mathcal{N}(\lambda)}{\lambda\ell}}\right)
$$

with probability greater than $1-\eta/3$. Since $\log(6/\eta) \ge 1$ and the spectral decomposition of T gives

$$
\mathcal{N}(\lambda) \ge \frac{\|T\|_{\mathcal{L}(\mathcal{H})}}{\|T\|_{\mathcal{L}(\mathcal{H})}+\lambda} \ge \frac{1}{2} \qquad \text{if} \quad \lambda \le \|T\|_{\mathcal{L}(\mathcal{H})},
$$

if (35) holds, then

$$
\begin{aligned}
\log(6/\eta)\left(\frac{2\kappa}{\lambda\ell} + \sqrt{\frac{\kappa\mathcal{N}(\lambda)}{\lambda\ell}}\right) &\le 4\frac{\log^2(6/\eta)\kappa\mathcal{N}(\lambda)}{\lambda\ell} + \sqrt{\frac{\log^2(6/\eta)\kappa\mathcal{N}(\lambda)}{\lambda\ell}} \\
&\le \tfrac{1}{16} + \tfrac{1}{8} \le \tfrac{1}{4},
\end{aligned}
$$

so that

$$
\Theta(\lambda, \mathbf{z}) \le \|(T+\lambda)^{-1}(T_{\mathbf{x}}-T)\|_{\mathcal{L}_2(\mathcal{H})} \le \tfrac{1}{2} \tag{41}
$$

with probability greater than $1-\eta/3$.

Step 2.2. Probabilistic bound on $\left\|(T-T_{\mathbf{x}})(f^\lambda - f_{\mathcal{H}})\right\|_{\mathcal{L}(\mathcal{H})}$. Now we apply Proposition 2 to the random variable $\xi_2 : X \to \mathcal{H}$,

$$
\xi_2(x) = T_x(f^\lambda - f_{\mathcal{H}}),
$$

so that

$$
\mathbb{E}[\xi_2] = T(f^\lambda - f_{\mathcal{H}}), \qquad \frac{1}{\ell}\sum_{i=1}^{\ell} \xi_2(x_i) = T_{\mathbf{x}}(f^\lambda - f_{\mathcal{H}}).
$$

Bound (13) and the definition of $\mathcal{B}(\lambda)$ give

$$\|\xi_2(x)\|_{\mathcal{H}} \leq \|T_x\|_{\mathcal{L}(\mathcal{H})} \left\| f^\lambda - f_{\mathcal{H}} \right\|_{\mathcal{H}} \leq \kappa \sqrt{\mathcal{B}(\lambda)} = \frac{L_2}{2}.$$

Since T_x is a positive operator

$$\langle T_x f, f \rangle_{\mathcal{H}} \leq \|T_x\|_{\mathcal{L}(\mathcal{H})} \langle f, f \rangle_{\mathcal{H}}, \qquad f \in \mathcal{H}, \tag{42}$$

so that

$$\begin{aligned}
\mathbb{E}[\|\xi_2\|^2_{\mathcal{H}}] &= \int_X \langle T_x T_x^{1/2}(f^\lambda - f_{\mathcal{H}}), T_x^{1/2}(f^\lambda - f_{\mathcal{H}}) \rangle_{\mathcal{H}} \, d\rho_X(x) \\
&\leq \int_X \|T_x\|_{\mathcal{L}(\mathcal{H})} \langle T_x(f^\lambda - f_{\mathcal{H}}), f^\lambda - f_{\mathcal{H}} \rangle_{\mathcal{H}} \, d\rho_X(x), \\
((13)) \quad &\leq \kappa \langle T(f^\lambda - f_{\mathcal{H}}), f^\lambda - f_{\mathcal{H}} \rangle_{\mathcal{H}} \\
&= \kappa \|\sqrt{T}(f^\lambda - f_{\mathcal{H}})\|^2_{\mathcal{H}} \\
&= \kappa \mathcal{A}(\lambda) = \sigma_2^2,
\end{aligned}$$

by definition of $\mathcal{A}(\lambda)$. So (33) holds and (32) gives

$$\|(T - T_{\mathbf{x}})(f^\lambda - f_{\mathcal{H}})\|_{\mathcal{H}} \leq 2 \log(6/\eta) \left(\frac{2\kappa \sqrt{\mathcal{B}(\lambda)}}{\ell} + \sqrt{\frac{\kappa \mathcal{A}(\lambda)}{\ell}} \right), \tag{43}$$

with probability greater than $1 - \eta/3$. Replacing (39), (43) in (37), if (35) holds, it follows that

$$\mathcal{S}_2(\lambda, \mathbf{z}) \leq 8 \log^2(6/\eta) \left(\frac{4\kappa^2 \mathcal{B}(\lambda)}{\ell^2 \lambda} + \frac{\kappa \mathcal{A}(\lambda)}{\ell \lambda} \right) \tag{44}$$

with probability greater than $1 - 2\eta/3$.

Step 3. Probabilistic bound on $\mathcal{S}_1(\lambda, \mathbf{z})$. Clearly,

$$\mathcal{S}_1(\lambda, \mathbf{z}) \leq \|\sqrt{T}(T_{\mathbf{x}} + \lambda)^{-1}(T + \lambda)^{1/2}\|^2_{\mathcal{L}(\mathcal{H})} \|(T + \lambda)^{-1/2}(g_{\mathbf{z}} - T_{\mathbf{x}} f_{\mathcal{H}})\|^2_{\mathcal{H}}. \tag{45}$$

Step 3.1. Bound on $\|\sqrt{T}(T_{\mathbf{x}} + \lambda)^{-1}(T + \lambda)^{1/2}\|_{\mathcal{L}(\mathcal{H})}$. Since

$$\sqrt{T}(T_{\mathbf{x}}+\lambda)^{-1}(T+\lambda)^{1/2} = \sqrt{T}(T+\lambda)^{-1/2}\{I-(T+\lambda)^{-1/2}(T-T_{\mathbf{x}})(T+\lambda)^{-1/2}\}^{-1},$$

reasoning as in Step 2.1, it follows that

$$\|\{I - (T + \lambda)^{-1/2}(T - T_{\mathbf{x}})(T + \lambda)^{-1/2}\}^{-1}\|_{\mathcal{L}(\mathcal{H})} \leq 2$$

provided that

$$\|(T + \lambda)^{-1/2}(T - T_{\mathbf{x}})(T + \lambda)^{-1/2}\|_{\mathcal{L}(\mathcal{H})} \leq \tfrac{1}{2}. \tag{46}$$

Moreover, the spectral theorem ensures that $\|\sqrt{T}(T+\lambda)^{-1/2}\|_{\mathcal{L}(\mathcal{H})} \leq 1$ so

$$\|\sqrt{T}(T_{\mathbf{x}}+\lambda)^{-1}(T+\lambda)^{1/2}\|_{\mathcal{L}(\mathcal{H})} \leq 2. \tag{47}$$

We will show that (46) holds for the training sets $\mathbf{z}$ satisfying (41). Indeed, if $B=(T+\lambda)^{-1/2}(T-T_{\mathbf{x}})(T+\lambda)^{-1/2}$, then

$$\begin{aligned}
\|B\|^2_{\mathcal{L}_2(\mathcal{H})} &= \operatorname{Tr}((T+\lambda)^{-1/2}(T-T_{\mathbf{x}})(T+\lambda)^{-1}(T-T_{\mathbf{x}})(T+\lambda)^{-1/2}) \\
&= \operatorname{Tr}((T+\lambda)^{-1}(T-T_{\mathbf{x}})(T+\lambda)^{-1}(T-T_{\mathbf{x}})) \\
&= \langle (T+\lambda)^{-1}(T-T_{\mathbf{x}}), ((T+\lambda)^{-1}(T-T_{\mathbf{x}}))^* \rangle_{\mathcal{L}_2(\mathcal{H})} \\
&\leq \|(T+\lambda)^{-1}(T-T_{\mathbf{x}})\|_{\mathcal{L}_2(\mathcal{H})} \|((T+\lambda)^{-1}(T-T_{\mathbf{x}}))^*\|_{\mathcal{L}_2(\mathcal{H})} \\
&= \|(T+\lambda)^{-1}(T-T_{\mathbf{x}})\|^2_{\mathcal{L}_2(\mathcal{H})}.
\end{aligned}$$

If (35) holds, then (41) ensures that (46) holds with probability $1-2\eta/3$.

Step 3.2. Bound on $\|(T+\lambda)^{-1/2}(g_{\mathbf{z}}-T_{\mathbf{x}} f_{\mathcal{H}})\|_{\mathcal{H}}$. Let $\xi_3 : Z \to \mathcal{H}$ be the random variable

$$\xi_3(x,y) = (T+\lambda)^{-1/2} K_x(y - f_{\mathcal{H}}(x)).$$

First of all, (22) gives

$$\mathbb{E}[\xi_3] = (T+\lambda)^{-1/2}(g - T f_{\mathcal{H}}) = 0$$

and (9) implies

$$\int_Y \|y - f_{\mathcal{H}}(x)\|_Y^m \, d\rho(y|x) \leq \tfrac{1}{2} m! \, \Sigma^2 M^{m-2} \qquad \text{for all } m \geq 2,$$

(see, e.g., [**31**]). It follows that

$$\begin{aligned}
\mathbb{E}[\|\xi_3\|^m_{\mathcal{H}}] &= \int_Z (\langle K_x^*(T+\lambda)^{-1} K_x(y - f_{\mathcal{H}}(x)), y - f_{\mathcal{H}}(x)\rangle_Y)^{m/2} \, d\rho(x,y), \\
\text{(Eq. (42))} &\leq \int_X \|K_x^*(T+\lambda)^{-1}K_x\|^{m/2}_{\mathcal{L}(\mathcal{H})} \left(\int_Y \|y - f_{\mathcal{H}}(x)\|_Y^m \, d\rho(y|x) \right) d\rho_X(x) \\
&\qquad (\|K_x^*(T+\lambda)^{-1}K_x\|_{\mathcal{L}(\mathcal{H})} \leq \operatorname{Tr}(K_x^*(T+\lambda)^{-1}K_x)) \\
&\leq \frac{m! \, \Sigma^2 M^{m-2}}{2} \sup_{x \in X} \|K_x^*(T+\lambda)^{-1}K_x\|^{(m-2)/2}_{\mathcal{L}(\mathcal{H})} \\
&\qquad \times \int_X \operatorname{Tr}(T+\lambda)^{-1} T_x \, d\rho_X(x) \quad ((13) \text{ and } \|(T+\lambda)^{-1}\|_{\mathcal{L}(\mathcal{H})} \leq \lambda^{-1}) \\
&\leq \tfrac{1}{2} m! \, \Sigma^2 M^{m-2} \left(\sqrt{\frac{\kappa}{\lambda}} \right)^{m-2} \operatorname{Tr}[(T+\lambda)^{-1} T] \\
&= \tfrac{1}{2} m! \, (\Sigma \sqrt{\mathcal{N}(\lambda)})^2 \left(M \sqrt{\frac{\kappa}{\lambda}} \right)^{m-2}.
\end{aligned}$$

Hence (31) holds with $L_3 = M\sqrt{\kappa/\lambda}$ and $\sigma_3 = \Sigma\sqrt{\mathcal{N}(\lambda)}$ and (32) gives that

$$\|(T+\lambda)^{-1/2}(g_{\mathbf{z}} - T_{\mathbf{x}} f_{\mathcal{H}})\|_{\mathcal{H}} \leq 2\log(6/\eta)\left(\frac{1}{\ell}\sqrt{M^2\frac{\kappa}{\lambda}} + \sqrt{\frac{\Sigma^2\mathcal{N}(\lambda)}{\ell}}\right) \quad (48)$$

with probability greater than $1-\eta/3$. Replacing (47), (48) in (45),

$$\mathcal{S}_1(\lambda, \mathbf{z}) \leq 32\log^2(6/\eta)\left(\frac{\kappa M^2}{\ell^2\lambda} + \frac{\Sigma^2\mathcal{N}(\lambda)}{\ell}\right) \quad (49)$$

with probability greater than $1-2\eta/3$.

Replacing bounds (44), (49) in (36),

$$\begin{aligned}\mathcal{E}[f_{\mathbf{z}}^{\lambda}] - \mathcal{E}[f_{\mathcal{H}}] \;\leq\; & 3\mathcal{A}(\lambda) \\ & + 8\log^2(6/\eta)\left(\frac{4\kappa^2\mathcal{B}(\lambda)}{\ell^2\lambda} + \frac{\kappa\mathcal{A}(\lambda)}{\ell\lambda} + \frac{4\kappa M^2}{\ell^2\lambda} + \frac{4\Sigma^2\mathcal{N}(\lambda)}{\ell}\right)\end{aligned}$$

and (34) follows by bounding the numerical constants with 32. □

The second step in the proof of the upper bound is the study of the asymptotic behavior of $\mathcal{N}(\lambda)$, $\mathcal{A}(\lambda)$, and $\mathcal{B}(\lambda)$ when λ goes to zero. It is known that

$$\lim_{\lambda\to 0}\mathcal{A}(\lambda) = 0,$$

$$\lim_{\lambda\to 0}\mathcal{B}(\lambda) = 0,$$

$$\lim_{\lambda\to 0}\mathcal{N}(\lambda) = N,$$

see, e.g., [**16**] and [**13**]. However, to state uniform rates of convergence we need some prior assumptions on the distribution ρ.

Proposition 3. *Let $\rho \in \mathcal{P}(b, c)$ with $1 \leq c \leq 2$ and $1 < b \leq +\infty$, then*

$$\mathcal{A}(\lambda) \leq \lambda^c \|T^{(1-c)/2} f_{\mathcal{H}}\|^2_{\mathcal{H}}$$

and

$$\mathcal{B}(\lambda) \leq \lambda^{c-1} \|T^{(1-c)/2} f_{\mathcal{H}}\|^2_{\mathcal{H}}.$$

Moreover, if $b < +\infty$ ($N = +\infty$),

$$\mathcal{N}(\lambda) \leq \frac{\beta b}{b-1}\lambda^{-1/b}.$$

Instead, if $b = +\infty$ ($N < +\infty$),

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

Proof. The results about $\mathcal{A}(\lambda)$ and $\mathcal{B}(\lambda)$ are standard in the theory of inverse problems, see, for example, [**16**], [**13**] and, in the context of learning, [**8**].

We study $\mathcal{N}(\lambda)$ under the assumption that $N = +\infty$ and $t_n \leq \beta/n^b$. Since the function $t/(t+\lambda)$ is increasing in t,

$$\mathcal{N}(\lambda) = \sum_{n=1}^{N} \frac{t_n}{t_n + \lambda} \leq \sum_{n=1}^{N} \frac{\beta}{\beta + n^b \lambda}.$$

The function $\beta/(\beta + x^b \lambda)$ is positive and decreasing, so

$$\begin{aligned}
\mathcal{N}(\lambda) \ &\leq \ \int_0^{\infty} \frac{\beta}{\beta + x^b \lambda}\, dx, \\
(\tau^b = x^b \lambda) \ &= \ \lambda^{-1/b} \int_0^{+\infty} \frac{\beta}{\beta + \tau^b}\, d\tau \\
&\leq \ \beta \frac{b}{b-1} \lambda^{-1/b},
\end{aligned}$$

since $\int_0^{+\infty} (\beta + \tau^b)^{-1} \leq b/(b-1)$. If N is finite, since $t/(t+\lambda)$ is a decreasing function of λ the claim follows. $\square$

We are now ready to prove the theorem.

Proof of Theorem 1. Let $1 < b \leq +\infty$ and $1 \leq c \leq 2$ be as in the statement of the theorem. For any $\rho \in \mathcal{P}(b, c)$, Proposition 3 and Theorem 4 imply that, given $0 < \eta < 1$, with probability greater than $1 - \eta$ it holds that

$$\mathcal{E}[f_{\mathbf{z}}^{\lambda}] - \mathcal{E}[f_{\mathcal{H}}] \leq C_\eta \left(R\lambda^c + \frac{\kappa^2 R \lambda^{c-2}}{\ell^2} + \frac{\kappa R \lambda^{c-1}}{\ell} + \frac{\kappa M^2}{\ell^2 \lambda} + \frac{\Sigma^2 \beta b}{(b-1)\ell \lambda^{1/b}} \right) \tag{50}$$

for all $\ell \in \mathbb{N}$ and $0 < \lambda \leq \|T\|_{\mathcal{L}(\mathcal{H})}$ satisfying

$$\ell \geq \frac{2C_\eta \kappa \beta b}{(b-1)\lambda^{(b+1)/b}}. \tag{51}$$

In the case $b = +\infty$ the above formulas hold adopting the formal identities $\lambda^{1/b} \equiv b/(b-1) \equiv 1$ and $\lambda^{(b+1)/b} \equiv \lambda$.

Assume now that $b < +\infty$ and $c > 1$. Given $\eta \in (0, 1)$, let

$$\ell_\eta \geq \left(\frac{2C_\eta \kappa \beta b}{(b-1)} \right)^{(bc+1)/b(c-1)}$$

(recall that $c > 1$). Then

$$\ell^{b(c-1)/(bc+1)} \geq \frac{2C_\eta \kappa \beta b}{(b-1)} \quad \text{for all } \ell \geq \ell_\eta,$$

so that, since $\lambda_\ell = \ell^{-b/(bc+1)}$,

$$\ell \geq \frac{2C_\eta \kappa\beta b}{(b-1){\lambda_\ell}^{(b+1)/b}} \quad \text{for all } \ell \geq \ell_\eta.$$

So, for any $\rho \in \mathcal{P}$, bound (50) holds with $\lambda = \lambda_\ell$. Since $b/(bc+1) < 1$, $\lambda_\ell \ell$ goes to $+\infty$, so that

$$\mathcal{E}[f_{\mathbf{z}}^{\lambda}] - \mathcal{E}[f_{\mathcal{H}}] \leq C_\eta D({\lambda_\ell}^c + \ell^{-1}{\lambda_\ell}^{-1/b}) = 2C_\eta D\ell^{-bc/(bc+1)} \quad \text{for all } \ell \geq \ell_\eta,$$

with probability greater than $1-\eta$, where D is a constant depending only on R, κ, M, Σ β, b, and c.

Let now $\tau = 2C_\eta D$ and solve this equation for τ, so that

$$\eta = \eta_\tau = 6e^{-\sqrt{\tau/64D}}.$$

Hence

$$\mathbb{P}_{\mathbf{z}\sim\rho^\ell}[\mathcal{E}[f_{\mathbf{z}}^{\lambda_\ell}] - \mathcal{E}[f_{\mathcal{H}}] > \tau\ell^{-bc/(bc+1)}] \leq \eta_\tau \quad \text{for all } \ell \geq \ell_{\eta_\tau}.$$

So that

$$\limsup_{\ell\to\infty} \sup_{\rho\in\mathcal{P}(b,c)} \mathbb{P}_{\mathbf{z}\sim\rho^\ell}[\mathcal{E}[f_{\mathbf{z}}^{\lambda_\ell}] - \mathcal{E}[f_{\mathcal{H}}] > \tau\ell^{-bc/(bc+1)}] \leq \eta_\tau.$$

Since $\lim_{\tau\to+\infty} \eta_\tau = \lim_{\tau\to+\infty} 6e^{-\sqrt{\tau/64D}} = 0$, the thesis follows.

Assume now $b < +\infty$ and $c = 1$. Then

$$\frac{2C_\eta \kappa\beta b}{(b-1){\lambda_\ell}^{(b+1)/b}} = \frac{2C_\eta \kappa\beta b\ell}{(b-1)\log \ell},$$

so there is ℓ_η such that

$$\ell \geq \frac{2C_\eta \kappa\beta b}{(b-1){\lambda_\ell}^{(b+1)/b}} \quad \text{for all } \ell \geq \ell_\eta.$$

Reasoning as above and taking into account that $1/\ell\lambda_\ell$ goes to zero faster than λ_ℓ, for any $\rho \in \mathcal{P}$ the bound (50) gives

$$\mathcal{E}[f_{\mathbf{z}}^{\lambda}] - \mathcal{E}[f_{\mathcal{H}}] \leq C_\eta D'{\lambda_\ell}^c = C_\eta D'\ell^{-b/(b+1)} \quad \text{for all } \ell \geq \ell_\eta$$

with probability greater than $1-\eta$, where D' is a constant depending only on R, κ, M, Σ, β, and b. The proof now follows reasoning as above.

The proofs for $b = +\infty$ ($N < +\infty$) are similar. Moreover, in this finite-dimensional case the semi-norms $\|T^{(1-c)/2}f\|_{\mathcal{H}}$ for different values of the parameter c are equivalent. Hence the final rates are not dependent on c. $\square$

5.3. *Minimax Lower Rate*

We assume now that Y is finite dimensional with $d = \dim Y$ and $N = +\infty$, we fix $1 < b < +\infty$, $1 \le c \le 2$, and $M, \Sigma, R, \alpha, \beta$ as in the definition of $\mathcal{P}(b, c)$.

To prove the lower bound we follow the ideas of [**10**]. The main steps are the following. First, we define a family of probability distributions $\rho_f \in \mathcal{P}(b, c)$ parametrized by suitable vectors $f \in \mathcal{H}$. Then, for all $0 < \varepsilon \le \varepsilon_0$, we construct a finite sequence of vectors $f_1, \ldots, f_{N_\varepsilon}$ such that $N_\varepsilon \ge e^{\gamma\varepsilon^{-1/bc}}$ and the Kullback–Leibler information

$$\mathcal{K}(\rho_{f_i}, \rho_{f_j}) \le C\varepsilon, \qquad i \ne j,$$

where γ and C depend only on $\mathcal{P}$. Finally, we apply a theorem of [**10**] to obtain the claimed lower bound.

We recall that the Kullback–Leibler information of two measures ρ_1 and ρ_2 is defined by

$$\mathcal{K}(\rho_1, \rho_2) = \int \log \varphi(z)\, d\rho_1(z),$$

where φ is the density of ρ_1 with respect to ρ_2, that is, $\rho_1(E) = \int_E \varphi(z)\, d\rho_2(z)$ for all measurable sets E.

In the following, we choose $\rho_0 \in \mathcal{P}(b, c)$ and we let ν be its marginal measure. Since ρ_0 satisfies Hypothesis 2, the operator T has the spectral decomposition

$$T = \int_X T_x\, d\nu(x) = \sum_{n=1}^{+\infty} t_n \langle \cdot, e_n\rangle e_n, \tag{52}$$

where $(e_n)_{n=1}^{+\infty}$ is an orthonormal sequence in $\mathcal{H}$ and, since $\rho_0 \in \mathcal{P}(b, c)$,

$$t_n \ge \frac{\alpha}{n^b}, \qquad n \ge 1.$$

The proposition below associates to any vector f belonging to a suitable subclass of $\mathcal{H}$, a corresponding probability measure ρ_f that belongs to $\mathcal{P}(b, c)$. In particular, ρ_f will have the same marginal distribution ν, so that the corresponding operator T_{ρ_f} defined by (14) with $\rho_X = (\rho_f)_X$ is, in fact, given by (52).

Proposition 4. *Let $(v_j)_{j=1}^d$ be a basis of Y. Given $f \in \mathcal{H}$ such that $f = T^{(c-1)/2} g$ for some $g \in \mathcal{H}$, $\|g\|^2 \le R$, let $\rho_f(x, y) = \nu(x)\rho_f(y|x)$ where*

$$\rho_f(y|x) = \frac{1}{2dL} \sum_{j=1}^{d} \big((L - \langle f, K_x v_j\rangle_{\mathcal{H}})\delta_{y+dLv_j} + (L + \langle f, K_x v_j\rangle_{\mathcal{H}})\delta_{y-dLv_j}\big)$$

with $L = 4\sqrt{\kappa^c R}$ and $\delta_{y\pm dLv_j}$ is the Dirac measure on Y at point $\mp dLv_j$. Then ρ_f is a probability measure with marginal distribution $(\rho_f)_X = \nu$ and regression function $f_{\rho_f} = f \in \mathcal{H}$. Moreover, $\rho_f \in \mathcal{P}(b, c)$ provided that

$$\min(M, \Sigma) \ge 2(4d + 1)\sqrt{\kappa^c R}. \tag{53}$$

Moreover, if $f' \in \mathcal{H}$ *such that* $f' = T^{(c-1)/2}g'$ *for some* $g' \in \mathcal{H}$, $\|g'\|^2 \leq R$, *then the Kullback–Leibler information* $\mathcal{K}(\rho_f, \rho_{f'})$ *fulfills the inequality*

$$\mathcal{K}(\rho_f, \rho_{f'}) \leq \frac{16}{15dL^2}\|\sqrt{T}(f - f')\|_{\mathcal{H}}^2. \tag{54}$$

Proof. The definition of f, (11), and (13) imply

$$|\langle f, K_x v_j\rangle_{\mathcal{H}}| \leq \|T^{(c-1)/2}g\|_{\mathcal{H}}\, \|K_x\|_{\mathcal{L}(Y,\mathcal{H})} \leq \kappa^{c/2}\sqrt{R} = \frac{L}{4}. \tag{55}$$

It follows that $\rho_f(y|x)$ is a probability measure on Y and

$$\int_Y y\, d\rho_f(y|x) = \sum_j \left\langle f, K_x v_j\right\rangle_{\mathcal{H}} v_j = K_x^* f = f(x).$$

So that ρ_f is a probability measure on Z, the marginal distribution is ν and the regression function $f_{\rho_f} = f \in \mathcal{H}$. In particular, condition (8) holds with $f_{\mathcal{H}} = f$ and items (ii) and (iii) of Definition 1 are satisfied.

Clearly, (7) is satisfied since $\rho_f(y|x)$ has finite support. Moreover, (53) ensures

$$\|y - f(x)\|_Y \leq \|y\|_Y + \left\|K_x^* f\right\|_Y \leq dL + \sqrt{\kappa^c R} = (4d+1)\sqrt{\kappa^c R} \leq M$$

and

$$\begin{aligned}
\mathbb{E}[\|y - f(x)\|_Y^2] \;&=\; \frac{1}{2dL}\sum_j (L - \left\langle f, K_x v_j\right\rangle_{\mathcal{H}})(dL + \left\langle f, K_x v_j\right\rangle_{\mathcal{H}})^2 \\
&\quad + (L + \left\langle f, K_x v_j\right\rangle_{\mathcal{H}})(dL - \left\langle f, K_x v_j\right\rangle_{\mathcal{H}})^2 \\
&\quad + \left(1 - \frac{1}{d}\right)\|f(x)\|_Y^2 \\
((55)) \;&\leq\; \tfrac{5}{4}L^2 + \left(1 - \frac{1}{d}\right)\kappa^c R \leq 4(4d+1)\kappa^c R \leq \Sigma^2,
\end{aligned}$$

so that (9) is satisfied.

The proof of (54) is the same as Lemma 3.2 of [**10**]. We only sketch the main steps. If $\varphi = d\rho_f/d\rho_{f'}$, clearly

$$\begin{aligned}
\log\varphi(x, \pm dLv_j) \;&=\; \log\left(\frac{L \pm \left\langle f, K_x v_j\right\rangle_{\mathcal{H}}}{L \pm \left\langle f', K_x v_j\right\rangle_{\mathcal{H}}}\right) \\
&=\; \log\left(1 \pm \frac{\left\langle f - f', K_x v_j\right\rangle_{\mathcal{H}}}{L \pm \left\langle f', K_x v_j\right\rangle_{\mathcal{H}}}\right) \\
&\leq\; \pm\frac{\left\langle f - f', K_x v_j\right\rangle_{\mathcal{H}}}{L \pm \left\langle f', K_x v_j\right\rangle_{\mathcal{H}}},
\end{aligned}$$

so that

$$\begin{aligned}
\mathcal{K}(\rho_{f'}, \rho_f) &\leq \frac{1}{2dL} \sum_{j=1}^{d} \int_X \left(\frac{\left\langle f - f', K_x v_j \right\rangle_{\mathcal{H}}}{L + \left\langle f', K_x v_j \right\rangle_{\mathcal{H}}} (L + \langle f, K_x v_j \rangle_{\mathcal{H}}) \right. \\
&\quad \left. + \frac{\left\langle -f + f', K_x v_j \right\rangle_{\mathcal{H}}}{L - \left\langle f', K_x v_j \right\rangle_{\mathcal{H}}} (L - \left\langle f, K_x v_j \right\rangle_{\mathcal{H}}) \right) d\nu(x) \\
&= \frac{1}{d} \sum_{j=1}^{d} \int_X \frac{(\left\langle f - f', K_x v_j \right\rangle_{\mathcal{H}})^2}{L^2 - (\left\langle f', K_x v_j \right\rangle_{\mathcal{H}})^2} \, d\nu(x) \\
((55)) \quad &\leq \frac{1}{2d} \sum_{j=1}^{d} \int_X \left\langle f - f', K_x v_j \right\rangle_{\mathcal{H}}^2 \frac{16}{15L^2} \, d\nu(x) \\
&= \frac{16}{15dL^2} \int_X \left\| K_x^*(f - f') \right\|_Y^2 \, d\nu(x) \\
&= \frac{16}{15dL^2} \int_X \left\langle T_x(f - f'), f - f' \right\rangle_{\mathcal{H}} \, d\nu(x) \\
&= \frac{16}{15dL^2} \left\langle T(f - f'), f - f' \right\rangle_{\mathcal{H}}. \qquad \square
\end{aligned}$$

Proposition 5. *There is an $\varepsilon_0 > 0$ such that for all $0 < \varepsilon \leq \varepsilon_0$, there exist $N_\varepsilon \in \mathbb{N}$ and $f_1, \ldots, f_{N_\varepsilon} \in \mathcal{H}$ (depending on ε) satisfying*:

(i) *for all $i = 1, \ldots, N_\varepsilon$, $f_i = T^{(c-1)/2} g_i$ for some $g_i \in \mathcal{H}$ with $\|g_i\|_{\mathcal{H}}^2 \leq R$*;
(ii) *for all $i, j = 1, \ldots, N_\varepsilon$,*

$$\varepsilon \leq \|\sqrt{T}(f_i - f_j)\|_{\mathcal{H}}^2 \leq 4\varepsilon; \tag{56}$$

(iii) *there is a constant γ depending only on R and α such that*

$$N_\varepsilon \geq e^{\gamma \varepsilon^{-1/bc}}. \tag{57}$$

Proof. Let $m \in \mathbb{N}$ such that $m > 16$ and $\sigma_1, \ldots, \sigma_N \in \{1, -1\}^m$ given by Proposition 6 so that

$$\sum_{n=1}^{m} (\sigma_i^n - \sigma_j^n)^2 \geq m, \tag{58}$$

$$N \geq e^{m/24}. \tag{59}$$

In the following we will choose m as a function of ε is such a way that the statement of the proposition will be true.

Given $\varepsilon > 0$, for all $i = 1, \ldots, N_\varepsilon$, let

$$g_i = \sum_{n=1}^{m} \sqrt{\frac{\varepsilon}{m t_n^c}} \sigma_i^n e_n,$$

(see (52)). Since $t_n n^b \geq \alpha$, then

$$\|g_i\|_{\mathcal{H}}^2 = \sum_{n=1}^{m} \frac{\varepsilon}{m t_n^c} \leq \frac{\varepsilon}{m} \sum_{n=1}^{m} \left(\frac{n^b}{\alpha}\right)^c \leq C \varepsilon m^{bc},$$

where here and in the following C is a constant depending only on R, α, b, and c. Hence $\|g_i\|^2 \leq R$ provided that

$$\varepsilon m^{bc} \leq \frac{R}{C}$$

and we let $m = m_\varepsilon \in \mathbb{N}$ be

$$m = \lfloor C' \varepsilon^{-1/bc} \rfloor \tag{60}$$

for a suitable constant $C' > 0$ (where $\lfloor x \rfloor$ is the greatest integer less than or equal to x.) Clearly, since m_ε goes to $+\infty$ if ε goes to 0, there is ε_0 such that $m_\varepsilon > 16$ for all $\varepsilon \leq \varepsilon_0$.

Let now $f_i = T^{(c-1)/2} g_i$, as in the statement of the theorem, then

$$\|\sqrt{T}(f_i - f_j)\|_{\mathcal{H}}^2 = \|T^{\frac{c}{2}}(g_i - g_j)\|_{\mathcal{H}}^2 = \sum_{n=1}^{m} \frac{\varepsilon}{m} (\sigma_i^n - \sigma_j^n)^2.$$

The conditions (58) and $(\sigma_i^n - \sigma_j^n)^2 \leq 4$ imply

$$\varepsilon \leq \|\sqrt{T}(f_i - f_j)\|_{\mathcal{H}}^2 \leq 4\varepsilon$$

and (59) and (60) ensure

$$N_\varepsilon \geq e^{m/24} \geq e^{\gamma \varepsilon^{-1/bc}}$$

for a suitable constant $\gamma > 0$. □

The proof of the above proposition relies on the following result regarding packing numbers over sets of binary strings.

Proposition 6. *For every $m > 16$ there exist $N \in \mathbb{N}$ and $\sigma_1, \ldots, \sigma_N \in \{-1, +1\}^m$ such that*

$$\sum_{n=1}^{m} (\sigma_i^n - \sigma_j^n)^2 \geq m, \qquad i \neq j, i = 1, \ldots, N,$$

$$N \geq e^{m/24},$$

where $\sigma_i = (\sigma_i^1, \ldots, \sigma_i^m)$ and $\sigma_j = (\sigma_j^1, \ldots, \sigma_j^m)$.

Proof. We regard the vectors $\sigma \in \{-1, +1\}^m$ as a set of m i.i.d. binary random variables distributed according to the uniform distribution $1/2(\delta_{-1} + \delta_{+1})$.

Let σ and σ' be two independent random vectors in $\{-1, +1\}^m$, then the real random variable

$$\mathrm{d}(\sigma, \sigma') = \sum_{n=1}^{m} (\sigma_i^n - \sigma_j^n)^2 = \sum_{n=1}^{m} \theta_n,$$

where θ_n are independent random variables distributed according to the measure $1/2(\delta_0 + \delta_4)$. The expectation value $\mathrm{d}(\sigma, \sigma')$ is $2m$ and the Hoeffding inequality ensures that, for every $\delta > 0$,

$$\mathbb{P}[|\mathrm{d}(\sigma, \sigma') - 2m| > \delta] \leq 2 \exp\left(-\frac{\delta^2}{8m}\right).$$

Setting $\delta = m$ in the inequality above, we obtain

$$\mathbb{P}[\mathrm{d}(\sigma, \sigma') < m] \leq 2 \exp\left(-\frac{m}{8}\right). \tag{61}$$

Now draw $N := \lceil e^{m/24} \rceil$ (where $\lceil x \rceil$ is the lowest integer greater than x) independent random points σ_i $(i = 1, \dots, N)$.

From inequality (61), by union bound it holds that

$$\begin{aligned}
&\mathbb{P}\left[\exists\, 1 \leq i, j \leq N,\ i \neq j, \quad \text{with } \mathrm{d}(\sigma_i, \sigma_j) < m\right] \\
&\qquad \leq (N^2 - N) \exp\left(-\frac{m}{8}\right) \leq \frac{N^2 - N}{(N-1)^3} = \frac{N}{(N-1)^2} < 1,
\end{aligned}$$

since the definition of N and the assumption $m > 16$ imply that $(N-1)^2 > N$ and $(N-1)^3 < \exp m/8$. It follows that there exists at least a sequence $(\sigma_1, \dots, \sigma_N)$ such that $\mathrm{d}(\sigma_i, \sigma_j) \geq m$ for all $i \neq j$ and $N > \exp m/8$. $\square$

The following theorem is a restatement of Theorem 3.1 of [**10**] in our setting.

Theorem 5. *Assume* (53) *and consider an arbitrary learning algorithm* $\mathbf{z} \mapsto f_{\mathbf{z}}^{\ell} \in \mathcal{H}$, *for* $\ell \in \mathbb{N}$ *and* $\mathbf{z} \in Z^{\ell}$. *Then, for all* $\varepsilon \leq \varepsilon_0$ *and for all* $\ell \in \mathbb{N}$, *there is a* $\rho_* \in \mathcal{P}(b, c)$ *such that* $f_{\rho_*} \in \mathcal{H}$ *and it holds that*

$$\mathbb{P}_{\mathbf{z} \sim \rho_*^{\ell}}\left[\mathcal{E}_{\rho_*}[f_{\mathbf{z}}^{\ell}] - \mathcal{E}_{\rho_*}[f_{\rho_*}] > \frac{\varepsilon}{4}\right] \geq \min\left\{\frac{N_\varepsilon^*}{N_\varepsilon^* + 1}, \bar{\eta}\sqrt{N_\varepsilon^*}\, e^{-4\ell\varepsilon/15d\kappa^c R}\right\},$$

where $N_\varepsilon^* = e^{\gamma \varepsilon^{-1/bc}}$ *and* $\bar{\eta} = e^{-3/e}$.

Proof. The proof is the same as in [**10**]. Given $\varepsilon \leq \varepsilon_0$, let N_ε and $f_1, \dots, f_{N_\varepsilon}$ be as in Proposition 5. According to Proposition 4, let $\rho_i = \rho_{f_i}$. Assumption (53) ensures that $\rho_i \in \mathcal{P}(b, c)$.

Observe that, since all the measures ρ_i have the same marginal distribution ν and $f_{\mathcal{H}} = f_{\rho_i} = f_i$,

$$\mathcal{E}_{\rho_i}[f] - \mathcal{E}_{\rho_i}[f_{\rho_i}] = \|\sqrt{T}(f - f_i)\|^2_{\mathcal{H}} = \int_X \|f_i(x) - f(x)\|^2_Y \, d\nu(x).$$

Given $\ell \in \mathbb{N}$, let

$$A_i = \left\{ \mathbf{z} \in Z^\ell \mid \|\sqrt{T}(f^\ell_{\mathbf{z}} - f_i)\|^2_{\mathcal{H}} < \frac{\varepsilon}{4} \right\}$$

for all $i = 1, \ldots, N_\varepsilon$. The lower bound of (56) ensures that $A_i \cap A_j = \emptyset$ if $i \neq j$, so that Lemma 3.3 of [**10**] ensures that there is $\rho_* = \rho_{i_*}$ such that either

$$p_* = \mathbb{P}_{\rho_*^\ell}[A_{i_*}] > \frac{N_\varepsilon}{N_\varepsilon + 1} \geq \frac{N^*_\varepsilon}{N^*_\varepsilon + 1}$$

(since $x/(x+1)$ is an increasing function and (57) holds) or, replacing the upper bound of (56), in Eq. 3.12 of [**10**],

$$\frac{4\ell\varepsilon}{15d\kappa^c R} \geq -\log p_* + \log(\sqrt{N}) - \frac{3}{e} \geq -\log p_* + \log(\sqrt{N^*_\varepsilon}) - \frac{3}{e}$$

since (57). Solving for p^*, the thesis follows. □

The proof of Theorem 2 is now an easy consequence of the above theorem.

Proof of Theorem 2. Since whenever a minimax lower rate holds over a prior, it holds a fortiori over a superset of it, without loss of generality we can assume

$$R \leq \frac{\min(M, \Sigma)}{2(4d+1)\sqrt{\kappa^c}},$$

hence enforcing condition (53).

Given $\tau > 0$ for all $\ell \in \mathbb{N}$, let $\varepsilon_\ell = \tau \ell^{-bc/(bc+1)}$. Since ε_ℓ goes to 0 when ℓ goes to $+\infty$, for ℓ large enough $\varepsilon_\ell \leq \varepsilon_0$, so Theorem 5 applies ensuring

$$\begin{aligned}\inf_{f_\ell} \sup_{\rho \in \mathcal{P}(b,c)} \mathbb{P}_{\mathbf{z} \sim \rho_*^\ell} \left[\mathcal{E}[f^\ell_{\mathbf{z}}] - \mathcal{E}[f_{\mathcal{H}}] > \frac{\tau}{\sqrt{2}} \ell^{-bc/(bc+1)} \right] \\ \geq \min \left\{ \frac{N^*_{\varepsilon_\ell}}{N^*_{\varepsilon_\ell} + 1}, \bar{\eta} e^{(C_1 \tau^{-1/bc} - C_2 \tau) \ell^{1/bc+1)}} \right\},\end{aligned}$$

where C_1, C_2 are positive constants independent of τ and ℓ. If ℓ goes to ∞, $N^*_{\varepsilon_\ell}/(N^*_{\varepsilon_\ell} + 1)$ goes to 1, whereas, if τ is small enough, the quantity $C_1\tau^{-1/bc} - C_2\tau$ is positive, so that

$$\lim_{\tau \to 0} \liminf_{\ell \to +\infty} \inf_{f_\ell} \sup_{\rho \in \mathcal{P}(b,c)} \mathbb{P}_{\mathbf{z} \sim \rho^\ell} \left[\mathcal{E}[f^\ell_{\mathbf{z}}] - \mathcal{E}[f_{\mathcal{H}}] > \frac{\tau}{\sqrt{2}} \ell^{-bc/(bc+1)} \right] = 1. \qquad \square$$

5.4. *Individual Lower Rate*

The proof of Theorem 3 is based on the similar result in [**17**, see Theorem 3.3]. Here that result is adapted to the general RKHS setting.

First of all we recall the following proposition, whose proof can be found in [**17**, Lemma 3.2, p. 38].

Proposition 7. *Let $\mathbf{g} \in \mathbb{R}^\ell$ and s a $\{+1, -1\}$-valued random variable with $\mathbb{P}[s = +1] = \mathbb{P}[s = -1] = \frac{1}{2}$. Moreover, let $\mathbf{n} = (n_i)_{i=1}^{\ell}$ be ℓ independent random variables distributed according to the Gaussian with zero mean and variance σ^2, independent of s. Set*

$$\mathbf{y} = s\mathbf{g} + \mathbf{n},$$

then the error probability of the Bayes decision for s based on $\mathbf{y}$ is

$$\min_{D:\mathbb{R}^\ell \to \{+1,-1\}} \mathbb{P}[D(\mathbf{y}) \neq s] = \Phi\left(-\frac{\|\mathbf{g}\|}{\sigma}\right),$$

where Φ is the standard normal distribution function.

Proof of Theorem 3. Let us reason for a fixed $B > b$, and let $\varepsilon := (B - b)c > 0$.

We first define the subset $\mathcal{P}'$ of $\mathcal{P}(b, c)$, then prove the lower rate on this subset. As in the proof of the minimax lower rate, we fix an arbitrary $\rho_0 \in \mathcal{P}(b, c)$ and let ν be its marginal measure. For every sequence $\mathbf{s} = (s_n)_{n\in\mathbb{N}} \in \{+1, -1\}^\infty$, we define a corresponding function in $\mathcal{H}$,

$$m^{(\mathbf{s})} := \sum_{n=1}^{+\infty} s_n \sqrt{t_n^{-1}\gamma_n} e_n = \sum_{n=1}^{+\infty} s_n g_n,$$

where

$$\gamma_n := n^{-(bc+\varepsilon+1)} \frac{\varepsilon}{\varepsilon + 1} \alpha^c R, \qquad g_n := \sqrt{t_n^{-1}\gamma_n} e_n,$$

where we recall that $(t_n)_{n\in\mathbb{N}}$ and $(e_n)_{n\in\mathbb{N}}$ are the eigenvalues and eigenvectors of the operator T defined by (52).

We define $\mathcal{P}'$ to be the set of probability measures ρ which fulfill the following two conditions:

- the marginal distribution ρ_X is equal to ν; and
- there is $\mathbf{s} \in \{+1, -1\}^\infty$ such that, for all $x \in X$, the conditional distribution of y given x,

  $$\rho(y|x) = \mathcal{N}(m^{(\mathbf{s})}(x), \sigma^2 \,\mathrm{Id}),$$

  that is, the multivariate normal distribution on Y with mean $m^{(\mathbf{s})}(x)$ and diagonal covariance σ^2 Id with

  $$\sigma^2 = \min\left(\frac{M^2}{2}, \frac{\pi^{d/2}\Sigma^2}{4S^d \int_0^{+\infty} e^{-z^2+z} z^{d+1}\, dz}\right),$$

and S^d is the volume of the surface of the d-dimensional unit radius sphere.

It is simple to check that $\mathcal{P}' \subset \mathcal{P}(b,c)$. Indeed, clearly $f_\rho = f_\mathcal{H} = m^{(\mathbf{s})}$ and

$$\begin{aligned}\|T^{-(c-1)/2}m^{(\mathbf{s})}\|_\mathcal{H}^2 &= \sum_{n=1}^{+\infty} t_n^{-(c-1)} t_n^{-1}\gamma_n = \sum_{n=1}^{+\infty}\left(\frac{\alpha}{n^b t_n}\right)^c n^{-(1+\varepsilon)}\frac{\varepsilon}{\varepsilon+1}R \\ &\le \sum_{n=1}^{+\infty} n^{-(1+\varepsilon)}\frac{\varepsilon}{\varepsilon+1}R \\ &\le \left(\int_1^{+\infty} t^{-(1+\varepsilon)}\,dt + 1\right)\frac{\varepsilon}{\varepsilon+1}R = R,\end{aligned}$$

where we used the lower bound (17). Moreover, $\mathcal{N}(0,\sigma^2\,\mathrm{Id})$ fulfills the moment condition (9) in Hypothesis 2. Indeed,

$$\begin{aligned}&\int_Y \left(e^{\|y-f_\mathcal{H}(x)\|_Y/M} - \frac{\|y-f_\mathcal{H}(x)\|_Y}{M} - 1\right)\rho(y|x) \\ &\quad= (2\pi\sigma^2)^{-d/2} S^d \int_0^{+\infty}\left(e^{z/M} - \frac{z}{M} - 1\right) e^{-z^2/2\sigma^2} z^{d-1}\,dz \\ &\quad= \pi^{-d/2} S^d \int_0^{+\infty} e^{-z^2} z^{d-1}\sum_{k=2}^{+\infty}\frac{1}{k}\left(\frac{\sqrt{2}z\sigma}{M}\right)^k dz \\ &\quad\le \frac{2\sigma^2}{M^2}\pi^{-d/2} S^d \int_0^{+\infty} e^{z(\sqrt{2}\sigma/M)} e^{-z^2} z^{d+1}\,dz \le \frac{\sigma^2}{2M^2}.\end{aligned}$$

We now are left with proving the lower bound on the reduced set $\mathcal{P}'$ by showing the inequality

$$\inf_{\{f_\ell\}_{\ell\in\mathbb{N}}} \sup_{\rho\in\mathcal{P}'} \limsup_{\ell\to+\infty} \frac{\mathbb{E}_{\mathbf{z}\sim\rho^\ell}(\mathcal{E}[f_\mathbf{z}^\ell] - \mathcal{E}[m^{(\mathbf{s})}])}{\ell^{-Bc/(Bc+1)}} > 0. \tag{62}$$

Since $(e_n)_{n\in\mathbb{N}}$ is an orthonormal sequence in $\mathcal{H}$, then for any $\mathbf{s}$ it holds that

$$\mathcal{E}[f_\mathbf{z}^\ell] - \mathcal{E}[m^{(\mathbf{s})}] = \|\sqrt{T}(f_\mathbf{z}^\ell - m^{(\mathbf{s})})\|_\mathcal{H}^2 = \sum_{n=1}^{+\infty}(c_{\mathbf{z},n} - s_n)^2\gamma_n, \tag{63}$$

with

$$c_{\mathbf{z},n} = \sqrt{\frac{t_n}{\gamma_n}}\left\langle f_\mathbf{z}^\ell, e_n\right\rangle_\mathcal{H}.$$

Now let $\tilde{c}_{\mathbf{z},n}$ be 1 if $c_{\mathbf{z},n} \ge 0$ and -1 otherwise. Because of the straightforward inequality

$$2|c_{\mathbf{z},n} - s_n| \ge |\tilde{c}_{\mathbf{z},n} - s_n|, \tag{64}$$

from (63) we get

$$\begin{aligned}\mathcal{E}[f_{\mathbf{z}}^{\ell}] - \mathcal{E}[m^{(\mathbf{s})}] &\geq \sum_{n=1}^{+\infty} \tfrac{1}{4}(\tilde{c}_{\mathbf{z},n} - s_n)^2 \gamma_n \\ &= \sum_{n=1}^{+\infty} I_{\{\tilde{c}_{\mathbf{z},n} \neq s_n\}} \gamma_n \geq \sum_{n \in \mathcal{D}_\ell} I_{\{\tilde{c}_{\mathbf{z},n} \neq s_n\}} \gamma_n,\end{aligned}$$

where the set $\mathcal{D}_\ell$ is defined by

$$\mathcal{D}_\ell := \{n \in \mathbb{N} \mid \ell \gamma_n \leq 1\}.$$

Note that due to (64) and defining the quantity

$$R_\ell(\mathbf{s}) := \sum_{n \in \mathcal{D}_\ell} \mathbb{P}_{\mathbf{z} \sim \rho^\ell}[\tilde{c}_{\mathbf{z},n} \neq s_n] \gamma_n \leq \sum_{n \in \mathcal{D}_\ell} \gamma_n, \tag{65}$$

from (62) we are led to prove the inequality

$$\inf_{\{f_\ell\}_{\ell \in \mathbb{N}}} \sup_{\mathbf{s} \in \{+1,-1\}^\infty} \limsup_{\ell \to +\infty} \frac{R_\ell(\mathbf{s})}{\ell^{-Bc/(Bc+1)}} > 0. \tag{66}$$

This result is achieved considering a suitable probability measure over the set $\{+1, -1\}^\infty$ (and hence over $\mathcal{P}'$ itself), and proving that the inequality above holds true not just for the worst $\mathbf{s}$, but also on average. Then let us introduce the sequence $\mathbf{S} = (S_i)_{i \in \mathbb{N}}$ of independent $\{+1, -1\}$-valued random variables with

$$\mathbb{P}[S_i = +1] = \mathbb{P}[S_i = -1] = \tfrac{1}{2} \quad \text{for all } i \in \mathbb{N}.$$

The plan is first to show that (66) is a consequence of the inequality

$$\mathbb{E} R_\ell(\mathbf{S}) \geq C \sum_{n \in \mathcal{D}_\ell} \gamma_n, \qquad C > 0, \tag{67}$$

and subsequently proving that indeed (67) is true for some $C > 0$.

Defining the constants $u := [\varepsilon/(\varepsilon+1)]\alpha^c R$ and $v := [(1/2Bc)]u^{-Bc/(Bc+1)}$, the definition of γ_n gives

$$\begin{aligned}\sum_{n \in \mathcal{D}_\ell} \gamma_n &= \sum_{n \geq (u\ell)^{1/(bc+1+\varepsilon)}} u n^{-(bc+1+\varepsilon)} \\ &\geq \int_{(u\ell)^{1/(bc+1+\varepsilon)}}^{+\infty} t^{-(bc+1+\varepsilon)}\, dt - \ell^{-1} \\ &= \frac{1}{Bc}(u\ell)^{-Bc/(Bc+1)} - \ell^{-1} \geq \frac{1}{2Bc}(u\ell)^{-Bc/(Bc+1)} \\ &= v \ell^{-Bc/(Bc+1)},\end{aligned} \tag{68}$$

where the last inequality holds for all $\ell \geq 2Bc(2Bcu)^{Bc}$.

Then using inequalities (67) and (68) we get

$$\begin{aligned}
\inf_{f_\ell} \sup_{\mathbf{s}\in\{+1,-1\}^\infty} \limsup_{\ell\to+\infty} \frac{R_\ell(\mathbf{s})}{\ell^{-Bc/(Bc+1)}} &\geq Cv \inf_{f_\ell} \sup_{\mathbf{s}\in\{+1,-1\}^\infty} \limsup_{\ell\to+\infty} \frac{R_\ell(\mathbf{s})}{\mathbb{E}R_\ell(\mathbf{S})} \\
&\geq Cv \inf_{f_\ell} \mathbb{E} \limsup_{\ell\to+\infty} \frac{R_\ell(\mathbf{S})}{\mathbb{E}R_\ell(\mathbf{S})} \\
&\geq Cv \inf_{f_\ell} \limsup_{\ell\to+\infty} \mathbb{E}\left(\frac{R_\ell(\mathbf{S})}{\mathbb{E}R_\ell(\mathbf{S})}\right) = Cv > 0,
\end{aligned}$$

where in the last estimate we applied the Fatou lemma, recalling that by inequalities (65) and (67) the sequence

$$\frac{R_\ell(\mathbf{s})}{\mathbb{E}R_\ell(\mathbf{s})}$$

is uniformly bounded for every $\mathbf{s} \in \{+1, -1\}^\infty$.

As planned we finally proceed proving inequality (67). Recall that by definition

$$\mathbb{E}R_\ell(\mathbf{S}) = \sum_{n\in\mathcal{D}_\ell} \mathbb{P}[\tilde{c}_{\mathbf{z},n} \neq S_n]\gamma_n,$$

where $\tilde{c}_{\mathbf{z},n}$ can be interpreted as a decision rule for the value of S_n given $\mathbf{z}$. The least error probability for such a problem is attained by the Bayes decision $\bar{c}_{\mathbf{z},n}$ which outputs 1 if $\mathbb{P}[S_n = 1 \mid \mathbf{z}] \geq \frac{1}{2}$ and -1 otherwise, therefore

$$\mathbb{P}[\tilde{c}_{\mathbf{z},n} \neq S_n] \geq \mathbb{P}[\bar{c}_{\mathbf{z},n} \neq S_n].$$

Since by construction S_n is independent of the X component of the data $\mathbf{z}$, we can reason conditionally on $(x_i)_{i=1}^{\ell}$. The dependence of the Y component of $\mathbf{z}$ on S_n has the form

$$y_i = m^{(\mathbf{s})}(x_i) + n_i = S_n g_n(x_i) + n_i + \sum_{k\neq n} S_k g_k(x_i), \qquad i = 1, \ldots, \ell,$$

with n_i independent Y-valued random variables distributed according to the Gaussian $\mathcal{N}(0, \sigma^2 \text{ Id})$. Hence it is clear that also the component of y_i perpendicular to $g_n(x_i)$ is independent of S_n. Consequently, the only dependence of $\mathbf{z}$ on S_n is determined by the longitudinal components

$$y_i' := \frac{\langle y_i, g_n(x_i)\rangle_Y}{\|g_n(x_i)\|_Y} = S_n\|g_n(x_i)\|_Y + n_i' + h_i, \tag{69}$$

where n_i' are real-valued random variables distributed according to the Gaussian $\mathcal{N}(0, \sigma^2)$ and

$$h_i = \sum_{k\neq n} S_k \frac{\langle g_k(x_i), g_n(x_i)\rangle_Y}{\|g_n(x_i)\|_Y}.$$

From equation (69) we see that the structure of the data available to the Bayes rule $\bar{c}_{\mathbf{z},n}$ for S_n is similar to that assumed in Proposition 7, except for the presence of the term $\mathbf{h} = (h_i)_{i=1}^{\ell}$. However, this term is independent of S_n, and it is clear that the Bayes error cannot decrease when such a term is added to the available data, in fact,

$$\begin{aligned}\min_{D:\mathbb{R}^{\ell}\to\{+1,-1\}} \mathbb{P}[D(\mathbf{g}+\mathbf{h}) \neq S_n] &= \min_{D:\mathbb{R}^{\ell}\to\{+1,-1\}} \mathbb{E}_{\mathbf{h}}\mathbb{P}\,[D(\mathbf{g}+\mathbf{h}) \neq S_n \mid \mathbf{h}] \\ &\geq \mathbb{E}_{\mathbf{h}} \min_{D:\mathbb{R}^{\ell}\to\{+1,-1\}} \mathbb{P}[D(\mathbf{g}+\mathbf{h}) \neq S_n \mid \mathbf{h}] \\ &= \mathbb{E}_{\mathbf{h}} \min_{D:\mathbb{R}^{2\ell}\to\{+1,-1\}} \mathbb{P}[D(\mathbf{g},\mathbf{h}) \neq S_n] \\ &= \min_{D:\mathbb{R}^{\ell}\to\{+1,-1\}} \mathbb{P}[D(\mathbf{g}) \neq S_n],\end{aligned}$$

where the last equality derives from the independence of $\mathbf{h}$ on S_n and we let $\mathbf{g} = (\|g_n(x_i)\|_Y)_{i=1}^{\ell}$.

Hence, by Proposition 7,

$$\begin{aligned}\mathbb{P}[\bar{c}_{\mathbf{z},n} \neq S_n \mid (x_i)_i] &= \min_{D:\mathbb{R}^{\ell}\to\{+1,-1\}} \mathbb{P}[D(\mathbf{g}+\mathbf{h}) \neq S_n \mid (x_i)_i] \\ &\geq \min_{D:\mathbb{R}^{\ell}\to\{+1,-1\}} \mathbb{P}[D(\mathbf{g}) \neq S_n \mid (x_i)_i] \\ &= \Phi\left(-\sqrt{\frac{\sum_i \|g_n(x_i)\|_Y^2}{\sigma^2}}\right).\end{aligned}$$

Moreover, since $\Phi(-\sqrt{x})$ is convex, by Jensen's inequality

$$\begin{aligned}\mathbb{P}\left[\bar{c}_{\mathbf{z},n} \neq S_n\right] &\geq \mathbb{E}\Phi\left(-\sqrt{\frac{\sum_i \|g_n(x_i)\|_Y^2}{\sigma^2}}\right) \\ &\geq \Phi\left(-\frac{1}{\sigma}\sqrt{\mathbb{E}\sum_i \|g_n(x_i)\|_Y^2}\right) = \Phi\left(-\frac{1}{\sigma}\sqrt{\ell\gamma_n}\right),\end{aligned}$$

where we used

$$\mathbb{E}\|g(x)\|_Y^2 = \int_X \left\langle K_x^* g_n, K_x^* g_n\right\rangle_Y \, d\nu(x) = \langle T g_n, g_n\rangle_{\mathcal{H}} = \gamma_n.$$

Thus

$$\begin{aligned}\mathbb{E}R_{\ell}(\mathbf{S}) &\geq \sum_{n\in\mathcal{D}_{\ell}} \Phi\left(-\frac{1}{\sigma}\sqrt{\ell\gamma_n}\right)\gamma_n \\ &\geq \Phi\left(-\frac{1}{\sigma}\right)\sum_{n\in\mathcal{D}_{\ell}} \gamma_n,\end{aligned}$$

which finally proves inequality (67) with $C = \Phi(-1/\sigma)$, and concludes the proof. □

6. Conclusion

We presented an error analysis of the RLS algorithm on RKHS for general operator-valued kernels. The framework we considered is extremely flexible and generalizes many settings previously proposed for this type of problem. In particular, the output space need not be bounded as long as a suitable moment condition for the output variable is fulfilled, and input spaces which are unbounded domains are dealt with. An asset of working with operator-valued kernels is the extension of our analysis to the multitask learning problem; this kind of result is, to our knowledge, new.

We also gave a complete asymptotic worst-case analysis for the RLS algorithm in this setting, showing optimality in the minimax sense on a suitable class of priors. Moreover, we extended previous individual lower rate results to our general setting.

Finally, we stress the central role played by the effective dimension in our analysis. It enters in the definition of the priors and in the expression of the non-asymptotic upper bound given by Theorem 4 in Subsection 5.1. However, since the effective dimension depends on both the kernel and the marginal probability distribution over the input space, our choice for the regularization parameter depends strongly on the marginal distribution. This consideration naturally raises the question of whether the effective dimension could be estimated by unlabeled data, allowing in this way the regularization parameter to adapt to the actual marginal distribution in a semisupervised setting.

Acknowledgments

The authors wish to thank T. Poggio, L. Rosasco, and S. Smale for useful discussions, and the two anonymous referees for helpful comments and suggestions.

This paper 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 Department of Brain & Cognitive Sciences, and which is affiliated with the Computer Sciences & Artificial Intelligence Laboratory (CSAIL), and also at the Dipartimento di Informatica e Scienze dell'Informazione (DISI), University of Genoa, Italy, as well as at the Dipartimento di Matematica, University of Modena, Italy.

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

Additional support was provided by: Central Research Institute of Electric Power Industry (CRIEPI), Daimler-Chrysler AG, Compaq/Digital Equipment Corporation, Eastman Kodak Company, Honda R&D Co., Ltd., Industrial Technology

Research Institute (ITRI), Komatsu Ltd., Eugene McDermott Foundation, Merrill-Lynch, NEC Fund, Oxygen, Siemens Corporate Research, Inc., Sony, Sumitomo Metal Industries, and Toyota Motor Corporation.

This research has also been funded by the FIRB Project ASTAA and the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778.

References

[1] Y. Altun, I. Tsochantaridis, and T. Hofmann, Hidden Markov support vector machines, in *20th International Conference on Machine Learning ICML*-2004, Washington, DC, 2003.
[2] N. Aronszajn, Theory of reproducing kernels, *Trans. Amer. Math. Soc.* **68** (1950), 337–404.
[3] J. Burbea and P. Masani, *Banach and Hilbert Spaces of Vector-Valued Functions*, Vol. 90, Pitman, Boston, MA, 1984.
[4] C. Carmeli, A. De Vito, and E. Toigo, Reproducing kernel Hilbert spaces and the Mercer theorem, *Anal. Appl.* (2006), to appear.
[5] F. Cucker and S. Smale, Best choices for regularization parameters in learning theory: On the bias-variance problem, *Found. Comput. Math.* **2** (2002), 413–428.
[6] F. Cucker and S. Smale, On the mathematical foundations of learning, *Bull. Amer. Math. Soc.* (*N.S.*) **39**(1) (2002), 1–49.
[7] E. De Vito and A. Caponnetto, Risk bounds for the regularized least-squares algorithm with operator-valued kernels, Technical report, Massachusetts Institute of Technology, Cambridge, MA, May 2005. CBCL Paper #249/AI Memo #2005-015.
[8] E. De Vito, A. Caponnetto, and L. Rosasco, Model selection for the regularized least-squares algorithm in learning theory, *Found. Comput. Math.* **5**(1) (2005), 59–85.
[9] E. De Vito, L Rosasco, A. Caponnetto, U. De Giovannini, and F. Odone, Learning from examples as an inverse problem, *J. Mach. Learn. Res.* **6** (2005), 883–904.
[10] R. DeVore, G. Kerkyacharian, D. Picard, and V. Temlyakov, Mathematical methods for supervised learning, *IMI Preprints* **22** (2004), 1–51.
[11] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Applications of Mathematics, Vol. 31, Springer-Verlag, New York, 1996.
[12] R. M. Dudley, *Real Analysis and Probability*, Cambridge Studies in Advanced Mathematics, Vol. 74, Cambridge University Press, Cambridge, UK, 2002. Revised reprint of the 1989 original.
[13] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Mathematics and its Applications, Vol. 375, Kluwer Academic, Dordrecht, 1996.
[14] T. Evgeniou, C. A. Micchelli, and M. Pontil, Learning multiple tasks with kernel methods, *J. Mach. Learn. Res.* **6** (2005), 615–637.
[15] T. Evgeniou, M. Pontil, and T. Poggio, Regularization networks and support vector machines, *Adv. Comput. Math.* **13** (2000), 1–50.
[16] C. W. Groetsch, *The Theory of Tikhonov Regularization for Fredholm Equations of the First Kind*, Research Notes in Mathematics, Vol. 105, Pitman, Boston, MA, 1984.
[17] L. Györfi, M. Kohler, A. Krzyzak, and H. Walk, *A Distribution-Free Theory of Non-parametric Regression*, Springer Series in Statistics, Springer-Verlag, New York, 2002.
[18] M. Kohler and A. Krzyżak, Nonparametric regression estimation using penalized least squares, *IEEE Trans. Inform. Theory* **47**(7) (2001), 3054–3058.
[19] S. Mendelson, On the performance of kernel classes, *J. Mach. Learn. Res.* **4** (2003), 759–771.
[20] C. A. Micchelli and M. Pontil, On learning vector-valued functions, *Neural Comput.* **17** (2005), 177–204.
[21] I. F. Pinelis and A. I. Sakhanenko, Remarks on inequalities for probabilities of large deviations, *Theory Probab. Appl.* **30**(1) (1985), 143–148.

[22] T. Poggio and F. Girosi, A theory of networks for approximation and learning, in C. Lau, ed., *Foundation of Neural Networks*, pp. 91–106, IEEE Press, Piscataway, NJ, 1992.
[23] T. Poggio and S. Smale, The mathematics of learning: Dealing with data, *Notices Amer. Math. Soc*. **50**(5) (2003), 537–544.
[24] B. Schölkopf and A. J. Smola, *Learning with Kernels*, MIT Press, Cambridge, MA, 2002.
[25] L. Schwartz, Sous-espaces Hilbertiens d'espaces vectoriels topologiques et noyaux associés (noyaux reproduisants), *J. Analyse Math*. **13** (1964), 115–256.
[26] S. Smale and D. Zhou, Shannon sampling II: Connections to learning theory, to appear in *Appl. Comput. Harmon. Anal*. **19** (2005), 285–302.
[27] S. Smale and D. Zhou, Learning theory estimates via integral operators and their approximations, Preprint, Toyota Technological Institute, Chicago, 2005.
[28] V. N. Temlyakov, Nonlinear methods of approximation, *Found. Comput. Math*. **3** (2003), 33–107.
[29] V. N. Temlyakov, Approximation in learning theory, *IMI Preprints* **5** (2005), 1–42.
[30] S. A. van de Geer, *Applications of the Empirical Process Theory*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge, UK, 2000.
[31] A. W. van der Vaart and J. A. Wellner, *Weak Convergence and Empirical Processes*, Springer Series in Statistics, Springer-Verlag, New York, 1996. With Applications to Statistics.
[32] V. N. Vapnik, *Statistical Learning Theory*, Adaptive and Learning Systems for Signal Processing, Communications, and Control, Wiley, New York, 1998. A Wiley-Interscience Publication.
[33] G. Wahba, *Spline Models for Observational Data*, CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. 59, SIAM, Philadelphia, PA, 1990.
[34] J. Weston, O. Chapelle, A. Elisseeff, B. Schoelkopf, and V. Vapnik, Kernel dependency estimation, in S. Thrun, S. Becker, and K. Obermayer, eds., *Advances in Neural Information Processing Systems*, Vol. 15, pp. 873–880, MIT Press, Cambridge, MA, 2003.
[35] V. Yurinsky, *Sums and Gaussian Vectors*, Lecture Notes in Mathematics, Vol. 1617, Springer-Verlag, Berlin, 1995.
[36] T. Zhang, Effective dimension and generalization of kernel learning, *NIPS* 2002, pp. 454–461.
[37] T. Zhang, Leave-one-out bounds for kernel methods, *Neural Comput*. **13** (2003), 1397–1437.